Australasian Association for Digital Humanities (aaDH)
Association for Computers and the Humanities (ACH)
Canadian Society for Digital Humanities / Société canadienne des humanités numériques (CSDH/SCHN)
centerNet
European Association for Digital Humanities (EADH)
Humanistica
Japanese Association for Digital Humanities (JADH)

Digital Humanities 2018
Puentes-Bridges

Book of Abstracts
Libro de resúmenes

Mexico City
26-29 June 2018

ALLIANCE OF DIGITAL HUMANITIES ORGANIZATIONS

EL COLEGIO DE MÉXICO

RedHD

PROGRAM COMMITTEE / COMITÉ PROGRAMA ACADÉMICO

Élika Ortega – Northeastern University (PC Co-chair)
Glen Worthey – Stanford University (PC Co-chair)
Sarah Kenderdine – aaDH
Chris Thomson – aaDH
Lisa Rhody – ACH
Alex Gil – ACH
Constance Crompton – CSDH/SCHN
Dan O'Donnell – CSDH/SCHN
Nancy Friedland – centerNet
Brian Rosenblum – centerNet
Bárbara Bordalejo – EADH
Elisabeth Burr – EADH
Björn-Olav Dozo – Humanistica
Emmanuel Chateau Dutier – Humanistica
Akihiro Kawase – JADH
Maki Miyake – JADH

LOCAL ORGANIZING COMMITTEE / COMITÉ LOCAL ORGANIZADOR

Isabel Galina – Universidad Nacional Autónoma de México (UNAM) (Co-chair)
Ernesto Priani – Universidad Nacional Autónoma de México (UNAM) (Co-chair)
Miriam Peña – Universidad Nacional Autónoma de México (UNAM)
Jonathan Girón Palau – Universidad Nacional Autónoma de México (UNAM)
Ernesto Miranda – Secretaria de Cultura
Micaela Chávez Villa – El Colegio de México (Colmex)
Alberto Santiago Martinez – El Colegio de México (Colmex)
Silvia Gutiérrez – El Colegio de México (Colmex)
Natalie Baur – El Colegio de México (Colmex)
León Ruiz – El Colegio de México (Colmex)

SPONSORS / PATROCINADORES

Agenda Digital de Cultura. Secretaría de Cultura
Consejo Nacional de Ciencia y Tecnología (Conacyt)
Gale, Cengage
Stanford University Press
Tecnológico de Monterrey. Escuela de Humanidades y Educación
The Association for Computers and the Humanities (ACH)
Universidad del Claustro de Sor Juana

We would like to thank the support of the Instituto de Investigaciones Sobre la Universidad y la Educación (IISUE) and the Instituto de Investigaciones Bibliográfica (IIB) of the Universidad Nacional Autónoma de México (UNAM). Also the generous funding from Conacyt, project number 293068 - Convocatoria 2018 del Programa de Apoyos para Actividades Científicas, Tecnológicas y de Innovación de la Dirección Adjunta de Desarrollo Científico.

La elaboración del libro de resúmenes fue posible gracias al apoyo del Instituto de Investigaciones Sobre la Universidad y la Educación (IISUE) y el Instituto de Investigaciones Bibliográfica (IIB) de la Universidad Nacional Autónoma de México. También fue posible gracias al financiamiento Conacyt proyecto número: 293068 - Convocatoria 2018 del Programa de Apoyos para Actividades Científicas, Tecnológicas y de Innovación de la Dirección Adjunta de Desarrollo Científico.

Digital Humanities 2018
Puentes-Bridges

Book of Abstracts
Libro de resúmenes

El Colegio de México

Universidad Nacional Autónoma de México

Red de Humanidades Digitales

26 - 29 June 2018
Mexico City

26 - 29 de junio 2018
Ciudad de México

Edited by / Editores
Jonathan Girón Palau
Isabel Galina Russell

DHConvalidator service
Aramís Concepción Durán
Christof Schöch

On-line abstracts / Resúmenes en línea
Reynaldo Crescencio

Design and typesetting / Diseño y maquetación
Yael Coronel Navarro
Juan Carlos Rosas Ramírez

Proof-reading / Revisión
Karla Guadalupe González Niño
Jessica América Gómez Flores

Online abstract available at: dh2018.adho.org/abstracts

Title: Digital Humanities 2018: Book of Abstracts / Libro de resúmenes.
Contributor (Corporate Author): Alliance of Digital Humanities Organizations.

Publisher: Red de Humanidades Digitales A. C.

Date of Publication: 2018

ISBN: 978-0-911221-62-6

Welcome to DH2018

Élika Ortega and Glen Worthey, Program Committee Co-chairs
Isabel Galina and Ernesto Priani, Local organizers, Co-chairs

As many old-timers and some newcomers know, this is the first time that the annual international Digital Humanities conference takes place in the Global South. This is a momentous achievement for an organization that has always strived to be truly global, diverse, and inclusive. The geographic movement of the conference has brought with it a renewed awareness of the differences among the numerous communities that constitute ADHO and the DH field at large. As we celebrate these differences, we have also made every effort for DH2018 to create meeting points, foster connections, and build bridges across the many Digital Humanities.

Making the conference bilingual, a tradition that we're following from DH2017, has been central to our work. Indeed, although English continues to be a powerful *lingua franca* in our field, about 20% of the presentations, posters, and panels this year are in another language. This development in the program is the result not only of the Program Committee's work; it was possible thanks to the 'backstage' volunteer labor of hundreds of reviewers who lent both their DH expertise, and their strong linguistic capacities. We also endeavored to make as much of the information and official communications of DH2018 bilingual, including its website, our email communications, the Convalidator tool, and this *Book of Abstracts*, to mention a few. There is still much left to do, and many interfaces are still available only in English, but we hope that our collective efforts will encourage all future ADHO conference organizers to continue in this tradition.

This year the conference includes twenty-two long paper sessions, twenty-two short paper sessions, thirty-three panel sessions, and sixteen workshops. Additionally, a two-part poster session will showcase the work of over 150 scholars. The topics and approaches represented span from linked data to digital ethnography; from classical antiquity to online activism; from pedagogy to theory; from indigenous languages to natural disasters. The broad scope of the program attests to the long-standing practices that first propelled the consolidation of the field of Digital Humanities, while making ample room for new approaches that increasingly bring us closer to the social, political, and natural challenges the world currently faces.

Our two DH2018 keynote speakers, Janet Chávez Santiago and Schuyler Esprit, bring our attention to the territories of the Central Valleys in Oaxaca in Mexico and the Caribbean island of Dominica. Impacted in distinct ways by colonial and neo-colonial powers, these sites are sources of *other* ways of seeing, weaving, and redesigning the world. They are also a locus sustaining the communities, academic and otherwise, that seek to utilize digital technologies for cultural, epistemological, and sometimes physical, survival.

Organizing DH2018 in Mexico City has been a challenge and a learning experience. Certain cultural assumptions have come to light simply by holding the conference in a different geographical location. We are sure that these experiences will be helpful as the conference continues to move to new and different locations. For us, Mexico's sociocultural diversity makes it an ideal location for converging digital humanists from distinct cultures, contexts, and socio-political realities. We believe that our steps towards bridging cultural, technological, political, and ideological borders will lead to the creation of a Digital Humanities community that is truly global, diverse, and inclusive.

Bienvenidos a DH2018

Élika Ortega y Glen Worthey, Co-presidentes del Comité Científico
Isabel Galina y Ernesto Priani, Co-presidentes del Comité Organizador Local

Como saben muchos veteranos y algunos novatos de DH, esta es la primera vez que la conferencia internacional Humanidades Digitales se lleva a cabo en el Sur Global. Se trata de un logro memorable para una organización que siempre se ha esforzado por ser verdaderamente global, diversa e incluyente. El cambio de ubicación de la conferencia ha aportado una conciencia renovada de las diferencias entre las diversas comunidades que forman ADHO y el campo de las HD, en general. Con el mismo entusiasmo con el que celebramos estas diferencias, nos hemos esforzado por crear puntos de encuentro en DH2018, establecer conexiones y construir puentes entre las muchas humanidades digitales.

Un aspecto central de nuestro trabajo ha sido preparar una conferencia bilingüe, una tradición que seguimos desde DH2017. Y si bien el inglés continúa siendo una importante *lingua franca* en nuestro campo, cerca de 20% de las presentaciones, pósters y paneles en el programa de este año están en otro idioma. Esta característica del programa no es el resultado solamente del trabajo del Comité Científico; fue posible gracias a la labor voluntaria "tras bambalinas" de cientos de dictaminadores que ofrecieron tanto su experticia en HD como sus habilidades lingüísticas. Asimismo, nos esforzamos para que gran parte de la información y las comunicaciones oficiales de DH2018 fueran bilingües, incluidos el sitio web, los correos electrónicos, la herramienta Convalidator, y este Libro de Resúmenes, por mencionar algunos. Aún falta mucho por hacer y muchas interfaces todavía se encuentran disponibles solamente en inglés, pero esperamos que el esfuerzo colectivo alentará a futuros organizadores de la conferencia de ADHO a continuar esta tradición.

Este año la conferencia incluye veintidós sesiones de presentaciones largas, veintidós sesiones de presentaciones breves, treinta y tres paneles y dieciséis talleres. También incluye una sesión doble de pósteres, que mostrará el trabajo de más de 150 académicos. Los tópicos y las aproximaciones presentados en el programa comprenden los datos conectados a la etnografía digital; de la antigüedad clásica al activismo en línea; desde la pedagogía a la teoría; de las lenguas indígenas a los desastres naturales. Este amplio rango de temas da cuenta de las prácticas que impulsaron la consolidación de las humanidades digitales y, al mismo tiempo, abre espacios para nuevas aproximaciones que, cada vez más, nos acercan a los desafíos sociales, políticos y naturales que el mundo encara actualmente.

Las dos ponentes magistrales para DH2018, Janet Chávez Santiago y Schuyler Esprit, nos transportan a los territorios de los Valles Centrales de Oaxaca, México y a la isla caribeña de Dominica. Impactados de formas distintas por las potencias coloniales y neocoloniales, estos sitios son la fuente de otras formas de ver, tejer y rediseñar el mundo. Son también los *loci* que sostienen comunidades, académicas y no académicas, que buscan utilizar las tecnologías digitales para la preservación cultural, epistemológica y, a veces, incluso la supervivencia física.

Organizar DH2018 en la Ciudad de México ha sido un reto y un aprendizaje. El simple hecho de que la conferencia se lleve a cabo en una región diferente ha sacado a la luz ciertas presuposiciones culturales y estamos seguros de que el aprendizaje se irá enriqueciendo en la medida en que la conferencia se realice en distintas ubicaciones. Consideramos que, por su diversidad sociocultural, México es un lugar ideal para la convergencia de humanistas digitales de culturas, contextos y realidades sociopolíticas particulares. Estamos convencidos de que, al encaminarnos hacia la creación de puentes entre fronteras culturales, tecnológicas, políticas e ideológicas nos acercaremos cada vez más a formar una comunidad de humanidades digitales verdaderamente global, diversa e incluyente.

Table of Contents

Plenary lectures

Panels

Long Papers

Short Papers

Posters

Preconference Workshops

Sig Endorsed

Plenary lectures

Weaving the Word

Janet Chávez Santiago
jazoula.10@gmail.com
Indigenous Languages Activist

The weft is a thread that is woven among the warp's yarns; these are our paper and pencil in the creation of a rug. Together, warp and weft are the bridge that unites the threads with our past and our present, and we weave the patterns of Mitla's friezes as a form of reading, or of interpreting, and of writing our ancestors, but also as a way to recount our dreams and our experiences. We weave in Zapotec. When we complete a rug, we share it with the world, and although the weave is in Zapotec, it can be interpreted in English, in Spanish, in Mixtec, or in Chatino.

Digital media can be seen as a warp on which the speakers of indigenous languages have an opportunity to weave their word and to share it within their own community and beyond. Although in our times digital media and social networks are a practical part of our daily lives and of our interactions with the world, we as speakers of indigenous languages must truly appropriate these spaces, to weave our word well, in order to liberate ourselves from the denial of the present.

Tramando la palabra

La trama es el hilo que se teje entre la urdimbre, son nuestro papel y lápiz para crear un lienzo. Juntos, trama y urdimbre, son el puente que unen los hilos con nuestro pasado y nuestro presente, tejemos las grecas de Mitla como una forma de leer o interpretar y escribir a nuestros ancestros, pero también para contar nuestros sueños y nuestras experiencias. Tramamos en zapoteco. Cuando terminamos un tapete lo compartimos con el mundo, y aunque el tejido está en zapoteco se puede interpretar en inglés, en español, en mixteco o en chatino.

Los medios digitales se pueden ver como una urdimbre en donde hablantes de lenguas indígenas tengan la oportunidad de tramar su palabra y compartirla dentro de su propia comunidad y más allá. Aunque hoy en día los medios digitales y las redes sociales son prácticamente parte de nuestra vida cotidiana y de nuestra interacción con el mundo, como hablantes de lenguas indígenas todavía nos hace falta apropiarnos realmente de estos espacios, tramar bien nuestra palabra para liberarnos de la negación del presente.

Digital Experimentation, Courageous Citizenship and Caribbean Futurism

Schuyler Esprit
schuyleresprit@gmail.com
Research Institute at Dominica State College

The violence and trauma of climate change have arrived. The Caribbean region is the unfortunate recipient of the impacts of climate change and, much like its inheritance of plantation slavery and colonialism, it is left with the infrastructural, social and cultural pillage of imperial and neocolonial imposition. My talk will consider whether and how the humanities, and digital humanities in particular, can produce the ideal intersection between planetary responsibility, community accountability and sustainable living.

In this talk I discuss Create Caribbean Research Institute's digital humanities praxis through the example of the environmental sustainability project, *Carisealand*. Through the exploration and discussion of theories, tools, methodologies and praxis of digital humanities applied to the project, I position Caribbean afrofuturism in the context of contemporary Caribbean digital environments and the lived experience of Caribbean people in the aftermath of climate change.

I apply discourses of afrofuturism to imagine an alternate Caribbean future represented in the redesign, digital imagination and representation of selected Caribbean communities. By offering models for rethinking, visualizing and rebuilding physical spaces, I hope to raise questions and offer insights about the power of digital humanities for social and environmental justice in the contemporary and future Caribbean. The goal is to also offer the model as a template for developing other mapping projects that can propose an alternate future for the Global South.

Experimentación Digital, Ciudadanía Valiente y Futurismo Caribeño

La violencia y el trauma del cambio climático ya comenzaron. La región del Caribe es la desafortunada receptora de los impactos del cambio climático y, al igual que con la herencia de esclavitud en las plantaciones y del colonialismo, sufre del saqueo infraestructural, social y cultural de la imposición imperial y neocolonial. Mi charla considerará si, y de qué forma, las humanidades, y las humanidades digitales en particular, pueden producir una intersección ideal entre la responsabilidad planetaria y comunitaria, y una vida sustentable.

Asimismo, en mi charla, discuto la práctica de las humanidades digitales en el Instituto de Investigación Create Caribbean utilizando como ejemplo el proyecto de sustentabilidad ambiental *Carisealand*. Por medio de la exploración y discusión de las teorías, herramientas, metodologías y prácticas de las humanidades digitales aplicadas en el proyecto, ubico el afrofuturismo caribeño en el contexto de los ambientes digitales contemporáneos del Caribe y la experiencia de los caribeños que viven con las repercusiones del cambio climático.

Finalmente, pongo en práctica los discursos del afrofuturismo para imaginar un futuro caribeño alternativo representado en el rediseño, la imaginación y representación digitales de ciertas comunidades caribeñas. Al ofrecer modelos para repensar, visualizar y reconstruir los espacios físicos, deseo despertar preguntas y ofrecer entendimiento acerca del poder que las humanidades digitales tienen para crear justicia social y ambiental en el Caribe contemporáneo y futuro. La meta es también ofrecer este modelo como una plantilla para desarrollar otros proyectos de mapeo que pueden proponer un futuro alternativo para el Sur Global.

Short Papers

Archivos digitales, cultura participativa y nuevos alfabetismos: La catalogación colaborativa del Archivo Histórico Regional de Boyacá (Colombia)

Maria Jose Afanador-Llach
mj.afanador28@uniandes.edu.co
Universidad de los Andes; Fundación Histórica Neogranadina, Colombia

Andres Lombana
alombana@cyber.law.harvard.edu
Berkman Klein Center for Internet and Society, Harvard University, United States of America

Este artículo explora las prácticas colaborativas de catalogación y creación de metadatos en archivos localizados en contextos de escasa conectividad, acceso tecnológico limitado, e incipiente desarrollo de nuevos alfabetismos. Tomando como ejemplo el proyecto de Catalogación Colaborativa del Fondo Notaría Segunda del Archivo Histórico Regional de Boyacá en la ciudad de Tunja, Colombia, analizamos cómo una plataforma digital y una comunidad de práctica pueden suplir las necesidades de acceso a tecnología, información y conocimiento a través de la "producción entre pares" o "*peer production*" (Benkler 2006; Benkler, Shaw, & Hill 2015) y la cultura participativa (Jenkins et al. 2006; Jenkins 2010). Dada la desigualdad de acceso a recursos tecnológicos, culturales y humanos para proyectos de digitalización y catalogación documental, en este artículo identificamos estrategias para acceder a tecnologías abiertas, y desarrollar nuevos alfabetismos (Lankshear and Nobel 2006, 2007; Dussel 2009; Jenkins et al. 2006; Jenkins 2010) que faciliten la producción colectiva de conocimiento y la construcción de culturas participativas desde el sur global.

En Colombia, la situación de numerosos archivos históricos regionales, se ha caracterizado por la carencia de una organización sistemática de sus colecciones, contribuyendo a que permanezcan subutilizados por parte de los investigadores y del público general (Marín 2004). Los procesos de digitalización presentan entonces una oportunidad no solamente para la preservación de archivos sino también para la catalogación y creación de metadatos de calidad que garanticen el acceso y usabilidad a futuro, y para el fomento de una cultura participativa. Sin embargo, existen numerosos archivos privados con colecciones patrimoniales que carecen de acceso a los recursos para llevar a cabo procesos de digitalización, catalogación y creación de metadatos. Tal es el caso del Archivo Histórico Regional de Boyacá (AHRB), en Tunja, Colombia, un archivo privado con colecciones que van desde 1539 hasta 1850, sin acceso a los recursos y apoyos de la red pública de archivos, y carente de catálogos para algunas de sus colecciones documentales.

A partir de un experimento de construcción colaborativa de catálogos para el AHRB en este artículo abordamos la siguiente pregunta: ¿De qué manera puede el uso de tecnologías digitales y en red por expertos y aficionados ampliar el alcance de la investigación en las humanidades digitales en contextos de escasa conectividad, acceso tecnológico limitado e incipiente desarrollo de nuevos alfabetismos? A través del análisis de las motivaciones y prácticas socioculturales desarrolladas por los participantes del proyecto AHRB, elaboramos una reflexión sobre los retos y oportunidades que la producción colaborativa de información y conocimiento, o „producción entre pares" (*peer production*), ofrece a los procesos de migración de materiales culturales a formatos digitales, particularmente en contextos donde el acceso a recursos tecnológicos es limitado. En dicho experimento, el proceso de catalogación del Fondo Notaría Segunda permitió a un grupo de expertos y aficionados conformar una comunidad de práctica (Wenger 1998), desarrollar nuevos alfabetismos relacionados a la paleografía y participar en un proceso de producción entre pares.

Existen diversos proyectos de *crowdsourcing* en las humanidades digitales que han sido objeto de análisis en el mundo angloparlante (Terras 2016). Sin embargo, los retos de la colaboración abierta distribuida en el sur global están conectados a factores culturales, económicos y de acceso a tecnología, que han sido poco estudiados. Nuestro análisis del proyecto del AHRB permite apreciar cómo la comunidad de práctica conformada para la catalogación de documentos históricos le ofrece a los participantes no solo la oportunidad de contribuir a la construcción de la memoria pública (Owens 2012) sino también desarrollar nuevos alfabetismos como el trabajo en red entre pares y la inteligencia colectiva (Jenkins et. al. 2006). A pesar de las brechas digitales existentes en algunos contextos locales, el proceso de catalogación colaborativa permite crear puentes de acceso a tecnología, tejer redes entre expertos y aficionados, y cultivar una cultura participativa, a la vez que contribuye a la conservación y promoción del patrimonio cultural.

References

Benkler, Y. (2006). *The Wealth of Networks: How Social Production Transforms Markets and Freedom*. New Heaven, Connecticut: Yale University Press.

Benkler, Y., Shaw, A and Hill, B.M. (2015) "Peer Production: A Form of Collective Intelligence." In *Hand- book of Collective Intelligence*, edited by Thomas Malone and Michael Bernstein. MIT Press, Cambridge, Massachusetts.

Dussel, I. (2009) "Los nuevos alfabetismos en el siglo XXI: desafíos para la escuela", *conferencia en Virtualeduca, 2009*. http://www.virtualeduca.info/Documentos/veBA09%20_confDussel.pdf

Jenkins, H. (2010) Afterword: Communities of readers, clusters of practices. In M. Knobel and C. Lankshear (Eds) *DIY Media: Creating, Sharing and Learning with New Technologies*. New York: Peter Lang, pp. 231–53.

Jenkins, H. et al. (2006) *Confronting the Challenges of a Participatory Culture: Media Education for the 21st Century*. Chicago: The MacArthur Foundation.
Lankshear, C., & Knobel, M (2007) "Sampling the New' in New Literacies." In Lankshear, C., & Knobel, M. *A new literacies sampler*. New York : P. Lang.
Lankshear, C., & Knobel, M. (2006). *New literacies: Everyday practices and classroom learning*. 2nd ed. Maidenhead, UK: Open University Press.
Marín, M. "Elementos de la archivística colombiana para la historia de los orígenes de la provincia." En *Theologica Xaveriana*, 152 (2004), 707-718.
Owens, T. (2012a). Crowdsourcing Cultural Heritage: The Objectives Are Upside
Down. http://www.trevorowens.org/2012/03/crowdsourcing-cultural-heritage-the-objectives-are-upside-down/.
Terras, M. (2016) "Crowdsourcing in the Digital Humanities," in Schreibman, S., Siemens, R., and Unsworth, J. (eds). *A New Companion in the Digital Humanities*, Blackwell Companions to Literature and Culture Series, Wiley
Wenger, E. (1998). *Communities of practice: Learning, meaning, and identity*. Cambridge, UK: Cambridge University Press.

The Programming Historian en español: Estrategias y retos para la construcción de una comunidad global de HD

Maria Jose Afanador-Llach
mj.afanador28@uniandes.edu.co
Universidad de los Andes, Colombia

The Programming Historian comenzó en el año 2008 como una publicación de acceso abierto que publica tutoriales revisados por pares dirigidos a humanistas para aprender una amplia gama de herramientas, técnicas computacionales y flujos de trabajo útiles para investigación y pedagogía. El proyecto está liderado por un equipo de doce editores voluntarios de seis países con el objetivo de crear una comunidad colaborativa y una audiencia de carácter global. Desde agosto de 2016, el equipo editorial de contenidos en español de PH comenzó el proceso de traducción de los más de 50 tutoriales publicados en el portal del proyecto en inglés. Al la fecha se han traducido alrededor de 30 tutoriales a partir de la participación de alrededor de 15 colaboradores de países como Argentina, España, Colombia y México.

La expansión de una comunidad de práctica de humanidades digitales en el mundo hispanoparlante plantea preguntas sobre acceso y diversidad. La brecha digital, en su dimensión de uso y aprovechamiento de tecnologías de la información y el desarrollo de competencias digitales, implica serios retos para la producción de conocimiento sobre HD en el sur global. PH, una publicación en línea de acceso abierto bilingüe, ha desarrollado un modelo para afrontar el problema del acceso global y lingüístico a recursos, metodologías y herramientas digitales para las humanidades. Este compromiso con la diversidad lingüística y geográfica en las humanidades digitales significa comprender los límites y posibilidades de los contextos institucionales, históricos, culturales y económicos en el mundo hispanoparlante.

Estamos en un momento de expansión del campo de las humanidades digitales en España y América Latina. Ya existen programas de posgrados en HD en universidades Latinoamericanas (Universidad de los Andes, Universidad del Claustro de Sor Juana), que se suman a las ofertas ya existentes en España (por ejemplo, LINHD). En este contexto de expansión, PH representa un proyecto colaborativo de servicio académico voluntario, que se sostiene en la conformación de redes globales de conocimiento abierto. El proyecto ha enfrentado los retos que suponen encontrar voluntarios que quieran revisar, traducir y crear tutoriales del inglés al español. Lo anterior, teniendo en cuenta la falta de reconocimiento y validez académica dada la carencia de mecanismos de evaluación de productos de investigación digital (Galina Russell 2016). De igual manera, ha resultado un reto garantizar la calidad de los contenidos desde un punto de vista lingüístico. Por último, el proyecto afronta el reto de combinar una aproximación global, que al mismo tiempo respete la diversidad local y que no reproduzca prácticas colonizadoras. Estos retos además se alinean con la misión de PH crear recursos sustentables con una prioridad por el Acceso Abierto y los recursos libres y de código abierto.

Esta presentación es una reflexión sobre la experiencia del equipo de contenidos en español de *The Programming Historian* en relación al panorama general de las humanidades digitales en el mundo hispanoparlante. En primer lugar, se pretende analizar las estrategias de divulgación del proyecto y evalúa las experiencias de uso de los tutoriales de PH en el salón de clase y en talleres. En segundo lugar, analizamos el comportamiento del tráfico de usuarios del portal de PH en español desde su lanzamiento en comparación con la evolución del tráfico en el portal en inglés. (Ver muestra de datos en las Figuras 1 y 2) Se analizará también cuáles han sido los tutoriales más visitados y los menos visitados, los lugares de mayor acceso y el tiempo promedio de los usuarios en los tutoriales. En tercer lugar, nos gustaría reflexionar, asimismo, sobre los retos de construir una comunidad de colaboradores que además de hacer traducciones, produzca contenidos sobre herramientas y metodologías de trabajo digital para las humanidades en español.

Las estrategias de divulgación del proyecto y de construcción de una comunidad de colaboradores se ha llevado a cabo mayoritariamente a través de redes sociales, encuestas en línea, listas de correos y ocasionalmente charlas presenciales. Mientras se consolidan espacios institucionales que apoyen la investigación desde

las humanidades digitales, consideramos que será difícil que los países de habla hispana produzcan contenidos y tutoriales en español. Sin embargo, los esfuerzos de traducción son esenciales para impulsar una comunidad de práctica en el sur global. A futuro, esperamos que los investigadores en el mundo hispanoparlante y otras partes del mundo contribuyan a la producción de nuevas metodologías, herramientas y flujos de trabajo digital que reflejen las particularidades sociales, culturales e históricas de las humanidades de los contextos de Latinoamérica y España, y el sur global.

País	Enero 2017	Mayo 2017	Octubre 2017
México	163	472	2100
Colombia	85	252	1200
España	454	912	2300
Argentina	94	188	891
Brasil	392	585	878

Figura 1. Número de sesiones en portal de PH desde países hispanoparlantes, 2017

País	Enero 2017	Mayo 2017	Octubre 2017
Estados Unidos	10,000	13,000	23,000
India	3,900	4,500	7,300
Alemania	1,500	1,900	2,100
Reino Unido	2,400	2,200	5,800

Figura 2. Número de sesiones en portal de PH desde países angloparlantes, 2017

La Sala de la Reina Isabel en el Museo del Prado, 1875-1877: La realidad aumentada en 3D como método de investigación, producto y vehículo pedagógico

Eugenia V Afinoguenova
eugenia.afinoguenova@marquette.edu
Marquette University, United States of America

Chris Larkee
christopher.larkee@marquette.edu
MarVL: Marquette University Visualization Laboratory, United States of America

Giuseppe Mazzone
gmazzone@nd.edu
School of Architecture, Notre Dame University, United States of America

Pierre Géal
pierre.geal@univ-grenoble-alpes.fr
Université Grenoble Alpes, France

http://prado.nfshost.com

Hacia 1875-1877, el fotógrafo francés Jean Laurent retrató la Sala de la Reina Isabel del Museo del Prado en Madrid (Fig. 1). Ocupando un espacio absidial en el centro del edificio que el arquitecto Juan de Villanueva había planeado un siglo antes como un salón de juntas, la Sala de la Reina Isabel reunía los cuadros que entonces se consideraban las "perlas" de la colección. De modo similar a la Tribuna de la Galería de los Uffizi o el Salon Carré del Louvre, la colocación de los cuadros propiciaba comparaciones estéticas. En 1893, el espacio fue reformado y en 1899 se convirtió en la Sala Velázquez.

En 2015-2017, a partir de la fotografía de Laurent, nuestro equipo interdisciplinar emprendió una reconstrucción digital en 3D de este espacio que todavía existe, pero ha sido profundamente transformado. Para reconstruir la estructura original, hemos utilizado las medidas que se encuentran en el proyecto de la reforma fechado en 1887 y las pruebas de color recientemente hechas en las paredes del museo. Un bosquejo original de Federico de Madrazo fue utilizado para reconstruir los banquillos.

Una vez que el modelo estaba hecho, había que "colgar" los cuadros. Pero la fotografía original solamente recogía una parte de la sala. La tarea de reconstruir la exposición transformó el trabajo de visualización en un proyecto de investigación. Al analizar el posicionamiento de la cámara fotográfica, se llegó a la conclusión de que la cámara no fue centrada y, además, tenía una inclinación. Este análisis permitió averiguar la superficie de las paredes en que se debía poner los cuadros restantes. Sabíamos qué obras eran debido al trabajo previo de Géal (2001 y 2005: 495-515), quien había utilizado las guías decimonónicas para establecer una lista de obras que se encontraban en la Sala de la Reina Isabel.

Nuestro plan era terminar la identificación de los cuadros en la foto, llegar a una hipótesis sobre los criterios subyacentes en la colocación de los cuadros y aplicar estos criterios para encontrar un sitio para las obras restantes. Para lograrlo, tuvimos que superar dos desafíos: 1) no sabíamos exactamente en qué orden estaban los cuadros y 2) no había sitio para todos los cuadros que, según las guías, estaban en la sala. La solución para el primer problema vino en forma de la guía de España publicada en 1878 (Ford 1878: 57-59), que menciona gran parte de los cuadros expuestos en la Sala. Infiriendo el movimiento de la descripción comparando el texto con la foto, extrapolamos el orden de la mención a otras paredes. La misma guía nos permitió establecer una lista mínima de las obras expuestas.

La atribución y los marcos han cambiado considerablemente desde 1875-1877. Nuestro proyecto reconstruye los

marcos de aquel entonces a base de las placas de cristal hechas por Laurent en la misma época. Mientras las tablillas reproducen la atribución decimonónica, los usuarios pueden activar las anotaciones que reflejan la atribución actual.

La resultante reconstrucción existe en tres versiones, cada una diseñada para un público y usos diferentes. La versión inicial fue ideada como espacio inmersivo interactivo para una "cueva" de proyección en 3D (Fig. 2). Este espacio, de 20 pies de ancho, se utiliza para clases y conferencias que crean una experiencia extremadamente detallada, en algunos aspectos superior a una visita al museo, generada a partir de los programas Blender e Unity. Para abrir la experiencia a un mayor número de usuarios, hemos creado una versión optimizada para teléfonos móviles Samsung Galaxy S6 y gafas de RV. Esto nos hizo buscar soluciones ingeniosas en cuanto a las texturas y la iluminación para reducir los requisitos técnicos sin sacrificar el detalle y el efecto. Dado el éxito de esta versión, decidimos buscar aún mayor accesibilidad, llevando la experiencia interactiva de "realidad aumentada" a cualquier ordenador o dispositivo móvil a través del buscador de la red: en una proyección en 2D para todos y en RV para los que tienen las gafas. Debido al gran volumen de datos necesario para exponer y anotar 104 cuadros y la gran variedad de dispositivos, se decidió rechazar la opción más obvia, Unity WebGL y usar, en su lugar, un nuevo instrumento A-Frame que se utiliza en juegos interactivos. A través de un código QR, los visitantes que acuden ahora al Museo del Prado podrán utilizar sus dispositivos para proyectar la reconstrucción sobre las paredes actuales de la sala y ver los cambios en la arquitectura y el uso del espacio sin tener que descargar ninguna aplicación adicional (Fig. 3).

Así, la reconstrucción permite reflexionar, a cualquier distancia de Madrid y 140 años después, sobre los criterios de "comparación estética" y las ideas museísticas, estudiar los cuadros y entender los fundamentos intelectuales de la exposición. Por ejemplo, nos hace preguntar si los criterios nacionalistas no formaban parte de la confrontación entre las obras incluso en esta sala, a pesar de haber sido diseñada para ofrecer un paréntesis en el recorrido por un museo ordenado por escuelas nacionales. O nos hace comprender la influencia que ejercían los patrones del ornato de los templos (en el ábside) y los retratos en las casas particulares (en las paredes a los dos lados de la entrada). Esto indica que, en un museo como el Prado, la exposición de obras maestras contribuía al pensamiento nacionalista mientras sugería paralelismos con la esfera pública confesional y, a la vez, la esfera privada.

Esta presentación demuestra que una reconstrucción en 3D a base de datos incompletos puede constituir un proyecto de investigación que no sólo permite cotejar diversas fuentes para producir, refinar y compartir hipótesis, sino también se convierte en una exposición visitable *in situ* y remotamente que, a su vez, genera otras hipótesis y abre nuevas líneas de investigación.

Fig. 1. Fotografía original de Jean Laurent, 1875-77

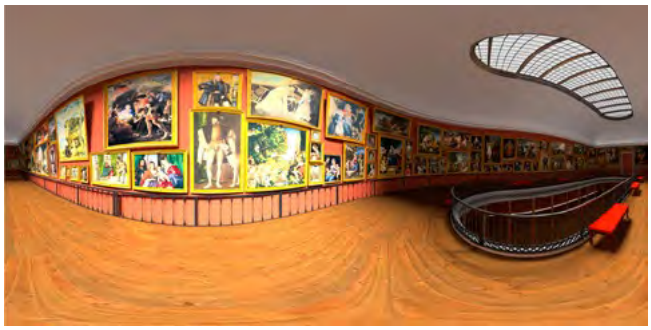


Fig. 2. Rendición de la cueva 3D

Fig. 3. Reconstrucción para cualquier dispositivo

Referencias

Ford, Richard (1878). *A Handbook for Travellers in Spain*. 1845. 5th edition. London, Murray.

Géal, Pierre (2001). "El Salón de la Reina Isabel en el Museo del Prado (1853-1899)." *Boletín del Museo del Prado*, XIX: 37, 143-72.

— (2005). *La Naissance des musées d'art en Espagne (XVIIIe–XIXe siècle)*. Madrid, Casa de Velázquez.

A Digital Edition of Leonhard Euler's Correspondence with Christian Goldbach

Sepideh Alassi
sepideh.alassi@unibas.ch
Digital Humanities Lab, University of Basel, Switzerland

Tobias Schweizer
t.schweizer@unibas.ch
Digital Humanities Lab, University of Basel, Switzerland

Martin Mattmüller
martin.mattmueller@unibas.ch
Bernoulli Euler Center, University of Basel, Switzerland

Lukas Rosenthaler
lukas.rosenthaler@unibas.ch
Digital Humanities Lab, University of Basel, Switzerland

Helmut Harbrecht
helmut.harbrecht@unibas.ch
Bernoulli Euler Center, University of Basel, Switzerland

Introduction

The edition of the works of Leonhard Euler (1707-1783), entitled *Leonhardi Euleri Opera omnia* (LEOO), is a monument of scholarship known to most historians of science. Leonhard Euler's *Opera omnia* consists of 81 volumes, 76 of which have already been published in paper format as four series of books. Volume IV, LEOO IV, of the fourth series contains the correspondence between Leonhard Euler and the German mathematician Christian Goldbach, encompassing 200 letters sent over 35 years(Martin Mattmüller, 2015). The aim of our project is to present this volume to researchers in science and history as a digital edition via the Bernoulli-Euler Online Platform, BEOL(Tobias Schweizer, 2017). BEOL is implemented using Knora(Benjamin Geer, 2017), a generic virtual research environment for the humanities. In this environment, scientists have access to all edited materials of LEOO IV, and can also annotate and edit material in their private workspace and share the results of their research with others. In Knora, the contents of the LEOO IV volume can be represented as a directed graph providing an overview of the network of different entities (letters, persons, bibliographic items, etc.). The tools provided in this environment are intended to facilitate research on the origin of ideas and findings.

Technical steps

LEOO IV consists of two parts: one with transcriptions of the letters in the original languages (Latin and German), and another with English translations of the letters. LaTeXis used to edit both text and mathematical formulas. The volume also contains an index of persons, a bibliography of cited works by Euler, and a general bibliography. The project aims to import all this content into Knora, which represents data as RDF graphs using OWL ontologies(Pascal Hitzler, 2012). Therefore, ontologies are created to describe the structure of the texts and entities of this edition. The data itself must then be converted to XML and imported into Knora.

Specifying the structure of the data

The data model specifying the structure of the data to be imported must be given in the form of OWL ontologies.[1] All bibliographical items, as well as persons in the name index of the edition, are represented internally as RDF triples. For example, every person is represented as an RDF resource belonging to the OWL class beol:Person, which has properties such as beol:hasFamilyName. The property beol:hasIAFIdentifier refers to the IAF/GND dataset maintained by German national library[2], and ensures the uniqueness of each person mentioned in the BEOL platform.

Figure.1 illustrates a part of the generic bibliography ontology, which we have defined to describe all the bibliographical information needed in the BEOL platform (publication types, manuscripts, publishers, etc.). The prefix biblio refers to this ontology, beol refers to the ontology of BEOL-specific entities, and knora-base is the standard Knora ontology, which defines the basic data structures that Knora works with. Ellipses represent types or classes of resources, arrows semantically defined properties attached to them, and rectangles their literal values.

In Knora, a text document (stored in a knora-base:TextValue) can contain markup as well as text. Internally, markup is stored separately from the text, using an RDF-based standoff format[3]. A project such as BEOL defines a mapping between XML and Knora's standoff/RDF markup; texts can then be imported from XML into standoff and exported from standoff back into identical XML[4]. Standoff/RDF markup can contain links to other resources, such as a person or a bibliographical entity mentioned in a text. The Knora API server ensures that the target of the link exists. Standoff links are directed statements, but can easily be queried as incoming links to a given resource.

1 A user interface for designing these ontologies is under development.
2 Integrated Authority File, Deutsche National Bibliothek, http://www.dnb.de/EN/Standardisierung/GND/gnd_node.html
3 Text with Standoff Markup, http://www.knora.org/documentation/manual/rst/knora-ontologies/knora-base.html#text-with-standoff-markup
4 Creating a Custom Mapping, http://www.knora.org/documentation/manual/rst/knora-api-server/api_v1/create-a-mapping.html#creating-a-custom-mapping

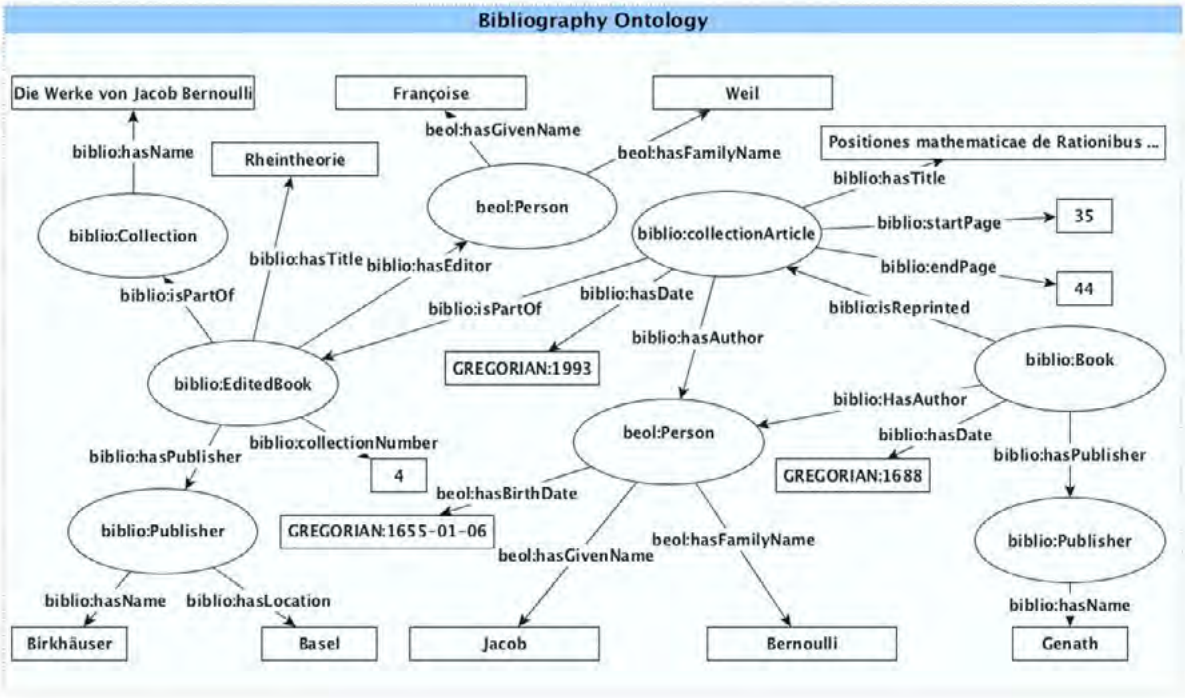


Figure 1. OWL ontology for bibliographical data

We have also defined a data model for letters and their metadata such as author, recipient, date, etc., which provides a network of the correspondence included in the edition. Figure.2 illustrates an excerpt from ontology of the whole LEOO IV project.

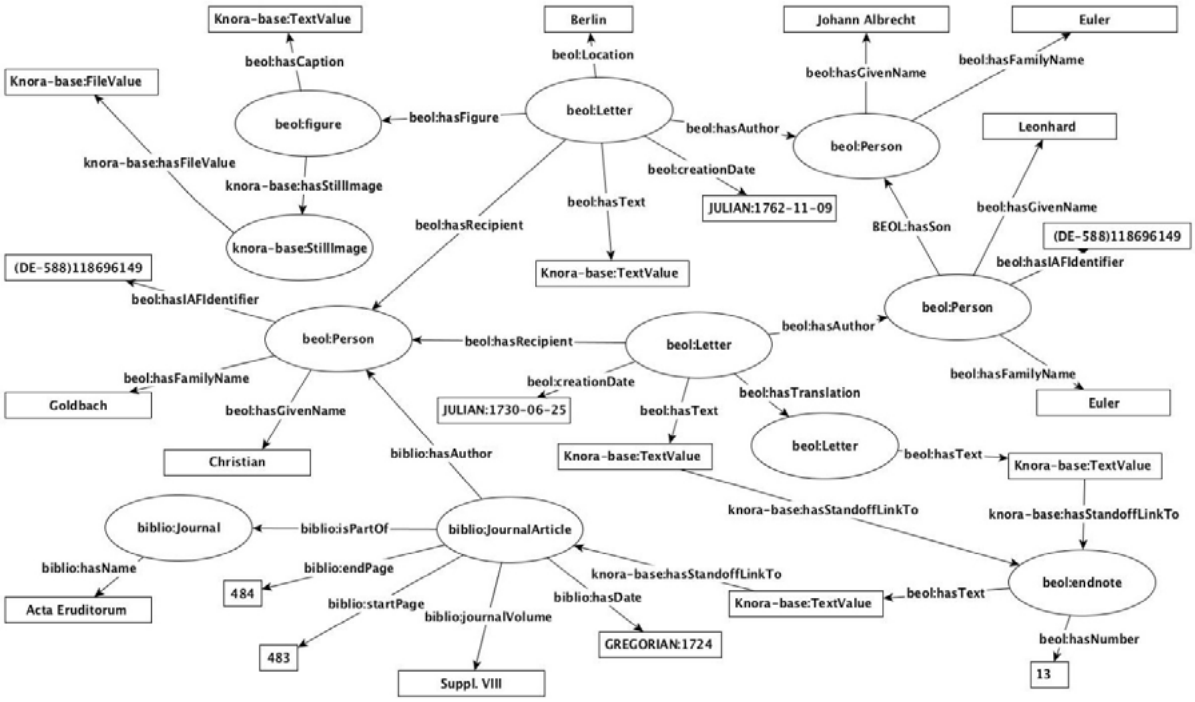


Figure 2. Excerpt from the LEOO IV project data model

Importing data into the BEOL platform

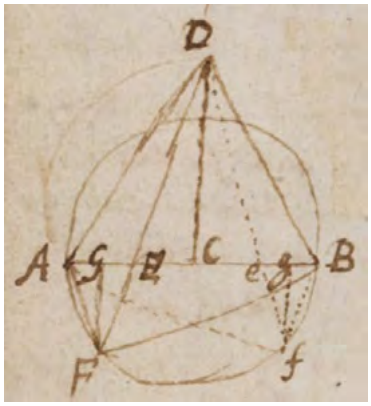
First, the index of persons and the bibliographical items of LEOO IV are written in XML format, using XML schemas that are automatically generated by the Knora API server, based on the ontologies defined for the project. This XML data is then validated against these schemas. After validation, the data can be imported in a single API request (an HTTP POST request to the Knora API server).

Second, the text of the letters is imported using a similar process. Although the text has been transcribed in LaTeX, these transcriptions are first converted to XML to ensure the homogeneity of texts from different editions, and to make it possible to present texts as TEI/XML by applying XSL transformations. The LaTeXML tool(Miller, 2017), with the addition of some BEOL-specific Perl scripts, is used to convert LaTeXto XML. All references to persons and bibliographical items within the text of the letters are replaced with references to the corresponding resources in BEOL, making them queryable via the Knora API. The XML representing the letters is then imported using the same process as for the bibliographical data.

Future work

Since we have developed the methodology for this type of digital edition in a generic way, we expect to be able to integrate all the other recent volumes of Leonhard Euler's *Opera omnia*, which have also been edited using LaTeX. The older volumes in printed form should be scanned, their text should be recognized via OCR, and their structure should be defined with markup.

Most of the older volumes contain figures that are reproduced from scanned letters. We are working on a machine learning algorithm to interpret these figures as well as their labels, so they can be automatically redrawn as vector graphics, see Figure. 3.



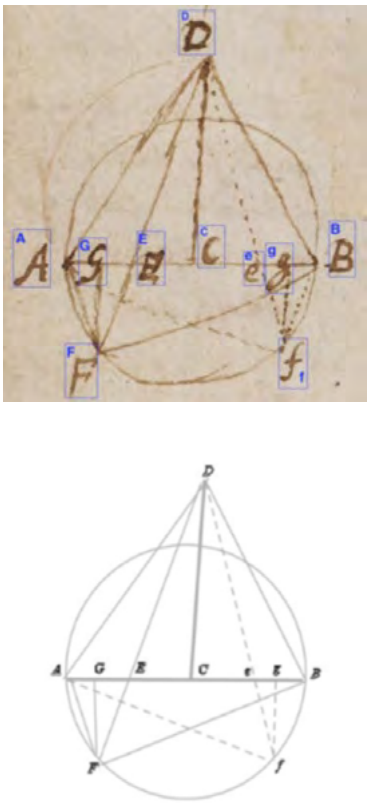


Figure 3. Original figure, detected labels, and reconstructed figure

References

Benjamin Geer, et al (2016). *Knowledge, Organization, Representation, and Annotation*. Digital Humanities Lab http://www.knora.org/.

Martin Mattmüller, F. L. (ed). (2015). *Leonhardi Euleri Opera Omnia: Correspondence of Leonhard Euler with Christian Goldbach*. Vol. IVA/4. Basel.

Miller, B. R. (2017). *LaTeXML: A Latex to Xml/Html/Mathml Converter*. http://dlmf.nist.gov/LaTeXML/.

Pascal Hitzler, et al (2012). *OWL 2 Web Ontology Language Primer (Second Edition)*. W3C http://www.w3.org/TR/2012/REC-owl2-primer-20121211/.

Tobias Schweizer, et al (2017). Integrating historical scientific texts into the Bernoulli-Euler online platform. *Digital Humanities 2017*. https://dh2017.adho.org/abstracts/147/147.pdf.

Bridging the Divide: Supporting Minority and Historic Scripts in Fonts: Problems and Recommendations

Deborah Anderson
dwanders@sonic.net
UC Berkeley, United States of America

Introduction

Today, users of many modern minority and historic scripts in Unicode are not able to reliably send text electronically, because Unicode-enabled fonts and software are not available.[1] In addition, some communities have access to Unicode fonts, but the fonts aren't used, because they do not provide features deemed necessary, such as positioning of characters (e.g., Egyptian Hieroglyphs [Richmond and Glass, 2016]) or variant glyphs (e.g., Old Italic [Anderson, 2017]). Instead, images are used, which are not searchable or, alternatively, "hacked" fonts are employed, which require each person to have the same, non-standard font to send text. Keyboards or other input mechanisms are also not available for many of these same scripts. As a result, the promise that Unicode will "enable people around the world to use computers in any language" (Unicode Consortium, 2018a), does not yet ring true for some communities.

This short paper will highlight font-related problems with specific examples and will provide suggestions on how to address them.

Problems

- Creating a Unicode-enabled font for a language is often not a simple task, especially when the script for the language includes combining marks (which require correct positioning), or if the script has special rendering behavior, such as the consonant clusters found in South Asian scripts (Evans, 2017).
- Font creation is made more challenging when typographic details on the script (and language) are not available. Since many recently approved scripts in Unicode are not well known, information on the typography is not readily available. Unfortunately, fine details are often not included in Unicode proposals for the scripts.
- Interaction with the user community is critical in developing a suitable font, but some communities are difficult to contact. In addition, there can be differing views on the preferred shapes of glyphs. For a set of 51 Tamil numbers and fractions, for example, the community took 8 years to come to agreement on the preferred representative shapes. Specific cases will be cited, based on the author's experience, including discussion of how to connect user communities with font providers.

Technical Issue: Glyph Variants

- For some script users, access to glyph variants is important. This is true, for example, for the Old Italic Unicode block which unified several related alphabets of Italy, dating from approximately the 8 until 1c BCE. In Old Italic, the glyph in a particular alphabet may vary from that shown in the Unicode Standard. The Old Italic block was encoded with the understanding that different fonts would be used for the different languages and alphabets (Unicode Consortium, 2017). How should the two forms of Faliscan (above) be handled in the same font then? How should a pan-Old Italic font handle the different alphabets (which use the same code points)?

This paper will describe the pros and cons of different options available, including use of:

- Code points in Unicode's Private Use Area (with the caveat that these code points would not be reliable for general interchange) (Unicode Consortium, 2018c).
- A Unicode variation sequence, when a distinction needs to be captured in plain-text (Unicode Consortium, 2018d).
- An OpenType font feature, such as character variants, stylistic alternates, stylistic sets, or localized forms (Microsoft Typography, 2018).
- Language-specific fonts (i.e., Faliscan1 and Faliscan2 fonts for the two forms above).

Suggested Solutions

- Incorporate font creation as a part of the overall script encoding effort, such as: including a font item in the budget to pay for a font designer to develop a font; provide information on how to create a font for users; fund a font-creation workshop within the community.
- Encourage user communities to submit a list of the basic repertoire of characters and auxiliary characters to the Common Locale Data Repository (Unicode Consortium, 2018b), since this information is used for by font and software developers worldwide. In addition, provide information on the shapes of the needed letters and variants, citing reference works (i.e., a book or website) on a publicly accessible webpage.
- For handling glyph variants, short-term and long-terms approaches should be considered:
  - If a given variant is deemed by users to be necessary in plain-text, submit a Unicode proposal
  - If OpenType features are used in a font, lobby software vendors to provide better support for the features in applications (as support for some features is still spotty [4])
  - For the short-term, PUA or separate fonts may be necessary.

For font designers:

- Use language tags from ISO 639 (SIL International, 2017), BCP 47 (Phillips and Davis, 2009), and

1 Especially true for scripts in Unicode versions 6.0 to 9.0 (2010 – 2016), where over 40% of the scripts have no fonts. (Unicode version 10.0 was released in June 2017, so support in fonts would not yet be expected). The Google Noto project aims to provide fonts for all approved scripts, but release of fonts is only up to fonts for Unicode version 6.2, released in 2012.

OpenType language/script tags (Microsoft Typography, 2017a; Microsoft Typography, 2017b) in the font internals. If a language (or script) is missing a tag, a new tag should be registered. According to Roozbeh Pournader, an expert at implementation of fonts, these tags are the way the fonts communicate with other software today.

- Encourage users to review the glyphs in alpha versions of any forthcoming or any released Noto fonts, and submit comments to the Noto project (Google.com, n.d.).

Conclusion

Access to a Unicode font is critical for users of lesser-used scripts, in order to participate more fully in the digital world. Unicode fonts make the user's text interchangeable, discoverable, and able to be preserved for the long-term in a stable format. Recognition of font-related issues is a small step towards addressing the problem. Input from the audience will be encouraged in order to identify other potential approaches.

Funding

This work was supported by the National Endowment for the Humanities [grant number PR-253360-17].

References

Anderson, D. (2017). Dealing with Variants in Historic Scripts. Presentation at *41st Internationalization and Unicode Conference*, Santa Clara, California, October, 2017.

Evans, L. (2017). Beyond Unicode Proposals: Encoding Characters and Scripts is Not Enough! Presentation at *41st Internationalization and Unicode Conference*, Santa Clara, California, October 2017.

Google.com. (n.d.). *Google Noto Fonts*. https://www.google.com/get/noto/ (accessed April 17, 2018).

Microsoft Typography. (2017a). *Language system tags*. https://www.microsoft.com/typography/otspec/languagetags.htm (accessed April 17, 2018).

Microsoft Typography. (2017b). *Script tags*. https://www.microsoft.com/typography/otspec/scripttags.htm (accessed April 17, 2018).

Microsoft Typography. (2018). *OpenType® specification*. https://www.microsoft.com/en-us/Typography/OpenTypeSpecification.aspx (accessed April 17, 2018).

Phillips, A., and Davis, M. (2009). *Tags for Identifying Languages*. https://tools.ietf.org/html/bcp47 (accessed April 17, 2018).

Richmond, B. and Glass, A. (2016). *Proposal to encode three control characters for Egyptian Hieroglyphs. Proposal submitted to the Unicode Technical Committee*. http://www.unicode.org/L2/L2016/16018r-three-for-egyptian.pdf (accessed April 17, 2018).

SIL International. (2017). *ISO 639-3: ISO 639 Code Tables*. http://www-01.sil.org/iso639-3/codes.asp (accessed April 17, 2018).

Unicode Consortium. (2017). Old Italic. In: Unicode Consortium, *The Unicode Standard, Version 10.0.0*. Mountain View, CA: The Unicode Consortium, 349-351. http://www.unicode.org/versions/Unicode10.0.0/ (accessed 24 April 2018).

Unicode Consortium. (2018a). *The Unicode Consortium website*. http://unicode.org/ [accessed April 17, 2018).

Unicode Consortium. (2018b). *CLDR - Unicode Common Locale Data Repository. Unicode Consortium website*. http://cldr.unicode.org (accessed April 17, 2018).

Unicode Consortium. (2018c). *Private-Use Characters, Noncharacters & Sentinels FAQ. Unicode Consortium website*. http://www.unicode.org/faq/private_use.html (accessed April 17, 2018).

Unicode Consortium. (2018d). *Variation Sequences. Unicode Consortium website*. http://www.unicode.org/faq/vs.html (accessed April 17, 2018).

Unwrapping Codework: Towards an Ethnography of Coding in the Humanities

Smiljana Antonijevic Ubois
smiljana@smiljana.org
The Pennsylvania State University, United States of America

Joris van Zundert
joris.van.zundert@huygens.knaw.nl
Royal Netherlands Academy of Arts and Sciences,
The Netherlands

Tara Andrews
tara.andrews@univie.ac.at
University of Vienna, Austria

Code and codework share many properties with text and writing, and code can be seen as an argument, corresponding to Galey and Ruecker's (2010) understanding of the epistemological status of graphical user interfaces as argument. From an epistemic point of view, the practice of a programmer is no different from the practice of a scholar when it comes to writing (Van Zundert, 2016). Both are creating theories about existing epistemic objects (e.g. text and material artifacts, or data) by developing new epistemic objects (e.g. journal articles and critical editions, or code) to formulate and support these theories. However, as expressions of a technē whose inner workings are opaque to most humanities scholars, code and codework are all too often treated as an invisible hand, influencing humanities research in ways that are not transparent. The software used in research is treated as a black box in the

sense of information science—expected to produce a certain output given a certain input—but at the same time often mistrusted precisely for this lack of transparency.

The digital humanities (DH) does not generally engage with the code and coding parts of programming in an explicit and critical manner, which is necessary for opening up black boxes of code. The invisibility and un-critiqued use of code means that the scholarly quality and contribution of codework goes both uncredited and unaccounted for. Black-boxing the code results in neglect of its epistemological contributions and imperils one of the key components of knowledge production in the DH. Much more insight into code and codework in the humanities is needed, including how coders approach their tasks, what decisions go into its production, and how code interacts with its environment.

The purpose of this paper is to provide some of those insights in the form of an ethnography of codework, wherein we observe the decisions that programmers make and how they understand their own activities. Our study follows in the footsteps of ethnographies of technoscientific practice (see: Forsythe, 2001; Coleman, 2013), Critical Code Studies (see: Marino, 2014), and reflections on coding and tool development in the DH (see: Schreibman and Hanlon, 2010; Ramsey and Rockwell, 2012). The study does not aspire to be representative of the DH coding practice, but to initiate a debate about some still overlooked elements of that practice.

This exploration applies Latour's (1998) first rule of method to the context of narrative creation through codework, looking at the practices, dilemmas, and decisions of programmers. To do that, we use analytical autoethnography (cf. Anderson, 2006), combined with collaborative ethnography (cf. Lassiter, 2005). In our methodological design, the team ethnographer first formulated a set of ten questions aimed at generating reflexive accounts and examples of DH coding in the making. Each of the team DH programmers then individually answered the questions in a written form, providing elaborate, semi-formal accounts of his or her DH programming practice. Thus generated written accounts became the basis for a series of team discussion, both written and oral, which eventually formed the results of this contribution. This methodological design enabled us to return from the final outputs of DH coding to scholarly uncertainties and resolutions that preceded them. Through such reconstruction, we were able to document some of the key phases in epistemological construction of coding artifacts, and to identify methodologically significant moments in stabilization of those artifacts. In other words, we relied on the experiences of scholars proficient in both humanities research and coding seeking to make explicit what DH coders themselves know, maybe tacitly, about why and how they code.

We have grouped our observations into the categories known as the five canons of rhetoric, proposed in Cicero's *De Inventione*. Although originally developed for public speaking, these canons have proven to be equally potent heuristic in analyzing written and, more recently, digital discourse (Gurak and Antonijevic, 2009). Our contribution sought to extend this heuristic to the analysis of coding as argumentation, not in an attempt to fit codework and its elements into a pre-defined ontology, nor to suggest that it fully conforms or matches classical rhetoric. Rather, it was a way of presenting our experiences and claims in a form that we expected to facilitate interpretation by scholars well versed in text production but likely less so in codework.

Our study showed that codework reflects humanistic discovery (*inventio*) in that humanities- specific research questions drive coding, and tasks specific to the humanities research motivate software development. Similarly, crafting and organizing code resonates with development and arrangement of a scholarly argument (*dispositio*)—a programmer writes lines of code and makes many decisions on how to arrange these pieces into larger, meaningful constructs that influence the epistemological and methodological structure of research. Our study also illustrated that, like any humanities scholar, an author of software has her own style (*elocutio*) in the aesthetics of code and in her way of working to create code, and this style develops through both individual and norms of coding communities. We also showed that, parallel to books or libraries, code and codework serve as memory systems (*memoria*) that embed theoretical concepts in order to augment research methodology and create new theory, where code can be regarded as a performative application or explanation of theory. Finally, our ethnography illustrated how codework *actio* compares to the publication and reception of the software, where DH programming is still not recognized as a locus of humanities expertise, and it is hard for humanities programmers to have their code academically evaluated as digital output.

The insights of our study demonstrate that both code as an epistemic object and coding as an epistemic practice increasingly shape research in the humanities and must be given a proper theoretical and methodological recognition in the DH, with both the consequences and the rewards that such a recognition bears. Therefore, a strategy for making code and codework visible, understandable, trustworthy and reputable within humanities scholarship is needed. Such a strategy should be comprehensive, both in the sense of accounting for the source code and the executed result of software. While we agree with Ramsay and Rockwell (2012) that providing source code is not sufficient for understanding the underlying theoretical assumptions, we disagree in viewing the "dependence on discourse" as a feature that relativises epistemic and communicative capacities of code and codework. We argue in contrast that interdependence of code and text should be embraced as a means of acknowledging their distinctive yet corresponding methods of knowledge production and communication. Just as code enhances text making it amenable to methodological and epistemological approaches of DH, text enhances

code making it more visible and intelligible for the humanities community. Evaluating code and DH programming in a disengaged way would thus be similar to the literary criticism enacted on a novel without reading it. Yet currently it is practice to "criticize" software and code based only on a journal article that derived from it. As much as possible, coders should support the involved evaluation of code as opposed to its disengaged criticism. We believe that theoretical discussions of codework should become an established trajectory in the humanities, along with the development of methods for documenting, analyzing, and evaluating code and codework.

One important element of that strategy is understanding codework as necessarily shaped by its social context, which influences the attitude and perception that both coders and other scholars hold towards their work. Too often, DH programmers are treated as service instead of research focused scholars, which results in a number of negative consequences. A necessary step in the direction of a real change in how codework is received into the humanities is recognition and reward for peer-reviewed digital outputs, including code, as research outputs (cf. Nowviskie, 2011; Presner, 2012; American Historical Association, 2015). A precondition for this is to start grassroots procedures for peer review of code (Fitzpatrick, 2011), and to regard the code as alternative expressions of research or epistemologies with equal research value and validity instead of subordinating code and codework to 'humanities proper' (cf. Burgess and Hamming, 2011 and Ramsay and Rockwell, 2012). There is a need for peer review and critical examination of actual code, which is hardly even present in DH (Zundert and Haentjens Dekker, 2017). Also, open publishing of code in verifiable ways can be easily facilitated through existing public code repositories or institutionally-run versions of the same repositories, but it is not common practice throughout the humanities to publish code. Finally, reflexive accounts on (digital) humanities codework and ethnographic studies of actual work can help us understand how code and codework are changing the humanities (Borgman, 2009). We believe that an important step in illuminating the process and results of DH programmers' codework is to develop and explicate reflexive insights into its key epistemological, methodological, and technical aspects. Explaining, for instance, what kind of research questions give impetus to one's codework and how new research insights co-evolve during code development can help both DH programmers and their traditionally trained colleagues recognize the important epistemological connections between humanistic theory and scholarly programming.

References

American Historical Association, A. H. C. on P. E. of D. S. by H. (2015). *Guidelines for the Professional Evaluation of Digital Scholarship in History*. Draft http://bit.ly/1PC1tDL (accessed 8 November 2017).

Anderson, L. (2006). Analytic Autoethnography. *Journal of Contemporary Ethnography*, 35(4): 373–95 doi:10.1177/0891241605280449.

Borgman, C. (2009). The Digital Future is Now: A Call to Action for the Humanities. *DHQ: Digital Humanities Quarterly*, 3(4) www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html.

Burgess, H. J. and Hamming, J. (2011). New Media in Academy: Labor and the Production of Knowledge in Scholarly Multimedia. *DHQ: Digital Humanities Quarterly*, 5(3) http://digitalhumanities.org/dhq/vol/5/3/000102/000102.html (accessed 2 September 2016).

Coleman, E. G. (2013). *Coding Freedom: The Ethics and Aesthetics of Hacking*. Princeton (US), Woodstock (UK): Princeton University Press http://gabriella-coleman.org/Coleman-Coding-Freedom.pdf (accessed 8 November 2017).

Fitzpatrick, K. (2011). Peer Review, Judgment, and Reading. *Profession*(6): 196–201 doi:-prof.2011.2011.1.196.

Forsythe, D. and Hess, D. J. (2001). *Studying Those Who Study Us: An Anthropologist in the World of Artificial Intelligence*. Stanford, CA: Stanford University Press.

Galey, A. and Ruecker, S. (2010). How a prototype argues. *Literary and Linguistic Computing*, 25(4): 405–424 doi:10.1093/llc/fqq021.

Gurak, L. and Antonijevic, S. (2009). Digital Rhetoric and Public Discourse. In Lunsford, A. A., Eberly, R. A. and Wilson, K. H. (eds), *The Sage Handbook of Rhetorical Studies*. London, Thousand Oaks: SAGE Publications, Inc., pp. 497–508.

Lassiter, L. E. (2005). *The Chicago Guide to Collaborative Ethnography*. (Chicago Guides to Writing, Edi). Chicago, London: University of Chicago Press http://bit.ly/2iLCmGY.

Latour, B. (1988). *Science in Action: How to Follow Scientists and Engineers Through Society*. Cambridge, MA, USA: Harvard University Press.

Marino, M. C. (2014). Field Report for Critical Code Studies, 2014. *Computational Culture—A Journal of Software Studies*(4) http://computationalculture.net/article/field-report-for-critical-code-studies-2014%E2%80%A8 (accessed 10 June 2015).

Nowviskie, B. (2011). Where Credit Is Due: Preconditions for the Evaluation of Collaborative Digital Scholarship. *Profession*(6): 169–181 doi:-prof.2011.2011.1.169.

Presner, T. (2012). How to Evaluate Digital Scholarship. *Journal of Digital Humanities*, 1(4) http://journalofdigitalhumanities.org/1-4/how-to-evaluate-digital-scholarship-by-todd-presner/.

Ramsay, S. and Rockwell, G. (2012). Developing Things: Notes toward an Epistemology of Building in the Digital Humanities. In Gold, M. K. (ed), *Debates in the Digital Humanities*. Minneapolis: University of Minnesota Press, pp. 75–84 http://dhdebates.gc.cuny.edu/debates/text/11.

Schreibman, S. and Hanlon, A. M. (2010). Determining Value for Digital Humanities Tools: Report on a Sur-

vey of Tool Developers. *DHQ: Digital Humanities Quarterly*, 4(2) http://digitalhumanities.org/dhq/vol/4/2/000083/000083.html (accessed 9 November 2017).
Zundert, J. J. van (2016). Author, Editor, Engineer — Code & the Rewriting of Authorship in Scholarly Editing. *Interdisciplinary Science Reviews*, 40(4): 349–375 doi:http://dx.doi.org/10.1080/03080188.2016.1165453.
Zundert, J. J. van and Haentjens Dekker, R. (2017). Code, Scholarship, and Criticism: When is Coding Scholarship and When is it Not? *Digital Scholarship in the Humanities*, 32(Suppl_1): i121–i123 doi:https://doi.org/10.1093/llc/fqx006.

Conexiones Digitales Afrolatinoamericanas. El Análisis Digital de la Colección Manuel Zapata Olivella

Eduard Arriaga
earriaga@alumni.uwo.ca
University of Indianapolis, United States of America

Las manifestaciones afrolatinoamericanas y sus conexiones con el mundo digital han comenzado a generar un creciente interés en diversos campos de estudio: las humanidades digitales, los estudios culturales, literarios y antropológicos entre otros. A pesar del interés, el estudio de tal intersección se encuentra en una etapa inicial debido a factores como a) las limitaciones de acceso a herramientas digitales por parte de algunos agentes y comunidades identificadas y auto-identificadas como afrolatinoamericanas/afrolatinas; b) limitaciones en la consecución de derechos de autor de algunas piezas y manifestaciones cuya distribución e intercambio digital se hace más difícil; y c) falta de innovación en la forma de clasificar piezas y manifestaciones que, en muchos casos, no coinciden con la tradición letrada que subyace al proceso de archivo ya sea digital o no. Tales limitaciones han hecho más difícil la consolidación de propuestas analíticas que, desde las humanidades digitales, den cuenta del estado y evolución de las culturas afrolatinoamericanas, así como de sus aportes a nivel de conocimiento en espacios locales, regionales y globales.

Algunas formas de revertir dichas limitaciones ha sido el desarrollo de iniciativas y colecciones digitales por parte las mismas comunidades afrolatinoamericanas en cooperación con entidades académicas, agencias multilaterales, gubernamentales, intergubernamentales y no gubernamentales. Tales iniciativas muestran la diversidad de manifestaciones generadas desde dichas comunidades; manifestaciones que son fundamentales para su identificación, visibilización y, sobre todo, consideración dentro de un modelo de justicia social que, como el contemporáneo, se centra en el reconocimiento de los derechos humanos. Asimismo, dichas adaptaciones tecnológicas se convierten en una forma de lo que Steve E. Jones determina como 'eversion" (Jones, 2016) o la consolidación de unas realidades híbridas entre lo digital, lo análogo y lo performático. Algunos de los proyectos más importantes en este ámbito son, entre otros, Digital Portobelo, Mueseu Afro Digital Río de Janeiro o Proyecto Afrolatin@, a partir de los cuales se hacen evidentes diversas formas de ser afrolatinoamericano, así como diversas formas de representación y expresión de sujetos cuya identificación intersecta varios espacios discursivos, políticos y de acción. Algunos de los puntos positivos de dichas plataformas y colecciones es que a) son espacios en constante construcción –actuales y constantemente actualizados- y b) permiten ver procesos de acceso, creatividad, justicia simbólico-social que las comunidades están persiguiendo y han perseguido por largo tiempo. Sin embargo, el carácter de construcción constante de dichas plataformas es, al mismo tiempo, un aspecto negativo dado que el flujo de información se convierte en un desafío para unas humanidades digitales cuyo modelo se ha centrado en la digitalización y análisis de información canónica, única, extraordinaria (Manovich, 2016). Las plataformas generadas por parte de esas comunidades afrolatinoamericanas, por el contrario, registran el flujo de la cultura en el presente que no ha sido propiamente abordado por las humanidades ya sean análogas o digitales. En el caso de la intersección entre estudios afrolatinoamericanos y estudios digitales, el proceso de análisis ha estado mucho más rezagado no solo por la falta de bases de datos o de construcción de archivos digitales, sino por la falta de interés y apoyo para construirlos y, a partir de allí, desarrollar metodologías innovadoras de análisis (Gomez, 2011).

De acuerdo con el panorama descrito, esta presentación corta dará cuenta del proceso de investigación e implementación metodológica llevado a cabo a partir de *Manuel Zapata Olivella Collections*, una colección digital desarrollada por la biblioteca de la Universidad de Vanderbilt. Manuel Zapata Olivella fue uno de los escritores y activistas afrolatinoamericanos más importantes del siglo XX, cuya obra y pensamiento han influido al movimiento afrolatinoamericano contemporáneo. Sus cartas, manuscritos y documentación personal como escritor, artista y activista habían quedado en un archivo personal manejado por su familia. Sólo hasta el 2008 la Universidad de Vanderbilt adquirió el fondo y desarrolló una colección digital en el cual se hacen visibles varios de sus documentos y proyectos tanto etnológicos como antropológicos. Entre los archivos digitalizados se encuentran los documentos –cartas, panfletos, memorias, comunicaciones personales, fotografías y audios- del *Primer Congreso de Cultura Negra de las Américas*, realizado en Colombia en 1978. El proyecto, llevado a cabo con apoyo de la Universidad de Indianápolis, consistió en el análisis digital de dicha documentación y del Congreso como uno de los nodos centrales de la acción política, literaria y cultural afrolatinoamericanas del siglo XX y XXI. El proyecto buscaba a) responder preguntas tales

como: ¿Cuáles fueron las redes artísticas y textuales que permitieron la emergencia del Congreso?, ¿Cuáles fueron los discursos socio-culturales latinoamericanos con los cuales el congreso desarrolló un diálogo y logró establecer su propio conjunto de valores y códigos para explicar lo afrolatinoamericano?, ¿Cuáles de los valores políticos y estrategias estéticas creadas y adoptadas por el Congreso devinieron patrones de acción y fueron transmitidas al movimiento afrloatinoamericano de la era digital?. Asimismo, el proyecto buscaba b) desarrollar propuestas metodológicas digitales para comenzar a entender la complejidad e interconexión –en tiempo y espacio- del movimiento afrolatinoamericano. Esta última actividad se desarrolló a través de la implementación de mapas de tópicos y el uso de plataformas digitales para visualizar la información de forma inter-relacional –Vg. Scalar, Wandora, Gephi, etc.-, considerando la diversidad de materiales en el ecosistema informativo de la tradición afrolatinoamericana.

La presentación entonces mostrará los resultados de esa investigación a través del mapeo de textos, de agentes, instituciones y sistemas de valores relacionados para, finalmente, conectarlo con las propuestas ideológicas fundamentales del movimiento afrolatinoamericano surgido de la Conferencia Mundial Contra el Racismo realizada en Durbán en 2001. A través de esta presentación se discutirán no solamente los hallazgos de la investigación en particular sino, sobre todo, las perspectiva de unas humanidades digitales afrolatinoamericanas que, aunque se incluyan en las discusiones regionales (RedHD, Humanidades digitales en Latinoamérica) intentan ir más allá, en busca de la conexión entre activismo e investigación académica con un objetivo claro: la justicia social y la descolonización del conocimiento.

References

Gómez F. P. (2011). La colección Manuel Zapata Olivella. *Revista de estudios colombianos*, 37-38: 117-118.

Jones E. S. (2016). The Emergence of the Digital Humanities. *Debates in the Digital Humanities,* University of Minnesota Press. http://dhdebates.gc.cuny.edu/debates/text/52

Manovich, L. (2016). The Science of Culture? Social Computing, Digital Humanities and Cultural Analytics. *Journal of Cultural Analytics*. Doi: *10.22148/16.004*

Dal Digital Cultural Heritage alla Digital Culture. Evoluzioni nelle Digital Humanities

Nicola Barbuti
nicola.barbuti@uniba.it
Dipartimento di Studi Umanistici DISUM - Università degli Studi di Bari Aldo Moro, Italy

Ludovica Marinucci
lud.marinucci@gmail.com
Scuola a Rete per la Formazione nel Digital Cultural Heritage, Arts and Humanities - DiCultHer, Italy

Introduction

Digital has transformed the way to produce, transmit and share knowledge. The increasingly widespread diffusion of digital methods and techniques in all the social and cultural levels of the communities, in fact, brings an unheard democratization of knowledge and culture, making the citizen a privileged and intelligent actor in the sustainable development of the new smart societies which are based on the process of digitization, digital co-creation and digital design.

The **art. 2** of the UE "**Council conclusions of 21 May 2014 on cultural heritage as a strategic resource for a sustainable Europe" (2014/C 183/08)** states: "Cultural heritage consists of the resources inherited from the past in all forms and aspects - tangible, intangible and digital (born digital and digitized), including monuments, sites, landscapes, skills, practices, knowledge and expressions of human creativity, as well as collections conserved and managed by public and private bodies such as museums, libraries and archives. It originates from the interaction between people and places through time and it is constantly evolving. These resources are of great value to society from a cultural, environmental, social and economic point of view and thus their sustainable management constitutes a strategic choice for the 21st century".

It is therefore inevitable to rethink digital and digitization as social and cultural expressions of the contemporary age. This implies the need to rethink data as cultural entities and no longer as mere tools for simplifying administration management, or as extemporary surrogates for enhancing the fruition of tangible and intangible cultural heritage.

The current process for archiving and storing data, although they generate from the awareness of the need to preserve them, don't solve the problem of their both current and historical reuse, because they are still strongly conditioned by the instrumental function that presides over their production and use.

Towards a first classification of Digital Culture

This paper aims to provide a new definition of methodological and technological approach to digital and digitization, with the goal to guarantee data stability, sustainability, usability and reusability so as to foster their long term preservation.

The research originates from observing that, in the human evolution, the survival, preservation and permanence over time of any entity has always been strictly linked to its identification as cultural heritage, because of its value of historical witness which conveys knowledge.

For several years, authoritative scientific voices have highlighted how long term digital preservation is the real emergency to be faced worldwide. In 2015, Vinton Cerf raised the alarm about the risk that the Twenty-First Century will become for posterity the first black hole in human evolution since the establishment of intelligent communication. The alarm resumed what was debated in the 2012 UNESCO Conference held in Vancouver with the significant title "The Memory of the World in the Digital Age: Digitization and Preservation".

In order to start a serious and shared process for cultural identification of digital and digitization, it is therefore essential to recognize data as cultural entities, defining a clear and regulated position in the contemporary cultural scene. In fact, several existing digital entities could be considered contemporary **Digital Cultural Heritage (DCH)**, expression of the **Digital Culture** of the Twenty-First Century smart societies.

A first useful identification could come out from a classification of **digital cultural entities**, which can be traced back to the following three basic categories in which the Digital Culture could be declined:

- **Digital FOR Cultural Heritage**: process, methods and techniques aimed at co-creation of digital artifacts reproducing in their contents tangible and intangible cultural heritage: e.g., digital objects, digital libraries, virtual museums, demo-etno-anthropological databases.
- **Digital AS Cultural Heritage**: approach, process, methods and techniques aimed at recognizing and preserving both digital artifacts reproducing intangible and tangible cultural heritage, and dematerialization as expression of contemporary cultural *facies* to be known, safeguarded, preserved and transferred in time as witness and memory of the current **Digital Age**.
- **Born Digital Heritage**: process, methods and techniques aimed at co-creating and managing digital entities that record the current activities of contemporary communities, to be safeguarded, preserved and transferred to future generations as witness and memory of Twenty-First Century culture and societies.

Digital Culture as identity of contemporary age

According to the above classification, Digital Culture could therefore be defined as implementation of integrated cultural and training approaches, processes, methods, and techniques aimed at co-creating an ecosystem of aware digital knowledge. This, in fact, will be enabled to trigger processes for the construction of networks to safeguard, preserve, sustain, transfer, reuse **DCH** through awareness of its identity as historical memory of the contemporary age and, therefore, as source of knowledge for future generations.

So, starting from the analysis and co-design of a digital entity, whatever it is – one digital artifact, a digital library, a management system for Public Administrations or an app for Augmented Reality –, the focus on preservation is primary to define it a digital cultural entity. It will determine and regulate both the co-creation process, and the methodological and technological approaches, systems, information, metadata schema, digital image content structures, data description, complex data set, and their any further development and sustainability. This approach can only exist in an ecosystem of aware digital culture, in which digital and digitization with their processes are recognized as DCH.

In this regard, our opinion is that what differentiates DCH from the non-cultural digital artifacts are the descriptive metadata for indexing digital object. Above all, it is primary the correct proportion between:

- **quantity**: it is the correct ratio among exhaustiveness of information, knowledge to be provided, number of metadata elements and attributes necessary to retrieve, reuse and store it;
- **quality**: it is the correct ratio among choice of the informative and cognitive level to be given both to each descriptor and to set of descriptors, and the variables of information and cognitive need of the users, according to whether they are current or future.

Descriptive metadata as sources of Digital Cultural Heritage

The issue is addressed with regard to the preservation of **Digital AS Cultural Heritage**. The case study object of the research is the metadata schema co-created for the digitization project "Historical Archive of the G. Laterza & Figli Publishing House", undertaken at the end of 2015 and today publishing in the Puglia Digital Library of the Puglia Region.

The metadata schema used for managing and indexing the digital artifacts scanned from the original documents has been co-created with reference to the Italian national METS-SAN standard structured by the National Archival System.

The preservation of both the process of digital co-creation and of the digital resources themselves has been the focus of the project. So, attention has been focused on descriptive metadata both of the project as a whole, and of each section of the original Archive (series, sub-series, etc.), and of each one digital artifact. The starting point was the awareness that, at the state of the art, the images present great difficulty for long term digital preservation. The planning and structuring of the metadata schema has therefore been focused not only on the needs of contemporary users, but above all on the cognitive and informative needs of future users about our

contemporary culture. So, metadata will be the only sources of knowledge on both the digital artifacts we produce today, and the processes by which we co-create them.

We preferred to use "granular" indexing, describing each digital document with its metadata.

In structuring the metadata schema, we considered the tag sequence as an organic structure composed of forms entities (elements and attributes) and descriptive information. The narrative contents have been articulated hybridizing methods and techniques of archival description with cataloguing solutions, and they have been written with stylistic criteria deduced from the storytelling methodology, providing information on both the whole project and the detail of each section and, inside the sections, of each partition.

In each metadata, the <header> section, after the namespaces (<xlmns: --->) enbeds the descriptors related to:

- project: body responsible for the project, owner of original Archive, editor of digital resources;
- history of the original Archive;
- structure of the original Archive;
- historical/biographical profile of the owner of the original Archive;
- rights that regulate the use of original documents.

The <desc> section has been divided into two sub-sections:

1. context: it enbeds the data relating to entities involved in the ownership and management of the original documents;
2. description: it describes the consistency of the subfund to which the resource described in the sub-section <File> belongs.

The <File> section dedicated to single document describes:

- the original document represented in the image: subject, text abstract, creator, contributors, chronic date, topical date, support, language;
- the physical position of the original in the Archive;
- the editor who creates the descriptions.

The section on rights follows, which describes:

- ownership of the digital artifact;
- accessibility and reuse of the digital artifact;
- ownership and accessibility of the original document.

The schema closes with the technical metadata describing the different image formats in which each digital objects relating to the respective pages of a document have been reproduced, with their structural components.

Conclusion

Starting from the art. 2 of the UE "Council conclusions of 21 May 2014 on cultural heritage as a strategic resource for a sustainable Europe" (2014/C 183/08), the paper focuses on the need to rethink digital and digitization process for long term digital preservation, aiming to redefine them as the new Cultural Heritage of the contemporary era.

This new way to observe digital artifacts and their co-creation process is the indispensable prerequisite for co-creating aware Digital Culture and for giving due importance to digitization and dematerialisation, whose process, from the planning stages, need an approach focused on data preservation and, to this goal, on the decisive role that the descriptive metadata play.

The case study was the digitization project of the "Historical Archive of the Giuseppe Laterza & Figli Publishing House". In particular, the attention to preservation focused on structuring the schema of metadata and, above all, on descriptive writing, with regard to the choice of tags, elements and attributes, and to draft descriptive information of each digital artefact. In fact, our opinion is that they constitute the main source for the knowledge of both the single digital artifact, and the full project and its evolution, thus configuring itself as fundamental elements to validate and certify the data, guaranteeing quality, authenticity and sustainability as witness and memory of the contemporary Digital Age, with the aim to increase the knowledge of future generations about Twenty-First Century.

References

https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52014XG0614%2808%29

http://www.interpares.org/

http://www.pugliadigitallibrary.it/

Agenzia per l'Italia Digitale (AgID), Presidenza del Consiglio dei Ministri, *Linee guide sulla conservazione dei documenti informatici*,, Versione 1.0 – dicembre 2015, pp. 45 ss. (http://www.agid.gov.it/sites/default/files/linee_guida/la_conservazione_dei_documenti_informatici_rev_def.pdf).

L. Bailey, *Digital Orphans: The Massive Cultural Black Hole On Our Horizon*, Techdirt, Oct 13th 2015 (https://www.techdirt.com/articles/20151009/17031332490/digitalorphans-massive-cultural-blackhole-our-horizon.shtml).

S. Cosimi, *Vint Cerf: ci aspetta un deserto digitale*, Wired.it, 16 febbraio 2015 (http://www.wired.it/attualita/2015/02/16/vint-cerf-futuro-medievale-bit-putrefatti/).

T. Di Noia, A. Ragone, A. Maurino, M. Mongiello, M. P. Marzocca, G. Cultrera, M. P. Bruno, *Linking data in digital libraries: the case of Puglia Digital Library*, in A. Adamou, E. Daga, L. Isaksen, "Proceedings of the 1st Workshop on Humanities in the Semantic Web,

co-located with 13th ESWC Conference 2016 (ESWC 2016)", Anissaras, Greece,May 29th, 2016 (http://ceur-ws.org/Vol-1608/paper-05.pdf).
L. Duranti, E. Shaffer (ed. by), *The Memory of the World in the Digital Age: Digitization and Preservation. An international conference on permanent access to digital documentary heritage*, UNESCO Conference Proceedings, 26-28 September 2012, Vancouver (http://ciscra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf)
V. Gambetta, *La conservazione della memoria digitale*, [Rubano], Siav, 2009.
P. Ghosh, *Google's Vint Cerf warns of 'digital Dark Age'*, BBC News, Science & Environment, 13 February 2016 (http://www.bbc.com/news/science-environment-31450389).
M. Guercio, *Gli archivi come depositi di memorie digitali*, "Digitalia", Anno III, n. 2, ICCU Roma, 2008, pp. 37-53.
M. Guercio, *Conservare il digitale. Principi, metodi e procedure per la conservazione a lungo termine di documenti digitali*, Roma-Bari, Laterza, ed. 2013.
M. Guercio, *Conservazione delle e-mail: le raccomandazioni del progetto InterPares* (http://www.conservazionedigitale.org/wp/wp-content/uploads/2014/12/Guercio-8-Conservare-documenti-email.pdf)
Joint Steering Committee for Development of RDA, *Resource Description and Access (RDA)* (http://www.iccu.sbn.it/opencms/export/sites/iccu/documenti/2015/RDA_Traduzione_ICCU_5_Novembre_REV.pdf)
W. Kool, B. Lavoie, T. van der Werf, *Preservation Health Check: Monitoring Threats to Digital Repository Content*, OCLC Research, Dublin (Ohio), 2014 (http://www.oclc.org/content/dam/research/publications/library/2014/oclcresearch-preservation-health-check-2014.pdf).
B. Lavoie, R. Gartner, *Preservation Metadata (2nd edition), DPC Technology Watch Report*, 03 May 2013, DPC Technology Watch Series (http://www.dpconline.org/docman/technology-watch-reports/894-dpctw13-03/file).
Library of Congress, *PREMIS – Preservation Metadata: Implementation Strategies*, v. 3.0 (http://www.loc.gov/standards/premis/v3/premis-3-0-final.pdf)
G. Marzano, *Conservare il digitale. Metodi, norme, tecnologie*, Milano, Editrice Bibliografica, 2011.
Mellon Foundation and Digital Preservation Coalition Sponsor Formation of Task Force for Email Archives, 1 November 2016 (https://mellon.org/resources/news/articles/mellon-foundation-and-digital-preservation-coalition-sponsor-formation-task-force-email-archives/).
OCLC. *PREMIS (PREservation Metadata: Implementation Strategies) Working Group*, 2005 (http://www.oclc.org/research/projects/pmwg/).
S. Pigliapoco, *Conservare il digitale*, Macerata, EUM, 2010.
David S. H. Rosenthal, *Emulation & Virtualization as Preservation Strategies*, Report commissioned by The Andrew W. Mellon Foundation, October 2015 (https://mellon.org/media/filer_public/0c/3e/0c3eee7d-4166-4ba6-a767-6b42e6a1c2a7/rosenthal-emulation-2015.pdf).
Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information, Final Report of the Blue Ribbon Task Force on Sustainable Digital Preservation and Access (F. Berman and B. Lavoie, co-chairs), La Jolla, February 2010 (http://brtf.sdsc.edu/biblio/BRTF_Final_Report.pdf).
F. Tomasi, M. Daquino, *L'uso delle ontologie per la preservazione dei dati*, in presentazione al Convegno AIUCD 2017, Roma, 26-28 gennaio 2017.
M. Zane, *Per una nuova pedagogia del patrimonio*, "Giornale delle Fondazioni" (http://www.ilgiornaledellefondazioni.com/content/una-nuova-pedagogia-del-patrimonio).

Mesurer Merce Cunningham: une expérimentation en «theatre analytics»

Clarisse Bardiot
clarisse_bardiot@mac.com
University of Valenciennes, Belgium

Theatre studies is a largely under-discussed topic in digital humanities research projects. It's lagging behind the first wave of digital humanities scholarship, « focus[ing] on large-scale digitization projects and the establishment of technological infrastructure» (Presner, 2010). Theatre studies remains on the fringe of a growing phenomenon: culture analytics. In the context of big and complex datasets, culture analytics «is the data-driven analysis of culture» (IPAM, 2016). I suggest the expression «theatre analytics» (Bardiot, 2017). To paraphrase the culture analytics definition, theatre analytics is the data-driven analysis of theatre, whether it concerns theatre history (Caplan, 2016), drama or mise-en-scène. To understand what quantitative methodologies can bring to the knowledge of theatre, I propose a case study of Merce Cunningham. What can we learn about Merce Cunningham, one of the most influential *choreographers* of the 20th century, thanks to theatre analytics? A leader of the American avant-garde throughout his seventy year career from 1938 to 2009, he establishes in 2000, in the twilight of his career, the Merce Cunningham Trust, in order to preserve the integrity of his work. At the same time, he decides to dissolve the Merce Cunningham Dance Company (MCDC) two years after his death and a legacy tour. This is an unprecedented initiative. On one hand, it demonstrates exceptional effort and dedication to document the works. On the other hand, it challenges the ephemeral nature of

performing arts : 86 out of 183 choreographies are documented with "digital Dance Capsules" "so that it may be performed in perpetuity"(Dance Capsules, n.d.). By the way, two groups of works are defined: the canon (key works with extensive documentation in order to perform them again and again); the auxiliary (minor works with no documentation available to the public and *de facto* impossible to replay).

The data was collected from the Merce Cunningham Trust website. It concerns theatre production and cast, Dances Capsules documentation and the history of the MCDC. The dataset contains 183 works from 1938 to 2009 (including 86 Dance Capsules) and 347 people. We can identify three main data categories: people, works and documentation. What can we infer from beyond the data about the MCDC history, Cunningham's aesthetics and documentation strategies?

Measuring means measuring instruments. I used various and complementary tools in order to vary the approaches and analysis of the same dataset : Gephi for network analysis; Palladio for geographic and temporal representation; spreadsheet (Excel, Open Office, Datamatic) for statistics analysis. This paper will present the first results of this research, part of it conducted with students during a graduate «introduction to digital humanities» course. Statistical diagrams show three different periods of Cunningham's work; a stylistic signature with a preference for pieces that are 30 minutes long, and for soli, sextets and works with 13 to 15 dancers; a general trend towards more dancers and more length; the special place of soli in order to articulate the canon and the auxiliary; the organization of documents in the Dance Capsules. Network analysis let me define two different ways of collaboration, the «star» and the «spiral», and raises awareness on pivotal dancers. Geographic representation highlights relations between Europe and the United States.

In a wider historical perspective, it would be interesting to compare these preliminary results with other datasets. One example : two patterns have been identified in the Cunningham collaborations network : the star (figure 1), with discontinuous, centralized collaborations and groups separated from each other; the spiral (figure 2), with continuous, collective collaborations and one group growing organically. The change from the star to the spiral takes place when the company is created. Do these patterns characterize other choreographers and directors careers ? Is the creation of the company the main factor causing the evolution from the first pattern towards the second one ? While a well-worn issue – we do know that the creation of a company plays a crucial role in a career – the fact remains that "theatre analytics" let us visualize the patterns this break constitutes (or maybe not) and define different ways of collaborations.

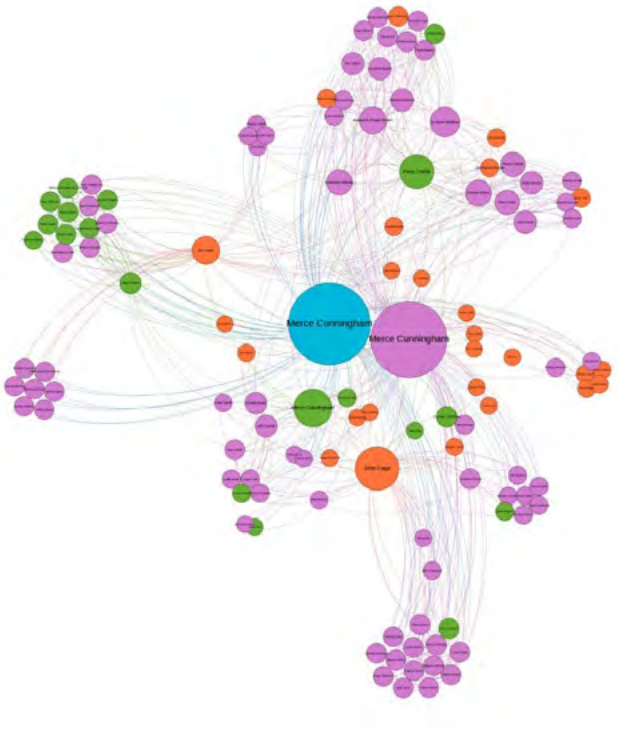


Figure 1 : Merce Cunningham's collaborations network before 1954. The star pattern.
Pink, dancers; orange, composers ; green, stage designers ; blue, choreographer.

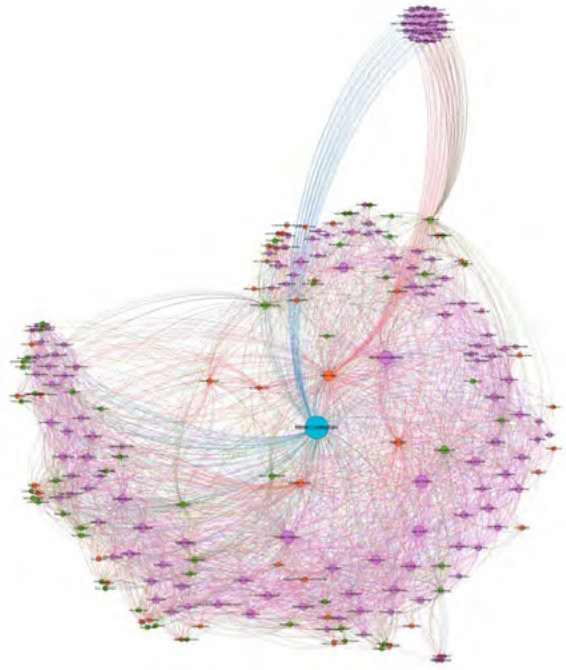


Figure 2 : Merce Cunningham's collaborations network after 1954. The spiral pattern.

References

Dance Capsules - Merce Cunningham Trust https://mercecunningham.org/film-media/dance-capsules/ (accessed 29 May 2018).

Merce Cunningham Dance Capsules http://dancecapsules.merce.broadleafclients.com/about.cfm (accessed 29 May 2018).

Bardiot, C. (2017). Arts de la scène et culture analytics. (Ed.) Galleron, I. *Revue d'historiographie du Théâdre. Etudes théâtrales et humanités numériques*(4): 11–20.

Caplan, D. (2016). Reassessing Obscurity: The Case for Big Data in Theatre History. *Theatre Journal*, 68(4): 555–73.

Tangherlini, T. R. (ed). (2016). *Culture Analytics : White Papers*. http://www.ipam.ucla.edu/wp-content/uploads/2016/09/Culture_Analytics_WhitePapers.pdf.

Presner, T. (2010). Digital Humanities 2.0: a report on knowledge. *Connexions Project*.

Is Digital Humanities Adjuncting Infrastructurally Significant?

Kathi Inman Berens
kathiberens@gmail.com
Portland State University, United States of America

The question of when "digital humanities" will drop the "digital" modifier and become "humanities" has special resonance for adjunct instructors. Digital humanities senior scholars might bridge the gap between tenured working conditions and adjunct working condition when crafting field infrastructures: not just because adjuncts merit both employment protections and what I call "microbenefactions" (more on that below), but because adjuncts are the invisible mass of humanities faculty buttressing every kind of institution, from community college to elite research-1 university. Adjuncts shoulder the humanities enterprise, teaching the general education classes that free researchers to pursue critical questions that advance the field.

This talk examines the infrastructural causes of DH adjunct invisibility and proposes two remedies: to motivate DH adjunct self-identification by convening DH adjunct-specific prizes and bursaries; and to invite senior DH faculty to perform "microbenefactions" that cost little effort and give adjuncts access to prize-worthy work opportunities or other benefits, such as renewable funding.

When "digital" humanities becomes just humanities, what's to stop "adjunctification" from converting DH tenure lines into part-time or other tenure-ineligible work, as has happened pervasively in other sub-specialties? In 2012, Stephen Ramsay problematized DH as "the hot thing." It's a skepticism shared by many in the field, including panelists of the DH 2017 Conference panel "Challenges for New Infrastructures and Paradigms in DH Curricular Program Development," which openly wondered whether graduate students were well served by DH certificate programs.[1] Miriam Posner notes that DH's "sexiness" today obscures the "widespread understaffing" of many DH initiatives"[2] This is an analog to adjunctification, the "shortsighted" boom/bust cycles of "soft" money quickly depleted which then require maintenance with a precarious budget. Amy Earhart has documented the unsustainability of early DH passion projects, websites whose hand-built archives rusticate when the faculty author retires or moves institutions.[3] Startups are sexy, but maintenance is not. When today's senior DH faculty retire in ten or twenty years, what infrastructures of care will be in place to stop those vacated tenure lines from being converted to part-time positions? The gender politics of "sexy," "hot" DH cast a pall over the field when one factors in that the majority of adjuncts are women. "As a woman of color," Liana M. Silva wonders, "I am especially interested to know what the women in contingent ranks look like. According to the Education Department's 2009 report, 51.6 percent of contingent faculty are women. The same report says 81.9 percent of contingent faculty are white. To what extent is contingent labor a problem for white women? Or, from another angle, to what extent is this a white labor issue, where class is meant to trump race?"[4] These questions about race, gender and contingent labor are digital humanities questions.

Awarding DH Adjuncts

In its mentoring, promotion, and awards structures, the humanities professoriate is legacy-bound, oriented to a tenure system that pertains to only one quarter of people working in the field.[5] If, as James F. English contends in

1 See the DH 2017 panel abstract here: https://dh2017.adho.org/abstracts/176/176.pdf. Ryan Cordell pointedly observes in published version of his DH 2017 talk that "completing the hours required for our robust [DH graduate] certificate program requires students to decide their path almost immediately upon admission, and the decision to pursue the certificate dictates very particular routes through the larger Ph.D. program." See Cordell's "Abundance and Usurpation While Building a DH Curriculum" posted to his blog: http://ryancordell.org/research/abundance/

2 Miriam Posner, "Money and Time," http://miriamposner.com/blog/money-and-time/

3 Earhart, *Traces of the Old, Uses of the New*.

4 Liana M. Silva: https://chroniclevitae.com/news/1017-how-many-women-are-adjuncts-out-there; National Center for Education Statistics 2009 report to which Silva refers: https://nces.ed.gov/pubs2011/2011150.pdf See also: "Women as Contingent Faculty: The Glass Wall," published by the American Association of University Professors http://archive.aacu.org/ocww/volume37_3/feature.cfm?section=1 and New Faculty Majority's "Women and Contingency" project: http://www.newfacultymajority.info/women-and-contingency-project/

5 "Adjunctification" is well documented by adjunct advocacy organization like New Faculty Majority and Adjunct Nation; professio-

The Economy of Prestige, the key indicator of any contemporary cultural phenomenon entering the mainstream is the creation of a prize (2), then perhaps it is time for digital humanists to create criteria of DH excellence specific to DH adjunct working conditions because adjuncting is the instructional mainstream. Doing so would motivate adjunct DHers to identify their work as DH and contribute recognizably toward DH research and pedagogy field development. Lack of access to an adjunct-specific DH prize reinforces adjunct invisibility, making it highly unlikely that even very good research will attain the recognition necessary to vault the scholar out of adjuncting. Most of the seven DH adjuncts I interviewed don't necessarily identify themselves as "digital humanists" because they are not hired specifically to teach DH, though their methods are consistent with DH pedagogy practices.[6] "Imposter syndrome" is intensified by employment insecurity and DH definitional heterogeneity.[7]

How to give adjuncts access to prize-worthy work opportunities? Senior scholars are key. In my talk, I will discuss microbenefactions senior scholars gave me when I adjuncted (2011-2014). Those invitations gave me access to nationally-visible projects and let me train myself in techniques that are now a core part of my tenure track job.

"Microbenefactions" is a term I invented to signify the opposite of microaggressions. They are small actions that shift the balance of power, the order of operations, that give adjuncts access to prestige or information otherwise inaccessible to them. Note that I use the singular here: "an" adjunct. These acts of inclusion are do-able as a one-off or in the course of a given term, not the Herculean efforts of adjunct advocates such as New Faculty Majority President Maria Maisto, Adjunct Nation, and the PrecariCorps collective who publish PrecariTales, 300-500 word anonymously authored adjunct stories.[8]

Unlike state-mandated employment protections, microbenefactions are individual and hyperlocal. They layer adjuncting's transactional dyad with the more branching, collegial conceptualization of value typical of tenure-track employment. This is human-centered DH infrastructure. We acknowledge that humans are not widgets, and that DH teaching is not a dissemination of knowledge. The medium is the message. If the medium is adjuncting, then the message our students imbibe is that learning is transactional. Microbenefactions disrupt neoliberal infrastructure that shrinks learning and collegiality to transactions.

What is a microbenefaction? It's action by a tenured or tenure-track scholar who

- writes funding for adjunct salary into grant proposals
- advises and mentors adjuncts
- seeks input from adjuncts about student-centered pedagogy
- aids adjuncts in finding university resources or paid extra work
- invites adjuncts to meetings
- co-authors with adjuncts
- doesn't eliminate adjunct applications when deciding awards and honors
- authorizes support for adjunct professional development, such as conference travel
- pays to license adjunct-authored course materials after the adjunct leaves the institution
- writes letters of recommendation for adjuncts

Microbenefactions enact DH's ethical ambit, which the Global Outlook::Digital Humanities special interest group articulates as a recognition "that excellent work is being done around the world,"[9] even in elite first-world institutions that rely on adjunct labor but largely eliminate that labor from tenure and promotion consideration.

References

Berens, Kathi Inman. "Judy Malloy's Seat at the (Database) Table: A Feminist Reception History of Early Electronic Literature Hypertext." *Literary and Linguistic Computing*, Volume 29, Issue 3, 1 September 2014, pages 340–348, https://doi.org/10.1093/llc/fqu037.

nal groups such the AAUP and the Modern Language Association (2014); intra-university studies such as George Mason's, which surveyed 240 GMU adjuncts and "has been hailed as the most comprehensive study of a university's contingent faculty working conditions to date" (2014); trade journals like *Inside Higher Education* and *The Chronicle of Higher Education*; and the popular press. I am struck by *The Atlantic Monthly*'s occasional series (2013-present) that features titles like "There's No Excuse for How Universities Treat Adjuncts" and "The Cost of an Adjunct." See also Kathi Inman Berens and Laura E. Sanders, "DH and Adjuncts: Putting the Human Back in the Humanities."

6 A note about method. My university's Human Subjects Research Review Committee determined an IRB was not required for me to conduct informational interviews with adjuncts. I used a common set of questions with each adjunct. The conversations veered to the specifics of their own particular cases.

7 The authors of the "Alternate Histories of DH" panel note in their abstract: "Matthew Kirschenbaum's identification of the digital humanities in 2014 as a 'discursive construction' that ignores the 'actually existing projects' of the field set the stage for scholars to rethink how the digital humanities conceptualizes its work and its history ('What Is' 48). More recently, in the introduction to *Debates in the Digital Humanities 2016*, Matthew Gold and Lauren Klein use the scholarship of Rosalind Krauss who, in 1979, described art history as emerging as 'only one term on the periphery of a field in which there are other, differently structured possibilities.'"

8 https://precaricorps.org/about/true-stories/ The pinned story at time of writing details an adjunct who's taught at the same university for ten years and has been hired to revise materials for a large enrollment course. One chair made sure she got paid the first lump sum; the replacement chair didn't with the second, and she's still waiting with "no recourse except to wait." The Twitter hashtags #AdjunctLife and #RealAcademicBios also gather adjunct (but don't curate) stories.

9 Global Outlook::Digital Humanities is a special interest group of the Alliance of Digital Humanities Organization. See http://www.globaloutlookdh.org/

_____. "Want to Save the Humanities? Pay Adjuncts to Learn Digital Tools" in *Disrupting the Humanities: Digital Edition*, 05 January 2015, http://www.disruptingdh.com/want-to-save-the-humanities-pay-adjuncts-to-learn-digital-tools/. Accessed 27 November 2017.

_____. "Sharing Precarity: Adjuncts, Global Digital Humanities, and Care," in *Debates in Digital Humanities 2017*, eds. Lauren F. Klein and Matthew K. Gold. Minneapolis: University of Minnesota Press. In press.

Berens, Kathi Inman and Laura E. Sanders. "Putting the Human Back in the Humanities: Adjuncts and Digital Humanities" in *Disrupting Digital Humanities: Print Edition*, eds. Dorothy Kim and Jesse Stommel. New York: Punctum Press. 2017.

Bessette, Lee Skallerup. *Adjunct Run*. https://adjunctrun.readywriting.org/ Accessed 27 November 2017.

Bretz, Andrew. "The New Itinerancy: Digital Pedagogy and the Adjunct Instructor in the Modern Academy." *Digital Humanities Quarterly* Vol. 11, No. 3 (2017). http://www.digitalhumanities.org/dhq/vol/11/3/000304/000304.html

Clement, Tanya [panel chair], Alison Booth, Ryan Cordell, Miriam Posner, Maria Sachiko Cecire. "Challenges for New Infrastructures and Paradigms in DH Curricular Program Development," panel at the 2017 Digital Humanities Conference in Montréal, Québec, Canada August 8-11, 2017. https://dh2017.adho.org/abstracts/176/176.pdf.

Cordell, Ryan. "Abundance and Usurpation While Building a DH Curriculum." http://ryancordell.org/research/abundance/ 23 August 2017.

Davis, Rebecca Frost, Matthew K. Gold, Katherine D. Harris and Jentery Sayers, eds. *Digital Pedagogy in the Humanities*, Digital Edition (peer editing version) https://digitalpedagogy.mla.hcommons.org/. Accessed 27 November 2017.

English, James F. *The Economy of Prestige: Prizes, Awards, and the Circulation of Cultural Value.* Cambridge: Harvard University Press, 2008.

Finley, Ashley. "Women as Contingent Faculty: The Glass Wall." *On Campus With Women* featured article of the Association of American Colleges and Universities. Vol. 37, No.3 (Winter 2009). http://archive.aacu.org/ocww/volume37_3/feature.cfm?section=1 Accessed 27 November 2017.

Gonzales, Andrea and Sophie Houser. *Tampon Run*. http://tamponrun.com/ Accessed 27 November 2017.

Higgen, Parker. Tweet dated 6 January 2015. https://twitter.com/xor/status/552456370629672960. Accessed 27 November 2017.

Honn, Joshua. "Never Neutral: Critical Approaches to Digital Tools Culture in the Humanities." https://figshare.com/articles/Never_Neutral_Critical_Approaches_to_Digital_Tools_Culture_in_the_Humanities/1101385 Accessed 21 November 2017.

Jacobs, Ken, Ian Perry, and Jenifer MacGillvary. "The High Public Cost of Low Wages: Poverty-Level Wages Cost U.S. Taxpayers $152.8 Billion Each Year in Public Support for Working Families." UC Berkeley Center for Labor Research and Education. April 13, 2015 Report. http://laborcenter.berkeley.edu/the-high-public-cost-of-low-wages/

Jasnik, Scott. "Humanities Job Woes." *Insider Higher Ed.* January 4, 2016.
https://www.insidehighered.com/news/2016/01/04/job-market-tight-many-humanities-fields-healthy-economics Accessed 27 November 2017.

Knapp, Laura G., Janice E. Kelly-Reid and Scott A. Ginder. "Employees in Postsecondary Institutions, Fall 2009, and Salaries of Full-Time Instructional Staff, 2009-10," a Report published by the U.S. Department of Education. November 2010. https://nces.ed.gov/pubs2011/2011150.pdf Accessed 27 November 2017.

Koseff, Alexei. "Part-time community college instructors to get job protections" [*sic*]. *Sacramento Bee*. 30 September 2016. http://www.sacbee.com/news/politics-government/capitol-alert/article105301086.html Accessed 27 November 2017.

Losh, Elizabeth, ed. *MOOCs and Their Afterlives: Experiments in Scale and Access in Higher Education.* Chicago: University of Chicago, 2017.

Manyika, James, Susan Lund, Jacques Bughin, Kelsey Robinson, Jan Mischke, and Deepa Mahajan. "Independent Work: Choice, necessity, and the Gig Economy." *McKinsey Global Institute*. October 2016. https://www.mckinsey.com/global-themes/employment-and-growth/independent-work-choice-necessity-and-the-gig-economy.

McGrail, Anne. "Whole Game: Digital Humanities at Community Colleges." *Debates in Digital Humanities 2016*, eds. Lauren F. Klein and Matthew K. Gold. Minneapolis: University of Minnesota Press, 2016. http://dhdebates.gc.cuny.edu/debates/text/53

McPherson, Tara. "Theory/Practice: Lessons Learned from Feminist Film Studies," on the panel "Alternate Histories of the Digital Humanities: a Short Paper Panel Proposal," convened by Roger Whitson and featuring Whitson, Amy Earhart, Steven Jones and Padmini Ray Murray, at the Digital Humanities 2017 conference July 8-11, 2017 in Montréal, Québec, Canada. https://dh2017.adho.org/abstracts/115/115.pdf Accessed 27 November 2018.

Molloy College DH Adjunct Job Advertisement. https://main.hercjobs.org/jobs/10389448/new-media-and-digital-humanities-adjunct. Accessed 18 November 2017. [The link will expire; see screenshot in Appendix.]

Nazer, Daniel and Elliot Harmon. Electronic Frontier Foundation, "Stupid Patent of the Month: Elsevier Patents Online Peer Review." August 31, 2016. https://www.eff.org/deeplinks/2016/08/stupid-patent-month-elsevier-patents-online-peer-review

New Faculty Majority "Women and Contingency Project." http://www.newfacultymajority.info/women-and-contingency-project/ Accessed 27 November 2017.

Pierazzo, Elena. "The Disciplinary Impact of the Digital: DH and ‚The Others'." Keynote at the Digital Humanities Summer Institute 2017, Victoria, B.C., Canada 16 June 2017. Abstract viewable here: http://dhsi.org/schedule.php
Precaricorps "True Stories." https://precaricorps.org/about/true-stories/ Accessed 27 November 2017.
Ramsay, Stephen. "The Hot Thing." (2012) https://github.com/sramsay/sramsay.github.com/blob/master/_posts/2012-04-09-hot-thing.markdown. Accessed 27 November 2017.
Risam, Roopika and Susan Edwards. "Micro DH: Digital Humanities at Small Scale." Conference talk at Digital Humanities 2017 Conference in Montréal, Québec, Canada August 8-11, 2017. Abstract viewable here: https://dh2017.adho.org/abstracts/196/196.pdf
Silva, Liana. "How Many Women Are Adjuncts Out There?" *Chronicle of Higher Education.* 27 May 2015. https://chroniclevitae.com/news/1017-how-many-women-are-adjuncts-out-there Accessed 27 November 2017.
Stanley, Sara Catherine. "Why DH?" (2017) http://scatherinestanley.us/2017/06/why-is-dh Accessed 21 November 2017.
Varner, Stuart. "Digital Humanities or Just Humanities?" https://stewartvarner.com/2013/11/digital-humanities-or-just-humanities/ Accessed 21 November 2017.
Warford, Erin. "StoryTelling with Digital Maps," a workshop at the 2017 Summer Digital Humanities Workshop Series at Canisius College. https://blogs.canisius.edu/digital-humanities/gis2017/ Accessed 27 November 2017.

Transposição Didática e atuais Recursos Pedagógicos: convergências para o diálogo educativo

Ana Maria Bosse
anahboss@hotmail.com
Universidade Federal de Santa Catarina, Brazil

Juliana Bergmann
jcfbergmann@gmail.com
Universidade Federal de Santa Catarina, Brazil

Resumo: Esta pesquisa, desenvolvida com alunos do 3º ao 5º ano do Ensino Fundamental brasileiro, tem como objetivo analisar a importância da renovação dos recursos pedagógicos no contexto educacional, da sociedade contemporânea, e refletir sobre as possibilidades e potencialidades destes recursos no processo de repensar o papel da escola nesta cultura digital, para assim atender as necessidades educacionais e favorecer o diálogo educativo.

Introdução

Atualmente, em nossa sociedade, as Tecnologias Digitais de Informação e Comunicação, associadas à internet, têm proporcionado mudanças constantes na circulação dos saberes, na produção e apropriação dos conhecimentos, passando a **informação** a ser o **bem** de maior valor social, e como já apontado por Pérez Goméz (2015:15), nesta era "a atividade principal dos seres humanos tem a ver com a aquisição, o processamento, a análise, a recriação e a comunicação da informação". As constantes inovações tecnológicas, desta cultura digital, vêm influenciando e interferindo nas relações interpessoais, despertando novas formas de gerenciar socialmente o conhecimento, de ensinar e aprender.

Nesta perspectiva, podemos destacar duas situações recorrentes no contexto educacional desta sociedade: 1) os recursos pedagógicos oferecidos nas escolas muitas vezes não levam em conta o uso potencial das novas mídias pelos alunos, ignorando todas as experiências cotidianas que eles desenvolvem e adquirem com essas novas tecnologias; 2) muitas vezes a escola dispõe de inúmeros recursos tecnológicos (midiáticos) de última geração, mas estes são subutilizados, sem que o educador os inclua em seu planejamento, seja por desconhecimento, seja por não acreditar que fará qualquer diferença ao aluno. Assim, se faz necessário pensarmos em elos que favoreçam esta aproximação.

Uma proposta para fomentar um maior diálogo educativo, conforme a presente pesquisa – realizada em uma escola de Ensino Fundamental onde atuo como Coordenadora Pedagógica –, dá-se através da compreensão do uso e da definição dos recursos pedagógicos no processo de ensino e aprendizagem, considerando que estes precisam ser constantemente reavaliados, de forma a beneficiar principalmente a transposição didática dos saberes, acreditando, assim, que o caminho para a construção de um novo pensar e de um novo fazer se edifica no questionamento, na pesquisa, no revisitar e analisar os modelos existentes para então propor novos indicativos. Para exemplificar, apontamos que o uso da internet, dos sites e dos aplicativos, através dos computadores, dos *tablets* e dos celulares, utilizados como recursos pedagógicos, podem proporcionar novas práticas para aproximar o conteúdo didático com a práxis da sala de aula, estabelecendo uma conexão concreta com a cultura cotidiana do aluno.

Recursos Pedagógicos na era digital

Dentro de todo este entrelaçar de mudanças advindas das novas tecnologias, é notório que a informação está à disposição em qualquer momento, a todo tempo, nos mais diversos locais; as tecnologias de comunicação trazem consigo esta potencialidade, fenômeno intitulado "*ubiquidade*" (Santaella, 2013); as potencialidades da co-

municação, principalmente com os dispositivos móveis e digitais, são inúmeras.

As novas gerações já nascem imersas nesse contexto da cultura digital. Desde muito cedo os sujeitos interagem com as mais diversas tecnologias de informação e comunicação e o mundo do ciberespaço já é parte constituinte do seu cotidiano. Assim, se adaptam a ele muito rapidamente e trafegam por entre essas novidades tecnológicas com desenvoltura e habilidade. Diante desta realidade, precisamos refletir, analisar e repensar o papel da escola e do ensino de modo que compreenda o contexto da sociedade atual.

Rivoltella (2007, *apud* Didonê, 2007), propõe que a mídia pode e deve permear os processos de ensino e aprendizagem, como acontece com a escrita, destacando que o papel assumido pelo professor que usa as novas tecnologias midiáticas não se limita a falar, mas sim, a direcionar o uso dos meios de comunicação pelos alunos.

A partir destas reflexões, podemos destacar que no atual contexto educacional nos encontramos diante de *"escolas analógicas e cabeças digitais"* (Petarnella, 2008), sendo pertinente e necessário trazer o mundo vivencial do aluno – tecnológico e midiático desta cultura que já faz parte do nosso cotidiano, para o ambiente escolar, e assim favorecer um verdadeiro diálogo educativo em que todos se beneficiem.

Recursos Pedagógicos: caminho para o diálogo educativo

Acreditando nas potencialidades dos recursos pedagógicos e na contribuição destes para aproximar e envolver o aluno no processo de ensino e aprendizagem, ponderamos também a importância destes como elementos que fazem parte da cultura do homem, que o colocam em contato com o seu tempo, com a sua historicidade.

Ao considerarmos que os recursos pedagógicos comportam em si a missão e o potencial, de se bem utilizados, de aproximar o aprendiz da sua aprendizagem, possibilitando maior entendimento na relação com o currículo pedagógico, mais interação na relação dos sujeitos envolvidos neste processo educacional, compreendemos que eles abrem para uma nova linguagem do aprender. De acordo com Eiterer e Medeiros (2010: 1), definimos como recursos pedagógicos "o entendimento daqueles lugares, profissionais, processos e materiais que visem assegurar a adaptação recíproca dos conteúdos a serem conhecidos aos indivíduos que buscam conhecer", e atendendo o importante papel que estes ocupam e desempenham no universo pedagógico, ainda compete destacar que sua abrangência está além da materialidade dos recursos em si.

Atualmente estamos diante de outro pensar pedagógico, que leva em consideração a importância da transposição didática nas relações de aprendizagem, nas relações entre aluno, professor, conhecimentos científicos, currículo, escola, prática pedagógica e recursos pedagógicos, e Almeida (2011: 11), enfatiza que "as nossas discussões acerca da transposição didática têm de ser entendidas dentro de uma concepção multiforme e ininterrupta". Pois, se é ao fazer pedagógico que compete tornar esta cultura transmissível e assimilável, ainda de acordo com o autor (Almeida, 2011), de algum modo é necessário transcender as diferenças e, através da interdisciplinaridade, rompermos com uma técnica homogeneizadora e homogeneizante de currículo, que engessa os conhecimentos, e que não compreende o valor da contextualização na prática educativa. Faz-se necessário pesarmos o fazer pedagógico através da prática reflexiva, e conforme Perrenoud (2002: 65), "a prática reflexiva é uma relação com o mundo: ativa, crítica e autônoma. Por isso, depende mais da postura do que de uma estrita competência metodológica". Nesse sentido, diante de todo o contexto apresentado sobre as condições da escola contemporânea e do aluno nesta sociedade da informação - da cultura digital, esta pesquisa, prima por investigar as possibilidades do uso de recursos pedagógicos e tecnológicos digitais (*tablets*, celulares, internet, sites, aplicativos), promover o diálogo educacional entre professor e aluno bem como favorecer a transposição didática, estimular no aluno o hábito da pesquisa e tornar mais significativo ao aprendiz o processo de ensino e aprendizagem, e analisar se estes recursos podem proporcionar uma nova relação no diálogo entre currículo, metodologia, professor e aluno.

References

Almeida, G P. (2011). *Transposição didática por onde começar?* São Paulo: Cortez Editora.

Didonê, D. Pier Cesare Rivoltella: F*alta cultura digital na sala de aula. Nova Escola*. Disponível em: <http://novaescola.org.br/formacao/formacao-continuada/pier-cesarerivoltella-falta-cultura-digital-sala-aula-609981.shtml>. Acesso em: 14 maio 2016.

Eiterer y Medeiros, C. L. *Recursos Pedagógicos*. Disponível em: http://www.gestrado.net.br/pdf/155.pdf. Acesso em 09/11/2017.

Freire, P. (2007). *Educação como Prática da Liberdade*. 29. ed. Rio de Janeiro: Paz e Terra.

Gentile, P. Antonio Nóvoa: *Professor se forma na escola. Nova Escola*. Disponível em: <https://novaescola.org.br/conteudo/179/entrevista-formacao-antonio-novoa>. Acesso em: 20 novembro 2017.

Moran, J. M., Masetto, M. T. y Behrens, M. A. (2003) *Novas Tecnologias e Mediação Pedagógica*. 7. ed. São Paulo: Papirus.

Pérez Gómez, Á. I. (2015). *Educação na era digital: a escola educativa*. Porto Alegre: Penso.

Perrenoud, P. (2002). *A Prática reflexiva no Ofício do Professor*. Porto Alegre: Artmed.

Petarnella, L. (2008). *Escola analógica: Cabeças digitais: o cotidiano escolar frente às tecnologias midiáticas e digitais de informação e comunicação*. Campinas, SP: Alínea.

Sacristán, J. G. (2000). *O currículo: uma reflexão sobre a prática*. Porto Alegre: Artmed.
Santaella, L. (2013). *Comunicação ubíqua: Repercussões na cultura e na educação*. São Paulo: Paulus.

Hurricane Memorial: The United States' Racialized Response to Disaster Relief

Christina Boyles
christina.boyles@trincoll.edu
Trinity College, United States of America

On September 20, 2017, Hurricane Maria made landfall in Puerto Rico. As the strongest hurricane to hit the island since 1928, the storm has caused significant damage—especially to infrastructure including roads, dams, communications networks, the electrical grid and the water supply. With much of the island still without power, and with limited aid coming from the United States, Puerto Rico is being left to deal with a humanitarian crisis on its own. The slow nature of the United States' response, coupled with Donald Trump's barrage of tweets, highlight the ways in which colonial narratives are feeding into disaster response efforts. For example, when San Juan Mayor Carmen Yulin Cruz requested an increase in federal aid, Trump replied, "Such poor leadership ability by the Mayor of San Juan, and others in Puerto Rico, who are not able to get their workers to help. They want everything to be done for them when it should be a community effort" (@realDonaldTrump). He later went on to claim that Puerto Rico's need for aid was hurting the federal budget and to claim that Hurricane Maria was not "a real catastrophe" for the island ("Trump compares Puerto Rico to Katrina"). Trump's victim-blaming behavior highlights both his lack of empathy for the citizens of Puerto Rico and the racial prejudice that undergirds the U.S. colonial enterprise. Although rarely so blatant, such behavior is not new; rather, the United States has an ongoing legacy of racialized disaster relief that is grounded in its colonial endeavors, particularly in Puerto Rico.

According to *El Nuevo Día*, the most widely-circulated newspaper in Puerto Rico, "El huracán María no superó a San Felipe II según un informe preliminar", or "Hurricane Maria did not surpass the strength of the San Felipe II Hurricane" (Ortega Marrero). Nevertheless, the two storms bear striking similarities. Both hit the island of Puerto Rico as category 5 hurricane, both crossed the island from the southeast corner and moved through the center of the island to the northwest corner, and both had significant long-term effects on the island.

While coverage of the 1928 storm's devastation in Florida is prominently displayed in novels and journalistic reports, coverage of the damage in Puerto Rico is almost non-existent in the mainland United States. I argue that the vulnerabilities created by the hurricane of 1928 were pivotal to the United States colonial agenda in Puerto Rico, resulting in a land grab by corporations and government entities that would impede the island's agricultural industry and economy for decades. This is made evident by the fact that the U.S. downplayed effects of the storm, the U.S. implemented policies to hurt small farmers & agricultural workers, and the U.S. denied that their actions caused economic and environmental harm to Puerto Rican citizens.

To make these connections clearer and to bring the stories of the storm's underrepresented victims back into our cultural memory, I have launched a digital work called the Hurricane Memorial project. This site includes my preliminary research, visualizations of my findings, and interviews with survivors and their family members.

As Florida and the Caribbean start to recover from Hurricane Maria, it is important to note that those living in economically disadvantaged communities will suffer the greatest from the storm's damage—just as they did in 1928. Aid quickly was rushed to Florida, while the federal government is "killing [Puerto Rico] with inefficiency" ("I Am Mad As Hell"). Such a response demonstrates the ways in which United States' racialized response to natural disasters is deeply rooted in its colonial enterprise. Failing to address these issues risks reinforcing harmful colonial narratives and causing irreparable harm to communities throughout the Caribbean and the world.

References

@realDonaldTrump. "Such poor leadership ability by the Mayor of San Juan, and others in Puerto Rico, who are not able to get their workers to help. They want everything to be done for them when it should be a community effort." Twitter, 30 September 2017, 5:26 A.M. https://twitter.com/realDonaldTrump/status/914089003745468417
Hurston, Zora Neale. *Their Eyes Were Watching God*. 1937. HarperPerennial, 2006.
"'I Am Mad As Hell': San Juan Mayor Carmen Yulin Cruz Criticizes Maria Response." *YouTube*, uploaded by NBC Nightly News, 29 September 2017. https://www.youtube.com/watch?v=41hl5RwfOVc
Ortega Marrero, Melisa. "El huracán María no superó a San Felipe II según un informe preliminar." *El Nuevo Dia* [Guaynabo, Puerto Rico], 29 September 2017.
Sharp, Deborah. "Storm's path remains scarred after 75 years."*USA Today*, 4 September 2003.
Sterghos Brochu, Nicole. "Florida's Forgotten Storm: the Hurricane of 1928." *South Florida Sun-Sentinel,* 2003.
"Trump compares Puerto Rico to Katrina, 'a real catastrophe." *YouTube*, uploaded by USA Today, 3 October 2017. https://www.youtube.com/watch?v=-J18rugiTxoU

Backoff Lemmatization as a Philological Method

Patrick J. Burns
pjb311@nyu.edu
Institute for the Study of the Ancient World, United States of America

Automated lemmatization, that is the retrieval of dictionary headwords, is an active area of research in Latin text analysis. Latinists have available web-based applications like Collatinus (Ouvard and Verkerk, 2014) and LemLat (Bozzi et al., 1992) and web services like Morpheus (Almas, 2015). LatMor (Springmann, 2016) and TreeTagger (Schmid, 1994) offer lemmatization as a byproduct of their primary tasks as morphological taggers. Recent work, to name a few developments, has seen lexicon-assisted tagging and rule induction (Eger et al., 2015; cf. Juršic, 2010) as well as neural networks (Kestemont and De Gussem, 2017) used as strategies for improving Latin lemmatization.

In this short paper, I describe the implementation of the Backoff Lemmatizer (https://github.com/cltk/cltk/blob/master/cltk/lemmatize/latin/backoff.py) for the Classical Language Toolkit, an open-source Python platform dedicated to developing natural language processing tools for historical languages (Johnson, 2017). The Backoff Lemmatizer is in fact not a single lemmatizer but rather a customizable suite of sub-lemmatizers, based on the Natural Language Toolkit's SequentialBackoffTagger. The SequentialBackoffTagger allows the user to "chain taggers together so that if one tagger doesn't know how to tag a word, it can pass the word on to the next backoff tagger" (Perkins, 2014, 92). While the backoff process was originally designed to handle part-of-speech tagging, and so, a task with a limited tagset, it works well for lemmatization (~90.34% accuracy compared to the 93.49% to 95.30% range reported in Eger et al., 2015).

A default class for sequential lemmatization, BackoffLatinLemmatizer, is available through the CLTK "Lemmatize" module using the following backoff sequence: 1. a dictionary-based lemmatizer for high-frequency, indeclinable vocabulary; 2. a unigram-model lemmatizer based on training data; 3. a rules-based lemmatizer based on regular expression patterns; 4. a variation on the previous regular-expression-based lemmatizer that factors in principal-part information; 5. another dictionary-based lemmatizer using the Morpheus lemma dictionary; and finally 6. an identity lemmatizer that returns the token as lemma.

Although currently available and tested only for Latin, the Backoff Lemmatizer is in theory language agnostic, since the sub-lemmatizers can be passed language-specific training data and models. So, for example, the UnigramLemmatizer requires training data in the form of a Python list of tuples of the form *[('token1', 'lemma1'), ('token2', 'lemma2'), ...]*. A Latin model with data in this form based on The Ancient Greek and Latin Dependency Treebank (Celano, Crane, and Almas, 2017) is available in the CLTK Latin corpora, but a similar model could be built for any language. Similarly, the RegexLemmatizer relies on a custom dictionary of regular expression patterns extracted from Latin morphological patterns. But again, a list of patterns could be written for any language and worked into this sub-lemmatizer. Furthermore, the sub-lemmatizers can be added or removed as necessary, and can be reordered based to optimize accuracy for a given language or language domain. Accordingly, the BackoffLemmatizer is particularly well-suited to less-resourced languages (Piotrowski, 2012, 85): a language without sufficient training data could build a backoff chain that ignores the UnigramLemmatizer and rely only on dictionary- and rules-based sub-lemmatizers.

Because of its multipass combination of probabilistic tagging based on existing Latin text, Latin lexical data, and a ruleset based on Latin morphology, the Backoff Lemmatizer can be described as following a philological method. By this, I mean that the process reflects the reading, decoding, and disambiguating strategies of the modern Latin reader (McCaffrey, 2006). For example, the process echoes the classroom process of Paul Diederich, who describes groups of students reading together and analyzing their text first through a combination of previous knowledge and dictionary lookups, but then "if no member of the group can clear up the difficulty, they resort to a formal analysis of the endings" (Hampel, 2014, 95).

One limitation of the current Backoff Lemmatizer setup is its binary sequential decision making; that is, a token is assigned a lemma based on the first match encountered in the backoff chain. By way of conclusion, I will discuss work in progress on a progressively scored Backoff Lemmatizer, or one that returns the lemma with the highest likelihood found after a token passes through and is assigned a score by every sub-lemmatizer in the chain.

References

Almas, B. (2013). *Morpheus-Wrapper*. https://github.com/PerseusDL/morpheus-wrapper (accessed 21 November 2017).

Bozzi, A., G. Cappelli, M. Passarotti, E. Pulcinelli, and P. Ruffolo. (1992). *LemLat*. http://www.ilc.cnr.it/lemlat/ (accessed 21 November 2017).

Celano, G. G. A., G. Crane, and B. Almas. (2017). *The Ancient Greek and Latin Dependency Treebank*. https://perseusdl.github.io/treebank_data/ (accessed 21 November 2017).

Eger, S., T. vor der Brück, and A. Mehler. (2015). Lexicon-Assisted Tagging and Lemmatization in Latin: A Comparison of Six Taggers and Two Lemmatization Methods, in Proceedings of the 9th SIGHUM Work-

shop on Language Technology for Cultural Heritage, *Social Sciences, and Humanities*: 105–13.
McCaffrey, D. (2006). Reading Latin Efficiently and the Need for Cognitive Strategies, in *When Dead Tongues Speak: Teaching Beginning Greek and Latin*, ed. J. Gruber-Miller. New York: Oxford University Press.
Hampel, R. L. (2014). *Paul Diederich and the Progressive American High School*. Charlotte, NC: Info Age.
Juršič, M., I. Mozetic, T. Erjavec, and N. Lavrac. (2010). LemmaGen: Multilingual Lemmatisation with Induced Ripple-Down Rules. *Journal of Universal Computer Science*: 1190–1214. https://doi.org/10.3217/jucs-016-09-1190.
Johnson, K. P. (2017). *CLTK: The Classical Language Toolkit*. https://github.com/cltk/cltk. (accessed 21 November 2017).
Kestemont, M., and J. De Gussem. (2017). Integrated Sequence Tagging for Medieval Latin Using Deep Representation Learning. *Journal of Data Mining & Digital Humanities*, Special Issue on Computer-Aided Processing of Intertextuality in Ancient Languages. https://arxiv.org/abs/1603.01597v2.
Loper, E., S. Bird, and T. Tresoldi. (2017). *NLTK 3.2.5 Documentation: nltk.tag.sequential*. http://www.nltk.org/_modules/nltk/tag/sequential.html (accessed 21 November 2017).
Ouvard, Y., and P. Verkerk. (2014). *Collatinus Web*. http://outils.biblissima.fr/en/collatinus-web/index.php (accessed 21 November 2017).
Perkins, J. (2014). *Python 3 Text Processing with NLTK 3 Cookbook*. Birmingham, UK: Packt Publishing.
Piotrowski, M. (2012). *Natural Language Processing for Historical Texts*. San Rafael, CA: Morgan & Claypool Publishers
Schmid, H. 1994. *Probabilistic Part-of-Speech Tagging Using Decision Trees*, In Proceedings of the Conference on New Methods in Language Processing, Manchester, UK.
Springmann, U., H. Schmid, and D. Najock. (2016). LatMor: A Latin Finite-State Morphology Encoding Vowel Quantity. *Open Linguistics* 2(1). https://doi.org/10.1515/opli-2016-0019. (accessed 21 November 2017).

Las humanidades digitales y el patrimonio arqueológico maya: resultados preliminares de un esfuerzo interinstitucional de documentación y difusión

Arianna Campiani
acampiani@ucmerced.edu
University of California Merced, United States of America

Rodrigo Liendo
rodrigo@liendo.net
Universidad Nacional Autónoma de México, Mexico

Nicola Lercari
nlercari@ucmerced.edu), University of California Merced, United States of America

El uso de tecnologías digitales para el registro y conservación del patrimonio ha demostrado ser de gran utilidad ya que permite contar con una documentación exacta que puede constituir la base para proyectos de restauración, pero también de investigación y difusión (De Reu et al., 2013; Forte et al., 2015). En las últimas dos décadas, el cambio climático, la creciente inestabilidad política y el saqueo han llevado al deterioro de numerosos sitios arqueológicos mesoamericanos (Juárez Cossío, 2000; Lario Villalta, 2000; Noriega y Quintana, 2002). En este escenario, la documentación del patrimonio digital y la difusión de datos en línea se convierten en recursos invaluables para registrar, monitorear y preservar el patrimonio cultural maya del sur de México. (Forte et al., 2015)

La Coordinación Nacional de Monumentos Históricos del Instituto Nacional de Antropología e Historia (INAH) ha implementado el Laboratorio de Imagen y Análisis Dimensional para integrar un acervo tridimensional del patrimonio arquitectónico, pero, en cuanto al patrimonio arqueológico la documentación digital se ha limitado a edificios específicos de pocos sitios arqueológicos. En la última década, universidades de los Estados Unidos y Canadienses que conducen investigaciones en la península de Yucatán han empleado tecnología LiDAR y otras herramientas digitales para la documentación de sitios arqueológicos, no obstante estas iniciativas raramente contemplan la participación de universidades mexicanas o estudiantes locales (Golden et al., 2016; Hare, 2014; Hutson 2015; Hutson et al., 2016; Magnoni et al., 2016; Reese-Taylor et al., 2016).

En 2018, gracias a una colaboración entre la Universidad Nacional Autónoma de México y la Universidad de la California- Merced hemos empezado los trabajos de levantamiento digital en el sitio arqueológico de Palenque, Chiapas, patrimonio de la UNESCO desde 1980. En paralelo con las actividades de excavación en el Grupo IV, al noroeste del núcleo cívico-ceremonial, hemos empleado un escáner láser terrestre (TLS) y dos drones con cámaras de alta resolución para producir mapas y modelos 3D de los edificios y de sus espacios asociados, con una precisión al centímetro. En un lugar de la importancia de Palenque, donde los edificios necesitan de constante mantenimiento, esta labor nos parece relevante y necesaria.

En cuanto al centro del asentamiento y a los edificios monumentales con ello asociados, los vuelos con drones permiten no solo tener un registro cuidadoso sino complementar el levantamiento hecho manualmente a través de los años. Además, la fotogrametría consiente situar los trabajos de restauración llevados a cabo y reflexionar

sobre la manera en que estos complementan y a la vez modifican la percepción de las construcciones, puestos que dejan a la vista una sobreposición de diferentes etapas constructivas.

En acuerdo con los arqueólogos y conservadores del INAH, se escanearon la Casa E y C del Palacio y el Templo de las Inscripciones con énfasis en la Tumba de Pakal, ya que a corto plazo el Instituto empezará un proyecto de investigación y restauración de dichos edificios. La documentación producida servirá para planear las excavaciones en el Palacio y a la vez constituye la base para el monitoreo de los edificios y de sus decoraciones en piedra y estuco, y para evaluar la eficacia de las técnicas empleadas para su conservación.

A mediano plazo esperamos contar con un dron con cámara LiDAR para hacer prospección más detallada, perfeccionar el mapa de la ciudad y planear las excavaciones de acuerdo a las preguntas de investigación de los diferentes investigadores y estudiantes involucrados.

Estas técnicas digitales de documentación arqueológica y de monitoreo del patrimonio que empezamos a emplear en Palenque han sido adoptadas por el equipo de UC Merced en otros proyectos. Por ejemplo, en el sitio patrimonio mundial de Çatalhöyük, Turquía, el registro se ha complementado de modelos predictivos para la conservación gracias a la comparación de los datos 3D (con el uso del software open source Cloud Compare) y su implementación en una plataforma GIS (ESRI) (Campiani, Lercari y Lingle, 2018).

A parte de contar con el equipo para el mapeo digital, y paralelamente a la documentación, el objetivo de las dos instituciones es formar estudiantes gracias a la experiencia en campo, la organización de talleres y el intercambio de estudiantes y profesores. A través de estas estrategias, los datos recolectados por el equipo interinstitucional pueden ser analizados por todos los usuarios mediante software abiertos. A la fecha se ha empezado con la formación de arqueólogos en la temporada 2018.

A la vez, con el programa Unity, tanto para Çatalhöyük como para el sitio histórico de Bodie, California, en UC Merced se han desarrollado tres apps con fines diferentes: una para la simulación de las excavaciones y la interpretación de la estratigrafía (Lercari et al., 2017), una para los restauradores para la comparación de los elementos arquitectónicos y su estado de conservación (Lingle y Seifert, 2017) y otra app para guiar al público en el parque de Bodie (Lercari et. al, 2018). Los códigos generados constituyen la base para los trabajos a implementar en Palenque en cuanto a estudio y difusión.

Con fundamento en estas premisas pensamos que nuestra colaboración interinstitucional pueda sentar las bases metodológicas para el estudio y monitoreo del patrimonio arqueológico maya, gracias a la participación interdisciplinaria, el intercambio y formación de estudiantes y profesores, el desarrollo de nuevos métodos para el estudio arqueológico, la conservación y la difusión.

En esta ponencia breve queremos presentar los resultados de la primera temporada de campo con el empleo de estas tecnologías y reflexionar sobre objetivos a futuro y buenas prácticas en cuanto a documentación, difusión y divulgación de conocimiento para un público especializado y el público en general, para que el uso de la tecnología digital aplicada a la documentación del patrimonio arqueológico maya se vuelva un puente entre investigación y sociedad.

Referencias

Campiani, A., Lercari, N. y Lingle A. (2018). Analytical models for at-risk heritage conservation and 3D GIS. *Society for American Archaeology Conference: Abstracts of the 83rd annual meeting*. Washington DC, p.83, http://www.saa.org/Portals/0/SAA/MEETINGS/2018%20Abstracts/Individual%20Level%20Abstracts_C_D_2018.pdf (consultado el 1 de mayo de 2018)

De Reu, J., Plets, G., Verhoeven, G., De Smedt, P., Bats, M., Cherretté, B. y De Maeyer, W. (2013). Towards a Three-Dimensional Cost-Effective Registration of the Archaeological Heritage. *Journal of Archaeological Science*, 40 (2): 1108–21.

Forte, M., Dell'Unto, N., K. Jonsson K. y Lercari, N. (2015). Interpretation process at Çatalhöyük using 3D. In Hodder I. y Marciniak, A. (eds), *Assembling* Çatalhöyük. Maney Publishing, pp. 43-57.

Golden, C., Murtha, T., Cook, B., Shaffer, D.S., Schroder, W., J. Hermitt, E., Alcover Firpi, O. y Scherer, A. K. (2016). Reanalyzing environmental lidar data for archaeology: Mesoamerican applications and implications. *Journal of Archaeological Science: Reports*, 9: 293-308.

Hare, T., Masson, M. y Russell, B. (2014). High-density LiDAR mapping of the ancient city of Mayapan. *Remote Sensing* 6 (9): 9064–85.

Hutson, S. R., Kidder, B., Lamb, C., Vallejo-Cáliz, D. y Welch, J. (2016). Small Buildings and Small Budgets. Making Lidar Work in Northern Yucatan, Mexico. *Advances in Archaeological Practice* 4(3): 268-83.

Hutson, S. (2015). Adapting LiDAR data for regional variation in the tropics: A case study from the Northern Maya Lowlands. *Journal of Archaeological Science: Reports*, 4: 252–63.

Juárez Cossío, D. (2000). El Proyecto Yaxchilán y las alternativas de conservación en la década de los setenta. *XXII Simposio de Investigaciones Arqueológicas en Guatemala: Sitios arqueológicos en el área Maya: un reto para la conservación*. The Getty Conservation Institute, pp. 27-37.

Lario Villalta, C.R. (2000). El reto de conservación Tikal, Guatemala. *XXII Simposio de Investigaciones Arqueológicas en Guatemala: Sitios arqueológicos en el área Maya: un reto para la conservación*. The Getty Conservation Institute, pp. 59-69.

Lercari, N., Jaffke, D., Aboulhosn, J., Baird, G. y Guillem, A. (2018). Citizen Science Archaeology at Bodie

State Historic Park. *Society for American Archaeology Conference: Abstracts of the 83rd annual meeting*. Washington DC, p. 283, http://www.saa.org/Portals/0/SAA/MEETINGS/2018%20Abstracts/Individual%20Level%20Abstracts_I_L_2018.pdf (consultado el 1 de mayo de 2018)

Lercari, N., Shiferaw, E., Forte M. y Kopper R. (2017). Immersive Visualization and Curation of Archaeological Heritage Data: Çatalhöyük and the Dig@IT App. Journal of Archaeologial Method and Theory: 1-25.

Lercari, N., Lingle, A. y Umurhan O. (2016). Catalhöyük Digital Preservation Project. Çatalhöyük *2016 Archive Report*. http://www.catalhoyuk.com/sites/default/files/media/pdf/Archive_Report_2016.pdf (consultado el 2 de febrero de 2017).

Lingle, A. y Seifert, J. (2017). Update on the Çatalhöyük Digital Preservation Project. Çatalhöyük *2017 Archive Report*. http://www.catalhoyuk.com/sites/default/files/Archive_Report_2017.pdf (acceso 1 Mayo 2018)

Magnoni, A., Stanton T., Barth, N., Fernandez-Diaz, J. C., Osorio León, J. F., Pérez Ruíz, F. y Wheeler, J. A. (2016). Detection Thresholds of Archaeological Features in Airborne Lidar Data from Central Yucatán. *Advances in Archaeological Practice* 4(3): 232-248.

Noriega, R. y Quintana, O. (2002). Programa de restauración: Proyecto Protección de Sitios Arqueológicos en Petén. In Laporte, J.P., Escobedo, H. y Arroyo B. (eds), *XV Simposio de Investigaciones Arqueológicas en Guatemala, 2001*. Museo Nacional de Arqueología y Etnología, pp. 228-238

Reese-Taylor, C., Anaya Hernández, A., Flores Esquivel, F. C. A., Monteleone, K., Uriarte, A., Carr, C., Geovannini Acuña, H., Fernandez-Diaz, J. C., Peuramaki-Brown M. y Dunning, N. (2016). Boots on the Ground at Yaxnohcah: Ground-Truthing Lidar in a Complex Tropical Landscape. *Advances in Archaeological Practice* 4(3): 314-338.

Cartonera Publishers Database, documenting grassroots publishing initiatives

Paloma Celis Carbajal
pceliscarbaj@wisc.edu
University of Wisconsin-Madison, United States of America

Starting in Buenos Aires with Eloísa Cartonera in 2003, Cartonera publishers emerged as a reaction to the over commercialization of the book industry and its ever-growing conglomerates. With their unique hand embellished covers and their peculiar aesthetics, these publishers have challenged how books and literature are produced and distributed. Their collective manual process is equal to the intellectual one, resulting in a more democratic mode of production.

For thirteen years, the Cartonera Publishers Database has been documenting and preserving the diverse initiatives that stem from these grassroots projects which use recycled cardboard as book covers. The database is comprised of more than 1,200 entries which include Dublin Core metadata, scanned images of the back and front covers, copyright pages, and title pages, and audio files of interviews of several members of Cartonera publishing houses. An electronic crosswalk connects these entries to local cataloging of the Cartonera Book Collection. The audio files and an online full-text book "Akademia Cartonera: A primer of Latin American Cartonera Publishers" are additionally indexed and marked using TEI. This database is the only digital reference tool on these multi-pronged publishing initiatives. The ultimate goal is to connect this locally focused digital humanities project with cartonera books held at other institutions around the world in an interinstitutional Cartonera Catalog.

In the past year, I have been studying the possibility of using crowd sourcing and/or folksonomies to supplement the current content with the goal of providing a deeper understanding of the variety of contexts in which these books are created while also offering a space for the Cartonera publishers to contribute other content created directly by them. My proposed papers addresses the database and initial efforts to expand our work.

References

Cartonera Publishers Database, http://digital.library.wisc.edu/1711.dl/Arts.EloisaCart (accessed 20 November 2017).

Integrating Latent Dirichlet Allocation and Poisson Graphical Model: A Deep Dive into the Writings of Chen Duxiu, Co-Founder of the Chinese Communist Party

Anne Shen Chao
mrsannechao@gmail.com
Rice University, United States of America

Qiwei Li
liqiwei2000@gmail.com
University of Texas Southwestern Medical Center, United States of America

Zhandong Liu
zhandonl@bcm.edu
Baylor College of Medicine, United States of America

Chen Duxiu (1879-1942) co-founded the Chinese Communist Party in 1920, and served as its secretary general

from 1921 to 1927. He was a prolific author and a cultural rebel whose writings transformed the intellectual and social landscape of 20th century China. Yet from 1904 to about 1919, Chen advocated Western democracy and Social Darwinism as solutions to save China. His turn to communism was an abrupt transition, and many historians credited this to the influence of his colleague, and co-founder of the CCP, Li Dazhao (1888-1927). Both Li and Chen had studied in Japan, and through their interaction with Japanese Socialists and fellow students, became acquainted with literature on socialism and anarchism. Some say that Li was the theoretician who understood Bolshevism and Marxism in depth, while Chen did not become well-versed in Marxism until he founded the CCP (Yoshihiro, 2013).

In this paper, we applied topic modeling (Blei et al., 2012) to a select number of Chen's and Li's published articles, in an attempt to detect the difference, if any, in their interpretation on the subject of socialism, Marxism, communism and Bolshevism. We integrated two well-developed statistical methodologies, the Latent Dirichlet Allocation (LDA) and the Poisson Graphical Model (PGM), to probe in finer detail the broad themes in the 892 pieces of Chen's essays, correspondences, and occasional poetry, comprising a total of 1,347,699 Chinese characters. Based on the word counts per topic, we then implemented the PGM method to study the association among different topics. The use of PGM minimizes any misleading inference caused by confounding variables, and it also leads to a more concise structure of the network of topics.

Specifically, we chose 263 articles written by Chen Duxiu and 53 written by Li Dazhao, containing words related to Marxism, socialism, Bolshevism, and communism (Ren, 2018; Li ,1984). (Both selections covered the length of the men's publishing career; Chen passed way at age 63, while Li was executed at age 39). A document-term matrix (bag-of-words data) was generated from the preprocessed text. Next, we carefully selected a set of seed words for each of *K* topics of interest. We then applied the topic modeling method LDA to the bag-of-words data to find the remaining mixtures of words associated with each topic. Consequently, we could interpret each estimated topic by abstracting the top ranking terms within that topic. We then generated a new document-topic matrix from the document-term matrix by calculating the counts of those top words from the same topic. Finally, we applied the Poisson Graphical Model to the document-topic matrix to infer the conditional independence between each pair of topics. The resulting graph is a network visualization where each node represents a topic, and each edge indicates the conditional dependencies among the topics, meaning the two topics that are linked by an edge are correlated even after adjusting for all the other topics in the corpus.

The results yield several initial observations: Chen used a smaller set of vocabulary words over and over again to emphasize a point, while Li adopted a more discursive style with fewer repeats of the same word. Chen used many more verbs (such as: "agitate," "struggle," "unite," "lead," "develop," "carry out"), thereby exhorting his readers to action, while Li tended to use descriptive words. Chen focused on the present by analyzing different political groups: "Guomindang," "warlords," "proletariat," "bourgeoisie," "military," "students," "masses" and "imperialists." Li painted a larger scenario by using words such as "world," "humanity," "philosophy," "phenomenon," "relationship," "history" and "religion." The general conclusion at this early stage of analysis is that Chen urged his readers to put into action his plans to bring China under communism, while Li tended to explain to his readers the nature of Bolshevism and Marxism.

More interestingly, these calculations yielded "orbits" of vocabulary for each man's important ideas. For instance, Chen's use of the word "revolution" appeared three times in the 8 topics that we studied. In the first sub-topic, "revolution" appeared with words such as "class," "bourgeoisie," "proletariat," "develop," "strength," and "movement." In the second sub-topic, "revolution" again appeared alongside "peasants," "bourgeoisie," "proletariat," "lead," "China," "masses," "movement," and "action." In the third sub-topic, "revolution" appeared with "bourgeoisie," "proletariat," "struggle," "China," "Guomindang," "movement." Li, when he discussed "revolution," which appeared twice in the four topics we studied, he often used words such as "people," "Russia," "movement," "government," "masses," "future," and "China." While the general trend of these two men's writing is clear by a casual browsing of all of these articles, but this method of calculation demonstrates in a quantitative manner the qualitative interdependence of topics, and diagrams in an easy to read manner the network configuration of the vocabulary of each man.

References

Blei, David M. (2012) *Probabilistic topic models: Communications of the ACM*, 55(4): 77-84.

Li D. (1984). *Li Dazhao Wenji* [A literary collection of Li Dazhao]. Beijing: Renmin chubanshe.

Ren J. ed. (2008), *Chen Duxiu zhuzuo xuanbian* [A selected collection of Chen Duxiu's writing]. Shanghai: Shanghai renmin chubanshe. 6 vols.

Yoshihiro, I., tr. by Fogel J. (2013) *Formation of the Chinese Communist Party.* New York: Columbia University Press.

Sensory Ethnography and Storytelling with the Sounds of Voices: Methods, Ethics and Accessibility

Kelsey Marie Chatlosh
kchatlosh@gradcenter.cuny.edu
The Graduate Center, CUNY, United States of America

In contemporary anthropology, nearly all of us work with sound – usually oral interviews – but its quality as such is often taken for granted. Audio files of interviews are often quickly transcribed or qualitatively coded into text, then analyzed and written into books. And the soundscapes of our fieldwork sites are often taken for granted as well. Their meanings and textures as sounds are thus erased. The small sounds of voices and places often invoke an intimacy that anthropologists may attempt to render in text through "thick description" (Geertz 1973) and hopefully also "sincerity" (Jackson 2005), drawing from hermeneutic and poetic approaches in literary studies (Clifford and Marcus 1986, Behar and Gordon 1995).

Meanwhile, a growing body of interdisciplinary scholarship on sound studies foregrounds sound as "a modality of knowing and being in the world," of creating a sense of place or a narrative (Feld 2000). Performance studies scholars have also provided many contributions towards thinking about "the hegemony of textuality" (Conquergood 2002: 147) and, conversely, the "repertoire" of manifestations of knowledge and memory that exist outside the written, institutionalized archive (Taylor 2003). Needless to say, ontologies, storytelling, memories, and place- and identity- making are canonical topics of study in anthropology. As Steven Feld, an anthropologist and one of the leading theorists of sound studies, has discussed, there are many possibilities in "doing ethnography through sound—listening, recording, editing, and representation" that will hopefully one day be more than just "mostly about words" (Feld and Brenneis 2004: 461, 471). Further, as anthropologist and sound studies theorist, Roshanak Kheshti, has argued: "considering sound through the critical genealogy of feminist or race theory forces you to move beyond sound as an object and think of sound instead as an analytic or a hermeneutical tool for understanding inequality…" and the "social worlds" that scholars study (Brooks and Kheshti 2011: 330).

I am interested in approaches to methods, ethics and accessibility when working with the sounds of voices that cross-cut anthropology – specifically sensory ethnography, or ethnographic methods that foreground the senses – sound studies (and sound arts), and digital humanities. Anthropologists are not that common in the realm of digital humanities. However, many of us, one could argue, do projects that could be construed as "digital humanities," that is: "digital methods of research that engage humanities topics in their materials and/or interpret the results of digital tools with a humanities lens" (Lexicon of DH Workshop, The Graduate Center Digital Initiatives, tinyurl.com/lexicondh). And the thing with DH is, once we (scholars) start paying more acute attention to the ways in which our research is digital this can open up new questions and also new methods for doing what it is that we do, in terms of both research and pedagogy. This is particularly true, I suggest, for sound studies – given the importance of digital tools and platforms for recording, mixing, sharing and listening to audio.

Yet, new methods, digital tools and projects emerging through DH and internet research in general open up an array of rather new ethical and accessibility concerns (see e.g. Barnes 2006, Markham and Buchanon 2012). What constitutes personhood or "human subjects" on the internet? What data is or should be "public"? When should consent protocols be required? Can images or audio files of people and their voices bely anonymity? Who has access to make digital projects or to engage them, particularly in relation to differences of class, ability, and language fluency? How is the internet – its structure, its users, its algorithms – racialized and gendered (e.g. McPherson 2012, Noble 2018)? In what ways may some DH projects follow a practice of extraction without reciprocity? Indeed, anthropologists wrestle a lot with that last question in particular when extracting stories of individuals that then advance our careers, while many DH-ers may be, e.g., web scraping.

This short paper presentation will examine the possibilities of cross-cutting methodological approaches to anthropology, sound studies and arts, and digital humanities, specifically when recording and sharing the sounds of peoples' as a mode of storytelling. I will focus on oral interviews in particular. Driven by the aforementioned anthropological and interdisciplinary concerns, this paper will discuss the interplays of method and theory when cross-cutting these approaches, and issues of ethics and accessibility when recording and sharing sound. This includes being wary of institutional compliance with Institutional Review Boards but also following a feminist ethics beyond compliance, that, for example, foregrounds consent as not a one-time signature but reiterated, negotiated and subject to change (see Davis and Craven 2016). I will also consider various levels of intrusiveness and impact that the recording and sharing of the sounds – especially the sounds of peoples' voices – may have, and the potential roles of shared sounds within larger networks of listeners and what their availability may foreclose (e.g. Sugarman 1997, Brooks and Kheshti 2011, Kunreuther 2014, Kheshti 2015). Lastly, I will discuss digital modalities for sharing research with sound and their (limited) possibilities for storytelling, specifically for doing and sharing anthropological and other research in a more accessible form – with the exceptions structured by access to technology, limited hearing ability and translatability across languages and contexts. I will highlight free and open-source resources, such as sound archives and editing and hosting technologies, as well as low-cost Do-It-Yourself (DIY) microphones and speakers.

While websites are often great platforms for sharing oral history projects and other sounds, I will also discuss examples of other modalities for sharing sounds, such as exhibits and events, as well as digital platforms for scholarly publishing (e.g. Manifold). I will include a brief survey of various free online platforms that seem to have high potential for use in scholarship and pedagogy. These in-

clude: the SoundCloud online streaming platform, the Oral History Metadata Synchronizer in coordination with Omeka, podcasting via iTunes, StoryMaps for sharing audio on a map, and Chirbit for sharing audio on social media or embedding audio on a website. I will also discuss examples for the in-person sharing of sounds during, for example, an exhibit or class, including a brief survey of different kinds of speakers and headphones and different ways of transferring pre-recorded or live sounds to them, as well as spatial considerations for sharing sound. For example, placing numerous speakers inside an enclosed space, such as a tent, may allow for a focused listening space that is still shared and not as individuated as when using headphones (an idea I learned from sound artist Grant Smith of Reveil Radio in London). While I do not plan to conduct a full comparative analysis of these platforms, I will briefly discuss what I find to be some of openings and limitations of each.

In sum, this presentation aims to bridge together a number of themes: sound studies, oral histories, ethics, accessibility, and modalities for sharing sounds. I emphasize the intention that motivates my attempt to bridge these various themes: In my opinion, when recording and sharing human voices the researcher must always be vigilant in their ethical considerations (beyond IRB approval) at every step of the research design and practice, and then the sharing of these sounds is what makes their collection most worthwhile and to do so requires considerations of accessibility and modalities and ethics for such sharing.

References

Barnes, S. (2006). "A Privacy Paradox: Social Networking in the United States," *First Monday* 11(9). http://firstmonday.org/article/view/1394/1312 (accessed 26 April 2018).

Behar, R. and Gordon, D. A. (1995). *Women Writing Culture*. Berkeley, CA: University of California Press.

Brooks, D. and Kheshti, R. (2011). The Social Space of Sound, Theatre Survey 52: 329-334.

Clifford, J. and Marcus, G. ed.s. (1986). *Writing Culture: The Poetics and Politics of Ethnography*. Berkeley: University of California Press.

Conquergood, D. (2002). Performance Studies: Interventions and Radical Research, *The Drama Review* 46: 145-156.

Davis, D.-A. and Craven, C. (2016). *Feminist Ethnography: Thinking through Methodologies, Challenges, and Possibilities*. Lanham, MD: Rowman and Littlefield.

Feld, S. (2000). Sound Worlds. In Sound, Kruth, P. and Stobart, H. (eds). Cambridge, England: Cambridge University Press.

Feld, S. and Brenneis, D. (2004). Doing Anthropology in Sound, *American Ethnologist* 31(4): 461-474.

Geertz, C. (1973). *The Interpretation of Cultures*. New York: Basic Books.

Jackson, J. Jr. (2005). *Real Black: Adventures in Racial Sincerity*. Chicago, IL: Chicago University Press.

Kheshti, R. (2015). *Modernity's Ear: Listening to Race and Gender in World Music*. New York: New York University Press.

Kunreuther, L. (2014). *Voicing Subjects: Public Intimacy and Mediation in Kathmandu*. Berkeley, CA: University of California Press.

Markham, A. and Buchanan, E. (2012). Ethical Decision-Making and Internet Research: Recommendations from the AoIR Ethics Working Committee (Version 2.0). Association of Internet Researchers (AoIR). http://aoir.org/reports/ethics2.pdf (accessed 26 April 2018).

McPherson, T. (2012). Why Are the Digital Humanities So White? or Thinking the Histories of Race and Computation. In *Debates in the Digital Humanities*, Gold, M. (ed). Minneapolis, MN: University of Minnesota Press with Manifold. http://dhdebates.gc.cuny.edu/debates/text/29 (accessed 26 April 2018).

Noble, S. U. (2018). *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York: New York University Press.

Sugarman, J. (1997). *Engendering Song: Singing and Subjectivity at Prespa Albanian Weddings*. Chicago, IL: University of Chicago Press.

Taylor, D. (2003). *The Archive and the Repertoire*. Durham, NC: Duke University Press

Seinfeld at The Nexus of the Universe: Using IMDb Data and Social Network Theory to Create a Digital Humanities Project

Cindy Conaway
cindy.conaway@esc.edu
SUNY Empire State College, United States of America

Diane Shichtman
diane.shichtman@esc.edu
SUNY Empire State College, United States of America

This Digital Humanities project is an interdisciplinary project effort that uses the lens of, and data from, the U.S. TV show*Seinfeld*to explore questions about television and other media.*Seinfeld* has significant cultural influence over other media, but what is its **reach**, meaning the many other media items cast and crew worked on, also known as the **overlap**? We are starting with data from the Internet Movie Database (IMDb). This makes this project somewhat different from other Digital Humanities projects as we're using an existing database rather than primary sources. An associate professor of media studies, accustomed to conducting critical analysis of television shows, and an associate professor of information systems, more used to working with non-media studies data, are wor-

king to populate a relational database, to use quantitative analysis, and a social science theory--social network theory, particularly "Small Worlds" theory--to explain trends in media industries, including questions of genre, gender, race, and age in entertainment businesses.

Seinfeld (NBC 1989-1998) was a US-based half-hour, multi-camera, situation comedy, one of several that featured stand-up comics in stories similar to their own lives. Although it ended nearly 20 years ago, it heavily influences TV shows of today, including "hangout" sitcoms, one-camera comedies featuring conversation and digression, and antihero dramas. Journalist Jennifer Keishen Armstrong writes in the bestselling *Seinfeldia* that the show "snuck through the network system to become a hit that changed TV's most cherished rules; from then on, antiheroes would rise to prominence, unique voices would invade the airwaves, and the creative forces behind shows would often gain as much power and fame as the faces in front of the cameras" (Armstrong, 2016). It's a singularly important show for a variety of reasons.

Clearly, *Seinfeld* hassignificant cultural impact on other shows and movies, but what we wanted to know is, what is its 'reach'? Reach is defined as other media that texts cast and crew from *Seinfeld* worked on before, during, and after their appearance(s) on the show. Such texts exist in every media type (movies, video games, web-based media). When two media items share cast/crew, we look for overlap.

Dr. Conaway worked on the project for two years, using cut and paste and Excel spreadsheets for items and people, before involving Dr. Shichtman, who has created a relational database that may be searched. We first used MySql and an Amazon Web Services server, have recently shifted to the college's virtual machine and the Oracle database management system. We involved two students in a grant funded practicum in the Fall term as well.

Our research revealed that the 1551 cast/crewhad worked on over 32,500 other discrete media texts, starting in 1936, and with many texts still on the air today, often with an overlap of more than one. Nearly every televison series, TV movie, and TV special we could think of included overlap. Only recently, in "peak TV"—in which there are over 500 scripted TV shows in production this year alone, in addition to reality, sports, and news shows (many of which also have overlap)—are we seeing well-known US TV series with no overlap. Our research found that although most were US-based, there were media items from over 60 countries.

Social network theory would help us answer some questions. As Duncan Watts writes in *Six Degrees: the Science of a Connected Age*, „Affiliation networks . . . are . . . networks of overlapping cliques, locked together via the comembership of individuals in multiple groups" (Watts, 2004). Small worlds theory discusses how networks of people influence each other, and each others' connections.

Questions include, what genres did the cast/crew, presumably chosen for a common comic sensibility, work on other than comedy? What genres included the most cast/crew? What genres have less overlap, none at all, and what might be some reasons for that? What is the importance of gender, race, and age?

We looked for other, similar projects that used IMDb and found that there were few that did. Some computer scientists had used IMDb to trace the overlaps among actors involved in 'adult' films in the database as an example of a 'small world' environment. Media History scholars had traced 'race films' that ended before our database started, and Digital Humanities scholars used it to look at patterns of exhibition of films or specifically how Australians worked together, but not to examine how cast circulated among media.

IMDb, it turns, out, is a challenging tool for this purpose. Deb Verhoeven, Associate Dean of Engagement and Innovation of the University of Technology Sidney, who has done a lot of Digital Humanities work on Australian films explained in 2012 that IMDb consists of "elaborated sets of lists" created by fans, writing:

Accordingly, the primary users of filmographic catalogues are not cinema historians, information managers, analytical filmographers, or cinema scholars, but members of the public, film buffs, students and so on who are content to navigate these databases using the small number of structured search fields provided. (Verhoeven, 2012)

IMDb, which started in the early 1990s, is very robust, and provides information for free download using Python, but is not usable 'as is.' Entries may be misleading, incomplete, or unclear, with genres in particular organized in unhelpful ways. The Downloadable information includes the full cast and some types of crew members, but not others. In addition, the fields of the two faculty members made shared vocabulary difficult, and getting complete and clean data that could be turned into tables and graphs meant conducting additional research outside of IMDb, and reorganizing the data significantly from the way Dr. Conaway initially tagged it. SUNY Empire State College also lacks the structures that many institutions have for conducting Digital Humanities work.

However, we have been able to create some early data visualizations that will show a microcosm of how the US entertainment industry works for various types of actors and crew members, using specifically the data from television programs. We've compared *Seinfeld*'s numbers of actors and crew to that of other shows, analyzed how the media items break down by genre, and visualized how women's careers wax and wane in different patterns from men's careers. In the future we will do the same for subgenres, actors of color, and actors of various age groups.

References

Armstrong, J.K., 2016. *Seinfeldia: How a Show about Nothing Changed Everything*. Simon and Schuster.

Bajak, A. 2017. Seinfeld, big data and measuring the Internet's emotional landscape. *Mediashift.*

Gold, M.K. 2012. *Debates in the Digital Humanities*. University of Minnesota Press.

Gold, M.K. and Klein, L.F., 2014. *Debates in the Digital Humanities*. University of Minnesota Press.

Lavery, D. And Dunne, S.L. 2006. *Seinfeld, Master Of its Domain*. New York: Continuum.

Verhoeven, D. New cinema history and the computational turn. Beyond Art, Beyond Humanities, Beyond Technology: A New Creativity, Proceedings Of the World Congress Of Communication and the Arts Conference, University Of Minho, Portugal. 2012

Watts, D.J. 2004. *Six Degrees: The Science Of a Connected Age*. WW Norton & Company.

Exploring Big and Boutique Data through Laboring-Class Poets Online

Cole Daniel Crawford
cole_crawford@fas.harvard.edu
Harvard University, United States of America

Though quantitative methods are becoming increasingly common within the humanities, few researchers readily describe their primary texts as data. Most prefer to see their objects of study as contextually situated and socially constructed entities with independent value that resist complete digital representation. Miriam Posner argues that for many humanities researchers, describing an artifact as data implies "that it exists in discrete, fungible units; that it is computationally tractable; that its meaningful qualities can be enumerated in a finite list; and that someone else performing the same operations on the same data will come up with the same results." Defined this way, digital artifacts and metadata seem to simultaneously insist on particular interpretations and to be bereft of deeper meaning outside of an aggregate state, thereby resisting the hermeneutic methodologies which form the core of humanistic inquiry.

This position stems from understanding data primarily through a big data mindset. As corporations, governments, and universities have increasingly addressed business problems by embracing data analytics, the essential qualities of big data (large volume, high velocity, and heterogenous variety) have created the illusion among many that such datasets can perfectly model an imperfect and unpredictable world, gaining credibility simply by increasing in volume. The computational authority of big data is persuasive because it presents a seemingly objective, number-driven way of knowing reality – an epistemology of the database, predicated on scale, comprehensiveness, and reproducibility.

While an immense and complete archive possesses an undeniable allure (Manovich, 2012; Kaplan, 2015), there is still value in examining individual records and investigating the intangible stories and datapoints that hide in database gaps or reside outside of databases entirely. I use Cheryl Ball et al's term "boutique data" to emphasize the ongoing importance of small, localized, partial, and qualitative datasets to the humanities research process. I frame boutique data as both a thing (a boutique dataset) and a theoretical approach to data-intensive work in the humanities. While big data are often automatically generated, boutique data are manually curated – subjective, created *capta* as opposed to given data (Drucker, 2011). Big data hides the work and decisions that drive data processing, while boutique data foregrounds the hidden labor and assumptions that shape data. Big data fits information into a predetermined mold, while boutique data models are built from the bottom up. Where a big data mindset treats gaps in data coverage as a corrupting null to be fixed, a boutique approach to data sees these gaps not as empty voids but as evocative absences worth further investigation. In this presentation, I will examine both the successes and failures of a boutique approach to data through a case study of *Laboring-Class Poets Online* and speculate about possible future improvements to the project.

The texts and histories studied by scholars of laboring-class culture are riddled with gaps. Since the publication of E. P. Thompson's *The Making of the English Working Class* over fifty years ago, researchers have increasingly viewed laboring-class poets and their writing as subjects worthy of scholarly inquiry. Rather than portraying proletarian writers as isolated anomalies or novelties, such as how George Thomson characterized Robert Burns as a "heav'n taught ploughman" in his famous obituary for the Scottish bard, modern critics acknowledge that working-class writing was a significant, widespread phenomenon. However, while some British laboring-class poets such as Burns or John Clare have achieved near-canonical status, most of these writers are still obscure figures. Information on their lives and access to their writing remains scarce and scattered, hindering research on both their personal histories and their poetry.

Laboring-Class Poets Online (*LCPO*) addresses this gap by aggregating biographical and bibliographical information about the more than 2,000 British laboring-class poets who published between 1700 and 1900 and the texts they produced. *LCPO* draws on collaborative research initially collected by an international distributed team of researchers over several decades and presented as biographical entries in *A Database of British and Irish Labouring-Class Poets and Poetry*. *LCPO* transforms these freeform biographical snippets into structured, web-accessible records. This structure facilitates a pro-

sopographic approach to British working-class literary studies. Lawrence Stone defines prosopography as "the investigation of the common background characteristics of a group of actors in history by means of a collective study of their lives." This methodological shift from the study of individual biographies to collective biographical and bibliographic patterns enables a more comprehensive understanding of laboring-class literary production at a time of great social and economic change. Users can ask questions about laboring-class literature holistically and map trends and themes, including the impact of industrialization; the role of religion as a vehicle for literacy and a source of aesthetic influence; the tension between increased urbanization and a celebration of regional identity, often demonstrated through writing in dialect; the transformation of the publishing industry and the role of patronage and subscription publishing; the growth of literary miscellanies and magazine publishing; and the influence of organized labor movements (e.g., Chartism or Christian Socialism) on laboring-class artistic expression. Scholars can investigate emigration patterns, education level, labor engagement, health outcomes, poet occupations, and interactions with the criminal justice and social relief systems. Publications can similarly be filtered and searched by typical facets such as publication date, author, or location, but also by subscription lists, patronage, cost, or print run size.

Users can interact with aggregate data through numerous data visualizations including geographic maps that show poet and publication locations; timelines of individual lives or major events which shaped the working classes; and network graphs that display connections between writers based on correspondence, personal relationships, or literary influence. Each of these visual forms encourages users to shuttle back and forth between individual records and aggregate analysis. Users can also create collections of content for further interpretation and analysis, correct mistakes in poet entries, or contribute new data to the website. All data presented through *Laboring-Class Poets Online* are freely available for download or access via a REST API.

This information is vital for scholars of working-class writing and culture, but it is also an instance of boutique humanities data (capta): a collaboratively and manually created and curated small dataset of several thousand entities extracted during ongoing research. While scholars often use context to interpret data points in historical documents, databases and computational methods typically lack this capability. Uncertainty is embedded in historical sources, but databases often strip away ambiguity to perform the computational functions that make their use worthwhile. By taking a boutique approach to historical and literary information, *LCPO* retains much of this ambiguity and offers insight into how humanities researchers can accommodate a complex understanding of space and time as continuously unfolding events.

References

Ball, C., Graban, T. S. and Sidler, M. (Forthcoming). The Boutique is Open: Data for Writing Studies. In Rice, J. and McNely, B. (eds), *Networked Humanities: Within and Without the University*. Parlor Press.

Drucker, J. (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly*, 5(1) http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html

Kaplan, F. (2015). A Map for Big Data Research in Digital Humanities. *Frontiers in Digital Humanities*, 2 doi:10.3389/fdigh.2015.00001. http://journal.frontiersin.org/article/10.3389/fdigh.2015.00001/abstract

Goodridge, J. (ed). (2017) *A Database of British and Irish Labouring-Class Poets and Poetry, 1700-1900.*

Manovich, L. (2012). Trending: The Promises and Challenges of Big Social Data. In Gold, M. (ed), *Debates in the Digital Humanities*. University of Minnesota Press.

Posner, M. (2015). Humanities Data: A Necessary Contradiction *Miriam Posner's Blog* http://miriamposner.com/blog/humanities-data-a-necessary-contradiction/

Stone, L. (1972). Prosopography. In Gilbert, F. and Graubard, S. (eds), *Historical Studies Today*. New York.

Organizing communities of practice for shared standards for 3D data preservation

Lynn Cunningham
lynncunningham@berkeley.edu
University of California Berkeley, United States of America

Hannah Scates-Kettler
hannah-s-kettler@uiowa.edu
University of Iowa, United States of America

Scholars are producing and using 3D content more than ever due the advancement and availability of 3D technology. How is this 3D content and its metadata being captured, disseminated, and preserved? How is this digital scholarship being made available and discoverable for pedagogical and research purposes?

Although there is great interest in 3D applications in research, there is currently little available guidance regarding the preservation of digital objects and associated information in perpetuity. The preservation and sharing of research data is a necessary, invaluable responsibility of libraries, museums, and other cultural heritage institutions, and although standards and best practices have been developed for many kinds of digital data to ensure assets can be accessed and reused in perpetuity, the applicability of these standards to 3D data is limited.

Building off the discoveries made during the 2015/2016 NEH Advanced Challenges in Theory and Practice in 3D Modeling of Cultural Heritage Sites, this paper explores one of the main threads of discussion throughout the NEH Summer Institute: research longevity and publication. Underpinning the issue was concerns of the preservation of 3D data and their overall discoverability and (re)use beyond their creation.

This paper investigates the current state of existing standards and schemas for 3D data and explores what more needs to be done (and is being done) by practitioners, librarians and curators to ensure that this digital content is preserved and disseminated, enabling further humanistic inquiry and advancing scholarship of our shared cultural heritage.

In 2017 the Institute for Museum and Library Services received several proposals regarding the advancement of 3D research and support. Two of these grants were funded which are working in tandem to discuss issues related to 3D and virtual reality, and preservation and best practices for 3D data curation. This paper will focus on the developments regarding the latter IMLS grant - the Community Standards for 3D Data Preservation (CS3DP). According to the CS3DP grant proposal (Moore et.al., 2017):

> The project team surveyed an international community including individuals involved in digital curation and 3D data acquisition and research, primarily at universities and museums. Of 104 respondents 70% said that they did not use best practices or standards for preservation, documentation, and dissemination of 3D data. Of those not using standards/best practices, 69% said that they did not use them because they were unaware of such standards.

In order to respond to the lack of consensus around 3D data standards, the grant team will develop "a community-developed plan to move 3D preservation forward [and] recommendations for standards and best practices" for data creators and preservation specialists alike (Moore et. al., 2017). By the time of the 2018 DH conference, the CS3DP grant will have convened around 70 data creators and professionals to address the issues of 3D data preservation. This paper will report on initial findings and ongoing discussions and areas of work, as well as solicit feedback from the DH conference goers about other areas of concern, development and needs.

References

3D-ICONS Guidelines and Case Studies. https://pro.europeana.eu/files/Europeana_Professional/Projects/Project_list/3D-ICONS/Deliverables/3D-ICONS%20Guidelines%20and%20Case%20Studies.pdf (accessed 27 April 2018).

Advanced Challenges in Theory and Practice in 3D Modeling of Cultural Heritage Sites. https://advanced-challenges.com/ (accessed 27 April 2018).

Alliez, P., Bergerot, L., Bernard, J.-F., Boust, C., Bruseker, G., Carboni, N., Chayani, M., et al. (2017). *Digital 3D Objects in Art and Humanities: Challenges of Creation, Interoperability and Preservation. White Paper : A Result of the PARTHENOS Workshop Held in Bordeaux at Maison Des Sciences de l'Homme d'Aquitaine and at Archeovision Lab. (France), November 30th - December 2nd, 2016.*

Cook, M., Hall, N., Laherty, J. (2017). Developing Library Strategy for 3D and Virtual Reality Collection Development and Reuse. IMLS grant proposal: https://www.imls.gov/sites/default/files/grants/lg-73-17-0141-17/proposals/lg-73-17-0141-17-full-proposal-documents.pdf (accessed 27 April 2018).

D'Andrea, A. and Fernie, K., Addison, A. c., De Luca, L., Guidi, G. and Pescarin, S.(2013). CARARE 2.0: A metadata schema for 3D cultural objects. *2013 Digital Heritage International Congress (DigitalHeritage)*, vol. 2. pp. 137–43 doi:10.1109/DigitalHeritage.2013.6744745.

Moore, J., Rountrey, A., Scates Kettler, H. (2017). Community Standards for 3D Data Preservation (CS3DP). IMLS grant proposal: https://www.imls.gov/sites/default/files/grants/lg-88-17-0171-17/proposals/lg-88-17-0171-17-full-proposal-documents.pdf (accessed 27 April 2018).

Guidi, G., Micoli, L. L., Gonizzi, S., Navarro, P. R. and Russo, M. (2013). 3D digitizing a whole museum: A metadata centered workflow. *2013 Digital Heritage International Congress (DigitalHeritage)*, vol. 2. pp. 307–10 doi:10.1109/DigitalHeritage.2013.6744768.

Legacy No Longer: Designing Sustainable Systems for Website Development

Karin Dalziel
kdalziel@unl.edu
University of Nebraska–Lincoln, United States of America

Jessica Dussault
jdussault@unl.edu
University of Nebraska–Lincoln, United States of America

Gregory Tunink
techgique@unl.edu
University of Nebraska–Lincoln, United States of America

Introduction

The Center for Digital Research in the Humanities (CDRH) at the University of Nebraska–Lincoln is home to digital collections such as *The Walt Whitman Archive*, *The Willa Cather Archive*, *The Journals of Lewis and Clark*, and *O Say Can You See*. These projects contain overlap between subjects, individuals, and locations, yet are siloed, and many

are built in aging, unsupported technologies with no interoperability or common search. In order to address this, the Center has developed an API ("Henbit") as part of a modular software stack to index and display data and content.

Challenge

Over the past twenty years, the Center has created over 30,000 TEI files in addition to other data sets such as VRACore documents, spreadsheets, and databases. Sites showcase the content and metadata of these files using a variety of technologies, many of which are no longer maintained. In addition, some sites used commercial software which became unsustainable when costs went up, cementing a commitment to open source. This experience informed and reinforced our adopted design philosophy, which can be summed up as:

- Keep it simple, stable, and sustainable
- Embrace modularity by writing software for one purpose
- Avoid over-engineering solutions (i.e. graphical interfaces where command-line will do)
- Provide comprehensive documentation

The Center has been inspired to think bigger about what can be accomplished by including existing data in a new framework. An exciting next step is creating a site to search all Center data, find commonalities between projects, and read materials across sites for comprehensive research. This approach will also help solve accessibility issues of older project sites which do not meet modern requirements. As projects become unsustainable, the Center may retire them while keeping all content available.

While having one place to view and search the Center's data is important, it's also critical to allow the creation of independent sites which utilize unique organization and include special features requested by principal investigators for new and evolving projects. Quickly creating bare bones sites to view in-progress TEI is essential, as it allows metadata experts and PIs to refine their data and arguments. Such sites should be written for ease of maintenance, freeing future developer time to work on new projects rather than sustaining old ones.

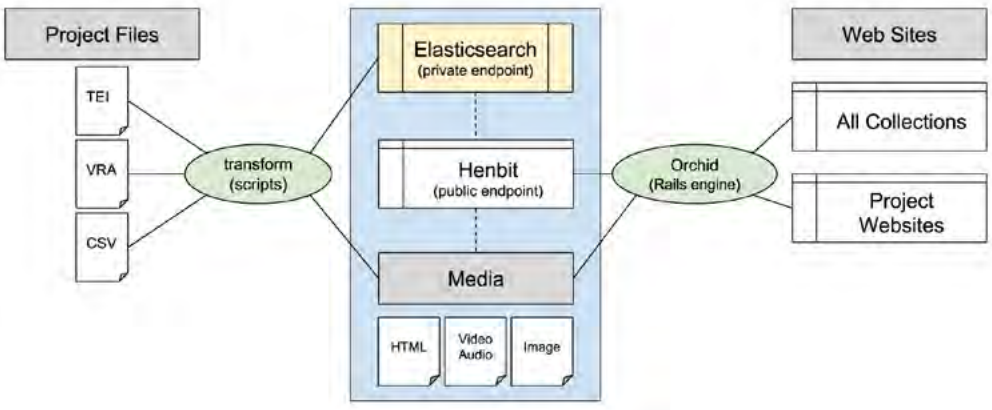
Solution

The Center explored the possibility of using existing software to address these challenges, such as XTF, Blacklight, and Fedora. These packages did not fit the Center's needs; though comprehensive, they were not flexible enough to accommodate the variety of document types and project site requirements. Additionally, many solutions would lock the API into using Solr instead of allowing an interchangeable search engine (Blacklight, 2017; DuraSpace, 2017).

Instead of heavily customizing existing software, The Center decided to create a modular solution. The system consists of several components:

- data repository for project files and scripts for transformation
- document datastore and search engine (Elasticsearch)
- Ruby on Rails (Rails) API to serve data (Henbit)
- media retrieval system for associated images, audio, and video
- template generator for rapid website creation (Orchid)

With a modular software stack, future changes in technology and project needs can be accommodated with independent upgrades rather than massive redesigns and rewrites.



Project Files and Scripts

The data repository houses original files for projects, such as TEI-XML, VRACore, CSV, and Dublin Core. The repository also contains CLI scripts which create HTML and populate search indexes with document content and metadata (CDRH, 2017a). New projects use generalized scripts, which are organized to allow overriding functionality in individual projects. Older websites may continue to use existing XSLT and populate legacy Solr indexes while their existing sites are supported, as well as populate Elasticsearch using the standardized script. Static HTML files derived from this process are used to create a document which can be viewed in a browser, regardless of the original data format.

Henbit (Public Endpoint)

Henbit is a Rails powered API (application program interface) which creates appropriate requests for the backend index, and returns JSON. Currently, Henbit uses Elasticsearch as a backend, but most of its features (sorting, filtering, aggregating on ranges, etc) could be ported to a different backend. The OpenAPI specification was used during Henbit's creation to fit current design practices (CDRH, 2017b).

Media Retrieval

In legacy sites, associated media lived inside the website directory. The Center has created a standard URL path for media files. It will be easier to optimize serving specific file types with this common retrieval structure. In the near future, the CDRH will be implementing a IIIF image server to serve images of varying sizes and resolutions.

Orchid (Rails Engine)

Orchid is a Rails engine which connects Rails 5 applications and Henbit. Orchid and a supporting gem, api_bridge, provide a template website that allows users to browse, search, filter, and view documents. This template is highly customizable, and can be altered to allow different URLs, search behavior, and anything possible in Rails (CDRH, 2017c).

Current Implementation and Future Plans

Beta versions of all components were released in 2017. In late 2017 the framework was used to build *The Complete Letters of Willa Cather* (launched January 2018). *The Complete Letters* demonstrates the customization which can be accomplished with this modular system. The CDRH is currently developing another project, *Family Letters*, which will also take advantage of the data repositories, scripts, Henbit, and Orchid template.

In the meantime, older websites are being converted for the new system. Updated documents and original XSLT have been reorganized into the structure required by the data repository scripts and are being posted to the Elasticsearch index. Once a site for Centerwide projects has been created, older sites can be retired as needed, replaced by content now available through the new API and supporting website.

The decision to use custom built software rather than an existing, out of the box solution, was not easy. Though at times it felt like reinventing the wheel, our highly customizable and flexible implementation prepares for future technological developments and enables flexibility in meeting project requirements.

Notes

https://cdrh.unl.edu
http://whitmanarchive.org, http://cather.unl.edu, https://lewisandclarkjournals.unl.edu, and http://earlywashingtondc.org
https://xtf.cdlib.org, http://projectblacklight.org, and http://fedorarepository.org
https://github.com/CDRH/data
https://github.com/CDRH/api
https://github.com/OAI/OpenAPI-Specification
http://iiif.io
https://github.com/CDRH/orchid
https://github.com/CDRH/api_bridge
http://cather.unl.edu/letters

References

Blacklight (2017). "Project Blacklight." http://projectblacklight.org.

CDRH (2017a). "CDRH Data Repository." *GitHub*. https://github.com/CDRH/data.

CDRH (2017b). "Henbit." *GitHub*. https://github.com/CDRH/api.

CDRH (2017c). "Orchid." *GitHub*. https://github.com/CDRH/orchid.

DuraSpace (2017). "Fedora Repository." http://fedorarepository.org

Histonets, Turning Historical Maps into Digital Networks

Javier de la Rosa Pérez
versae@stanford.edu
Center for Interdisciplinary Digital Research, Stanford University, United States of America

Scott Bailey
scottbailey@stanford.edu
Center for Interdisciplinary Digital Research, Stanford University, United States of America

Clayton Nall
nall@stanford.edu
Department of Political Science, Stanford University, United States of America

Ashley Jester
ajester@stanford.edu
Center for Interdisciplinary Digital Research, Stanford University, United States of America

Jack Reed
pjreed@stanford.edu
Digital Library Systems and Services, Stanford University, United States of America

Drew Winget
awinget@stanford.edu
Digital Library Systems and Services, Stanford University, United States of America

Introduction

The study of communication networks, specifically road networks, is a topic of broad interest to the scholarly community. It allows researchers to draw conclusions that range from historical events (Antrop, 2004; Trombold, 1991) to transit and traffic (Bash et al., 2017; Yang and Yagar, 1995), while adding a tangible and understandable dimension to their work. The appearance of Geographical Information Systems (GIS) made it possible to perform such analysis efficiently and accurately. It is just recently that the study of topological and growth properties of road networks are giving us the chance of understanding the bigger picture of cities (Antrop, 2005; Kasanko et al., 2016).

In the American landscape, network analysis of road networks has shown evidence that the construction of interstate highways affected the political and geographic polarization of cities, undermining representation and posing a threat to democracy itself (Nall, 2015; Ejdemyr et al., 2005). Most of these studies, however, rely on "the only rigorous year-to-year record of the construction of interstate highways and the incorporation of existing freeways into the system" (Nall, 2018), the Federal Highway Administration PR-511 database (FHWA PR-11). While the FHWA PR-11 is the most complete database available, it is based on highway construction records, which oftentimes misrepresent the complexity of turning political promises into reality, and does not include data on the development of road networks before the interstates. One way to approach this lack of data is to resort to roadmap collections, which might be a better proxy to understand the reality of transportations networks. Unfortunately, despite the number of digitized and scanned map collections, the lack of their availability in standard network data formats still represents a burden for the study of historical road networks. Although network analysis tools exist, we are not able to fully leverage their potential regarding historical datasets without a huge amount of manual work to generate network data.

As an alternative, modern approaches of road extraction from maps promise fully automated methods that rarely generalize well (Mena, 2003, Sharma et al., 2013), or rely on good quality labeled data (Isola et al., 2016), which is non existent or very difficult and costly to gather. We are then left to semi-automated methods where the researcher is guided to enter some crucial information needed for the automated process to start. However, these methods are usually conceived for satellite imagery or raster images of maps, lacking proper support for the variety of style and format found when dealing with collections of historical maps, and producing vector information not in network format. In order to fill this gap, we are presenting Histonets, a web-based platform to assist in the conversion of historical maps into digital networks, turning intersections into nodes and roads into edges.

Methodology

The platform begins with a login screen, after which each researcher can create a number of collections of images of maps by linking them from IIIF-compliant repositories. Furthermore, researchers are able to create settings for similar images (according to their criteria). Once images are selected, the pipeline for the Histonets platform is comprised of 4 steps: image preparation and cleaning, pattern matching, pathfinding, and graph correction. Cleaning can be fully automated or fine-tuned by adjusting the parameters of several actions to be applied. Once clean, image color depth is reduced by an automatic color clustering algorithm that only needs the final number of colors (defaults to 8).

With the image clean and posterized, the pattern matching step begins. In order to identify intersections and corners that will eventually become the nodes of the graph, researchers must circle around them, and, with a couple of samples, Histonets will try to find other instances in the images, taking into account rotation and orientation of the templates. Identifying roads is done by selecting their colors and a threshold. Areas under a certain threshold are removed as well. A final preview of the resulting graph is shown for the whole image. If the graph complies with the expectations the researcher can start a batch process to apply the same parameters to the whole collection. The tasks can be monitored and canceled. The final result of the process for each image map is a downloadable file in a compatible graph format, including Gephi and GraphML (sees Figure 1).

Figure 1. Sample of image input (upper left), internal output (upper right), and final graph as produced by Histonets (lower)

Discussion

Although in early stages, Histonets has already proved to reduce substantially the amount of hours of manual labour, cutting down the time needed to process an entire collection. Moreover, the easy parallelization built-in in Histonets is only limited by the computational resources available, making it easier for cloud or high performance computing center deployments to further boost its performance. However, without a proper benchmarking framework it is still difficult to assess its accuracy and completeness. One of our goals moving forward is to test and measure these factors, and adjust the platform for greater reliability.

While Histonets, as a whole pipeline, is focused specifically on extracting road networks from historical maps, collaborators have already identified uses outside of Political Science or History. As a general low-barrier and user friendly computer vision application, we have shown it to be useful for identifying capital letters in Medieval manuscripts, counting glyphs in Egyptian hieroglyphs, or even identifying architectural features. With its balance between meeting specific research needs and generalizable applicability, Histonets has a bright future as an adaptable tool in the Digital Humanities.

References

Antrop, M. (2004) Landscape change and the urbanization process in Europe. *Landscape and urban planning* 67.1, pp. 9-26.

Antrop, M. (2005) Why landscapes of the past are important for the future. *Landscape and urban planning* 70.1, pp. 21-34.

Bast, H., et al. (2017) Fast routing in road networks with transit nodes. *Science* 316.5824, pp. 566-566.

Champion, T. (2001) Urbanization, suburbanization, counterurbanization and reurbanization. *Handbook of urban studies* 160: 1.

Ejdemyr, S., Nall, C., and O'Keefe, Z. (2015) Building Inequality: The Permanence of Infrastructure and the Limits of Democratic Representation.

Isola, P., et al. (2016) Image-to-image translation with conditional adversarial networks. *arXiv preprint* arXiv:1611.07004.

Mena, J. B. (2003) State of the art on automatic road extraction for GIS update: a novel classification. *Pattern Recognition Letters* 24.16, pp. 3037-3058.

Nall, C. (2015) The political consequences of spatial policies: How interstate highways facilitated geographic polarization. *The Journal of Politics* 77.2, pp. 394-406.

Nall, C. (2018) The Road to Inequality: How the Federal Highway Program Polarized America and Undermined Cities. Cambridge University Press.
Kasanko, M., et al. (2016) Are European cities becoming dispersed?: A comparative analysis of 15 European urban areas. *Landscape and urban planning* 77.1, pp. 111-130.
Sharma, N, Bedi, R., and Dogra, A. K. (2013) A Survey on Road Extraction from Color Image using Vectorizaton. *IJRET: International Journal of Research in Engineering and Technology* 2.10.
Trombold, C. D. (1991) ed. *Ancient road networks and settlement hierarchies in the New World.* Cambridge University Press.
Yang, H., and Yagar, S. (1995) Traffic assignment and signal control in saturated road networks. *Transportation Research Part A: Policy and Practice* 29.2, pp. 125-139.

Alfabetización digital, prácticas y posibilidades de las humanidades digitales en América Latina y el Caribe

Gimena del Rio Riande
gdelrio.riande@gmail.com
CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas), Argentina

Paola Ricaurte Quijano
ricaurte.paola@gmail.com
Tecnológico de Monterrey, México

Virginia Brussa
virbrussa@gmail.com
Universidad Nacional de Rosario, Argentina

Atraviesan al concepto étnico-geográfico definido como Latinoamérica distintos procesos regionales en los que se observan políticas para impulsar estrategias de acceso a internet, incorporación de las tecnologías digitales al sistema educativo y/o implementación de programas de alfabetización digital. Desde la Cumbre Iberoamericana de San Salvador en el año 2008 se viene sosteniendo, por ejemplo, la necesidad de "impulsar políticas, que incluyan el marco de la colaboración público-privada, encaminadas a facilitar la integración plena de las y los jóvenes en la Sociedad de la Información y del Conocimiento a través del acceso universal a las Tecnologías de la Información y de la Comunicación (TIC) y el desarrollo de contenidos digitales, mediante programas de alfabetización digital que reduzcan la brecha existente y con la mira puesta en facilitar el acceso al empleo, el emprendimiento y la realización personal" (INTEF, 2013).

Pero ¿qué es la alfabetización digital en el marco de un campo científico como el de las Humanidades Digitales, un nuevo espacio de producción académica nacido bajo el amparo de las Digital Humanities del norte global? Como es sabido, las Humanidades Digitales (HD) se han consolidado como un campo académico en franca expansión, principalmente en países de habla anglosajona. Así y todo, su recepción ha sido diferente para nuestra región y, al día de hoy, no se han absorbido en el currículo universitario o actividades de investigación del mismo modo. Las HD dan cuenta de un diálogo entre las humanidades y la informática entendida como digitalidad, y también de la posibilidad de crear nuevos objetos de estudio y líneas de investigación mixta, aunque, tal vez la apuesta más interesante y menos apreciada de las HD sea los puentes interdisciplinares que tienden y ofrece a las distintas disciplinas humanísticas (del Rio Riande, 2016).

Si bien la alfabetización digital y el desarrollo de competencias supone mucho más que infraestructuras, la posibilidad de acceso físico, real y efectivo a las tecnologías, así como el desarrollo de políticas institucionales relativas a su impulso siguen siendo un desafío para el crecimiento de las HD como campo científico. Algunos de los elementos que es necesario considerar tienen que ver con la implantación de una cultura digital que no sea únicamente instrumental sino que implique una reflexión crítica acerca de la relación entre tecnología, humanidades y producción de conocimiento.

Por otra parte, las investigaciones recientes y las políticas educativas nos alertan sobre la importancia de lo que podríamos denominar "multiliteracies" o multialfabetizaciones (Cope & Kalantzis, 2000), incluyendo, en ese sentido, no sólo aspectos de manejo de herramientas, sino del impacto en cómo "leer", "traducir" aquello computacional desde un aspecto crítico. Clave, por ejemplo, es el proceso que contiene a los datos de investigación u objetos intensivos en datos digitales.

Con el fin de indagar acerca del estado de las prácticas digitales en la región, diseñamos una encuesta abierta, orientada a estudiantes, profesores, investigadores, bibliotecarios y documentalistas en América Latina en el marco del proyecto *Prácticas digitales en América Latina y el Caribe* (http://openlabs.limequery.com/954661?lang=es-MX). La encuesta buscó medir el conocimiento y las prácticas de estos agentes de producción en el ámbito académico sobre recursos para la investigación, la publicación científica y la preservación (desde los procesadores de texto, pasando por los repositorios, hasta las infraestructuras digitales). El proyecto se desarrolla en conjunto con Humanidades Digitales CAICYT (Centro Argentino de Información Científica y Tecnológica del CONICET-Argentina), +Datalab del Centro de Investigación en Mediatizaciones (Facultad de Ciencia Política y Relaciones Internacionales) de la Universidad Nacional de Rosario (Argentina) y Openlabs de la Escuela de Humanidades y Educación del Tecnológico de Monterrey (México). Se

recogieron, hasta el momento, más de 300 respuestas de diversos países de América Latina. Una primera versión de esta encuesta se realizó en Argentina en 2015-2016 en el marco del convenio entre CIM-CAICYT de CONICET.

Presentaremos en esta ocasión los resultados obtenidos respondiendo a estos imperativos, discutiendo los hallazgos clave sobre las interacciones entre la investigación, el acceso a la tecnología entre estudiantes cultural y lingüísticamente diversos, como parte del estado de la cuestión en espacios académicos y su incidencia sobre el desarrollo del campo científico de las Humanidades Digitales en América Latina así como de políticas y currículos universitarios más reales y democráticos.

References

Arellano, A. (2007). "De la epistemología de la ecología política latouriana a una epistemología de sustento antropológico". *Convergencia*. *Revista de Ciencias Sociales*, 44 (mayo-agosto).

Cope, B. & Kalantzis, M. (2000). Introduction. In Cope, B. & Kalantzis, M. (eds.), *Multiliteracies: Literacy learning and the design of social futures*. South Yarra, VIC: MacMillan.

Kreimer, P., Vessuri, H., Velho, L. & Arellano, A. (2014). Perspectivas Latinoamericanas en el estudio social de la Ciencia y la Tecnología. México, Siglo XXI.

del Rio Riande, G. (2016). Humanidades Digitales: estándares para su consolidación en el campo científico argentino. *SEDICI*. Repositorio Institucional de la UNLP. http://sedici.unlp.edu.ar/handle/10915/62008

Instituto Nacional y del Profesorado (INTEF) (2013). Declaración sobre innovación y TIC del Foro de Ministros de Educación de las Américas. [Blogpost] *Educalab.* http://blog.educalab.es/intef/2013/06/26/declaracion-sobre-innovacion-y-tic-del-foro-de-ministros-de-educacion-de-las-americas/

Listening for Religion on a Digital Platform

Amy DeRogatis
derogat1@msu.edu
Michigan State University, United States of America

What does religion in the United States sound like, and where should one go to listen for it? What are the different ways that religious individuals and communities make themselves heard--to each other, to their gods, and to others? How is religious pluralism reshaping the sounds and spaces of North American religious life? How might we begin to reconceptualize religion and its place in North American life if we begin by using our auditory perception as a source of knowledge? And how might this knowledge be represented and transformed through the use of new digital media?

I co-direct "The American Religious Sounds Project," a collaborative initiative of Ohio State and Michigan State Universities to leverage opportunities afforded by the new digital environment to consider what religion sounds like in the United States. The project centers on (1) the construction of a unique sonic archive, documenting the diversity of everyday American religious life through newly produced field recordings, interviews, oral histories, and related materials; and (2) the development of a new digital platform and website, which draws on materials in our archive to engage users in telling new stories about religious diversity in the U.S. This multi-modal platform includes a searchable archive, database-driven visualizations, which invite users to explore, discover, and listen for surprising connections among our materials, and a curated gallery of multimedia exhibits, which allow for greater interpretation and contextualization. Future phases include plans for museum installations, traveling exhibits, and community-based workshops.

It has become commonplace (if arguably inaccurate) to describe the United States as the most religiously diverse country in the world. Scholars of North American religions have recognized the pressing need for new approaches to documenting and making sense of this diversity. Our approach stems from our particular interests in the material and sensory cultures of American religions and in the varied ways that religion has become newly visible and audible in American life, confounding once dominant assumptions about secularization and privatization. Rather than retreating quietly into an interiorized or immaterial realm of personal belief, religion has remained an integral feature of the modern world, and religious communities have inscribed themselves on urban landscapes and soundscapes in a variety of ways.

The working we are doing through the American Religious Sounds Project also has been stimulated by a "sensory turn" in scholarship across the humanities and social sciences. Historians, anthropologists, geographers, and others have been attending to the cultural values and social ideologies expressed through different ways of sensing the world and to the multi-sensorial modes through which modern culture was constituted. The nascent field of sound studies, defined broadly as the cultural study of sound and listening, has proven particularly generative, giving rise to new ways of thinking about critical questions that have long animated humanistic inquiry, including the legacies of industrialization and urbanization, the role of technological production and mediation, and the construction of ethnic, racial, religious, sexual, gendered, and class-based differences. Research on sound and through sound provides a rich medium for understanding religious groups, people, events, and conflicts.

Religious studies scholars, however, have paid far more attention to visual and material culture than to audi-

tory culture. In part, this can be attributed to the limitations of the textual media through which scholars have traditionally presented their research, including published monographs and journal articles. Such media have not readily lent themselves to engagement with sonic materials, for sound can be difficult to represent in such formats. Acutely sensitive to this problem, many ethnomusicologists and sound artists have begun experimenting with digital tools and platforms, like soundmapping, but such approaches have not yet made their way into the discipline of religious studies. Scholars of religion should take greater advantage of the opportunities afforded by the new digital environment, while also reflecting critically on its limitations. The American Religious Sounds Project is designed to do both.

Our sound selections are robustly multi-religious, including a wide range of Christian and non-Christian traditions. We include the formal sounds of religious institutions, such as prayer, chanting, and hymns, as well as the informal, and often unintentional, sounds that arise during relaxed coffee hours and spontaneous conversations, ambient and incidental noises like laughter and crying, clapping and shouting, and the shuffling and movement of lived community during worship. We record regular weekly and daily services, as well as seasonal festivals and other special events. We move outside of formal religious institutions to capture the sounds of devotion in homes and schools, public parks and interfaith chapels, coffee shops and workplaces, as well as at ostensibly "secular" gatherings such as a school graduation, public arts festival, or college football games. For example, our researchers recently recorded the sounds of a public Christmas tree lighting, an interfaith prayer vigil against violence, a neo-Pagan brewing mead in his home kitchen, an anti-Islam protest rally, a (secular) Sunday Assembly meeting in a coffee shop, and a Bhutanese Nepali Hindu festival. By casting our net widely, we aim to build a resource that is broadly comprehensive, comparative, and even a bit provocative. We do not intend to answer definitively the question of what counts as religious, but to invite critical reflection on what is at stake in that designation and to consider the role that auditory perception plays in its constitution.

In this paper, I will introduce the project and present our website, which we expect to launch in March 2018. I will solicit critical feedback and offer reflections of my own on the capabilities and limits of new digital methods for enhancing our research of the varied sonic cultures of North American religious life. One of the goals of the American Religious Sounds Project is to provide a bridge between our academic settings and our local communities. That work must be done carefully and respectfully in the present political and religious climate of the United States. I will end with some thoughts on the precarious work of the public presentation of religious sounds and communities on an open accessible digital platform.

Words that Have Made History, or Modeling the Dynamics of Linguistic Changes

Maciej Eder
maciejeder@gmail.com
Institute of Polish Language (Polish Academy of Sciences), Poland; Pedagogical University in Kraków, Poland

Introduction

In the last decades, quantitative linguistics (following exact and social sciences) has developed a considerable number of statistic methods providing an insight into measurable phenomena of natural language. Although to a lesser extent, it also applies to the analysis of diachronic changes. The basic tool used to assess the chronology of linguistic changes is a rather effective yet simple method of trend search: the examined features are analyzed by mapping the frequency of the described phenomenon on a timeline (Ellegård, 1953). This timeline-centric visualization has become a standard in several studies and corpus tools. The most spectacular example is the corpus of several dozens of million of documents (mainly in English) accompained by the service Google Books Ngram Viewer http://books.google.com/ngrams, which, according to its authors, enables to examine changes taking place not only in the language, but also in culture (Michel et al., 2011).

A significant drawback of simple graphic representation of the trend, and hence of mapping the frequency of the examined phenomenon on a timeline, is a tacit assumption that the researcher knows in advance which elements of the language are subject to change. In other words, the method of plotting and inspecting the trend may be applied only to verify hypotheses stipulated earlier by traditional diachronic linguistics. For example, knowing in advance that Polish underwent the gradual replacement of the inflected ending *-bychmy* with *-byśmy*, one might draw the trendline and capture the dynamics of that change. Although many prominent diachronic works were based upon such an approach (Biber, 1988; Hilpert and Gries, 2009; Hu et al., 2007; Reppen et al., 2002; Smith and Kelly, 2002; Can and Patton, 2004), one might be interested in trend search without any *a priori* selection of the analyzed linguistic changes to be traced.

Needles to say, *some* selection of potential language change predictors (e.g.a predefined set of words, certain collocates, etc.) will always be the case. The strategy followed in this study was to analyze a considerably large set of 1,000 most frequent words without any further filters, with the assumption that some of them will turn out stronger than others. Arguably, in such a big set one should find a few dozen of function words, and a vast majority of content words. Another remark that has to be formulated here is that the language change cannot be reliably separated

from the stylistic drift (e.g.in literary taste of the epoch). This fact is well known in stylometric approaches to style ("stylochronometry"), where the actual changes in the system and stylistic signals of, say, the predominant genres are usually difficult to be told apart.

Supervised classification and the timeline

The most natural strategy to assess the discriminative power of numerous features at a time is to apply one of the multivariate methods. Since none of the out-of-the-box techniques is suitable to analyze temporal datasets, some tailored approaches have been proposed, e.g.using a variant of hierarchical clustering (Hilpert and Gries, 2009; Hulle and Kestemont, 2016). These methods, however, share a common drawback, namely their results are by no means stable. Also, no cross-validation can be considered a downside.

To assess these issues, an iterative procedure of automatic text classification was applied (Eder and Górski, 2016). Its underlying idea is fairly simple: first, we formulate a working hypothesis that a certain year – be it 1835 – marks a major linguistic break. The procedure randomly picks *n* text samples written before and after the assumed break; the samples then go into the *ante* and *post* subsets. In this study, a period of 20 years before and after the assumed break was covered (with an additional gap of 10 years), 500 text samples of 1,000 tokens were harvested into each of the subsets. To give an example: for the year 1835, 500 random samples covering the time span 1810–1830 were picked into the first subset, and another 500 samples from the years 1840–1860 into the second subset. Next, the both subsets are randomly divided into two halves, so that the training set and the test contain 500 samples representing two classes (*ante* and *post*). Then we train a supervised classifier – in this case, Nearest Shrunken Centroids – and record the cross-validated accuracy rates. Then we *dismiss* the original hypothesis, in order to test new ones: we iterate over the timeline, testing the years 1836, 1837, 1838, 1839, … for their discriminating power. The assumption is simple here: any acceleration of linguistic change will be reflected by higher accuracy scores.

Data and results

The above procedure has been applied to the Corpus of Historical American English (COHA), containing ca. 400 million tokens and covering the years 1810–2009 (Davies, 2010). The corpus provides the original word forms, part-of-speech tags, and the base word forms (lemmata). The results reported below were obtained using the lemmatized version of the corpus.

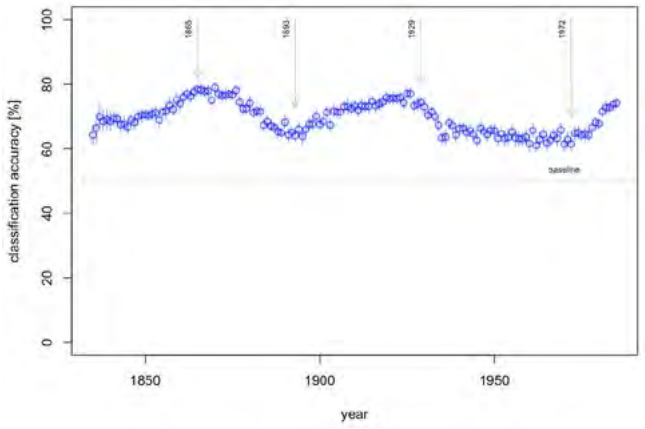


Fig. 1: Language change accelleration in the American English corpus: classification accuracy over the years 1835–1985.

In Fig. 1, the classification accuracy rates for the COHA corpus were shown (1,000 most frequent lemmata, NSC classifier). As one can observe, the scores obtained for each period are higher than the baseline, suggesting the existence of a temporal signal. Obviously, the higher the scores the faster the evolution of language, since the distinction between the period before and after the tested breakpoint is simpler for the classifier. More important, however, is the fact that the scores are not even: the signal becomes stronger in some periods, clearly indicating an acceleration of the language change. One of the stylistic breaks takes place in the 1870s (i.e.after the Civil War), the other in the 1920s (in the period of prosperity before the Great Depression); the third peak is not fully formed yet, even if one can observe an acceleration of language change at the end of the 20th century. Needless to say, any attempts at finding direct correlations between historical events and stylistic breaks are subject to human prejudices, and therefore might introduce substantial bias to the results. Even though, the coincidence of the three observed peaks and a few major changes in the American culture is rather striking.

Distinctive features

The results obtained in the above experiment seem to be rather promising. However, from the perspective of historical linguistic even more interesting is the question which features (words) were responsible for an given change observed in the dataset. It has been reported in several stylometric studies that attributing authorship relies, in most cases, on many features of individually very weak discriminative power. In the context of language change, a similar question can be asked: is it but a few characteristic words that trigger the change, or, alternatively, is the stylistic drift spread across dozens of tiny changes in word frequencies?

To answer the above question, one has to extract the features that played a prominent role in telling apart the *ante* and *post* periods as described above. The features exhibiting the biggest variance (that is, the overall impact on the results) are shown in Fig. 2. An important caveat needs to be formulated here: the plot shows the outputted weights from the classifier, rather than direct word frequencies. The underlying assumption is that the features' weights (to be precise: the *a posteriori* probabilities returned by the classifier) reflect the changes in actual word frequencies as combined with all the other frequencies being analyzed.

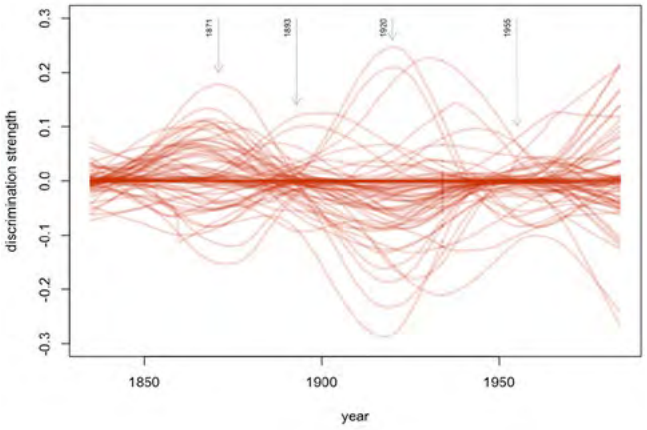


Fig. 2: Seventy-six linguistic features (words) that contributed considerably to the stylistic drift.

The main stylistic breaks form, again, three peaks that culminate roughly in the same years as presented in Fig. 1. What is counterintuitive, however, it is the fact that the features tend to form sinusoidal waves of their periodical discrimination power. Interestingly, these high impact features turned to be very frequent words that usually occupy the top positions on the frequency list. The 25 words of the highest discrimination strength are as follows:

the, *and*, *week*, *that*, *'s*, *last*, *is*, *be*, *of*, *it*, *we*, *i*, *to*, *was*, *mr.*, *our*, *my*, *been*, *not*, *u.s.*, *you*, *new*, *upon*, *there*, *has*

Even more interesting are individual trajectories of the high-impact words. In Fig. 3, one can observe a collinearity of function words: *the*, *and*, *that*, *is*, *been*, as opposed to the possessive *'s*. These function words seem to have impacted the language change at the turn of the 19th century. The possessive, in turn, contributed to the evolution of language roughly at the times of the Prohibition. (Again, this is not to say that any direct links between function words and actual events in history should be drawn).

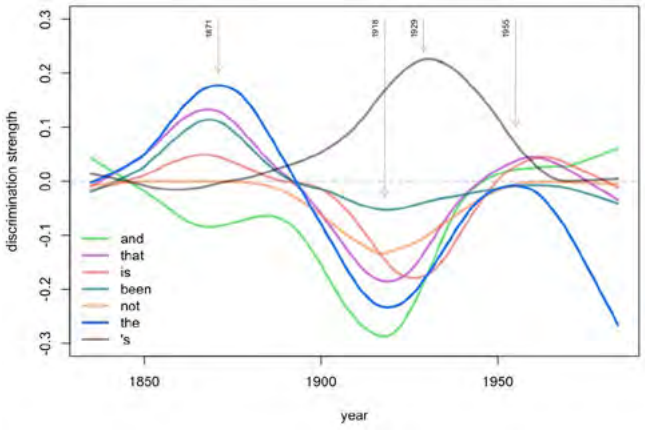


Fig. 3: Function words of the highest impact on the stylistic drift.

A different pattern is revealed by the "social" words, especially personal pronouns. It has been shown that these words, e.g. *I*, play prominent role in betraying someone's personality (Pennebaker, 2011). Certainly, traces of such individual profiles will hardly be noticeable at the

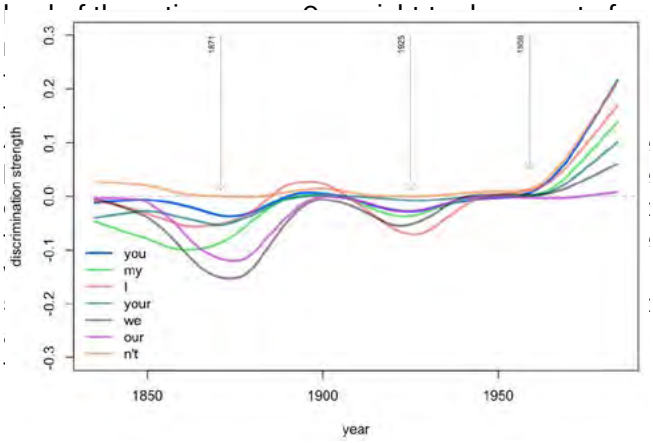


Fig. 4: High impact personal pronouns and contractions.

Conclusions

In this paper, we used a tailored stylometric method to assess the question of language change over time. Our chosen technique proved to be useful indeed, especially when one focuses on tracing the very linguistic features that were responsible for the observed change. The results were counterintuitive, since the set of strongly discriminative features contained common function words, which formed sinusoidal trajectories of their impact over time. One of the most interesting aspects of language development – overlooked in numerous existing studies – is the question of the dynamics of linguistic changes. Our study corroborated the hypothesis that epochs of substantial stylistic drift are followed by periods of stagnation, rather than forming purely linear trends.

Acknowledgements

This research is part of project UMO-2013/11/B/HS2/02795, supported by Poland's National Science Centre.

References

iber, D. (1988). *Variation Across Speech and Writing*. Cambridge University Press.

Can, F. and Patton, J. M. (2004). Change of writing style with time. *Computers and the Humanities*, 38(1): 61–82.

Davies, M. (2010). The Corpus of Historical American English (COHA): 400 million words, 1810-2009 https://corpus.byu.edu/coha/.

Eder, M. and Górski, R. L. (2016). Historical linguistics' new toys, or stylometry applied to the study of language change. In, *Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University & Pedagogical University, pp. 182–84 http://dh2016.adho.org/abstracts/398.

Ellegård, A. (1953). *The Auxiliary Do: The Establishment and Regulation of Its Use in English*. Stockholm: Almquist & Wiksell.

Hilpert, M. and Gries, S. T. (2009). Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4): 385–401.

Hu, X., McLaughlin, J. and Williamson, N. (2007). Syntactic Positions of Prepositional Phrases in the History of Chinese: Using the Developing Sheffield Corpus of Chinese for Diachronic Linguistic Studies. *Literary and Linguistic Computing*, 22(4): 419–34.

Hulle, D. van and Kestemont, M. (2016). Stylochronometry and the Periodization of Samuel Beckett's Prose. In, *Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University & Pedagogical University, pp. 393–95 http://dh2016.adho.org/abstracts/70.

Michel, J.-B., Shen, Y. K., Aiden, A. P., Veres, A., Gray, M. K., Pickett, J. P., Hoiberg, D., et al. (2011). Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014): 176–82.

Pennebaker, J. W. (2011). *The Secret Life of Pronouns: What Our Words Say About Us*. New York: Bloomsbury Press.

Reppen, R., Fitzmaurice, S. M. and Biber, D. (eds). (2002). *Using Corpora to Explore Linguistic Variation*. Amsterdam: John Benjamins.

Smith, J. A. and Kelly, C. (2002). Stylistic Constancy and Change Across Literary Corpora: Using Measures of Lexical Richness to Date Works. *Computers and the Humanities*, 36(4): 411–30.

The Moral Geography of Milton's Paradise Lost

Randa El Khatib
elkhatib.randa@gmail.com
University of Victoria, Canada

John Milton's *Paradise Lost* creates an extraordinarily rich and complex sense of space. The epic poem elegantly captures the cartographical leap of the sixteenth and seventeenth century that owes to advancements in navigation techniques and rapid colonial expansion. The world image was rapidly changing and gaining a more distinct contour as newly colonized lands were becoming better described and known. Maps in this time could often be considered prototypes since cartographers were still experimenting to find a more accurate mimesis of the world. At the same time, the strong foundation of *Paradise Lost* and many other retellings of the Genesis captures the saturation of the seventeenth century in religious tradition and references to sacred places. In this way, *Paradise Lost* can be seen as a prototype of its own that brings together spatial traditions, new and old, real and imaginary, into a single medium. To date, Milton's spatial allusions – spanning biblical, classical, and contemporary temporalities – have predominantly been studied in relation to the textual sources that had influenced them. However, *Paradise Lost* was written at a time when the visual tradition of mapping places of the bible with cartographic exactitude had reached its peak, seen in the King James Bible, which was also Milton's family Bible – a tradition that, in retrospect, is an early example of a geospatial, text-to-map project. Milton construed his spatiality on the existing framework of this visual tradition, and consolidated the geographies of classical antiquity and of his contemporary world. These temporalities were conceived to have progressed on a linear spectrum of geographical continuity, according to the prevalent notion of historical sequence of a seventeenth-century audience. By superimposing these layers, Milton uses textual sources to assign moral valence to geographical points; these inform the readers' understandings of the epic and of the space of human history that it encompasses. The GIS-based digital project, "A Map of the Moralized Geography of Paradise Lost," explores the multi-temporal complexity of Milton's spatial allusions through an open access map depicting the moralized geography of *Paradise Lost*. These multiple temporalities are delineated by various layers of georectified historical maps, including the map that supplies the visual paratext of the King James Bible, as well as John Speeds map of "The Turkish Empire" (1626). The interactive dimensions of the map permit users to recover and evaluate nuance (by resituating geographical names in their poetic contexts) even as they seek to apprehend and deduce larger patterns.

The most powerfully apparent pattern is the concentration of Milton's spatial allusions on the Mediterranean

world, forming a thick chain around the Mediterranean basin. Sites of biblical or classical significance were, in the seventeenth century, in territories almost entirely controlled by the Ottoman Empire; this superimposition creates a polarized dynamic of moral valence. Additionally, Milton's map is coordinated with a map based on place names extracted from the Book of Genesis in order to investigate the scope of influences of the biblical book itself on the epic poem. The extraction of geo-coordinates from both works was carried out manually for the sake of accuracy, since the limitations of present geoparsing techniques with variant and historical place names remain a methodological sticking-point. The Genesis map is less complex than the initial one, making it clear that it was literary and exegetical writings, and religious culture more broadly, that built thick association. This condition reinforces the status of geographical references in Milton's epic as references, as vectors that import or apply associations established through cultural tradition or poetic technique. In this way, *Paradise Lost* functions like an early modern chorography that contextualizes place names at use. The fruit of this project is a navigable visual network that invites users to trace contextualized recurring patterns in multiple temporalities.

References

Galey, A. and Ruecker, S. (2010). How a prototype argues. *Literary and Linguistic Computing*, 25(4), 405–24.

Gillies, J. (1994). *Shakespeare and the Geography of Difference*. Cambridge: Cambridge University Press.

Gregory, I. and Murrieta-Flores, P. (2016). Geographical information systems as a tool for exploring the spatial humanities. In Crompton C., Lane, R.J., and Siemens, R. (eds), *Doing Digital Humanities: Practice, Training, Research*, pp. 177–92.

Hill, L. *Georeferencing*. (2014) Cambridge, MA: The MIT Press.

Jacobson, M. (2014). *Barbarous Antiquity: Reorienting the Past in the Poetry of Early Modern England*. Philadelphia: University of Pennsylvania Press.

Jessop, M. (2008). Digital visualization as a scholarly activity. *Literary and Linguistic Computing*, 23(3), 281–93.

Lim, W. (2010). John Milton, orientalism, and the empires of the east in Paradise Lost. In Johanyak D. and Lim W. (eds), *The English Renaissance, Orientalism, and the Idea of Asia*. Palgrave Macmillan, pp. 203–235.

McLeod, B. (1999). The 'Lordly eye': Milton and the strategic geography of empire. In Rajan B. and Sauer E.(eds.), *Milton and the Imperial Vision*. Duquesne University Press, pp. 48–66.

Milton, J. (2007). *The Complete Poetry and Essential Prose of John Milton*, Kerrigan W., Rumrich, J. and Fallon S. (eds). The Modern Library.

Ng, M. (2014). Milton's maps. *Word and Image*, 29(4), pp. 428–442.

Locative Media for Queer Histories: Scaling up "Go Queer"

Maureen Engel
mengel@ualberta.ca
University of Alberta, Canada

This paper reports on the completion and launch of the locative media app "Go Queer." Taking the theorization, iteration, and development of "Go Queer" as a model and case study, the paper argues that locative media is uniquely suited to re/mediating queerness. It then proposes that these findings can be used as a framework and set of best practices for developing a variety of queer history applications.

Go Queer is a ludic, locative media experience that occurs on location, in the city, on the playful border between game and story, the present and the past, the queer and the straight, the normative and the *slant*. The app takes the city of Edmonton's queer history as its text, and produces a locative, spatialized narrative of that history by displaying text, images, video and audio in place at the actual locations where they occurred, thus creating what Richardson and Hjorth (2014, 256) call "the hybrid experience of place and presence." The app invites its users to drift queerly through the city, discovering the hidden histories that always surround us, yet somehow remain just beyond our apprehension. It compiles these traces into a media layer that augments quotidian city space, juxtaposing the past onto the present, creating a deep, queer narrative of place. By bringing together the physical navigation of the contemporary city with the imaginative navigation of its queer past, the app enacts a praxis that I characterize as a *queer ludic traversal*, one that renders the navigation itself as queer as the content that it presents. In so doing, the app produces the experience of *place*, in Lucy Lippard's (1997) formulation that

> Place is latitudinal and longitudinal within the map of a person's life. It is temporal and spatial, personal and political. A layered location replete with human histories and memories, place has width as well as depth. It is about connections, what surrounds it, what formed it, what happened there, what will happen there. (7)

The app proposes that a productive and underrepresented setting for queer play is the space of the city itself, and that the hybrid reality of locative media provides specific affordances to enable particularly queer navigations, occupations, and constructions of urban space.

The app arises from, and takes shape in relation to, a range of theoretical inspirations. First are the contributions queer theories of space, the urban, and community, such as David Bell's (2001) observation of "the special relationship between the city and the deviant" (84) and Theories recognizing the very public-ness of the formation, circu-

lation, and inhabiting of queer identities (D'Emilio, 1983; Berlant and Warner, 1998); central here is Sara Ahmed's theorization of "orientation" and her contention that "orientations are about the directions we take that put some things and not others in our reach" (552). New theorizations of space and place that have come to be called *the spatial turn* have similarly mobilized our thinking, challenging us to imagine space as a complex social production (Lefebvre, 1992) and asking us to think through how we move in space as either *tactical* or *strategic* (deCerteau, 2011). Praxis-based interactivity, which I draw principally from the field of Game Studies, has introduced concepts like the fidelity context (Galloway 2004) and ambient experience (Flanagan 2009). Deep mapping offers new possibilities for modeling space, particularly historical space, by bringing together the explanatory and critical capacities of both narrative and mapmaking (Bodenhamer 2007). These theoretical methods intersect in locative media itself, the vehicle for "Go Queer" and a platform, I argue, that holds significant promise for queer scholarship and expression.

By exploring how each of these theoretical arenas is literalized in the app itself, this paper aims to provide a framework and method for other practitioners interested in deploying locative media technologies to engage queer subjects, histories, and cultural productions.

References

Anthropy, Anna. *Rise of the Videogame Zinesters: How Freaks, Normals, Amateurs, Artists, Dreamers, Drop-Outs, Queers, Housewives, and People Like You Are Taking Back an Art Form*. New York: Seven Stories Press, 2012.

Bell, David, and Jon Binnie. "Authenticating Queer Space: Citizenship, Urbanism and Governance." *Urban Studies* 41.9 (2004): 1807–1820. usj.sagepub.com. Web. 12 Feb. 2015.

Bell, David, and Gill Valentine. *Mapping Desire: Geographies of Sexualities*. 1 edition. London ; New York: Routledge, 1995.

Berlant, Lauren, and Michael Warner. "Sex in Public." *Critical Inquiry* 24.2 (1998): 547–566.

Binnie, Jon et al. *Pleasure Zones: Bodies, Cities, Spaces*. Syracuse: Syracuse Univ Pr, 2001.

Bodenhamer, David J. "Creating a Landscape of Memory: The Potential of Humanities GIS." *Journal of Humanities & Arts Computing: A Journal of Digital Humanities* 1.2 (2007): 97–110.

Cabiria, Jonathan. "Virtual World and Real World Permeability: Transference of Positive Benefits for Marginalized Gay and Lesbian Populations." *Journal of Virtual Worlds Research* 1.1 (2008): n. pag. Web.

Certeau, Michel de. *The Practice of Everyday Life. Trans*. Steven F. Rendall. 3rd Revised edition edition. University of California Press, 2011.

Chisholm, Dianne. *Queer Constellations: Subcultural Space In The Wake Of The City*. 1 edition. Minneapolis: Univ Of Minnesota Press, 2004.

Chrisman, Nicholas R. "Design of Geographic Information Systems Based on Social and Cultural Goals." *Photogrammetric Engineering & Remote Sensing* 53.10 (1987): 1367.

Craig, William J., and Sarah A. Elwood. "How and Why, Community Groups Use Maps and Geographic Information." *Cartography & Geographic Information Systems* 25.2 (1998): 95.

Crampton, J.W. "Maps as Social Constructions: Power, Communication and Visualization." *Progress in Human Geography* 25.2 (2001): 235–252.

Crang, Mike. "Public Space, Urban Space and Electronic Space: Would the Real City Please Stand Up?" *Urban Studies* (Routledge) 37.2 (2000): 301–317.

Danielson, Laura. "An Exploration of Deep Maps." N.p., thepoliscenter.iupui.edu.

Désert, Jean-Ulrick. "Queer Space." *Queers in Space: Claiming the Urban Landscape*. Ed. Gorden Brent Ingram, Gordon B. Ingram, and Yolanda Retter. Seattle, Wash: Bay Pr, 1997. 17–26.

Dodge, Martin, Rob Kitchin, and Chris Perkins, eds. *The Map Reader: Theories of Mapping Practice and Cartographic Representation*. Chichester, West Sussex, UK; Hoboken, NJ: Wiley, 2011.

Flanagan, Mary. *Critical Play: Radical Game Design*. The MIT Press, 2013.

Gieseking, Jen Jack et al., eds. *The People, Place, and Space Reader. People, Place, and Space: A Reader*: Routledge, 2014.

Goodchild, Michael F., and Donald G. Janelle. "Toward Critical Spatial Thinking in the Social Sciences and Humanities." *GeoJournal* 2010: 3.

Gregory, Derek. *Geographical Imaginations*. Cambridge, MA : Blackwell, 1994.

Gregory, Ian, and Paul S. Ell. Historical GIS [electronic Resource] : *Technologies, Methodologies and Scholarship* / Ian N. Gregory, Paul S. Ell. Cambridge, UK ; New York : Cambridge University Press, 2007. Cambridge Studies in Historical Geography: 39.

Halberstam, Judith. "What's That Smell? Queer Temporalities and Subcultural Lives." International Journal of Cultural Studies 6.3 (2003): 313–333.

Hall, Stuart. *"Encoding/Decoding." Media and Cultural Studies: KeyWorks*. Ed. Meenakshi Gigi Durham and Douglas M. Kellner. Revised Edition. Malden MA: Blackwell, 2006. 163–173. KeyWorks in Cultural Studies.

Harris, Trevor M., John Corrigan, and David J. Bodenhamer. *The Spatial Humanities : GIS and the Future of Humanities Scholarship*. Bloomington: Indiana University Press, 2010.

Harris, Trevor, and Daniel Weiner. "Empowerment, Marginalization, and 'Community-Integrated' GIS." *Cartography and Geographic Information Systems* 25.2 (1998): 67–76.

Hjorth, Larissa, and Sun Sun Lim. "Mobile Intimacy in an Age of Affective Mobile Media." *Feminist Media Studies* 12.4 (2012): 477–484.

Johnston, Lynda, and Robyn Longhurst. *Space, Place, and Sex: Geographies of Sexualities*. Lanham: Rowman & Littlefield Publishers, 2009.

Juliano, Linzi. "Digital: A Love Story; Bully; Grand Theft Auto IV; Portal; Dys4ia (review)." *Theatre Journal* 64.4 (2012): 595–598.
Kirkpatrick, Graeme. Computer Games and the Social Imaginary. *Polity*, 2013.
Knowles, Anne Kelly. *Past Time, Past Place : GIS for History* / Edited by Anne Kelly Knowles. Redlands, Calif. : ESRI Press, 2002.
Lefebvre, Henri. *The Production of Space*. 1 edition. Wiley-Blackwell, 1992.
Lippard, Lucy. *Lure of the Local: Senses of Place in a Multicentered Society*. 1 edition. New York: The New Press, 1998.
Lynch, Kevin. *The Image of the City*. Cambridge, Mass.: The MIT Press, 1960. Mattern, Shannon. Deep Mapping the Media City. Univ Of Minnesota Press, 2015.
McLafferty, Sara. "Mapping Women's Worlds: Knowledge, Power and the Bounds of GIS." *Gender, Place and Culture* 9.3 (2002): 263–269.
Murphy, Kevin. "Walking the Queer City." *Radical History Review* 62 (1995): 195–201.
Paglen, Trevor, and John Emerson. *An Atlas of Radical Cartography*. Ed. Lize Mogel and Alexis Bhagat. Slp edition. Los Angeles: Journal of Aesthetics and Protest Press, 2008.
Pavlovskaya, Marianna. "Theorizing with GIS: A Tool for Critical Geographies?" *Environment and Planning A* 38.11 (2006): 2003–2020.
Presner, Todd, David Shepard, and Yoh Kawano. *HyperCities: Thick Mapping in the Digital Humanities*. Cambridge, Massachusetts: Harvard University Press, 2014.
Retter, Yolanda, Anne-Marie Bouthillette, and Gordon Brent Ingram, eds. *Queers in Space: Communities, Public Places, Sites of Resistance*. Seattle, Wash: Bay Press, 1997.
Ridge, Mia, Don Lafreniere, and Scott Nesbit. "Creating Deep Maps and Spatial Narratives through Design." *Journal of Humanities & Arts Computing: A Journal of Digital Humanities* 7.1/2 (2013): 176–189.
Rundstrom, Robert A. "GIS, Indigenous Peoples, and Epistemological Diversity." *Cartography and Geographic Information Systems* 22.1 (1995): 45.
Shaw, Adrienne. "Putting the Gay in Games: Cultural Production and GLBT Content in Video Games." *Conference Papers -- International Communication Association* (2008): 1–29.
Skeggs, Beverley et al. "Queer as Folk: Producing the Real of Urban Space." *Urban Studies* 41.9 (2004): 1839–1856.
Soja, Edward W. *Postmodern Geographies: The Reassertion of Space in Critical Social Theory*. 2nd edition. London; New York: Verso, 2011.
Warf, Barney, and Santa Arias. The Spatial Turn : *Interdisciplinary Perspectives* / Edited by Barney Warf and Santa Arias. London : Routledge, 2009
Wood, Denis, John Fels, and John Krygier. *Rethinking the Power of Maps*. New York: The Guilford Press, 2010.

Analyzing Social Networks of XML Plays: Exploring Shakespeare's Genres

Lawrence Evalyn
lawrenceevalyn@gmail.com
University of Toronto, Canada

Susan Gauch
segauch@gmail.com
University of Arkansas, United States of America

Manisha Shukla
mshukla@email.uark.edu
University of Arkansas, United States of America

Introduction

Our inquiry considers the speech interactions of characters within plays as a proxy for broad narrative structures. We analyze computationally-generated social networks of 37 plays by Shakespeare to see whether, and how, they can be used to distinguish between Shakespeare's comedies, tragedies, and histories.

Because dramatic performances enact social encounters, social network analysis translates surprisingly well to fictional societies. Stiller et al. have shown that social networks in Shakespeare's plays mirror those of real human interactions, particularly in size, clustering, and maximum degrees of separation (2003). However, as fictions, these networks are shaped not only by sociological principles, but also by narrative structures. Moretti uses social networks to examine the plots of three Shakespearean tragedies, and to contrast the structure of chapters in English and Chinese novels (2011). Alberich et al. (2002) and Sparavigna (2013) also discuss the interplay between social and narrative constraints on networks. We emphasize this distinction to look for specifically literary features of our networks.

Recent papers presented at DH2017 sought ways to richly quantify the details of one or two plays (Fischer et al., 2017; Tonra et al., 2017). At another scale, Algee-Hewitt examined 3,439 plays by looking only at the Gini Coefficient of each play's eigenvector centrality (2017). With our three dozen plays, we attempt to strike a fruitful middle ground in the inevitable balancing act between detail and scale. Each play is considered individually, but at a level of abstraction which allows rapid and direct comparisons.

Creation of social network graphs

Our parser tracks characters present on stage during speech. This approach is highly extensible: it can parse any play that follows TEI P5 guidelines for performance texts. Each speaking character is connected to all cha-

racters currently present on stage. These connections are recorded in a network graph, with characters as nodes and shared speech as edges. Edges are directional, and weighted based on the number of lines spoken. In future, we plan to extend our parser to identify the specific addressees of a character's speech, allowing us to model more detailed interactions.

To verify that our parser is accurate, we compare our generated network of Hamlet to Moretti's well-known handmade model of that play (2011). Despite some minor differences in peripheral characters like "Servant", and our less-minor difference of including the play-within-the-play, the two networks are highly similar. Our network graph supports Moretti's reading. Our tool also improves on Moretti's model by adding direction and weight to each connection. Although this level of detail turned out not to be necessary for the basic task of using network graphs to distinguish between Shakespeare's genres, it may be useful in future work examining a less homogenous corpus of plays, or in work asking different questions about this corpus.

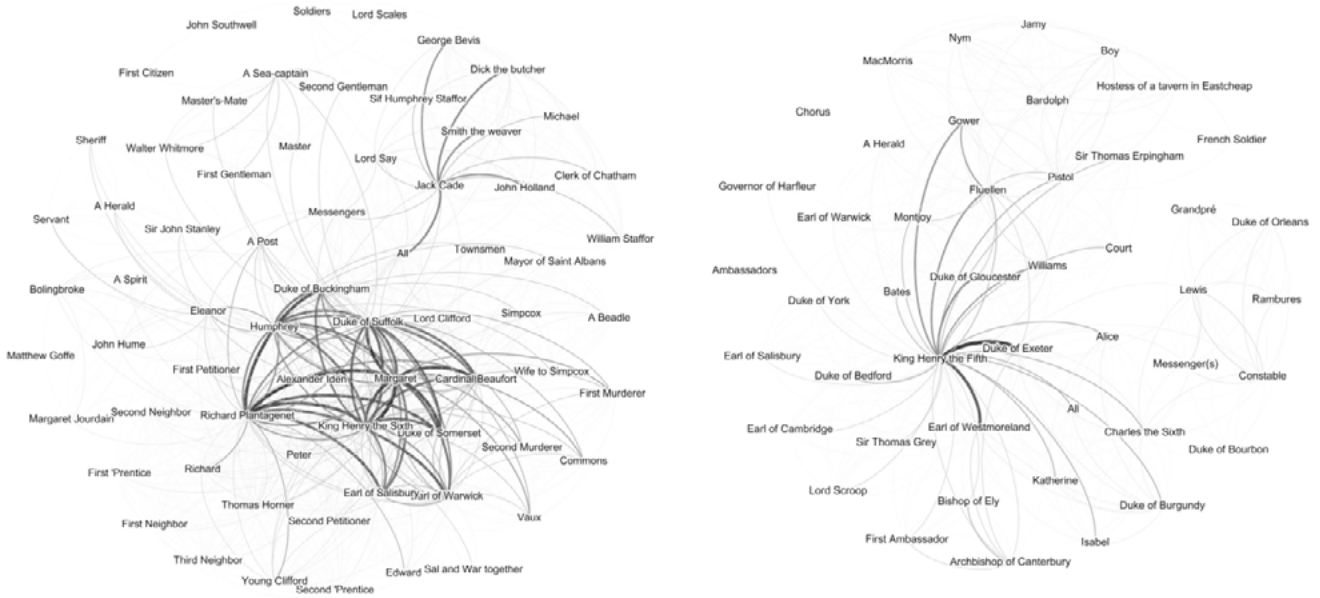
Using networks to identify genre

We then use our generated network graphs to test our central question: whether the social network enacted by a play's characters can be used as a proxy for features of the play's narrative content. More specifically, we ask whether social networks can be used to distinguish between the dramatic genres of tragedy, comedy, and history. Using a support vector machine with fivefold validation, we tested 17 different mathematical features of the networks. No single feature was independently sufficient to identify the genre, though graph density came closest (83% accuracy). However, if features are used in combination, the network graphs can indeed achieve full accuracy. One combination of features which does achieve 100% accuracy is edges, words, and degree. We are currently exploring other combinations that might also be capable of accurately identifying genres.

Discussion

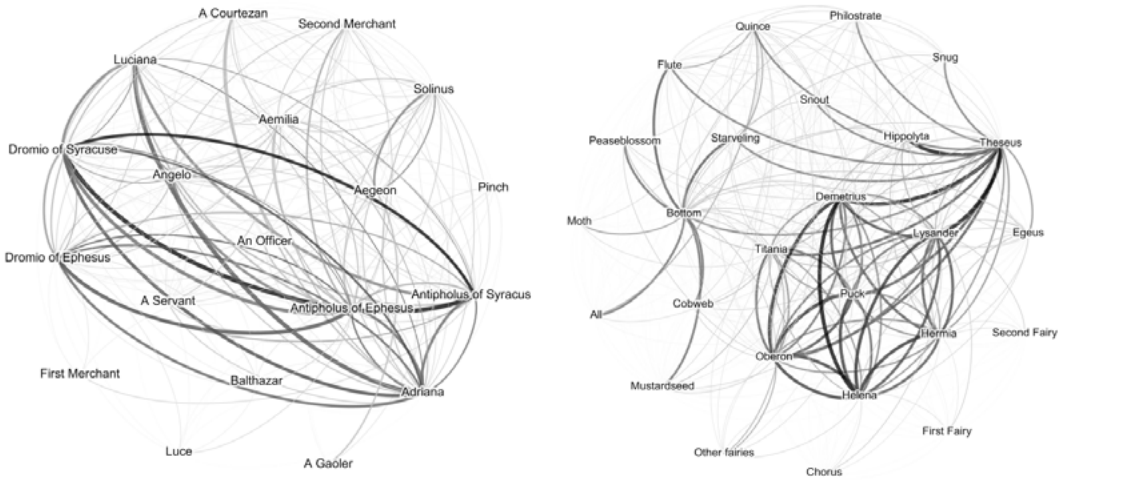
History, comedy, tragedy

The potential utility of graph density in distinguishing genres is visually obvious when individual comedy and history networks are compared. Histories feature highly dispersed networks, with large numbers of very minor characters, such as "First," "Second," and "Third" members of groups like soldiers and ambassadors, who each interject briefly in a single scene. Connections form chains of acquaintance with little overlap, so even the monarchs have low eigenvector centrality.

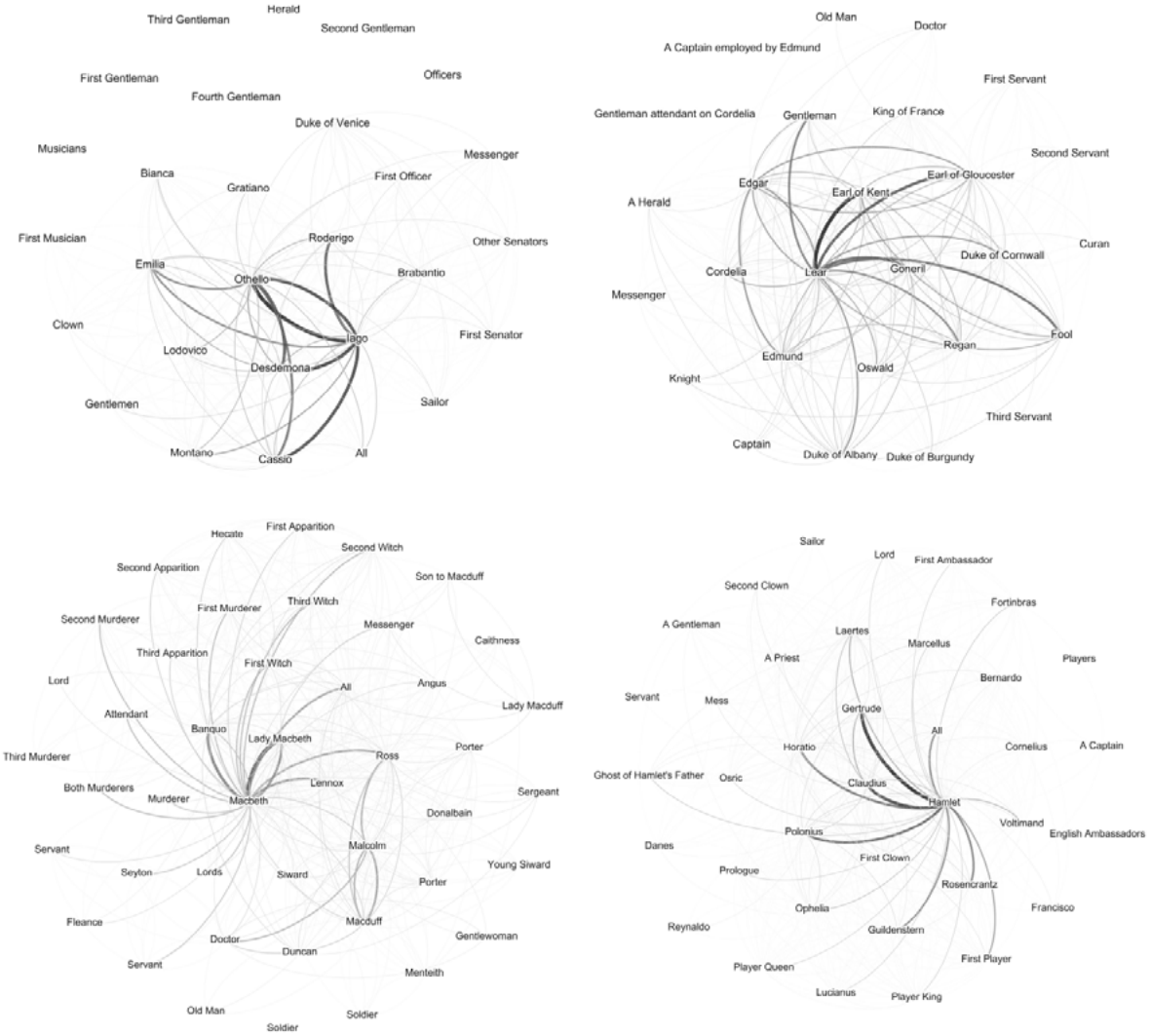


Social network graphs of the histories *Henry VI, Part 2* and *Henry V*.

Comedies, in contrast, feature networks with far fewer characters, in which nearly everybody speaks to nearly everybody else at some point. Although comedies often have multiple subplots, these separate stories do not result in highly-separated networks. We theorize that comedic networks are strongly shaped by the plays' final "resolution" scenes, which bring together the full cast. The average eigenvector centrality of the characters in comedies is much higher than in tragedies or histories; this suggests that many more of the characters in a comedy are "important," reflecting a focus on ensemble stories.



Social network graphs of the comedies *The Comedy of Errors* and *A Midsummer Night's Dream*.



Social network graphs of the tragedies *Othello*, *King Lear*, *Macbeth*, and *Hamlet*.

Graph density is insufficient, however, to fully distinguish the tragedies, which feature networks somewhere between history and comedy in their density. They often have a dense core with a secondary ring of more peripheral characters. What seems to distinguish them is the existence of the central tragic hero, whose influence directly touches more of the network than the protagonists of histories, but whose connections are less interconnected than the ensembles of comedies. These subtleties are better captured, it seems, by the combined metric of "edges, degree, and words."

The "problem plays"

We then use our preliminary identification of each genre's features to examine Shakespeare's various contested genres. Training our model only on the plays for which there is strong consensus, we applied it to the "Roman plays," the "problem plays," and the "romances" in turn. Of the Roman plays, all but *Antony and Cleopatra* are identified as tragedies by every metric; *Antony and Cleopatra* is identified by "edges, words, and degree" as a history and by "degree, modularity, and density" as a comedy. Of the problem plays, *All's Well that Ends Well* is always identified as a comedy; *Troilus and Cressida* and *Measure for Measure* are both identified as a comedy by all metrics except for "edges, criticality, and degree", which identify them as tragedies. The four romances, despite visually unusual networks which support literary arguments that Shakespeare's writing had grown more experimental at the end of his career, are identified as comedies by every mathematical metric. We treat none of these identifications as definitive declaration of the plays' "real" genres, but use them to distinguish between plays whose generic ambiguity lies in their subject matter, and plays whose ambiguity lies in their structure.

Conclusion

Our parser successfully and rapidly produces sophisticated social network graphs of TEI plays that can be used to computationally identify theatrical genre in Shakespeare's plays. Thirty-seven plays is a small scale for this approach: since the parser is highly extensible and can be used with any plays encoded in TEI, future work need not be restricted to the Early Modern period. It need not even be restricted to works written in English. Our networks of the well-studied works of Shakespeare can provide a baseline against which to contextualize analysis of these elements in works for which there is far less critical consensus.

References

Alberich, R.,Miro-Julia, J., and Rosselló., F. (2002). Marvel Universe Looks Almost Like a Real Social Network. arXiv:cond-mat/0202174v1

Algee-Hewitt, M. (2017). Distributed Character: Quantitative Models of the English Stage, 1500-1920. *Digital Humanities 2017: Book of Abstracts*. Montreal: McGill University and Université de Montréal, pp. 119–21.

Fischer, F., Göbel, M., Kampkaspar, D., Kittel, C., and Trilcke, P. (2017). Network Dynamics, Plot Analysis: Approaching the Progressive Structuration of Literary Texts. *Digital Humanities 2017: Book of Abstracts*. Montreal: McGill University and Université de Montréal, pp. 437–41.

Moretti, F. (2011). Network Theory, Plot Analysis. *New Left Review,* 68: 80–102.

Sparavigna, A. C. (2013). On Social Networks in Plays and Novels. *International Journal of Sciences*, 2: 20–25.

Stiller, J., Nettle, D., and Dunbar, R. I. M. (2003). The Small World of Shakespeare's Plays. *Human Nature*, 14(4): 397–408.

Tonra, J., Kelly, D., and Reid, L. (2017). Personæ: A Character-Visualisation Tool for Dramatic Texts. *Digital Humanities 2017: Book of Abstracts*. Montreal: McGill University and Université de Montréal, pp. 627–30.

Resolving the Polynymy of Place: or How to Create a Gazetteer of Colonized Landscapes

Katherine Mary Faull
faull@bucknell.edu
Bucknell University, United States of America

Diane Katherine Jakacki
diane.jakacki@bucknell.edu
Bucknell University, United States of America

In working with British colonial records and German church manuscripts of colonized and missionized landscapes in the North American mid-Atlantic, the authors have grappled with the problem of polynymy in their attempt to create a gazetteer of places. As Presner and Shepard (2016) have argued, unlike conventional positivistic approaches to mapping, DH and geohumanities have developed a rich vocabulary with which to describe and analyze the human perception of place. Whether through "deep maps" that recount the stories of place and experience or through the multiple layers of temporally inflected information, the spatial turn has revealed the need to see the practice of mapping as "arguments or propositions that betray a state of knowledge." (Presner and Shepard 2016, 207). However, whereas there are sophisticated models of temporal-spatial mapping now available to DHers working with historical materials, to date little critical attention has been paid to the place/person variable. The work of Ann Knowles (and her students) has paved the way for sophisticated representations of the experien-

ce of place (Knowles 2008; 2015). In her arguments for a nonpositivistic geo-practice within the humanities, Knowles has opened up the field to the "fuzzy data" of critical humanistic inquiry. Privileging design over data, Knowles' prize-winning visualizations of the Holocaust challenge us to reconsider in sophisticated ways the experience of landscapes. (Knowles 2014) On a similar path, as Presner and Shepard conclude, virtual reality and gaming allow for an experiential and avatar-based investigation of dynamic, embodied, albeit presentist, multiple perspectives of place. Students at Bucknell have already produced sophisticated critical cartographical visualizations of the Susquehanna river in the Colonial period that draw in part on Knowles' perspectives. This paper will explore the problem of creating a gazetteer of colonized landscapes, specifically those of the mid-Atlantic in the 18th century, in which the name of a place (toponym) changes depending on the person or political entity who is describing that place. In colonized landscapes, there can be multiple names for one place. Maps of this period are veritable palimpsests of conquests and defeats; and travel diaries, mission records and letters contain accounts of human experience of places that are multiply identified. The task is made more complicated still when one factors time into the equation: when competing spatial identities persist across generations. Using the case study of the research project "Moravian Lives" we will ask how we can create a gazetteer of places using authority IDs, when that very authority is itself the product of apolitical-historical struggle. "Moravian Lives" is an international collaborative DH project that aims to make available to the scholarly and lay community the vast corpus of life writings of members of the Moravian Church from the mid-18th century to today (http://moravianlives.org). Facing the simultaneity of multiple names for a place, can we create a system of "triples" that satisfactorily reflects the multiple perspectives and presence or absence of agency of those who name place? Drawing on the substantial cultural-historical GIS of the Susquehanna river produced by Faull and a team of Bucknell staff and students that supported the Department of the Interior designation of the Susquehanna River as a National Historic Water Trail in 2012, the Moravian Lives gazetteer aims to provide the most comprehensive place-name resource for researchers in many fields. The construction of an historical gazetteer for Moravian Lives involves complexities that arise from not only the naming of places but also how their spatial identities reflect respective, concurrent relationships to those places by Native American peoples, Moravian missionaries,and colonial representatives. There are multiple names for a single place as well as multiple understandings of place names, and these differences depend on who it was who did the naming. An example of this challenge is 18th-century Shamokin in Pennsylvania. Shamokin was at that point an Iroquois settlement at the confluence of the north and west branches of the Susquehanna River, encompassing the shores of both branches and an island at the river's fork. To Shikellamy, an Oneida emissary of the Six Nations of the Iroquois or Haudenosaunee, who oversaw the Algonquin-speaking nations of the Lenni Lenape, Shawnee, and Mahican in Iroquoia (present-day Pennsylvania and New York), and who lived in the town in the 1740s, "Shamokin" would have constituted the whole area of the rivers' confluence. To Count Nikolaus von Zinzendorf, the founder of the Moravian Church who visited Shikellamy in 1742, "Shamokin" represented an opportunity for Moravian missionaries offered to them by Shikellamy in the form of space for a blacksmith's shop and mission. While the location of that mission was small, it loomed large in Zinzendorf's interest in founding "Heiden-Collegia", or colleges of the "heathen", in Pennsylvania. To Conrad Weiser, a German settler and negotiator between the colonial government in Philadelphia and the Indian nations, and who worked with Shikellamy on several treaties between the Iroquois and the Colonial government, "Shamokin" would have represented a strategic and ultimately military outpost that would become the site of Fort Augusta during the French and Indian War. These "Shamokins" co-existed, with Native American, Moravian, and Colonial inhabitants and visitors relating to it in discrete yet overlapping ways. One byproduct of our work on the gazetteer could thus be the proposition of authority lists to the OCLC's VIAF council, thereby introducing and linking our information where there is currently no match. In compiling a gazetteer we realize that there is already a VIAF authority ID for Shamokin that is recognized by the Library of Congress/NACO but refers to another (modern) place called Shamokin some 18 miles to the east. (Shamokin, PA VIAF ID: 146606881 (Geographic). We cannot therefore "re-mint" an authority name for these Shamokins. Furthermore, a part of the 18th century Shamokin is now Sunbury (the site of Fort Augusta and Shikellamy's grave) also has its own VIAF ID, (3 Sunbury, PA VIAF ID: 123181256 (Geographic) but, for the historical and cultural studies scholar, it might be inaccurate, misleading, and in some ways irresponsible to equate Sunbury with or consider it as a variant for the historic Shamokin. How can we recognize spatial multivalence (or "polynymy") in the Moravian Lives gazetteer? How does the scholar act responsibly while acknowledging their own potential complicity in political-historical renegotiations and multiple cultural understandings of place? In effect, must we not push back at the idea of *an* authority, and work toward a system that recognizes and synchronizes multiple authorities? We propose a two-phased approach to developing the Moravian Lives gazetteer, which will expand geographically to places beyond North America and will need to resolve polynymic complexities in Central Europe, the Arctic areas of Greenland and Newfoundland, the Caribbean, South Africa and Australia. The first phase involves "stabilizing" all of the place names without giving primacy to any one of them. Each would be assigned a unique HTTP URI offering information about each toponym pertinent to its

own cultural relationships and link to its siblings. In this way we can push back against the need to choose one authority (whether it be restoring an indigenous name or opting among European ones) and demonstrate that these names are not "same as" or "variants" of the others. This, in turn, allows us to reflect upon colonial places in a much more nuanced way that takes into account geographical features and proximity (viz. 'Peace huts on the Susquehanna', 'an der Höhle bei Bethel'). It also enriches the companion personography under development for Moravian Lives. In the visualization already available through Moravian Lives, each person is associated with place using a single-point Google location (see Figs. 2 and 3); but by integrating the cultural historical mapping already completed for the Susquehanna river project, we can now connect these people with better suited vector data referencing each unique place's footprint or range at the same time acknowledging that our identification involves a consideration of certainty (or "fuzziness") by the editor. Through this process, we will strengthen the interlinking of temporo-spatial data within the Moravian Lives project, weaving together the text-based gazetteer with the mapped data. The second phase is to submit our set of authority files to the OCLC and its VIAF council through a member advocate (such as the Moravian Archives). Our work will then be reviewed, assessed against existing identified geographic places in the VIAF database, and where appropriate we hope that new VIAF IDs will be minted. In this way we will make these places discoverable to other researchers considering similarly complex cultural landscapes.

References

Faull, Katherine. ""Digital Lives: Reading Moravian Memoirs in the Age of the Internet" Short paper, *DH 2017*. Montreal, Canada.

---. "Charting the Colonial Backcountry: Joseph Shippen's Map of the Susquehanna River." *The Pennsylvania Magazine of History and Biography*, vol. 136, No. 4, 2012, 461-465.

---. "Writing a Moravian Memoir: The Intersection of History and Autobiography" in *Life Writing and Lebenslauf: Pillars of an Invisible Church*, eds. Christer Ahlberger and Per van Wachtenfeld, Artos Publishers, 2017.

Grumet, Robert. *Manhattan to Minisink. American Indian Place Names in Greater New York and Vicinity*. Norman, OK: University of Oklahoma Press, 2013.

Horsman, Stuart. "The Politics of Toponyms in the Pamir Mountains." *Area*, vol. 38, no. 3, 2006, pp. 279–291. JSTOR, www.jstor.org/stable/20004545.

Jakacki, Diane and Janelle Jenstad. "Mapping Toponyms in Early Modern Plays with the Map of Early Modern London and Internet Shakespeare Editions Projects." *Early Modern Studies and the Digital Turn*. Laura Estill, Diane Jakacki, Michael Ullyot, eds. Malden, MA: ITER. 2016

Meredyk, Steffany, Bethany Dunn, under the supervision of Katherine Faull. "A Corridor of Fear: Stories along the Susquehanna River, 1754-1768". *Stories of the Susquehanna Valley project*. 2013. Stable URL: http://bit.ly/2iXbJzA

Knowles, Ann. "Inductive Visualization: A Humanistic Alternative to GIS" (2015). *GeoHumanities* 1(2), pp. 233-265. DOI 10.1080/2373566X.2015.1108831.

---. *Placing History: How Maps, Spatial Data, and GIS Are Changing Historical Scholarship* (2008), edited by Knowles, digital supplement edited by Amy Hillier. Redlands, Cal.: ESRI Press Knowles, Ann, Tim Cole, and Alberto Giordano, eds. Geographies of the Holocaust. Bloomington, IN: U of Indiana Press. 2014.

Moravian Lives project website: http://moravianlives.org/

Oetelaar, Gerald A., and David Meyer. "Movement and Native American Landscapes: A Comparative Approach." *Plains Anthropologist*, vol. 51, no. 199, 2006, pp. 355–374. OCLC Virtual International Authority File webpage: http://www.oclc.org/en/viaf.html

Presner, Todd and David Shepard, "Mapping the Geospatial Turn" in *A New Companion to Digital Humanities*, eds. Susan Schreibman, Ray Siemens, and John Unsworth, (Oxford: Wiley Blackwell, 2016) 201-212.

Radding, Lisa, and John Western. "What's In A Name? Linguistics, Geography, And Toponyms." *Geographical Review*, vol. 100, no. 3, 2010, pp. 394–412. JSTOR, www.jstor.org/stable/25741159.

Audiences, Evidence, and Living Documents: Motivating Factors in Digital Humanities Monograph Publishing

Katrina Fenlon
kfenlon2@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Megan Senseney
mfsense2@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Maria Bonn
mbonn@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Janet Swatscheno
jswatsc2@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Christopher R. Maden
crism@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Introduction

How humanities scholars communicate their research - with one another, with interdisciplinary communities, and with diverse publics - continues to shift with the emergence of new publishing models. We do not understand enough about why scholars choose to publish in different modalities, or what the implications of their choices are for the use, evaluation, and sustainability of research. Thus, publishing systems and services lag behind the advance of digital methods and modes of communication.

This paper presents selected results of a multimodal study of humanities scholars' digital publishing needs. Building on national survey of humanities scholars in the United States, initially reported at DH2017 (Senseney et al., 2017), this paper describes preliminary outcomes of a series of interviews with humanities scholars who have a manifest interest in experimental digital publishing. This study seeks to deepen our understanding of scholarly goals for digital publication.

Outcomes of this study are guiding the development of a service model for library-based humanities publishing, as part of the Publishing Without Walls (PWW) project (http://publishingwithoutwalls.illinois.edu/). Funded by the Andrew W. Mellon Foundation, the University of Illinois Library is leading the PWW initiative in partnership with the Graduate School of Library and Information Science, the Illinois Program for Research in the Humanities, and the African American Studies Department at the University of Illinois. PWW aims to develop a scalable, shareable model for monograph publishing within libraries, with the goal of bridging gaps in current publishing systems, such as gaps between the complex materials scholars want to publish and what existing systems can accommodate, between scholarly practices and existing publishing tools, and between publishing opportunities at resource-rich and under-resourced institutions.

This paper focuses on humanities scholars' motivations for publishing digital, open access, and multimedia monographs. We explore three central motivations for digital publishing: (1) the desire to reach diverse audiences; (2) the desire to integrate interactive, multimedia, and linked evidence; and (3) the desire to publish "living" documents. These factors have implications for digital humanities scholars in understanding the impact of different modes of sharing, for libraries seeking to support digital scholarship, for data models underlying enhanced publications, and for publishing service models.

Methods

This study comprised a set of semi-structured interviews with humanities scholars. Interview participants were self-selected from among scholars who had already participated in the PWW initiative in some way, whether by attending publishing workshops or submitting to the new series. Nineteen interviews have been conducted to date; more are planned for summer 2018. All interviews are recorded and transcribed, and a formal analysis of resulting transcripts is underway. Participants are all affiliated with academic institutions. They include faculty, postdoctoral research associates, and academic professionals with backgrounds in humanities disciplines, information science, and communications.

Three motivations for enhanced digital publishing

Multiple audiences

Scholars turn to open access (OA) monograph publishing to increase impact by reaching more readers, not only within their disciplines but also cross-disciplinary peers and the general public. Visibility and broad dissemination are established motivations for OA book publishing; evidence suggests that these motivations are rewarded, as OA books receive significantly more usage and citation than non-OA counterparts (Emery et al., 2017). Yet, our study indicates that humanities scholars want more than to reach large audiences. They want to reach diverse audiences, ranging from peers in other disciplines to practitioners, policymakers, and the public. Despite potential impact, participants acknowledged that certain prevalent models of OA monograph publishing suffer from a lack of "institutional weight" and "automatic audiences." However, participants described leveraging their own social and research networks to promote their work directly.

Interactive, multimedia, and linked evidence

Authors pursue opportunities for representing new kinds of evidence in new contexts. The potential benefits of multimedia publishing are largely unrealized in publishing practice due to the challenges of managing complex digital publications (Jankowski et al., 2012). Scholars want to integrate or actionably link to more kinds of evidence, including multimedia sources, interactive visualizations, data sets, and curated collections. They also want to make their sources interactive, to allow readers opportunities to visualize, explore, and assess bodies of evidence while anchoring them to narrative descriptions and interpretations. One participant described his primary goal for multimedia publishing as making evidence "come alive in a narrative history."

Living documents

Some humanities scholars want to publish what participants call "living," evolving documents —works-in-progress that are subject to indefinite change. Participants value immediacy of entrance into ongoing scholarly dialogue, both for obtaining rapid feedback from peers and for flag-planting. Some participants see self-publication as a route toward obtaining high-quality peer review more quickly than through the conventional publication; the complexity of peer review in interdisciplinary settings — like the digital humanities — can lead to dilatory, frustrating review processes, which one participant compared to "the phenomenon of too many cooks in the kitchen," and which may yield "diluted" end work. The ultimate manifestation of a "living" document is a publication that facilitates ongoing co-authorship, annotation, interlinking, and revision. One participant described an ideal publication as an online document that "people can comment on, that can directly link to its sources and other people can link to it, that has an attached data set of results that other people can make use of and check," and which is subject to versioning. He described this as an evolving or living document and noted that, "at the moment, most of our research papers are dead documents."

Future work

While openness is a core value of digital humanities scholarship (albeit with qualifications see, e.g., Spiro, 2012), it is not clear how different modes of publication can most effectively open humanities research: to the stratified audiences identified in this study, to deep interaction with sources, and to ongoing evolution. This paper describes outcomes of our study on what humanities scholars need from the next generation of publishing systems and services, and how this study is guiding development of a new model for library-based publishing that can support and sustain highly diverse and broadly impactful research products.

References

Emery, C., Lucraft, M., Morka, A., and Pyne, R. (2017). *The OA effect: How does open access affect the usage of scholarly books?* Springer Nature.

Jankowski, N., Scharnhorst, A., Tatum, C., and Tatum, Z. (2012). Enhancing Scholarly Publications: Developing hybrid monographs in the humanities and social sciences. *Scholarly and Research Communication*, 4(1).

Senseney, M. F., Velez, L., Maden, C. R., Swatscheno, J., Bonn, M., Green, H., and Fenlon, K. (2017). Informing library-based digital publishing: A survey of scholars' needs in a contemporary publishing environment. Presented at the Digital Humanities (DH2017), Montréal, Canada.

Spiro, L. (2012). "This Is Why We Fight": Defining the values of the digital humanities. *Debates in the Digital Humanities*, 16.

Mitologias do Fascínio Tecnológico

Andre Azevedo da Fonseca
azevedodafonseca@gmail.com
Universidade Estadual de Londrina (UEL), Brazil

A cultura digital do século XXI tem sido marcada pela ascensão de um imaginário mágico em relação ao poder das tecnologias. Por meio de uma produção monumental de símbolos, as indústrias culturais e a publicidade das mais diversas empresas de tecnologia têm veiculado mensagens a fim de relacionar o consumo tecnológico à conquista progressiva da autonomia, da liberdade, da felicidade e, em última instância, da transcendência. Este imaginário que induz à devoção das tecnologias parece seduzir as novas gerações com a promessa da elevação dos seres humanos à condição de semidivindades a partir do consumo físico e simbólico de produtos e marcas.

No entanto, sob o brilho deste deslumbre, o Estado e as corporações têm se movimentado no sentido de empregar recursos tecnológicos de forma sistemática para aprofundar o controle social de natureza tecnocrática, de modo que cidadãos e consumidores são observados e analisados em sua intimidade. Ofuscados pelo brilho mágico das tecnologias, usuários entregam voluntariamente informações detalhadas de suas personalidades e experiências pessoais para delegar aos algoritmos de inteligência artificial decisões cada vez mais importantes de suas experiências humanas, tornando-se mais vulneráveis a estímulos publicitários e propagandas ideológicas cada vez mais personalizadas e eficientes.

Entre os vários elementos para que o capitalismo informacional lograsse legitimar essa sociedade de controle tecnocrático, observamos uma intensa produção simbólica nas indústrias culturais no sentido de instrumentar a cultura digital com um fabuloso repertório iconográfico para, primeiramente, exorcizar os temores apocalípticos que as tecnologias sem limites haviam inspirado na humanidade – sobretudo após o advento da bomba atômica e da chamada crise da razão – e, em seguida, substituir os antigos temores por uma nova devoção aos mitos tecnológicos. Nesse contexto, mitologias ancestrais que expressavam as maldições divinas decorrentes da desobediência de homens e mulheres que ousaram ultrapassar os limites do conhecimento foram esvaziadas e invertidas, de modo que os consumidores contemporâneos, mais do que apenas perder o medo, passaram a cultuar esses mitos: da maçã proibida do Éden à mação mordida da Apple, do terrível Big Brother de George Orwell ao sedutor Big Brother da Endemol, da maldição do monstro de

Frankenstein à celebração do gênio do cientista impetuoso no imaginário do Vale do Silício.

O objetivo desta pesquisa é compreender essa dinâmica de subversão de mitologias empregadas para superar os temores, atribuir uma conotação religiosa às experiências com tecnologias e, enfim, ofuscar o controle tecnocrático do ecossistema digital. Para isso, sob a perspectiva da Comunicação, da História Cultural e dos estudos de mitologia e imaginação social, analisamos um conjunto de símbolos evocados na imprensa, no cinema e na publicidade de empresas de tecnologia contemporâneas, situando-as no contexto histórico da utilização de arquétipos e mitologias na publicidade a partir do final do século XX. Como resultado, identificamos um conjunto de mitos e imagens arquetípicas manipuladas nas mídias para associar o consumo tecnológico ao imaginário sagrado da superação do pecado original, da reconquista do paraíso e da transcendência da condição humana.

Referencias

Barthes, R. (2009). *Mitologias*. 4 ed. Rio de Janeiro: Difel.

Chartier, R. (1985). *A história cultural: entre práticas e representações*. Rio de Janeiro: Difel/Bertrand Brasil.

ELLUL, J. (1964). *The technological society*. New York: Vintage Books.

JUNG, C. G. (2000). *Os arquétipos e o inconsciente coletivo*. 2 ed. Petrópolis: Vozes.

ROSZAK, T. (1972), *A contracultura: reflexões sobre a sociedade tecnocrática e a oposição juvenil*. 2 ed. Petrópolis: Vozes.

TURNER, F. (2006). *From Counterculture to Cyberculture: Stewart Brand, the Whole Earth Network, and the Rise of Digital Utopianism*. Chicago: The University of Chicago Press.

Latin@ voices in the Midwest: Ohio Habla Podcast

Elena Foulis
foulis.5@osu.edu
OSU, United States of America

In recent years, the use and development of Podcasts has significantly grown. Podcasts allow us to listen to topics we are interested in and learn more about an issue or community. Podcasts like *This American Life*, *Radio Ambulante* and *Latino USA*, put at the center of their stories experiences of people and places. Indeed, using audio as a medium to tell the larger stories of our community has proven successful as signaled by all top 10 iTunes podcasts—5 of which are documentary style. Creating university based podcast like, *Ohio Habla*, will allow us to connect and learn more the Latin@/Hispanic experiences locally, while amplifying the voices of the community everywhere. Language and cultural studies are in a unique position to utilize this medium to advance the understanding of how culture and language is both transmitted and analyzed.

The *Ohio Habla* podcast is primarily produced by students in advanced Spanish language and Latin@ studies classes together with their professor. Each student plans, researches, secures a podcast guest and carries out the interview. Students are able to continue to develop their written, reading, speaking and listening skills and they are responsible to produce one whole 30-45 minute podcast.

Ohio Habla is an extension of the digital oral history project, ONLO (oral narratives of Latinos/as in Ohio), however, it focuses topics, rather than life history. On the other hand, in the case of Latin@ students, they collect family stories, instead of interviewing a member of the community. Podcasting can help document issues that are of interest to our community and potentially be able to share it more widely. Finding new and real ways to use language and storytelling is of great benefit to our students, and podcast in the foreign language classroom can accomplish this. Our own teaching methodology here at Ohio State encourages second language learners to use the language communicatively and in real situations that are as authentic as possible. Podcasting is a great way to use the language in real and creative ways, and most importantly, in community—an element that is often left out the foreign language classroom for various reasons (mainly, time).

Teaching methodologies

As a pilot project, this use of podcasting in the classroom may pave the way for further research opportunities about the benefit of podcasting in advanced language courses, service-learning and heritage language learners. This course enhancement will also provide students with a structured opportunity to make deeper connections with Latino/a campus community, to reflect on that experience, and to gain interviewing skills that will serve them in the future. Additionally, students will be instructed in (1) Language and intercultural skills 2) Organizational and professional skills through the research of a topic, securing a guest that can speak about the topic, preparing the guest with agreed up points of conversation, and practicing before recording the interview (3) Technical skills through using recording equipment and editing software.

Spotting the Character: How to Collect Elements of Characterisation in Literary Texts?

Ioana Galleron
ioana.galleron@univ-paris3.fr
U. Sorbonne Nouvelle Paris 3, France

Fatiha Idmhand
fatihaidmhand@yahoo.es
U. de Poitiers, France

Cécile Meynard
cecile.meynard@gmail.com
U. d'Angers, France

Pierre-Yves Buard
pierre-yves.buard@unicaen.fr
U. de Caen

Julia Roger
julia.roger@gmail.com
U. de Caen, France

Anne Goloubkoff
anne.goloubkoff@unicaen.fr
U. de Caen, France

What is a literary character made of? To this question, a pragmatic answer is to say that it exists as a result of a chain of different linguistic elements, scattered throughout the text. The aim of this paper is to propose a digital method for collecting these elements, so as to analyse their nature, to observe their repartition in texts, and, ultimately, to contribute to a deeper understanding of the functions the literary device called "character" assumes in a text.

Projects dedicated to named-entity recognition put a great deal of effort into using Natural Language Processing (NLP) techniques for identifying names of people, places and organisations mentioned in various types of discourses, especially political ones, as well as the co-referential chains built on the basis of these names. However, in spite of important advances in the field, much remains to be done in order to train the computer to link correctly various phrases referring to the same entity, as well the pronouns pointing to it (see Schnedeker and Landragin, 2014). In our case, we are interested in such elements of a co-referential chain that bear characterization features, and this is, inevitably, a supplementary complication. In addition, we are interested in certain elements (eg. "his brother" in the phrase "John is his brother") that are often left aside in named entity recognition, as performing another functions than strictly pointing towards an entity. Therefore, NLP techniques did not appear adapted to our needs.

We will therefore resort to "crowd-reading", as another means, offered by the explosion of the digital sphere, to make sense from texts. Very similar to the crowdsourcing, the crowd-reading asks to benevolent contributors to annotate a document, bringing in their own view and understanding, instead of transcribing, or adding in information based on a (sometimes external) form of authority. Considering the nature of the work to be done, the crowd-reading appeared as a valid technique in our case.

In a first stage, we submitted a short text (Julio Cortazar "Continuidad de las parques") to the manual annotation of a hundred students from our universities. This brought to the fore the sheer variety of elements considered to be participating to the characterization of a literary "person" (nouns and adjectives, of course, but verbs and adverbs too), as well as the need to dispose of a controlled vocabulary allowing to understand what kind of characterization each respondent attached to the various strings of characters selected as participating to this function.

In a second phase, we have decided to build an interface, offering a more ergonomic experience to our respondents, and allowing us to extract automatically the linguistic elements selected, as well as to group them by categories. Built with XML Mind, this interface is in fact based on a text lightly encoded with TEI tags, in which our respondents add, every time they select a string of characters, an <rs> tag, bearing in addition two attributes:

a @key attribute, defined by each respondent every time he or she encounters a new character. The keys are subsequently available for reuse in the rest of the text. We expect the number of keys to vary considerably from a reader to another.

an @ana attribute, with a set of constrained values. Based on another project of character analysis, these values have been defined in Galleron, 2017, and cover aspects such as the ontological type of a character, its sex, age, family position, nationality, occupation, and so on.

The text submitted to annotation has been changed for this second experience: it concerns now the "Jardin aux sentiers qui bifurquent" ("Jardín de los senderos que se bifurcan") by Jorge Luis Borges. At the date of this proposal, the second campaign of crowd-reading has not started yet. We'll have a significant number of answers before the DH conference. Our respondents will be recruited again amongst the students enrolled in literary studies in our universities: while they have a certain level of training in linguistics, literature and poetics, so as to be able to recognise the type of linguistic elements we look for, their reading still remains close of the "non-informed", "amateur" reading of the "man in the street" (see Dufays, 205).

The results will be analysed so as to observe what kind of linguistic units have been identified most often, and what kind of values of the @ana attribute have been mobilised most often. We will further discuss the divergences between the selected elements, and those we were expecting to be selected. This will allow us, on the one hand, to suggest a possible use of our interface as a remediation tool in literary studies, for students with difficulties in extracting pertinent information from a text, so as to respond a specific task. On the other hand, we will advance an hypothesis about the observed distribution of the most frequent elements of characterization, that are far to appear where, intuitively, one would expect them to

be grouped together (so as to "introduce" the character) as shown by our first campaign of crowd-reading, and by our own annotation endeavours.

References

Dufays, Jean-Louis, Gemenne, Louis, et Ledur, Dominique (2005). *Pour une lecture littéraire. Histoire, théories, pistes pour la classe*, Bruxelles: De Boeck – Duculot.

Galleron, Ioana (2017). Conceptualisation of theatrical characters in the digital paradigm: needs, problems and foreseen solutions. *Human and Social studies*, De Gruyter. 6: 1 (Published Online: 2017-04-18 | DOI:https://doi.org/10.1515/hssr-2017-0007).

Schenedeker, Catherine; Landragin, Frédéric (2014). Les chaînes de référence: présentation. *Langages*, 3:145, 3-22.

Archivos Abiertos y Públicos para el Postconflicto Colombiano

Stefania Gallini
sgallini@unal.edu.co
Universidad Nacional de Colombia, Colombia

El 26 de noviembre de 2016 el Congreso colombiano votó su aprobación al "Acuerdo final para la terminación del conflicto y la construcción de una paz estable y duradera", firmados por el presidente de la República Santos y el comandante del Estado Mayor central de la guerrilla de las FARC-EP Jiménez. El fin legal del largo conflicto armado interno puso en marcha la creación de una nueva institucionalidad estatal para transitar hacia los necesarios procesos de, que lentamente y entre muchas resistencias ha ido tomando cuerpo en el año que siguió a la aprobación parlamentaria.

Dos de los institutos más significativos creados por los Acuerdos de paz son la Jurisdicción Especial para la Paz y la Comisión para el Esclarecimiento de la Verdad, la Convivencia y la no Repetición. Estas instancias se suman a los esfuerzos de la Comisión Nacional para la Memoria Histórica (creada en 2011) y de distintas iniciativas (regionales y locales, públicas y privadas, académicas, cívicas y gremiales) por recopilar, organizar, preservar y a menudo hacer público un complejo acervo de información acerca de la historia y la memoria del conflicto colombiano.

Estos archivos y repositorios – los que ya existen y los que la implementación del Acuerdo creerá a partir de los hallazgos de la justicia transicional, la Comisión de la Verdad y las iniciativas de la sociedad civil y organizaciones de derechos humanos – son las fuentes con las cuales el país apuesta reconstruir las bases de justicia, reparación y no repetición que deberán sostener el nuevo pacto social de la nación.

La situación no es nueva en el escenario global. Durante el siglo XX y lo que va corrido del XXI, muchas veces se han constituido archivos de derechos humanos, de comisiones de la verdad, de memoria de las víctimas al finalizar un conflicto armado interno o una dictadura. Los ejemplos van desde el Cono Sur latinoamericano á Irlanda, Suráfrica y Guatemala, para citar algunos.

Sin embargo, a diferencia de los casos anteriores, los archivos del conflicto armado interno de Colombia se construyeron, consolidarán e interrogarán en pleno auge de la era digital. Esta circunstancia influye de manera radical en cuestiones de adquisición, preservación, seguridad y acceso a la información, pero también implica dos consecuencias importantes: la oportunidad que la dimensión colaborativa y abierta de los archivos en la era digital brinda para alcanzar los objetivos de esclarecimiento de la verdad histórica y judicial, y el protagonismo que la adopción de técnicas y herramientas digitales y de la informática humanística puede jugar para permitir la efectiva apropiación social de los datos.

Los archivos de la era digital son intrínsecamente distintos a sus antepasados. Estos son repositorios participativos, de-institucionalizados, de acceso abierto, de contenidos digitales, de-localizados, que funcionan en red con otros archivos y repositorios documentales, capaces de generar y actualizar continuamente sus formas de hacerse accesible y apropiable por parte de un público heterogéneo.

El conflicto armado interno dejó detrás suyo un enorme volumen de datos que, junto con la complejidad de la gestión de esta información sensible, requiere pensar en metodologías y herramientas tanto archivísticas como informáticas que aseguren la interoperabilidad de los datos, la seguridad de la preservación y no obsolescencia tecnológica de la información, el procesamiento automatizado (incluyendo la georeferenciación) de metadatos, el tratamiento, la transcripción automatizada y codificación de fuentes orales, entre otros aspectos.

El Pensamiento Archivístico crítico y las Humanidades Digitales ofrecen una matriz epistemológica y técnica para pensar los archivos del conflicto y gestionar su información propiciando su visibilidad ante la opinión pública, visualización adaptada a las necesidades de distintos actores, interoperabilidad, traducibilidad en evidencia judicial, entre otros. Se trata de pensar de qué manera tanto las herramientas como la perspectiva cultural de las DH pueden contribuir a esta tarea colectiva.

La ponencia presentará los avances del proyecto de investigación que, con asentamiento en el Laboratorio de Cartografía Histórica e Historia Digital de la Universidad Nacional de Colombia en Bogotá, un grupo de docentes y estudiantes está desarrollando sobre las temáticas descritas, que tiene además el propósito de ofrecer lineamientos de políticas públicas en el tema de los archivos de la historia y la memoria del conflicto armado interno colombiano en la era digital.

Respuestas a evaluaciones – enero 2018

Dos de las evaluaciones sugieren mayor precisión en la propuesta. Agradezco esta oportunidad para poder aclarar que la ponencia presentará los primeros avances de un proyecto de investigación que está apenas empezando y que discute una materia – la nueva institucionalidad de Verdad, Justicia y No Repetición del conflicto colombiano - que también ha sido formalizada hace tres semanas (enero 2018). La participación en DH2018 durante la fase inicial del proyecto justamente apunta a encontrar en el congreso aquella retroalimentación de pares que no es posible siempre encontrar en el ámbito nacional, donde las HD se encuentran en un estadio todavía embrionario, aunque acelerado y entusiasta.

El objetivo principal de la ponencia es por ende presentar críticamente el caso colombiano como ocasión de construcción (a veces) y organización (a veces, cuando los repositorios ya existan) de archivos para la reconstrucción de la memoria, la historia y a menudo la verdad judicial del conflicto colombiano, en un momento histórico en el cual la revolución digital y la expansión de sus consumidores/actores abre el escenario a posibilidades, pero también a desafíos no antes conocidos.

Se tendrán a la vista archivos que ya existen (i.e. Centro Nacional de Memoria. "Archivo Virtual de Los Derechos Humanos Y Memoria Histórica." http: //www.archivodelosddhh.gov.co/, los de matriz periodística como Verdabierta, los de ongs y asociaciones de víctimas, los de historia oral de los movimientos, ver por ej. Suárez Pinzón, Ivonne. "El Archivo Oral de Memoria de Las Víctimas AMOVI-UIS: Un Archivo de Derechos Humanos." UIS y Corporación Compromiso, 2014. https://www.uis.edu.co/webUIS/es/amoviUIS/documentos/presentacionAMOVI-UIS.pdf.), pero también archivos los que se van a levantar a partir de las nuevas indagaciones e instituciones (p. ej. las instancias de la JEP y de la Comisión de la Verdad). En la ponencia será posible referirse a archivos o datasets más en detalle, pero sería apresurado indicarlos en esta fase que es todavía exploratoria. Igualmente, aunque me interesan especialmente algunos problemas "técnicos" (la interoperabilidad y la georeferenciación), la intención de la ponencia es presentar problemas y discutir desafíos, a la luz de una discusión que es álgida en Colombia (Centro Nacional de Memoria Histórica. "Política Pública de Archivos de Graves Violaciones a Los Derechos Humanos, Infracciones a Los Derechos Humanos, Infracciones Al DIH, Memoria Histórica Y Conflicto," February 2015. http://www.centrodememoriahistorica.gov.co/descargas/mesasRegionalesArchivos/Politica-publica-archivos-integrada-20-2-1.pdf).

Los referentes teóricos los he encontrado – como se indica en el Abstract – en el Pensamiento Archivístico crítico (MacNeil, Heather, and Terry Eastwood, eds. *Currents of Archival Thinking, 2nd Edition*. Santa Barbara, California: Libraries Unlimited, 2017; Schwartz, Joan M., and Terry Cook. "Archives, Records, and Power: The Making of Modern Memory." *Archival Science*, no. 2 (2002): 1–19; Weld, Kirsten. *Paper Cadavers: The Archives of Dictatorship in Guatemala*. American Encounters/Global Interactions. Durham: Duke University Press, 2014; Centro Nacional de Memoria Histórica. "Seminario Internacional Archivos Para La Paz." Centro Nacional de Memoria Histórica, 2014. http://www.centrodememoriahistorica.gov.co/centro-audiovisual/videos/seminario-internacional-archivos-para-la-paz) y las Humanidades Digitales.

References

Sanmiguel, Lahdy Diana del Pilar Novoa, and Diego Andrés Escamilla Márquez. "Archivos orales y memoria del conflicto armado interno colombiano: retos y posibilidades." *Advocatus* 14, no. 27 (March 1, 2017): 153–73. http://www.unilibrebaq.edu.co/ojsinvestigacion/index.php/advocatus/article/view/732.

Centro Nacional de Memoria Histórica. "Política Pública de Archivos de Graves Violaciones a Los Derechos Humanos, Infracciones a Los Derechos Humanos, Infracciones Al DIH, Memoria Histórica Y Conflicto," February 2015. http://www.centrodememoriahistorica.gov.co/descargas/mesasRegionalesArchivos/Politica-publica-archivos-integrada-20-2-1.pdf.

Centro Nacional de Memoria Histórica. "Seminario Internacional Archivos Para La Paz." Centro Nacional de Memoria Histórica, 2014. http://www.centrodememoriahistorica.gov.co/centro-audiovisual/videos/seminario-internacional-archivos-para-la-paz.

"Algunas Notas Sobre Los Repositorios Institucionales (Parte I) -." *Infotecarios* (blog), August 22, 2017. http://www.infotecarios.com/algunas-notas-los-repositorios-institucionales-ri-parte-i/.

Colectivo de Historia Oral. "Colectivo de Historia Oral (Colombia)." *Colectivo de Historia Oral* (blog). Accessed November 28, 2017. https://colectivohistoriaoral.wordpress.com/category/historia-oral/.

Jelin, Elizabeth. *Los Trabajos de La Memoria*. Memorias de La Represión 1. Madrid: Siglo XXI, 2002.

Suárez Pinzón, Ivonne. "El Archivo Oral de Memoria de Las Víctimas AMOVI-UIS: Un Archivo de Derechos Humanos." Universidad Industrial de Santader y Corporación Compromiso, 2014. https://www.uis.edu.co/webUIS/es/amoviUIS/documentos/presentacionAMOVI-UIS.pdf.

Brodsky, Marcelo. "Buena Memoria," 1997. http://v1.zonezero.com/exposiciones/fotografos/brodsky/defaultsp.html.

Histórica, Centro Nacional de Memoria. "Archivo Virtual de Los Derechos Humanos Y Memoria Histórica." Accessed November 28, 2017. http://www.archivodelosddhh.gov.co/.

Humanidades Digitales en Cuba: Avances y Perspectivas

Maytee García Vázquez
maytee.garcia.vazquez@gmail.com
Cubaliteraria, Cuba

Sulema Rodriguez Roche
sulema1985@gmail.com
Universidad de La Habana, Cuba

Ania Hernández Quintana
aniahdez@fcom.uh.cu
Universidad de La Habana, Cuba

Hasta hace pocos años, la Web proporcionaba información unilateralmente. Por un lado, estaban las grandes empresas e instituciones, que eran las que poseían espacio en la red, y por el otro, los usuarios, en actitud receptora y pasiva. Esa tendencia está siendo modificada por el movimiento denominado Web 2.0 que propugna que todos somos potenciales surtidores de contenidos y creadores de los registros del conocimiento. La evolución natural en la sociedad de la información se expresa en la metáfora del paso del ciudadano 1.0, consumidor de recursos, al ciudadano 2.0, creador de recursos, evidenciando la horizontalización y democratización de las fuerzas que rigen la red. Las Humanidades Digitales son un resultado de esas transfiguraciones digitales y como un ámbito disciplinar de convergencia cultural e investigativa, dejó de ser una moda para convertirse en una urgencia para cultura y memoria del mundo.

En Cuba, ya se notan de forma clara las comprensiones sobre la necesidad de fomentar este ámbito de teoría y práctica; aprovechando un contexto de crecimiento tecnológico en el país, en el que empieza a notarse la presencia digital en casi todos los sectores de la sociedad, con demandas infocomunicacionales y culturales crecientes y multilaterales. El presente trabajo tiene la finalidad de compartir los avances más visibles, así como las proyecciones hacia lo profesional, lo académico, lo investigativo y lo institucional, como parte de la agenda de las Ciencias Sociales y Humanísticas en Cuba.

Se abordan los conceptos de partida de las Humanidades Digitales para Cuba, como un campo interdisciplinar dispuesto para dar espacio a las reflexiones y prácticas suscitadas por los cambios que produce la introducción de las tecnologías digitales en el universo de la cultura y la información; con énfasis en el desafío epistemológico y metodológico para la articulación de conocimientos y prácticas profesionales y de investigación que enfrentan las ciencias humanas en el cibermundo. Se abordan como una oportunidad de transformación sinergética del consumo cultural, cada vez más urgente, en tanto se demanda mayor conocimiento de investigadores y usuarios, que a su vez demandan información, de forma activa, en espacios colaborativos.

La manera en que se puede trabajar en las Humanidades Digitales, partiendo de los límites casi precarios del desarrollo tecnológico en Cuba, ha creado proyectos sui géneris. Estos son inimaginables en Europa o en Estados Unidos. Varios ejemplos: la circulación de USB, tan común en el paso de los archivos, crea un sistema de distribución de conocimiento muy diferente. Dentro de esos sistemas de distribución, hay una relación política diferente hacia los derechos de autor que cambia la manera en que el conocimiento fluye. Se crean también proyectos digitales que pueden pasarse de máquina en máquina por USB, muy diferentes a aquellos que se colocan en un servidor. Se pueden desarrollar nuevas pistas para el análisis textual, por ejemplo.

En Cuba esto tiene un aspecto político que no siempre se verbaliza, pero son reconfiguraciones de trabajo en equipo que transforma la manera en que la investigación se ha hecho hasta ahora. Eso cambia la relación hacia la investigación a nivel social y su papel en la formación de grupos sociales para el trabajo cultural, ahora más equitativas. De la misma forma, existen jerarquías que vienen de la organización tradicional del trabajo de investigación, en la cual el personal técnico se encuentra separado del investigador y ambos de los bibliotecarios; transitando hacia un modelo colaborativo e interdisciplinar.

Los primeros pasos en Cuba proceden de los años 90 del siglo XX, marcado por un período social complejo en el país, con la publicación del primer libro digital. De forma aislada, varias instituciones académicas y de investigación han realizado proyectos de Humanidades Digitales y finalmente en mayo de 2017, se avanzó hacia una estrategia de articulación, con el primer curso de Humanidades Digitales impartido por profesoras del Laboratorio de Innovación de Humanidades Digitales (LINHD). Uno de los resultados de ese encuentro fue la disposición de crear iniciativa profesional que articule y visibilice el trabajo en Humanidades Digitales que se ha venido realizando en el país.

Se han identificado algunos focos muy visibles, en especial el de la carrera Ciencias de la Información, de la Universidad de La Habana, cuyo propósito es transversalizar las Humanidades Digitales en el campo de las Ciencias de la Información en Cuba. Ese proyecto, con una vocación claramente pedagógica, comenzó a trabajar en noviembre de 2016. Las investigaciones resultantes del primer año tienen como principio que las Humanidades Digitales se interesan por el estudio, preservación y acceso a la información registrada, objetivos que disciplinariamente enfrenta también la comunidad científica, académica y profesional del campo de esta carrera, y que para ello las Humanidades Digitales se distinguen por el uso intensivo de métodos de procesamiento automático y semiautomático, expresados científicamente a través de contribuciones en congresos, experiencias en laboratorios de I+D+I y en programas de formación universitaria.

Asimismo, el equipo se preocupa por las colecciones digitales, y en consecuencia por el requerimiento de protocolos de preservación de sus contenidos, determinados por funciones y estructuras más sofisticadas que modifican los procesos de gestión de esos recursos electrónicos a través de métodos globales como open data, linked data, linguistic link data y TEI.

Es un hecho que las bibliotecas digitales clasifican como uno de los sistemas de información más complejos por la multidisciplinariedad que implican; además, por la convergencia de conocimientos que supone organizar, difundir y usar información en este tipo de repositorios; y especialmente por las complicadas e interdependientes multirrelaciones que activa, que llegan a la autotransformación y a la construcción de contrahegemonías emancipatorias.

Las primeras siete investigaciones del grupo exploraron los conceptos de las humanidades digitales en su multiplicidad y complejidad, las redes profesionales y los currículos de humanidades digitales. Además, se realizaron indagaciones más enfocadas a la solución de problemas como el procesamiento de una revista infantil con carácter patrimonial con el método linked data y la creación de un espacio de aprendizaje colaborativo para estudiantes de Ciencias de la Información.

En Feria del Libro de la Habana, a realizada en el mes de febrero de 2018, se desarrolló un programa especial denominado "Cuba Digital". Libros digitales, aplicaciones móviles, conferencias de investigadores nacionales y extranjeros y proyectos cubanos, entre otros, integraron las propuestas de ese espacio, que contó con la coordinación de la Editorial Cubaliteraria.

La lista de proyectos e instituciones cubanas involucradas en proyectos que apuntan a las humanidades digitales, crece. Un levantamiento preliminar en la capital destaca los siguientes: Instituto de Historia de Cuba; la Fundación Fernando Ortiz con su proyecto Archivo de la palabra; el Instituto de Literatura y Lingüística, dedicado al estudio y descripción del español de Cuba; el proyecto www.postdata.club, del Centro Martin Luther King; la Biblioteca Nacional de Cuba con su catálogo digital, y el proyecto Mirador, en colaboración con Infomed, la red nacional de información en salud en Cuba, enfocado en el rescate de colecciones patrimoniales en Cuba y también aliado del Grupo de Investigación de Humanidades Digitales para las Ciencias de la Información, de la Universidad de La Habana. Convocados por la editorial digital Cubaliteraria, del Instituto Cubano del Libro, que lidera la producción de ebooks y multimedias sobre literatura cubana.

En el futuro cercano, se proyecta una postura sinérgica que aproveche los aprendizajes de las Humanidades Digitales del Sur, con un enfoque parecido a la realidad cubana y se articule en una iniciativa nacional o proyecto de Asociación de Humanistas Digitales.

Corpus Jurídico Hispano Indiano Digital: Análisis de una Cultura Jurisdiccional

Víctor Gayol
vgayol@colmich.edu.mx
El Colegio de Michoacán, A.C., Mexico

El proyecto *Corpus de derecho castellano-indiano / digital* es una propuesta colectiva e interdisciplinaria que abarca la compilación, digitalización, procesamiento, macroanálisis y publicación anotada en línea del conjunto de los textos jurídicos vigentes en el marco de la monarquía castellana entre el siglo XIII y principios del XIX. El núcleo principal del proyecto es la construcción de un modelo para el macroanálisis de estos textos jurídicos y, en consecuencia, la generación de herramientas analíticas y de consulta del corpus que permitan comprender la interrelación entre sus distintos elementos semánticos y conceptuales y su transformación a través de los siglos y así proponer una interpretación de cómo es que posiblemente funcionaban en el contexto del discurso y la práctica en el orden jurídico tradicional de la cultura jurisdiccional, tanto en el ámbito de la doctrina, del ejercicio de la potestad normativa como en el del actuar cotidiano del aparato de gobierno e impartición de justicia.

El proyecto implica diversas conexiones y diálogos en diversos ámbitos. En el ámbito interdisciplinario, entre los historiadores de la corriente crítica (cultural) del derecho, lingüistas, humanistas digitales y programadores; en el ámbito teórico y metodológico, entre dos posturas acaso antagónicas en apariencia: la lectura densa y cercana de los textos jurídicos hecha por la historia cultural del derecho a lo largo de varias décadas y la lectura distante. Lo anterior nos obliga a discutir ciertos principios teóricos, como lectura densa, tomada por la historia cultural del derecho de la idea de descripción densa (Geertz, 1973), como sistema capaz de ser leído como texto en relaciones contextuales, o un nivel más complejo (Genette, 1992) y su noción de transtextualidad. Varios historiadores del derecho han aplicado incluso algo parecido a la lectura cercana del criticismo literario (Clavero, 1991). Esto interesa al estudiar el derecho de antiguo régimen frente a la posibilidad de aplicación de metodologías computacionales enfocadas, generalmente, a una lectura distante (Moretti, 2013) en la búsqueda de estructuras formales mediante el análisis de grandes cantidades de texto/data. Es justamente necesario pensar en la posibilidad de ensayar no sólo una minería de texto cuantitativa sino en aspectos más cualitativos, modelando campos semánticos que se transforman históricamente.

Cabe aclarar que el criterio de selección de fuentes para la conformación del corpus es complejo y presenta muchos problemas. Responde a una historiografía que ha definido el campo de lo jurídico en el antiguo régimen

hispánico como algo más allá del texto jurídico normativo (entendido como ley). Incluye la doctrina de los juristas y de los teólogos por considerarse que la cultura jurídica tiene una estrecha relación con la doctrina católica. El corpus completo abarcaría tanto normas como doctrina y costumbre y se consideran textos jurídicos producidos tanto en Castilla como en los territorios americanos de la monarquía. Por lo tanto, no se trata de un corpus reunido de antemano en su propia época, sino de un corpus compuesto por el conjunto de la comunidad de historiadores dado que se ha analizado su utilización práctica a lo largo de los siglos y en un contexto cultural determinado (Castilla y sus dominios ultramarinos entre los siglos XIII y XIX). Tener claro cómo suponemos que se definía un texto jurídico en el antiguo régimen es de suma importancia ya que el interés del proyecto es generar una comunidad colaborativa de investigación interdisciplinario que determine sus elementos semánticos necesarios para poder caracterizar digitalmente este tipo de textos. Esto es primordial puesto que son textos completamente distintos de los literarios o de otra índole que se han considerado, por ejemplo, en la iniciativa TEI. Dicho de otra forma, el nodo fundamental del problema es cómo se construye un corpus histórico jurídico particular para que sea útil en las humanidades digitales.

Como la reunión del corpus completo es un proyecto a muy largo plazo, en una etapa piloto consideramos que trabajar con los textos normativos puede ser suficiente para ensayar la propuesta de un modelo flexible y escalable. Además, para el caso de los textos normativos ya existe un ordenamiento y un proceso de digitalización previo de esa parte del corpus. De unas 35,355 normas referenciadas se han puesto en línea, de manera digital básica, 26,831 por un grupo de académicos españoles que viene trabajando al respecto desde la década de 1970 y en el que se han ya recogido la mayor parte de las normas legisladas entre el año 1020 y 1868.

Por tanto, el objetivo de esta ponencia es discutir los diferentes ejes de nuestra propuesta teórica: 1) el aspecto de su realidad digital, es decir, cuáles son los requisitos para una digitalización óptima de fuentes jurídicas que se presentan en la realidad de maneras diversas –manuscritas, impresas, cuyos contenidos varían ortográfica y semánticamente a lo largo de los siglos-, 2) el problema de qué se concibe como texto propio de la cultura jurisdiccional en el orden jurídico tradicional –no sólo los obviamente jurídicos en apariencia-, y, en consecuencia, 3) los retos que implica el diseño de herramientas digitales propias que permitan el macroanálisis de los textos como datos masivos. Esto, a su vez, implica un problema mayor y de fondo que es el de la conexión entre un necesario abordaje hermenéutico de los textos jurídicos (lectura densa) en una perspectiva de larga duración –desde la baja edad media hasta el fin de la edad moderna– para entender su contexto cultural de sentido, y el reto de procesar dichos textos entendidos como corpus y en forma de datos masivos mediante computadora (lectura distante), no sólo en procesos de segmentación del corpus para su visualización (nubes de palabras, frecuencias relativas y absolutas, KWIC), sino la posibilidad de ensayar, sobre todo, un modelado tópico semántico con objeto de reflexionar sobre cuál sería un modelo de macroanálisis adecuado para este tipo de corpus. Finalmente, proponer un modelo particular para la edición digital del corpus de los textos jurídicos propios de la cultura jurisdiccional del orden jurídico tradicional.

Referencias

Clavero, B. (1991). *Antidora: antropología católica de la economía moderna*. Milano: Giuffrè.

Geertz, C. (1973). *The Interpretation of Cultures: Selected Essays*. New York: Basic Books.

Genette, G. (1992). *The architext : an introduction*. Berkeley: University of California Press.

Moretti, F. (2013). *Distant Reading*. London: Verso.

Designing writing: Educational technology as a site for fostering participatory, techno-rhetorical consciousness

Erin Rose Glass
erglass@ucsd.edu
UC San Diego, United States of America

In the past ten years, advancements in computing technology have lent themselves to diverse applications in teaching and learning such as seen with MOOCS, learning managements systems, networked collaborative pedagogy, virtual/augmented reality course modules, and algorithmic-driven approaches to personalized learning. While these engagements represent a variety of exciting (though often controversial) new directions for educational technology, the changing socio-technological conditions of our information landscape call for new critical approaches towards its development and use. Information communication technology (ICT) in educational settings should not only be evaluated according to the way it supports intended learning goals, but also according to the type of technological consciousness it produces in students. In this paper I will draw from methods and values in participatory design (Simonsen), critical pedagogy (Freire; Shor), and the digital humanities (Drucker & Svensson; Rockwell & Sinclair) to outline a way that academic technological practices and infrastructure might be re-engineered to foster more critical and participatory relationships to digital technology within higher education. I will focus specifically on how this approach has particular value for the teaching and use of writing in un-

dergraduate and graduate education in that it enables a praxis-oriented approach to analyzing and designing digitally-mediated rhetorical situations within and beyond academia. I will then describe KNIT, a digital commons at UC San Diego that aims to develop a participatory model of educational technology, and describe the challenges and opportunities experienced in its development.

Participatory approaches to ICT

The general user has little expectation or ability of being able to understand or modify the code of ICTs that mediate their everyday communicative activities, such as email, social media, Internet searching, or text editing. While this lack of critical user participation in software oversight and production may appear as natural, inevitable, and relatively inconsequential, I will argue that it has been normalized through corporate technical policies, cultural myths regarding programming, and the use of technology in educational settings. To demonstrate the range of alternatives to passive relationships to software, I will point to a number of software cultures, projects, and visions in which the everyday user has greater opportunity to democratically participate in shaping the technical functionality and policy of their digital tools. I will argue that examples such as the Free Software community, the Platform Coop movement, and Alan Kay's 1968 vision for Dynabook represent promising alternative software models that foster participatory design consciousness in the general user that could be fruitfully applied in educational settings. By implementing tools in the classroom that allow for participatory design and oversight, students would have the opportunity to experience greater forms of creative and critical control over ICTs that might lead them to question the lack of similar rights with regard to ICTs in everyday life. In this way, fostering participatory design approaches to digital technology stands as one promising approach to fostering critical and practical resistance in students to exploitative practices inherent in everyday ICTs such as dataveillance and algorithmic influence and manipulation. It also offers the possibility of turning educational technology into a site for producing open source ICT alternatives for general public use.

Techno-rhetorical consciousness

Participatory design approaches towards educational technology also have direct application for writing-intensive courses in the humanities in that they can help foster "techno-rhetorical consciousness," or a sensitive understanding of the way digital technology mediates rhetorical situations. By providing students with the perspective and control over ICTs normally only afforded by corporate or administrative entities, students have the opportunity to study more directly the way ICTs mediate their intellectual activities and communities, and explore how technical modifications might help support personal and collective intellectual goals and values. For example, access to data produced and transmitted through ICTs would enable students to use text analysis techniques from the digital humanities to study patterns in their individual and collective intellectual activities for the purpose of understanding the social dynamics of knowledge production and transmission. It would allow them to gain basic familiarity with algorithmic techniques that have increasing power in everyday life. And it would also provide students with the opportunity to experiment with how different aesthetic and algorithmic design features might better support individual cognitive activities related to writing process or productive intellectual exchange among students. These opportunities would not only have rich potential for the use and development of educational technology itself, but would also help students consider the way digital technology mediates the production and transmission of knowledge and power in everyday life.

KNIT, a digital learning commons

To explore some of these ideas in practice, we have launched KNIT, a digital commons for UC San Diego and institutions of higher education in the San Diego area. KNIT uses the free and open source software package Commons in A Box and thus, unlike many forms of proprietary software in educational settings, remains open to critical study and modification by the user community. In the final portion of my talk, I will discuss how we are using KNIT to test-drive participatory design practices for educational technology at UC San Diego and how we envision using it to give students a leadership position in its development and governance. I will also discuss the institutional, technical, and educational challenges of this approach and provide recommendations and resources for those interested in experimenting with this method at their home institution.

References

Drucker, Johanna, and Patrik BO Svensson. *The Why and How of Middleware*. Vol. 10, no. 2, 2016. *Digital Humanities Quarterly*.

Freire, Paulo. *Pedagogy of the Oppressed*. Continuum, 1993.

Rockwell, Geoffrey, and Stéfan Sinclair. *Hermeneutica: Computer-Assisted Interpretation in the Humanities*. MIT Press, 2016.

Shor, Ira. *Critical Teaching and Everyday Life*. University of Chicago Press, 1980.

Simonsen, Jesper, and Toni Robertson. *Routledge International Handbook of Participatory Design*. Routledge, 2012.

Expanding the Research Environment for Ancient Documents (READ) to Any Writing System

Andrew Glass
asg@uw.edu
Microsoft Corp., University of Washington, United States of America

The Research Environment for Ancient Documents (READ) is an integrated Open Source web platform for epigraphical and manuscript research. The original goal of the READ platform was to support scholars in researching and presenting studies of handwritten documents and inscriptions preserved in Gāndhārī language using the Kharo hī script. Since many of the workflows supported by READ are common to epigraphic and manuscript studies in other textual traditions we wanted to investigate how READ could be generalized to support other writing systems. This presentation will share the results of that investigation with examples from English, Aramaic, Chinese, and Mayan.

Three core components of the READ data model depend on the writing system used by the source material:

1. The link between physical and textual data
2. The constraint mechanism that allows a user to edit text without disrupting links
3. The sort weight API that allows data in the model to be displayed in an expected sort order

Part One. The database model underlying READ was designed to reflect the separate components and layers of interpretation which manuscript scholars and epigraphers typically use in their research (letter forms => paleography; graphemic units => phonology; inflectional forms => morphology, etc.). Furthermore, the model recognizes a continuum of factual confidence beginning from statements of fact (e.g., the name of a collection in which an item is kept), to data which may have multiple or variant interpretations (e.g., the transcription of a sample of writing). Such variant data is linked back through the model to original facts. At the crux of this system of links are the references between segments on an image each containing an orthographic unit in the writing system and the transcription of that unit. Because READ was originally developed for Kharo hī, an alphasyllabary or Abugida-type writing system, this link maps image segments to syllable clusters. Other writing systems can be supported by mapping the syllable cluster to the appropriate orthographic units. This has been tested by mapping syllable clusters as follows: English letters, Aramaic syllables, Chinese logographs, and Mayan syllables and logographs.

Part Two. READ is intended to be a working environment for born-digital text editions. A critical feature of the model is that links created within the system must be preserved during repeated editing. The editing interface allows users to modify linked syllable clusters. By constraining edits to valid transcriptions of a syllable cluster defined for the language, READ can keep track of user edits and prevent links from being broken. Other writing systems can be supported by defining the valid transcription forms for the orthographic units. In most cases this is less complex than for ak ara-based writing systems. This has been tested by defining valid orthographic units as follows: English – Consonants, Vowels; Aramaic syllables - Consonants, Vowels, Consonant with modifier; Chinese – Logograph; Mayan – Logograph, CV syllable. All systems also permit orthographic units to be Digits and Punctuation signs.

Part Three. READ uses custom sort tables to weight the orthographic units and subunits used by the model. Having custom sort tables allows correct sorting of Romanized transcription when the expected sort order is not equal to standard 'ABC' order. Other writing systems represented in Romanized transcription with non-standard sorts require dedicated sort tables. Alternatively, writing systems represented in native script via Unicode may be sorted via their Unicode weights. This has been tested using standard ABC weights for English, custom weights for Mayan transcription, Unicode weights for Hebrew transcription of Aramaic, and Pinyin sort weights for Chinese logographs.

The outcome of these investigations has been that the READ architecture is generalizable, and that the READ platform could be employed by projects with a focus on documents in any writing system.

The Latin American Comics Archive: An Online Platform For The Research And Teaching Of Digitized And Encoded Spanish-Language Comic Books Through Scholar/Student Collaboration

Felipe Gomez
fgomez@andrew.cmu.edu
Carnegie Mellon University, United States of America

Scott Weingart
scottbot@cmu.edu
Carnegie Mellon University, United States of America

Daniel Evans
djevans@andrew.cmu.edu
Carnegie Mellon University, United States of America

Rikk Mulligan
rikk@cmu.edu
Carnegie Mellon University, United States of America

Overview

This short paper looks into the process of developing the Latin American Comics Archive (LACA), a project created by our team at Carnegie Mellon University. LACA combines ongoing research in the Humanities with digital technologies as a tool for enhancing access and analysis capabilities for both scholars and students of these materials.The curated digital archive includes representative samples of Latin American comics digitally encoded in Comic Book Markup Language (CBML), while a technical foundation combining the open source content management system Omeka with TEI Boilerplate offers a customizable front-end for public or restricted access to the individual items and curated collections of the comics. This allows students and researchers access to source materials and possibilities to collaborate in their exploration, definition, tagging, and annotation for the analysis of visual and verbal language, cultural and linguistic characteristics or themes, and a variety of formal categories.

Statement of the problem

Despite the overdue growing recognition of the genre of comics in academia, the study of foreign/second language comics within the United States has encountered specific obstacles. Primary-source research of Spanish-language comics has often proved to be challenging. Among other difficulties, collections are most often housed in the source countries, and a desired piece of documentation may sometimes be in libraries hundreds or thousands of miles away. Items may be both in public and private hands, and access to certain items is often highly restricted due to their fragility, rarity, and value. Oftentimes, specific documents aren't cataloged in the archives' container lists, making the identification, location, and access of relevant materials problematic. When using traditional research methods, these challenges have to be confronted and resolved by the researcher, who works in isolation with the source documents. Many of these issues also generate constraints in the realm of teaching, where the limitations to the access of sources restricts course conceptualization and implementation, and where students don't usually have much agency or opportunities to engage in larger debates and conversations with other students or scholars of Latin American comics.

Digital tools have the potential to facilitate or solve many of these issues for research and teaching of this important cultural and literary medium. Indeed, they have the ability to address precisely the core values that Spiro (2012) associates with work in the Digital Humanities -- openness, collaboration, diversity, experimentation, collegiality, and connectedness. These tools can, for instance, create optimal opportunities to view and use some of these sources online, thus granting access to an audience who may never have had the chance to see them in the "analog" era, and opening and expanding the possibilities for a richer and deeper type of collaborative research. Our goal is to expand the possibilities of using Spanish-language comics by identifying and piloting the use of digital tools with which digital copies of representative Latin American comics can be made, accessed, and annotated in collaboration with students and scholars. Our focus is on developing an archive of sources that scholars and students can use for analysis, interpretation, and research employing digital tools.

Critical Context

LACA seeks to insert itself in the broad scholarly landscape created at the intersection of comics scholarship (e.g. Priego, 2016; Walsh, 2012), visual ontologies and comics (e.g. Bateman et. al., 2017; Turton, 2017), work done to encode comics elements (e.g. Dunst et. al., 2016; Haidar and Ganascia, 2016; Kuboi, 2014), and work on the value of comics as a pedagogical tool (e.g. Brooks, 2017).

Methodology

Given the team's expertise in Digital Humanities (DH) and Digital Scholarship, and with the support of an institutional Mellon DH seed grant, the project was initiated in the summer 2016. LACA was modeled after existing specialized collections such as MIT's Comics and Popular Culture archive, UNAM's specialized online resource http://www.pepines.unam.mx/, and the Grand Comics Database (GCD) with the purpose of combining the PI's ongoing research and teaching experience on Spanish-language Latin American comics with the use of DH tools to create an environment enabling students and scholars to have access to and collaborate in the analysis of the digital materials. At the current stage, LACA includes a small digital sample of Latin American comics produced throughout the last century, provided through a combination of previously digitized materials, materials we scanned, and those provided by authors themselves.

The presentation will detail three parts of the project:

1. Curating the comics and creating the archive;
2. Creating the online Metamedia platform to house digitized sources for the research and teaching of Spanish-language comics through student/scholar collaboration;
3. Piloting and implementing the digital archive for research and teaching.

Insights/Results

LACA was piloted at CMU over the past year as an instrument in courses for undergraduate students of Spanish

language and culture. Students and faculty collaborate in the analysis and CBML coding of the comics. In the process, students learn the basics of TEI and CBML, as well as critical approaches to Spanish-language comics, and their work contributes to the availability of comics on the site. Students are also able to develop integrated textual and visual competence, knowledge, and skills. The pilot courses provide initial evidence that coding the comics facilitates students' attention to details, notice of patterns, and, in general, collaborative advancement in the analysis and understanding of the linguistic and cultural elements contained in the comics. At the same time, it also helps students keep in mind communication to a wide public audience. The PI has benefitted from the additional opportunities afforded to glean information about students' progress toward cultural, linguistic, visual, and digital literacy. Thus, it is suggested that LACA could be of use and applicable to other courses in Hispanic studies, Modern Languages, and the Humanities.

We intend to make LACA publicly available for use as a hub where students and scholars interested in experimenting with the inquiry of Latin American comics can interact. This would help transform and expand the scale of traditional research methods used, and could open new modes and possibilities for text analysis that can be employed into the realm of student agency and learning. However, as we advance in the process to attain this goal, we acknowledge that IP/copyright permissions remain a challenge. Some creators have granted permission to distribute their works; others will only be used as part of course materials. Despite this, we think it is important to keep in mind Walsh's (2012) point that "nothing prevents a scholar from applying CBML markup to any text as part of a strategy for reading, interpretation, and analysis. The end goal of markup is not and should not always be publication of a digital surrogate. The encoding of a text may be a rigorous intellectual activity that has great value as process, not just as product."

References

Bateman, J. A., Veloso, F. O. D., Wildfeuer, J., Cheung, F. H. and Guo, N. S. (2017). An open multilevel classification scheme for the visual layout of comics and graphic novels: Motivation and design. *Digital Scholarship in the Humanities*, *32*(3), 476–510. https://doi.org/10.1093/llc/fqw024

Brooks, M. (2017). Teaching TEI to undergraduates: A case study in a digital humanities curriculum. *College and Undergraduate Libraries*, *0*(0), 1–15. https://doi.org/10.1080/10691316.2017.1326331

Dunst, A., Hartel, R., Hohenstein, S. and Laubrock, J. (2016). Corpus Analyses of Multimodal Narrative: The Example of Graphic Novels. *Digital Humanities 2016: Conference Abstracts*. Krakow, Poland. Retrieved from http://dh2016.adho.org/static/data-copy/387.html

Haidar, A. and Ganascia, J. (2016). Automatic Detection of Characters in Case Insensitive Text in Comics. In *Digital Humanities 2016: Conference Abstracts* (pp. 425–426). Jagiellonian University & Pedagogical University, Kraków.

Kuboi, T. (2014). Element Detection in Japanese Comic Book Panels. *Master's Theses and Project Reports*. https://doi.org/10.15368/theses.2014.141

Priego, E. (2016). Comics as Research, Comics for Impact: The Case of *Higher Fees, Higher Debts*. The Comics Grid: Journal of Comics Scholarship. 6, p.16. DOI: http://doi.org/10.16995/cg.101

Spiro, L. (2012). "This Is Why We Fight": Defining the Values of the Digital Humanities. In M. K. Gold (Ed.), *Debates in the Digital Humanities*. Minnesota: University of Minnesota Press. Retrieved from http://dhdebates.gc.cuny.edu/debates/text/13

Turton, A. (2017). Towards Feminist Data Production: A Case Study from Comics. In *Digital Humanities 2017: Conference Abstracts*. Montreal, Canada. Retrieved from https://dh2017.adho.org/abstracts/493/493.pdf

Verba Volant, Scripta Manent: An Open Source Platform for Collecting Data to Train OCR Models for Manuscript Studies

Samuel Grieggs
sgrieggs@nd.edu
University of Notre Dame, United States of America

Bingyu Shen
bshen@nd.edu
University of Notre Dame, United States of America

Hildegund Muller
hmuller@nd.edu
University of Notre Dame, United States of America

Christine Ascik
cascik@nd.edu
University of Notre Dame, United States of America

Erik Ellis
erik.z.ellis.67@nd.edu
University of Notre Dame, United States of America

Mihow McKenny
mihow.p.mcKenny.5@nd.edu
University of Notre Dame, United States of America

Nikolas Churik
nchurik@nd.edu
University of Notre Dame, United States of America

Emily Mahan
emahan@nd.edu
University of Notre Dame, United States of America

Walter Scheirer
walter.scheirer@nd.edu
University of Notre Dame, United States of America

Introduction

The transcription of handwritten historical documents into machine-encoded text has always been a difficult and time-consuming task. Much work has been done to alleviate some of that burden via software packages aimed at making this task less tedious and more accessible to non-experts. Nonetheless, an automated solution would be a worthwhile pursuit to vastly increase the number of digitized documents. As part of a continuing effort to expand the footprint of digital humanities research at our institution, we have embarked on a project to automatically transcribe and perform automated analysis of Medieval Latin manuscripts of literary and liturgical significance. Optical Character Recognition (OCR) is the process of converting images containing text into a machine encoded document. Recent advances in artificial neural networks have led to software that can transcribe printed documents with near human accuracy (LeCun et al., 2015). However, this level of accuracy breaks down when working with handwritten, and especially cursive, documents except when applied to restrictively specific domains.

Neural networks that are trained for this task require thousands of labeled examples so that their millions of parameters can be optimized. While there are thousands of high-quality scans of manuscripts available on the Internet, very few of these documents have been annotated for OCR tasks, and there is only a limited selection of ground-truth data which is annotated and segmented at the word-level (Fischer et al., 2011; Fischer et al., 2012) . There is no data available that provides annotations at the character-level. Normally, machine learning researchers would outsource the production of this ground-truth data to a platform such as Amazon's Mechanical Turk service, which allows crowdsourcing of human intelligence tasks. This is not an option for transcribing Medieval manuscripts, because it requires domain specific expertise. We put together a team of expert Medievalists and Classicists to generate the ground-truth data, and we have been developing a software platform that breaks the tedious task of producing pixel-level training data into more tractable jobs. The goals of this software go beyond just Latin manuscripts: it can be used to generate source data for any machine learning task involving document analysis. We are releasing it publicly, as free and open source software, in hopes that others can also use it to generate data, and help bring further advances in machine learning for handwritten text recognition.

Related work

State-of-the-art solutions to handwritten digit recognition on the MNIST dataset have achieved accuracies greater than 99% and have led some to declare handwritten OCR a solved problem (Wan et al., 2013). However, Cohen et al. have shown that adding the English alphabet to the dataset drops accuracy by more than 20% even when using the same methodology (Cohen et al., 2017). Some of the difference can be attributed to the fact that characters like "I", "l", and "1" are often ambiguous without context --- especially when handwritten. To combat this, many handwritten text recognition algorithms will often use recurrent neural networks that look at the whole word and utilize a language model to overcome ambiguities (Fischer et al., 2009; Sánchez et al., 2016). Additionally, Convolutional Neural Networks (CNN) have been shown to have promise in segmenting biomedical images, which are also difficult to ground truth (Ronneberger et al., 2015). A similar approach could be used to segment individual letters in manuscripts. Incorporating human performance information into the machine learning process has been shown to improve the accuracy of tasks like face detection (Scheirer et al., 2014). We hypothesize that incorporating a human weighted loss function will lead to similar improvements in this task as well.

Workflow

Currently the software runs in Google App Engine using high-resolution source images. We are in the process of setting up the software to be run in a vagrant environment to make it available for local environments. The vagrant script will provision a Virtual Machine, either locally or to the cloud to serve the software and configure it to work with a user-provided library of documents. In either case, transcribers can access the software via a web browser. The user then proceeds to segment the document by lines and words by drawing bounding boxes, and characters by drawing over them. It also collects text annotations of the text at the word- and character-level. It stores all the information in a MySQL database.

Line and word level

Our process starts by having experts segment the document into lines. Transcribers use a modified version of the Image Citation Tool from the Homer Multitext Project to quickly break the document down into CITE URNs representing each line by drawing boxes around them (Blackwell and Smith, 2014). After all the lines are selected the process is repeated for each word. A screenshot of these processes is shown in Fig. 1.

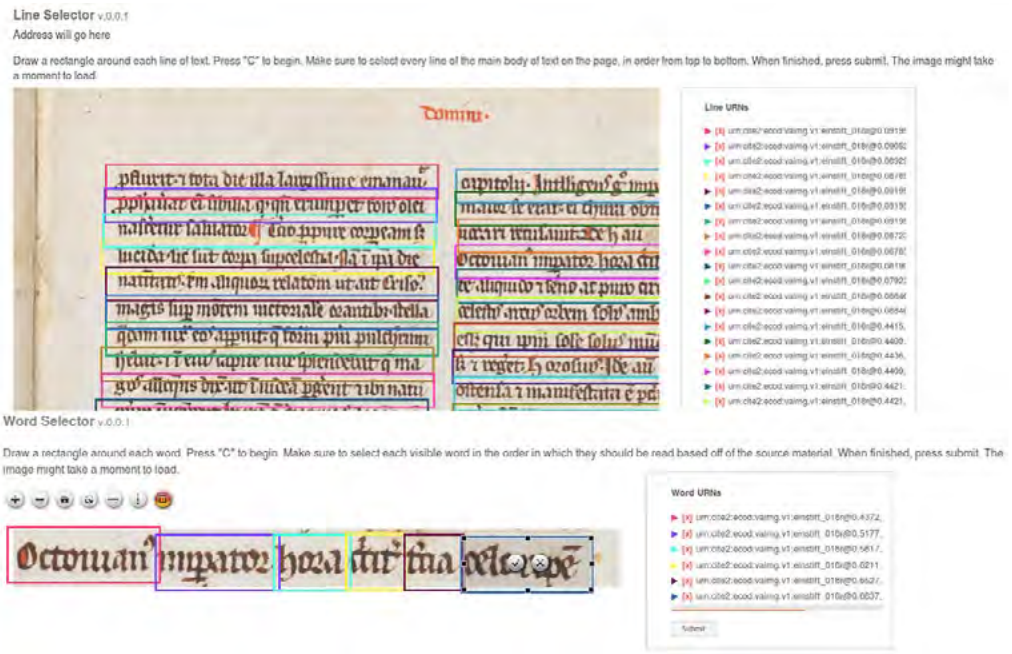


Figure 1: An example of the interface for selecting lines and words. Manuscript: Einsiedeln, Stiftsbibliothek, Codex 629(258), f. 4r – [Jacobus de Voragine] Legenda aurea sive lombardica (http://www.e-codices.unifr.ch/en/list/one/sbe/0629)

Pixel level annotation

After segmenting the document into words, our software prompts the expert to segment and annotate each word letter by letter. Instead of using a bounding box, we have the user trace over each character in the word using a pen tool. This gives us a pixel-by-pixel segmentation of the image that can be used to train a CNN to segment the characters automatically, much in the same way segmentation models are trained for other computer vision tasks (Ronneberger et al., 2015). At this stage the expert will also select which letter best represents each character from an array of buttons, as shown in Fig. 2.

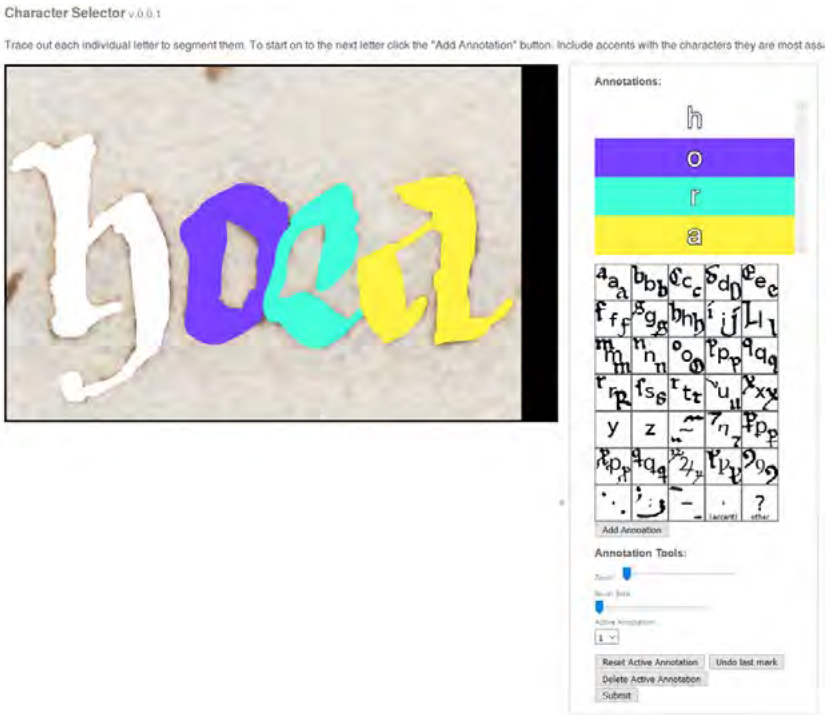


Figure 2: An example of the tool used to collect pixel level ground-truth at the character level.

Psychophysical measurements

The final stage collects psychophysical measurements of the human process of reading. The software brings up individual characters, as shown in Fig. 3, and asks the transcriber to pick an annotation for a character without word context. They will also be asked to select the difficulty of each character. The software also records how long it takes for the user to submit an answer and compares whether the user selected the same character that was selected during the word-level annotation.

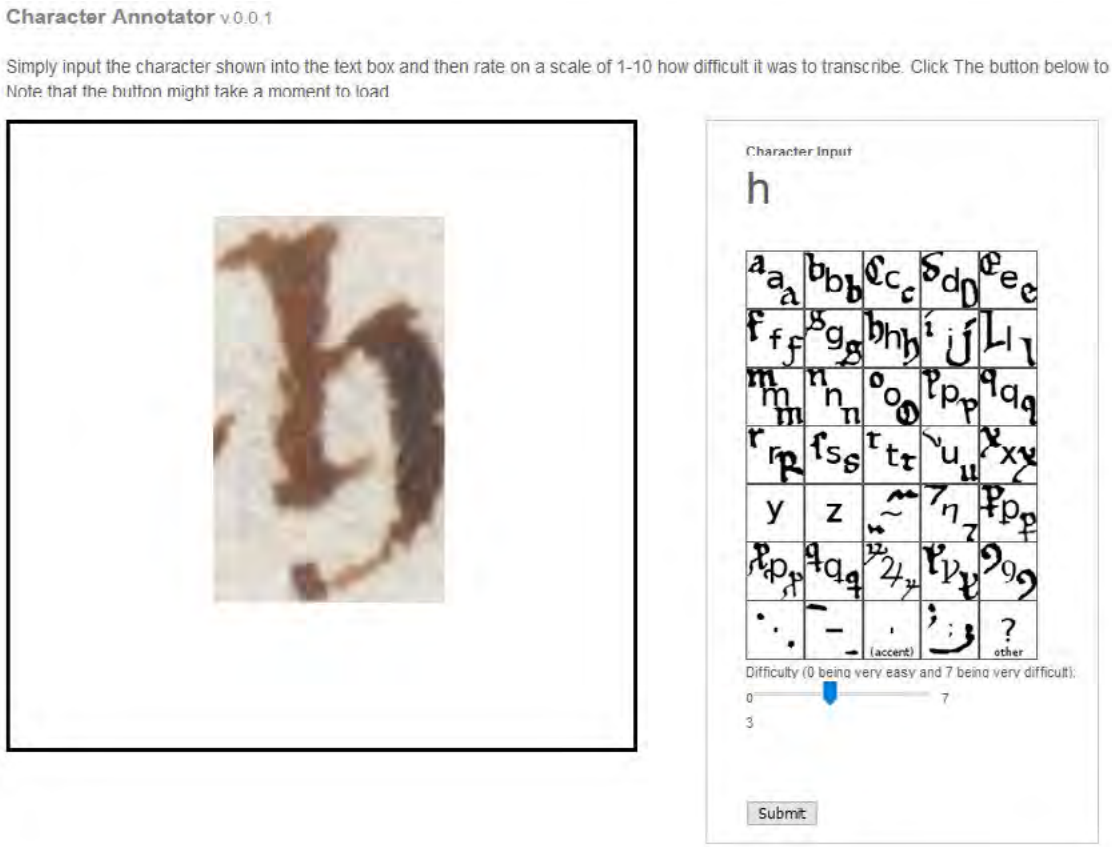


Figure 3: A screenshot of the psychometric data collection stage.

Outcomes

The software produces a segmented image for each document that can be used as training data for machine learning-based segmentation. Furthermore, it provides the pyschophysical measurements on the reading difficulty of each character. We also designed it to produce word-level segmented data in a similar format to the IAM Historical Document Database (Fischer et al., 2012; Fischer et al., 2011). Finally, the user will be able to export the transcribed document into a standard markup language such as TEI.

References

Blackwell, C. W. and Smith, D. N. (2014). The Homer Multitext and RDF-Based Integration. *Papers of the Institute for the Study of the Ancient World*, 7.

Cohen, G., Afshar, S., Tapson, J. and Schaik, A. van (2017). EMNIST: an Extension of MNIST to Handwritten Letters. *CoRR*, abs/1702.05373.

Fischer, A., Frinken, V., Fornés, A. and Bunke, H. (2011). Transcription Alignment of Latin Manuscripts Using Hidden Markov Models. Proceedings of the 2011 Workshop on Historical Document Imaging and Processing. *ACM*, pp. 29–36.

Fischer, A., Keller, A., Frinken, V. and Bunke, H. (2012). Lexicon-free Handwritten Word Spotting Using Character HMMs. *Pattern Recognition Letters*, 33(7): 934–942.

Fischer, A., Wuthrich, M., Liwicki, M., Frinken, V., Bunke, H., Viehhauser, G. and Stolz, M. (2009). Automatic Transcription of Handwritten Medieval Documents. Virtual Systems and Multimedia, 2009. VSMM'09. *15th International Conference on. IEEE*, pp. 137–142.

LeCun, Y., Bengio, Y. and Hinton, G. (2015). Deep Learning. *Nature*, 521(7553): 436–444.

Ronneberger, O., Fischer, P. and Brox, T. (2015). U-net: Convolutional Networks for Biomedical Image Segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 234–241.

Sánchez, J. A., Romero, V., Toselli, A. H. and Vidal, E. (2016). ICFHR2016 Competition on Handwritten Text Recognition on the READ Dataset. Frontiers in Handwriting Recognition (ICFHR), 2016 *15th International Conference on*. IEEE, pp. 630–635.

Scheirer, W. J., Anthony, S. E., Nakayama, K. and Cox, D. D. (2014). Perceptual Annotation: Measuring Human Vision to Improve Computer Vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(8): 1679–1686.

Wan, L., Zeiler, M., Zhang, S., Cun, Y. L. and Fergus, R. (2013). Regularization of Neural Networks Using Dropconnect. Proceedings of the *30th International Conference on Machine Learning* (ICML-13). pp. 1058–1066.

Indagando la cultura impresa del siglo XVIII Novohispano: una base de datos inédita

Víctor Julián Cid Carmona
vjcid@colmex.mx
El Colegio de México, México

Silvia Eunice Gutiérrez De la Torre
segutierrez@colmex.mx
El Colegio de México, México

Guadelupe Elisa Cihuaxty Acosta Samperio
cihuaxtysamperio@gmail.co
Universidad Nacional Autónoma de México

El objetivo del sistema es facilitar el estudio sistemático de los libros anunciados en la *Gazeta de México*, primera publicación periódica de América. Esta publicación, que imprimió su primer fascículo en 1722, ofrecía en algunos números información sobre las novedades bibliográficas de la época.

Estas notas incluían menciones de títulos, nombres de autores e impresores, ubicación de las imprentas y lugares de venta de los libros, entre otros (imagen 1).

La información de las novedades fue enriquecida con un proceso de investigación documental en el que se identificaron 32 campos distintos tales como: catálogos en los que se encuentra registrada la obra anunciada, disponibilidad en pdf, cargo o profesión de los autores, archivo de autoridad virtual internacional de los mismos (VIAF), página exacta del anuncio, precio de la obra, idioma, asignación temática sistematizada, entre otros.

Libros nuevos.

VNO en octavo cuio titulo es: *Razones para convencer à el Pecador para que salga de pecado, y se ponga en gracia de Dios*; su Autor, el R. P. M. Ignacio Fiol, de la Compañia de Jesvs, impresso en la Imprenta de D. Joseph Bernardo de Hogal, en la Calle nueva de la Monterilla.

Otro en lo mismo, intitulado: *Epitome de la vida de la V. Virgen Mariana de JESUS, Flores, y Paredes, Beata Jesuita, nombrada la Azucena de Quito, ilustre en Santidad, Profecias, y Milagros*, sacado de la Vida, q̃ latamente escriviò el P. Jacintho Morân de Butrõ, de la misma Compañia. Vendese en la Libreria de los Herederos de la Viuda de Miguel de Rivera Calderon, en el Empedradillo.

Otro en doze, intitulado: *Sol del Alma, cuias luces alumbran la razõ, y encienden los afectos. Meditaciones Santas del Señor San Bernardino de Sena*; escritas por el R. P. Fr. Francisco de la Concepcion Barbosa, Hijo de la Provincia del Santo Evangelio de Mexico, y Predicador Mayor en su Convento Grande. Hallarâse en la Sacristia del V. Orden Tercero de Penitencia de N. S. P. San Francisco, de esta Corte.

Otro en diez y seis, intitulado: *Devocion à Señor San Joseph, Esposo de la Virgen Maria N. Señora*, por el P. Francisco Garcia de la Compañia de JESVS.

Vn Sermon predicado en la Iglesia del Convẽto de Señoras Religiosas de la Seraphica Virgen Santa Catharina de Sena, de esta Corte, el dia primero de Abril del año de 1729. en la nueva Festividad de las Llagas de dicha Seraphica Virgen; por el M. R. P. M. Fr. Juan de Alvarado, del Sagrado Orden de Predicadores, Hijo de esta Provincia de Santiago de Nueva-España.

☞ *Impressos donde esta Gazeta.*

Imagen 1. Anuncio de libros en la *Gazeta de México*, 16 de septiembre de 1807

La propuesta tiene dos propósitos principales. Por un lado, representará uno de los catálogos digitales más completos de las obras novohispanas del siglo XVIII. El *Catálogo de impresos Novohispanos (1563-1766)* coordinado por Guadalupe Rodríguez, por ejemplo, únicamente contiene 505 registros. Por otro lado, el estudio de las gacetas de México se ha abordado, hasta ahora, desde dos perspectivas: tratarlas en su generalidad (Drwall, 1980; Ruíz Castañada, 1969, 1970, 1971); o bien, referirse a la representación de algún tema específico en sus páginas (ver Guedea, 1989, 1991). En este sentido, el proyecto **Ciudad letrada: la *Gazeta de México* y la difusión de la cultura impresa durante el siglo XVIII**, permite un nuevo acercamiento a las gacetas de México como fuente de la cultura impresa de la época.

A esto se aúna el hecho de que ha sido diseñada como una herramienta que apoye en las tareas del investigador interesado en los impresos de aquel siglo. El modelo de la base de datos, es del tipo Entidad-Relación. Para acercar a las personas a esta información se utilizó la plataforma Omeka 2.0, donde los vínculos se construyeron con el complemento llamado ItemRelations y la prueba de concepto puede ser consultada en: http://sandbox.colmex.mx/~silvia/omeka25/.

El procedimiento para integrar los datos que contiene la base, implicó la revisión de cada uno de los 1370 fascículos durante los 42 años de edición de la *Gazeta* con el fin de identificar los anuncios de libros nuevos. La información de cada uno de los anuncios se complementó con datos bibliográficos obtenidos de bibliografías especializadas y catálogos de bibliotecas con el propósito de enriquecer la información original y hacerla más útil. Esto dio como resultado la identificación de 1872 anuncios de libros y folletos, publicados entre los años de 1657 y 1809; es decir, libros en un rango de centuria y media que

jamás habían sido identificados sistemáticamente, por lo cual hablamos de una base de datos inédita.

Entre las características especiales de este desarrollo caben destacar las múltiples formas de explorar y acceder a los registros. Entre ellas: la exploración por etiquetas, por índices, por mapa de ubicación de imprentas o lugares de venta, navegación hipervinculada de los resultados y de cada registro.

La búsqueda por etiquetas ofrece una vista de pájaro sobre los temas más frecuentes, los cuales fueron desagregados de su forma clásica (es decir en triadas) para permitir exploraciones más granulares.

Imagen 2. Fragmento de la nube de etiquetas

Por otro lado, los índices fueron generados con el complemento 'Reference' desarrollada para Omeka por Daniel Bertherau. Estos índices permiten una navegación exploratoria de los temas, autores, impresores, años de publicación, lugares de impresión y de venta, etc., ordenados alfabéticamente junto con sus ocurrencias (imágenes 3 y 4).

Imagen 3. Índices disponibles para búsqueda sistemática

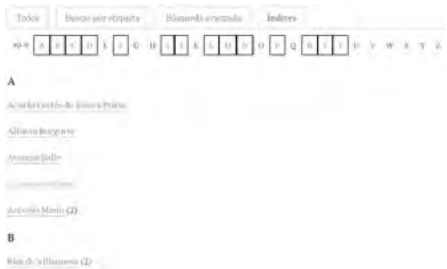


Imagen 4. Ejemplo de un índice (impresores)

Los mapas se crearon utilizando Carta, un complemento de AcuGis. En ellos se puede observar la distribución y concentración de imprentas (imagen 5) y lugares de venta (imagen 6) y de esta forma identificar los espacios clave de la Ciudad Letrada.

Imagen 5. Ubicación de imprentas en la Ciudad de México, siglo XVIII

Imagen 6. Librerías y otros lugares de venta

Por último, con navegación hipervinculada nos referimos a que sobre cada metadato se puede pulsar (imagen 7) para desplegar otros registros con esa misma característica (imagen 8).

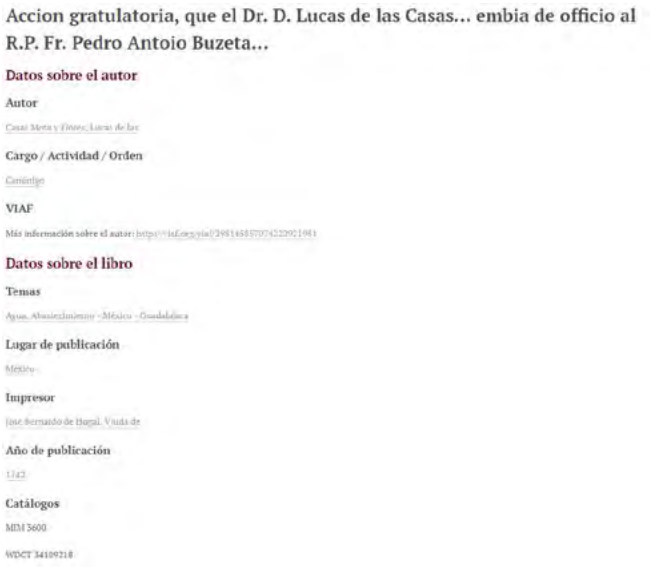


Imagen 7. Despliegue de registro con datos hipervinculados

Imagen 8. Despliegue de registros coincidentes (mismo autor)

Por mencionar un ejemplo de uso, imaginemos el siguiente escenario: digamos que el usuario explora el mapa de lugares de venta y abre la ubicación del Colegio de San Ildefonso (como en la imagen 6), al pulsar sobre el hipervínculo que dice 'Libros impresos en esta ubicación', el sistema despliega la lista completa de registros de la base que se vendían en ese lugar; en el despliegue de estos datos (imagen 9), se observa que todos son libros seculares y relacionados con las ciencias y la educación. Esto es interesante, si se considera que la mayoría de los registros son de contenido religioso y casi todos los puntos de venta ofrecían en mayor cantidad obras de esta naturaleza y podría guiar al estudioso del tema hacia nuevas preguntas como ¿todos los colegios vendían libros seculares?, ¿qué otros puntos distribuían obras de esta índole?, ¿por qué es más difícil encontrar el nombre de los autores de estas publicaciones?, etc.

Imagen 9. Libros a la venta en el Colegio de San Ildefonso

Cabe mencionar que se identificó un conjunto considerable de obras de carácter técnico o científico, algunos diccionarios generales y especializados, varios textos literarios y, principalmente, obras de contenido religioso, histórico, biográfico, dogmático y devocional.

Para concluir, consideramos que este desarrollo posibilitará a los interesados en impresos del siglo XVIII nuevas vetas de investigación a partir de la información sistemática que incluye. En particular, resultará útil para tratar asuntos relacionados con autores, impresores y comerciantes del libro en México durante el siglo XVIII. Además, ofrece la posibilidad de indagar sobre los mecanismos de propaganda de este bien cultural, así como saber en qué lugares se conservan ejemplares de estos documentos actualmente o, tener acceso a versiones electrónicas de ellos, en varios casos.

Referencias

AcuGis (s.f.). *"Carta 2.1.1". Github*. https://github.com/AcuGIS/Carta.

Adank, P. A. D. (1980). Accommodation and innovation: the Gazeta de México, 1784 to 1810 Arizona: Arizona State University Doctorado.

Bertherau, D. (s.f.). *"Reference 2.4.2". Github*. https://github.com/Daniel-KM/Reference.

Castera, I. (1785). Plano Geométrico de la Imperial, Noble y Leal Ciudad de México, teniendo por extremo la Zanxa y Garitas del Resguardo de la Real Aduana

Madrid, en la Calle de Atocha, frente de la Aduana vieja, Manzana 159, N.o 3.
Domínguez Rodríguez, G. (2012). Introducción. *Repertorio de impresos novohispanos (1563-1766)*, vol. 12. Xalapa, Veracruz: Biblioteca Digital de Humanidades. Universidad Veracruzana.
Guedea, V. (1989). La medicina en las gacetas de México. *Mexican Studies/Estudios Mexicanos*, 5: 175–99.
Guedea, V. (1991). *Las gacetas de México y la medicina: un índice*. México: Universidad Nacional Autónoma de México, Instituto de Investigaciones Históricas.
Ruíz Castañeda, M. del C. (1969). La Gaceta de México de 1722 primer periódico de la Nueva España. *Boletín del Instituto de Investigaciones Bibliográficas*, 1(1): 39–59.
Ruíz Castañeda, M. del C. (1970). La segunda Gazeta de México (1728-1739, 1742). *Boletín del Instituto de Investigaciones Bibliográficas*, 3(2): 23–42.
Ruíz Castañeda, M. del C. (1971). La tercera Gaceta de la Nueva España. Gaceta de México (1784-1809). *Boletín del Instituto de Investigaciones Bibliográficas*, 3(6): 137–150.

Puesta en mapa: la literatura de México a través de sus traducciones

Silvia Eunice Gutiérrez De la Torre
segutierrez@colmex.mx
El Colegio de México, Mexico

Jorge Mendoza Romero
enciclopedia.flm@gmail.com
Fundación para las Letras Mexicanas, Mexico

Amaury Gutiérrez Acosta
agutierrez@conabio.gob.mx
CONABIO, Mexico

Para responder cuáles han sido las tendencias de la circulación de la literatura de México hacia otros espacios lingüísticos, se partió de los datos disponibles en la *Enciclopedia de la literatura en México* (ELEM, www.elem.mx) para realizar un estudio de las traducciones de obras de escritores mexicanos, escritas en español y traducidas a 33 lenguas (incluidos los 7 idiomas indígenas del país de los que hubo al menos un registro). En esta presentación breve, daremos cuenta de los resultados de una investigación en curso sobre el modelado y puesta en mapa de estos datos.

En México, el estudio cuantitativo de las traducciones de la literatura nacional tiene un antecedente emblemático en la obra pionera de José Ignacio Mantecón, Índice de las traducciones impresas en México, de 1959. En este trabajo se recopilaron 544 traducciones hechas en México en ese año, y se registraron aspectos tales como el género al que pertenecían, lo que permitió derivar conclusiones como el hecho de que el grupo más representativo de traducciones lo constituían las obras literarias (35%), de las cuales un 13% eran libros infantiles (Mantecón, 1959: 14, 18). Sin embargo, además de que este trabajo no ha sido replicado, este estudio sólo da cuenta de las traducciones al español como lengua meta.

Otra referencia, en la que se perfila el objetivo de nuestra investigación –el estado de la traducción de la literatura de México– se encuentra en la introducción que hace Rosenzweig del intercambio epistolar entre Alfonso Reyes y el traductor al checo Zdeněk Šmíd. En ésta se lee lo siguiente:

> Salvo excepciones, la literatura mexicana, al igual que la latinoamericana, se comenzó a traducir a comienzos de los años treinta del siglo XX. Inicialmente se hicieron traducciones al inglés y al francés; en un segundo momento, impulsadas por el francés, a otras lenguas europeas como el alemán, neerlandés, checo e italiano. Las primeras novelas mexicanas que se tradujeron fueron Los de abajo y Mala yerba de Mariano Azuela; *El águila y la serpiente* y *La sombra del Caudillo*, de Martín Luis Guzmán; El indio, de Gregorio López y Fuentes; y *¡Vámanos con Pancho Villa!* de Rafael F. Muñoz. (Rosenzweig, 2014: 13)

No obstante, este extracto carece de referencias numéricas exactas y tampoco responde quiénes fueron esos primeros traductores al inglés, cuándo comenzaron exactamente las traducciones al francés o cuándo a otras lenguas europeas. Y es que, a excepción de algunas listas de idiomas específicos –como los 327 registros de obras de la literatura mexicana traducidas al inglés en Estados Unidos (Boyd, 2012); el catálogo análogo de 99 registros de obras traducidas al alemán (Küpper, s.f.); o la lista de las traducciones al italiano (Tedeschi, s.f.) – no existe ningún compendio que ofrezca el panorama completo de la proyección de la literatura mexicana en un sentido global. Por tal razón, la bibliografía de más de 1500 traducciones de la ELEM es una base de datos única en su tipo de la que es necesario expandir sus posibilidades heurísticas. Pero antes, algunas palabras sobre esta enciclopedia.

La ELEM comenzó a organizar el conocimiento en torno a la cultura literaria de México (oral y escrita) desde 2011, cuando fue creada. Cuenta con los registros de 13,040 personas (autores, traductores, investigadores literarios) y más de 40,000 obras impresas (primeras ediciones), que conforman una bibliografía general de la literatura en México, la cual abarca casi V siglos de cultura literaria. Entre sus prioridades se encuentra el registro de las obras traducidas a otros espacios lingüísticos con el propósito de observar, a través de las lenguas meta y los países del mundo en que son impresas, el grado de recepción de la literatura del país.

Por esto, emprendimos un trabajo colaborativo y transdisciplinario en el que se planteó un modelado de los datos disponibles en la enciclopedia (ver Imagen 1)

bajo el concepto de puesta en mapa (en analogía de la puesta en página del mundo editorial) y en consonancia con la línea de las Humanidades Digitales denominada spatial humanities. En este caso específico, designa al desarrollo de una interfaz que permite captar geopolíticamente la circulación de la literatura de los autores mexicanos que escriben en español (con algunas traducciones indirectas) hacia 19 lenguas indoeuropeas, 7 lenguas indígenas de México, además de estonio, euskera, finés, hebreo, húngaro, japonés y turco. El corpus del que partimos contempla un universo de 1658 primeras ediciones que se desdobla, a partir de las reimpresiones y reediciones de muchos títulos, en un total de 2088 objetos.

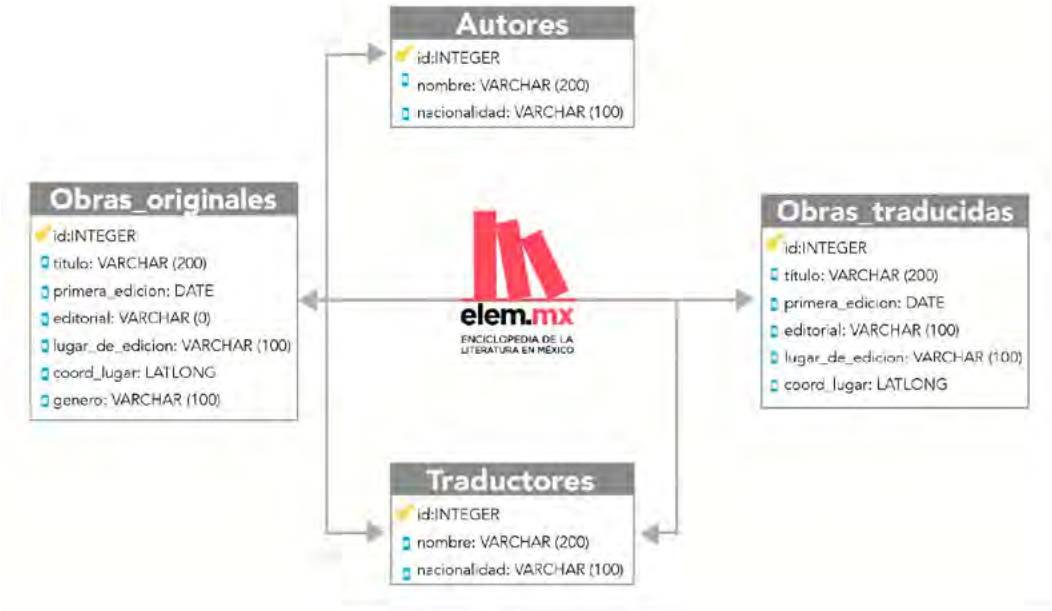


Imagen 1. Estructura de la base de datos

En un primer acercamiento, nos interesó indagar las relaciones espacio-temporales de las obras traducidas para responder las siguientes preguntas:

- ¿Qué autores o géneros han sido los más traducidos?
- ¿En qué años?
- ¿En qué geografías?
- ¿A qué idiomas?

Imagen 2. Perspectiva general de la puesta en mapa

Para facilitar la exploración de estas relaciones, se creó un prototipo de interfaz interactiva que permite iniciar investigaciones a partir de la puesta en mapa de los datos. El código en desarrollo de este prototipo se encuentra disponible en GitHub (Gutiérrez, 2017) y su versión para consulta estará en: http://elem.mx/estgrp/datos/1335.

Se describen las etapas de desarrollo a continuación. A partir de una consulta SQL a la base de la ELEM se creó un archivo separado por comas (csv) usando un script de Python (parser.py en el repositorio de GitHub). Estos insumos fueron transformados para obtener un formato adecuado para el consumo en Javascript: JSON. Para la arquitectura de la aplicación web se usó una herramienta para hacer empaques o bundles llamada Webpack (https://webpack.js.org/). La biblioteca usada para la creación del mapa es una herramienta de código abierto llamada Leaflet en su última versión 1.2.0 (http://leafletjs.com/). El desarrollo de la aplicación se puso en marcha en Javascript para la interfaz ya que la información, por el momento, existe de manera estática. En el futuro, cuando se integre con la base de datos con la dorsal final o backend, será deseable que las consultas de datos se realicen desde este punto y se exponga un end-point adecuado para el consumo.

La interfaz pretende facilitar la visualización e interacción con los datos de la base, así como el análisis exploratorio de los mismos (Behrens, 1997). Los usuarios podrán elegir filtros tales como: lengua meta, género literario, año de la traducción y, explorar los registros por ubicación geográfica. Además se provee de la siguiente información sobre los objetos: título de la traducción, autor/a, traductor/a, editorial de la traducción y título original de la obra.

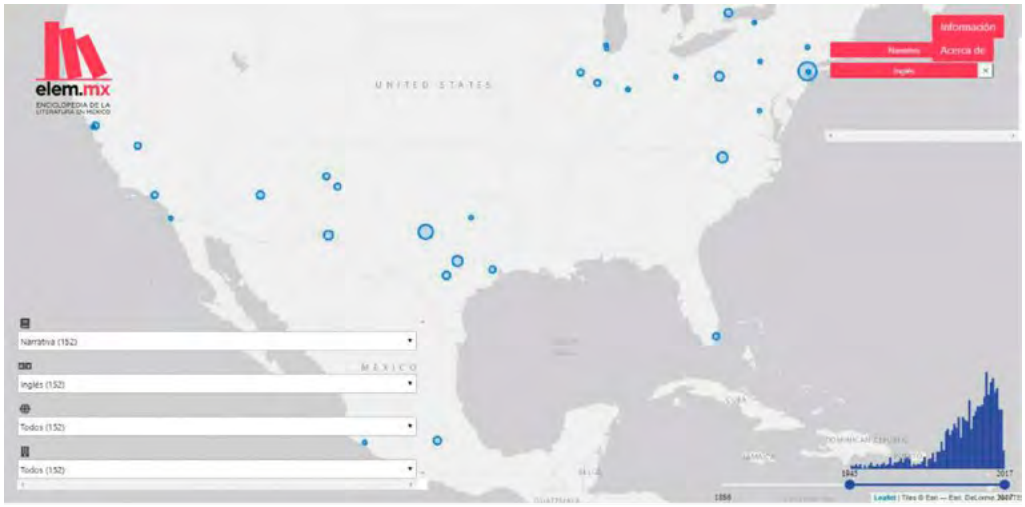


Imagen 3. Perspectiva del filtro: narrativa/inglés/1945-2017

Uno de los potenciales usos de esta herramienta puede ilustrarse a partir del siguiente ejemplo en el cual se usó el filtro de idioma (inglés), el de género literario (narrativa) y el rango de años de edición (1945-2017). La vista de los datos nos permitió observar un comportamiento no previsible. El título *Kill de Lion!* fue editado en México, D. F., en inglés. Es decir, el espacio geográfico no corresponde necesariamente con el espacio lingüístico, como se hubiera podido suponer en un principio.

Los especialistas e interesados en la cultura de la traducción literaria podrán contar con una visión de conjunto para realizar análisis e interpretaciones más minuciosas sobre la circulación de la cultura literaria de México a través de sus traducciones. Además, la puesta en mapa se irá actualizando conforme a las actividades de catalogación de la enciclopedia, lo que permitirá un acercamiento a las traducciones hacia otras lenguas aún no contempladas hasta ahora. Asimismo, los interesados en los contactos entre lenguas contarán con los insumos para poner en perspectiva las relaciones diglósicas, trasladadas a la cultura impresa, entre lenguas hegemónicas y lenguas minorizadas a partir de la traducción.

Referencias

Behrens, J. (1997), Principles and procedures of exploratory data analysis. *Psychological Methods* 2 (2): 131.

Boyd, M. (2012). *A Conflict of Narratives: The Influence of US Ideological Constructions of Mexican Identity in the Translation of Mexican Literature into English*. Universidad de York, Toronto, Canadá.

Burns, P. et al. (2017). Mapping Linked Data Subject Headings in the Library Catalog. *DH2017*. Montreal, Canadá.

Gutiérrez, A. (2017). Literature Translation. *GitHub*. https://github.com/amaurs/literature-translation

Küpper, K. (s.f.) Mexiko / Mittelamerika. *Archiv für übersetzte Literatur aus Lateinamerika und der Karibik*.http://www.lateinamerikaarchiv.de/antiquariat/mittelamerika-mexiko.html

Flexibility and Feedback in Digital Standards-Making: Unicode and the Rise of Emojis

S. E. Hackney
s.hackney@pitt.edu
University of Pittsburgh, United States of America

Background

The infrastructures that we use to navigate the world often become invisible as they become indispensable (Bowker and Star, 2000). However, critical examination of information systems is necessary to understand their implicit biases, and the ways that they invite some types of engagement and restrict others. Structures of power continue to be replicated in the ways that technologies are deployed in our lives (Noble, 2016; Tufekci, 2016), and the inability to access and assess the standards which make digital communication possible risks the uncritical perpetuation of those power structures (Drabinski, 2013). The moments of rupture, when an established system takes on a new facet with unintended consequences, can be an important moment of visibility, where we are able to reveal its ideological foundations, and the ways that its users adapt their own behaviors to it, or push back against its uncomfortable constraints (Raley, 2006; Marino, 2007). The introduction of emojis to the Unicode Standard, and their widespread adoption over the decade from 2006-2017 is one such moment of transition.

Scholars of standards and standardization argue that the input of users is necessary for a standard to meet the needs of those users (Foray, 1994), and while the process of adding content to the Unicode Standard remains rigid, the unicode.org website provides an explicit record of the development and evolution of the face that Unicode presents to its users, and is able to be read as a text which reveals the contemporary state of Unicode and the cultural ideologies which shape it.

Methodology

While major language- and script-based additions are made with each update to the Unicode Standard, my analysis focuses on changes to the unicode.org website, and its role as an intermediary document between the Consortium, the Standard itself, and everyday users. The introduction of emojis in various updates to the Standard has resulted in changes to the content and structure of the unicode.org website that reflect an increased engagement with end users, which I argue is the result of increased semantic value of emoji characters for the user[1], as compared to an individual character in a language's written script. It is my intention, through this analysis, to describe the types of changes that happen to the governing body and public documents of Unicode as major changes happen to the Standard itself.

A timeline was created of the dates of major updates to the Unicode Standard since its introduction in 1991, using the official release dates for updates to the Unicode Standard as maintained by the Unicode Consortium. I cross-reference this document with the rollout of each new version by the major platforms[2], with a particular emphasis on updates featuring new emoji characters, beginning with Unicode 6.0 in 2010[3].

With this timeline in mind, I scraped the unicode.org domain using Python and the Beautiful Soup[4] library to collect the URLS of all the unique pages under the parent domain, as well as a table of links between those pages. This serves as a source-target list for the creation of a network visualization of the unicode.org domain, using the network visualization software Gephi.[5] This process is repeated using archived versions of the unicode.org site, available from the Internet Archive's Wayback Machine[6], resulting in several structural snapshots of the unicode.org website over time, which can then be overlaid and compared to one another to note particular areas of change within the site.

Additionally, using points of change within the site structure as a guide, I also collect and code page content data to reflect the type of changes made to those pages during each major update. This coding is done on two axes: The first labels each change as being content- or structure-based (eg. adding text or links to a page, respectively), and the second designates which aspect of the Standard and/or Consortium is being addressed by the change. Examples of this second type of labelling would be "Emoji," "Membership," "Meta-Documentation," or "Language Scripts." This coding is done in two phases— an initial survey of this data in order to formally create labelling categories, and then a closer examination of the updates to apply those labels.

1 A notable exception to this semantic shift is written Chinese, which is already a semantic-character-based language, as opposed to syllable- or alphabet-based, as are the rest of the world's major languages. Thomas S. Mullaney gives a thorough historical analysis of the implication of this on text-encoding technologies in *The Chinese Typewriter* (MIT Press, 2017).

2 https://unicode.org/emoji/format.html#col-vendor lists the major "vendors" of emojis, or platforms with proprietary visual displays of emojis. These vendors are Apple, Google, Twitter, Facebook, Facebook Messenger, Windows, and Samsung.

3 While the first major batch of emojis were incorporated into Unicode in 2010, and the first official "Emoji 1.0" release was in 2015, work has been done within the standard since late 2006 to consider the addition and management of emoji-like characters within Unicode— hence the specific 2006-2017 emphasis of this research. (https://www.unicode.org/reports/tr51/#Introduction)

4 https://www.crummy.com/software/BeautifulSoup/

5 http://gephi.io

6 https://web.archive.org/

Discussion and next steps

This research project addresses issues of digital infrastructure from a unique angle: one that considers the socially-constructed nature of technology, as well as the meta-narrative of maintenance and upkeep of a system that has become crucial to our ability to communicate in a digital world. Through analysis of the secondary documents relating to the Unicode Standard, it is possible to gain invaluable insights into the ways that knowledge is organized collectively and continuously, as well as the embedded values that shape who can access and influence that knowledge.

This case study will provide a foundation for more expansive examination of systems of digital infrastructure. It is a beginning point both for further analysis of the adoption and adaptation of Unicode (and emojis in particular), but also as a framework for examining other forms of scaffolding which uphold the content of digital spaces.

References

Bowker, G. C., and Star, S. L. (2000). *Sorting things out: Classification and its consequences*. Cambridge: MIT Press.

Drabinski, E. (2013). Queering the catalog: queer theory and the politics of correction. *The Library Quarterly* 83(2): 94-111. doi:10.1086/669547

Foray, D. (1994). Users, standards and the economics of coalitions and committees. *Information Economics and Policy*, *6*(3): 269-293.

Marino, M. C. (2007, December 4). Critical code studies. *Electronic Book Review.* Retrieved from http://electronicbookreview.com/thread/electropoetics/codology

Noble, S.U. (2016). A future for intersectional black feminist technology studies. *The Scholar & Feminist Online.* 13.3 - 14.1. Retrieved from: http://sfonline.barnard.edu/traversing-technologies/safiya-umoja-noble-a-future-for-intersectional-black-feminist-technology-studies/0/

Raley, R. (2006). Code.surface || Code.depth, *Dichtung Digital*. Retreived from http://www.dichtung-digital.org/2006/01/Raley/index.htm

Tufekci, Z. (2016, June). *Machine intelligence makes human morals more important.* [Video file]. Retrieved from https://www.ted.com/talks/zeynep_tufekci_machine_intelligence_makes_human_morals_more_important

The Digital Ghost Hunt: A New Approach to Coding Education Through Immersive Theatre

Elliott Hall
elliott.hall@kcl.ac.uk
King's College London, United Kingdom

Figure 1 Heather Agyepong, disrupting an ordinary school day in KIT Theatre's Alfred the Great Time Travel Adventure

Introduction

The Digital Ghost Hunt combines coding education, Augmented Reality and live performance into an immersive storytelling experience. Students ten to eleven years old (Key Stage 2 in the UK) will explore the haunted Battersea Arts Centre with devices they've learned to program themselves. The key objective of The Digital Ghost Hunt is to present technology to students as an empowering tool, where coding emerges as – and fuses with – different forms of storytelling. It seeks to shift the context in which students see coding and engage groups who may be uninterested in or feel excluded by digital technology, opening up an imaginative space through play for them to discover the creative potential of technology on their own terms.

The Digital Ghost Hunt has been awarded funding through the UK Arts and Humanities Research Council (AHRC) New Generation of Immersive Experiences call, as part of an application led by Mary Krell, Senior Lecturer in Media Practice at the Centre for Material Digital Culture in the University of Sussex. A 'scratch' – a prototype of the experience – will be developed by Elliott Hall of King's Digital Lab and Tom Bowtell of Kit Theatre. It will be performed at the historic Battersea Arts Centre with a two-form entry of students from local schools.

Structure of the experience

The Digital Ghost Hunt is split into two parts. The first part begins with a regular coding class that suddenly goes haywire. While the teacher is trying to restore order, the lesson will be interrupted by Ms. Quill, Deputy Undersecretary of Paranormal Hygiene (Ghost Removal Section). She will enlist their help in the Ministry's work as apprentice ghost hunters. Students will use a simplified Python library to program their ghost hunting devices, which are based on two microcomputers: the Raspberry Pi and the BBC Micro:bit.

The coding in the project will focus in particular on two learning goals of the UK's National Curriculum: "Design, write and debug programs that accomplish specific

goals, including controlling or simulating physical systems; solve problems by decomposing them into smaller parts," and "Use sequence, selection, and repetition in programs; work with variables and various forms of input and output."It will teach students to take the overall goal of their devices – detecting 'paranormal' phenomena – and break it down into the discrete input, analysis and output tasks required, aided by the project's abstracted libraries.How they combine the functions of these libraries will be up to them, and will rely on their understanding of the fundamental logical structuresof programming to analyse sensor data, apply it to an algorithm, and debug when things go wrong.The project will also introduce students to embedded computing through the devices themselves. The emphasis will be on students taking ownership of their devices, deciding which of the ghost detectors they want to build and how it will work.

The second part is a ghost hunt, an immersive experience combining Augmented Reality (AR) and live theatre. Students will work together in small teams, using their devices to find objects and areas touched by the ghost. These traces will be both virtual objects in Augmented Reality, and actual physical phenomena such as radio waves, ultraviolet paint, and high-frequency sound. Each device will have different capabilities, forcing the students to work together to get all the clues. The ghost will in turn communicate with them, given life by actors, practical effects and the poltergeist potential of the Internet of Things. Only by using the devices they have programmed and working together can students unravel the mystery of why the ghost is haunting the building and set it free.

Coding, play and performance

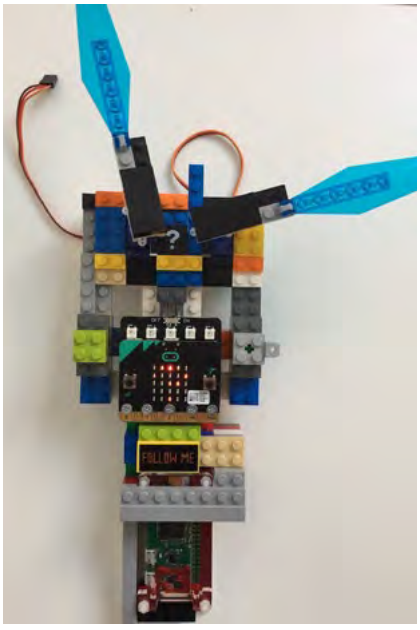


Figure 2 A proof of concept ghost hunting device using Lego and the Micro:Bit

Young people's familiarity with digital products are increasing, but their interest in learning the technology behind it is not, as evidenced in the UK by the low take up of the new GCSE in Computer Science(BCS, 2017). Teaching coding in schools is promoted by the UK Department of Education (DOE, 2014), but students often experience coding education as a classroom assignment, divorced from their intuitive and creative experiences with commercial digital applications.

There are several applications now using AR as a teaching tool (for example, The Battle of Mount Street Bridge (Schreibmen et al., 2017) and Virtual Roman Frontiers (Wilson at al.)) and initiatives to teach children coding, from commercial apps to coding clubs and the work of the Raspberry Pi and Micro:Bit foundations. These applications all seek the increase in engagement and experimentation that can occur when 'work' is reframed as 'play.' (Pellegrini, 2009)

However, these applications all take place within a screen, an approach that creates its own problems. A screen can shift a user's attention to the digital environment to the exclusion of the physical one. (Chrysanthi, 2012) The Ghost Hunt's approach is to bring AR interaction fully into the physical space without the mediating influence of a screen, reconnecting audiences to the world around them.

The addition of immersive theatre reframes the experience again, from 'play' to 'performance.' This second shift is important to reach groups not engaging with existing digital resources. In 2016, girls made up just 20% of entrants for the computer science exam, while pupils on free school meals made up just 19% of GCSE entrants even though they are 27% of the population (Cellan-Jones, 2016). Performance may draw in groups who would otherwise be uninterested in or feel excluded from traditional Computer Science education.

However, the performance should not be seen as secondary in any way to the coding elements of the project. The aim of the project is to expand the imaginative possibilities of digital technology through play; the coding elements are the means to that end, not the other way around. The Ghost Hunt seeks to shift how the context of computer science is perceived, from a skill intended only for a narrow group to a tool of creativity and play available to all.

As part of its evaluation, the project will use the student's code and feedback from educators on how the software libraries are used, as well as video, audio and device logging during the experience. It will be direct engagement with participants through formal and informal methods such as interviews, questionnaires and the creative material they create as part of the experience that will provide the crucial method of evaluation. The only way to assess the pedagogical value of the project in terms of creating a new and sustained interest in the possibilities of digital technology will be if students create new things on their own initiative, independent from the project's se-

tting and materials. This metric is beyond the scope of the pilot project but is something the project team are eager to explore in subsequent phases in collaboration with the educational partners.

Beyond the hunt

The lessons of the Digital Ghost Hunt scratch funded by the AHRC will direct refinement of the existing tools towards developing a technical and conceptual framework that can be adapted and implemented for different locations, stories and audiences. This short paper aims to present the practice-based collaborative framework of the Digital Ghost Hunt as conceived by its creators in its first funded iteration to elicit feedback from the Digital Humanities 2018 participants and integrate it into future development.

References

British Chartered Institute for IT (BCS). (2017) [online] *BCS deeply concerned over stagnation of number of Computer Science GCSE applicants*. Available at: http://www.bcs.org/content/conWebDoc/57904 *[Accessed 23/11/2017]*

Cellan-Jones, Rory. *(2017).* Computing in schools - alarm bells over England's classes. *BBC News*. [online.] Available at: *http://www.bbc.co.uk/news/technology-40322796 [Accessed 23/11/2017]*

Department of Education (DOE). (2014.) Teaching children to code. In: *D5: London*. London. Available at: https://www.gov.uk/government/publications/d5-london-summit-themes/d5-london-teaching-children-to-code [Accessed 23/11/2017]

Pellegrini, A. (2009) *The role of play in human development*. Oxford: Oxford University Press.

Chrysanthi, A., Papadopoulos, C., Frankland, T., and Earl, G. (2013). 'Tangible Pasts': User-centred Design of a Mixed Reality Application for Cultural Heritage. In: *Conference of Computer Applications and Quantitative Methods in Archaeology*. Southampton: Amsterdam University Press, pp. 31-41.

Schreibman, S, Papadopoulos, C., Hughes, B., Rooney, N., Brennan, C., Fionntann, M., Healy, H. Phygital Augmentations for Enhancing History Teaching and Learning at School. In: *Digital Humanities 2017*. Montreal. [online] Available at: https://dh2017.adho.org/abstracts/401/401.pdf [Accessed 23/11/2017]

Wilson, L., Weeks, P, Rawlinson A., Dobat, E., Fluegel, C., Hermann, C. (2017). Virtual Roman Frontiers: 3D Visualisation and Innovative Technology Applications for the Antonine Wall. In: *3D Imaging in Cultural Heritage*. London: The British Museum. Available at: https://www.3dimaginginculturalheritage.org/resources/3D_Imaging_in_Cultural_Heritage_Abstracts.pdf [Accessed 23/11/2017]

Exploration of Sentiments and Genre in Spanish American Novels

Ulrike Edith Gerda Henny-Krahmer
ulrike.henny@uni-wuerzburg.de
Universität Würzburg, Germany

Background, aims, and hypotheses

In 19th century Spanish American novels, the expression of emotionality is an essential characteristic of the texts belonging to different subgenres.[1] Especially during the Romantic period in the first half of the century, many sentimental novels have been written (Zó, 2015). But emotions also play an important role in other types of novels: a love story is often a basic plot element for example in historical or costumbrista novels. Also, there are novels characterized more by negative emotions, like Cuban anti-slavery novels (Rivas, 1990), Argentine anti-Rosas novels (Molina, 2011: 285-312, García Ardeo, 2006), or sociopolitical novels in general.

In text mining, a common method to analyze emotions is Sentiment Analysis (Pang and Lee, 2008). Sentiment Analysis is the computational treatment of sentiment, opinion, or emotion in text. Sentiments are usually modelled in terms of polarity values (positive, negative, neutral) or emotion values (such as trust, fear, joy, etc.).

The aim of this proposal is to test several hypotheses about sentiments in subgenres with an explorative analysis of a corpus of Spanish American novels. To this end, sentiment values are used as features in a text classification task. A secondary objective of this contribution is to compare the results of two different sentiment lexica for Spanish to see how well they perform.

The first hypothesis of this proposal is that the degree and kind of emotionality in the novels differs for different subgenres. The second hypothesis here is that not just emotions in general matter, but also whether they are expressed in the direct speech of the characters of the novels or in narrated text.[2]

State of the Art

Two recent examples for the usage of Sentiment Analysis with literary texts are Zehe et al., 2016 for the prediction of happy endings in German novels and Kim et al., 2017 for the analysis of prototypical emotion developments in literary genres with English texts. Sentiment Analysis has been used with Spanish texts, as well, mainly for the analysis of reviews and tweets (see Henríquez Miranda and Guz-

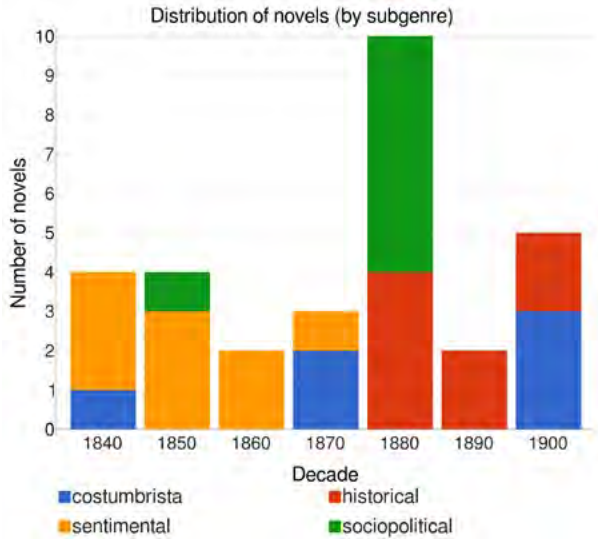
1 This contribution is concerned with the linguistic manifestation of emotions in literary texts on the textual surface. See Winko, 2003 for a discussion of how emotional meaning and literary texts are related

2 The anti-slavery novel, for example, has been defined in terms of its atmosphere of fear, but also by vigorous interferences of the narrator. Cf. Rivas, 1990.

mán, 2017 for an overview). To the best of my knowledge, there are no applications of Sentiment Analysis on Spanish novels yet, and the distinction of direct speech and narrated text has not previously been used in combination with the analysis of sentiments in literary texts.

Data

For this analysis, a corpus of 30 Spanish American novels has been selected. The collection has the following characteristics: The novels have been published between 1840 and 1910 (13 before 1880 and 17 after 1880), are from three countries (Argentina: 16, Cuba: 9, Mexico: 5), and have been written by 16 different authors.[3] Fig. 1 shows the distribution of novels per decade and subgenre:



Distribution of novels per decade and subgenre

As the texts at hand are not easily distinguishable genre fiction but more general literary fiction, the assignment of subgenre labels is a non-trivial task. For the assignment of subgenre labels to the novels, the subgenres as given in titles and subtitles of the novels were collected and subgenre assignments made in secondary literature were considered. Both types of information were used to derive four kinds of interpretive[4] subgenre labels corresponding to four broad types of novels: costumbrista (6 novels),[5] historical (8), sentimental (9), and sociopolitical (7) novels.[6]

Methods

In general, Sentiment Analysis can be done with a machine learning approach and a lexicon-based approach. Here, two sentiment lexica were used: (1) SentiWordNet 3.0, an adaptation of WordNet 3.0 for sentiment analysis (Miller, 1995, Baccianella et al., 2010) and (2) the NRC Emotion Lexicon (Saif and Turney, 2013). The two lexica differ in how sentiments are modelled and also in their volume. SentiWordNet has polarity values (positivity, negativity, neutrality) for WordNet synsets which range between 0 and 1 and sum up to 1. The NRC lexicon, in contrast, has only binary values (0 or 1), but those are provided for positivity and negativity as well as eight basic emotions (Trust, Fear, Joy, Sadness, Anger, Disgust, Anticipation, Surprise). SentiWordNet contains 117,653 entries, the NRC lexicon just 14,182.[7]

In order to use the sentiment lexica, the texts had to be lemmatized (for NRC) and annotated with WordNet synsets (for SentiWordNet) which was done with the NLP library FreeLing (Padró and Stanislovsky, 2012). To be able to use the distinction between direct speech and narrated text as a feature, the texts were annotated semi-automatically in their TEI master files (see Fig. 2):

```
<p><said>—Parece que duerme</said>, dijo examinando
atentamente las facciones de la viejecita, <said>¡quiera Dios que
este sueño alivie sus dolencias y reponga en un tanto sus ya
gastadas fuerzas!</said></p>
```

Example of a paragraph with annotated direct speech, from „Camila o la virtud triunfante“ (1856) by Estanislao del Campo

Each paragraph was split into sentences. Each sentence was annotated with FreeLing and the words with sentiment values were determined using the lexicons. The sentiment values for the words were summed up for each sentence.[8] For the eight basic emotions of the NRC (Trust, Fear, etc.), a sentence is assigned the emotion with a highest value in the sentence. Besides the sentiment features that come directly from the lexicons, the following features were determined for each sentence:

A Decision Tree classifier was used for the classification of the novels by subgenre, using the above-mentioned features (see Manning and Schütze, 1999: 578-589 on this method). The advantage of Decision Trees is that

3 This is a subcollection of a larger corpus of Spanish American novels being prepared in the context of the junior research group Computation Literary Genre Stylistics (CLiGS), see https://cligs.hypotheses.org/sprachen/english.
4 Because the many variations found had to be normalized for this computational analysis, an interpretive step was unavoidable.
5 Novels of manners in the context of the Costumbrismo movement.
6 The distribution of novels shows that there is a tendency for sentimental novels to belong to the first half and for non-sentimental novels to the second half of the century. This observation may be relevant for future tasks with a bigger corpus and interested in the development of genres over time. More detailed metadata for the novels can be found at https://github.com/cligs/projects2018/blob/master/sentgenre-dh/metadata.csv.
7 SentiWordNet can be used for Spanish because the synset IDs can be mapped to the Spanish version of WordNet. The NRC lexicon has been translated into Spanish automatically. See Baccianella et al., 2010 for evaluation reports for SentiWordNet. The authors of the NRC lexicon state that the translated versions may contain errors. An orthographic check on the NRC lexicon returned 409 entries that were not recognized as Spanish words. A further evaluation and improvement of the translated lexica is desirable.
8 The Sentiment Analysis could be refined further by considering the sentence structure (and negation), which is a future task.

they can be interpreted. This is desirable in an explorative analysis interested in the kind of sentiment-based features that are relevant to differentiate novels of different subgenres. When compared to other types of classifiers, Decision Trees do not necessarily yield the best results in terms of accuracy, but their interpretability is valued higher here in order to gain insight into how sentiments, the opposition of direct speech vs. narrated text, and subgenres are related.

Feature name	Description
emotional	Proportion of emotional sentences in the text. To determine emotionality, a threshold of 1 was set: all sentences with a positive value > 1 or a negative value < -1 were considered emotional.
neutral	Proportion of neutral sentences in the text, with a sentiment value between -1 and 1.
positive	Proportion of positive sentences in the text, with a sentiment value > 1.
negative	Proportion of negative sentences in the text, with a sentiment value < 1.

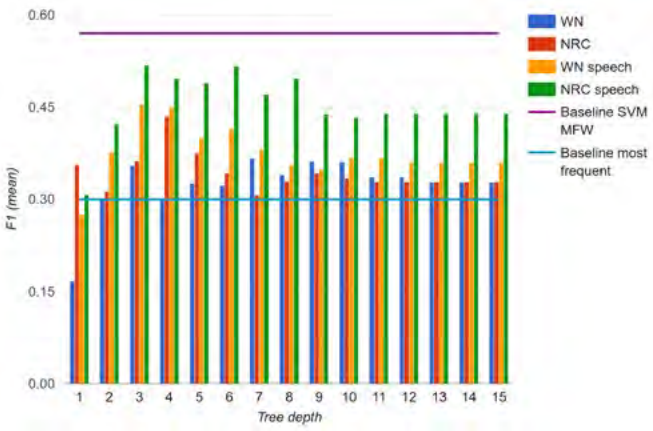
Additional features for the Sentiment Analysis

To generate data for the machine learning task, the values of the single sentences were aggregated into five sections and divided by the section length (number of sentences contained in the section), resulting in 150 data points for the 30 novels. 60 different experiments were run, varying the sentiment features and lexicon used, and the depth of the decision tree. A 5-fold cross-validation was applied.

Results and Discussion

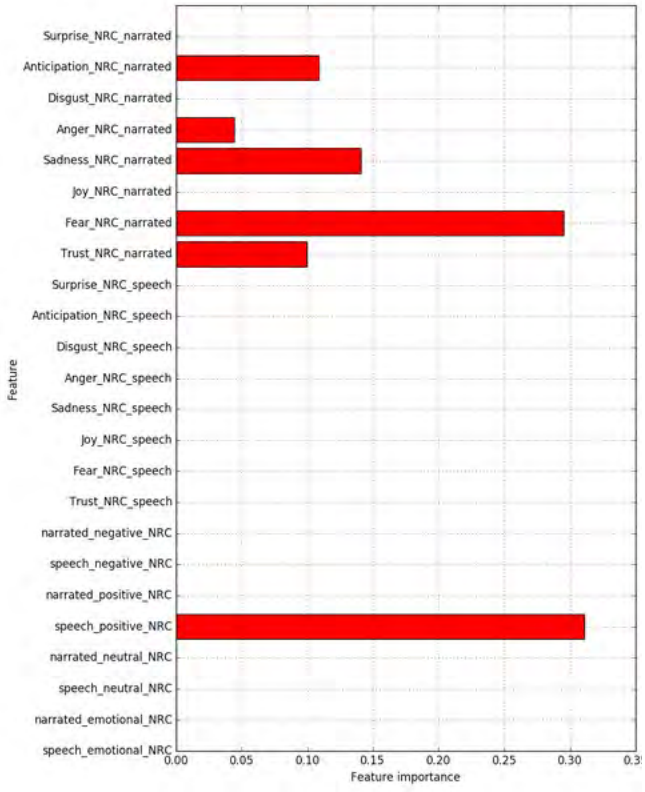
The results of the experiments are summarized in Fig. 4 below. The depth of the Decision Tree was varied between 1 and 15.[9] The accuracy is given as the mean F1 score obtained from the cross-validation. Four different sets of sentiment features were used: Features from the SentiWordNet lexicon (WN) and from the NRC lexicon (NRC), both without differentiating between direct speech and narrated text, as well as WN- and NRC-features with separate sentiment values for direct speech and narrated text (WN speech and NRC speech). The results of all experiments are compared to the "most frequent"-baseline and to a baseline obtained with an SVM classifier, using the 5,000 most frequent words.

Although the F1 scores are not very high (the highest mean value being at 0.52), almost all of them outperform the "most frequent"-baseline (0.3) which confirms that sentiment features are relevant for subgenre classification. Still, the results do not reach the best mean score of the MFW classification (0.57).[10] In terms of classification accuracy, a next step will be to combine both sentiment features and MFW to see if the sentiment features can contribute to improve the overall results.



Results for subgenre classification with sentiment features

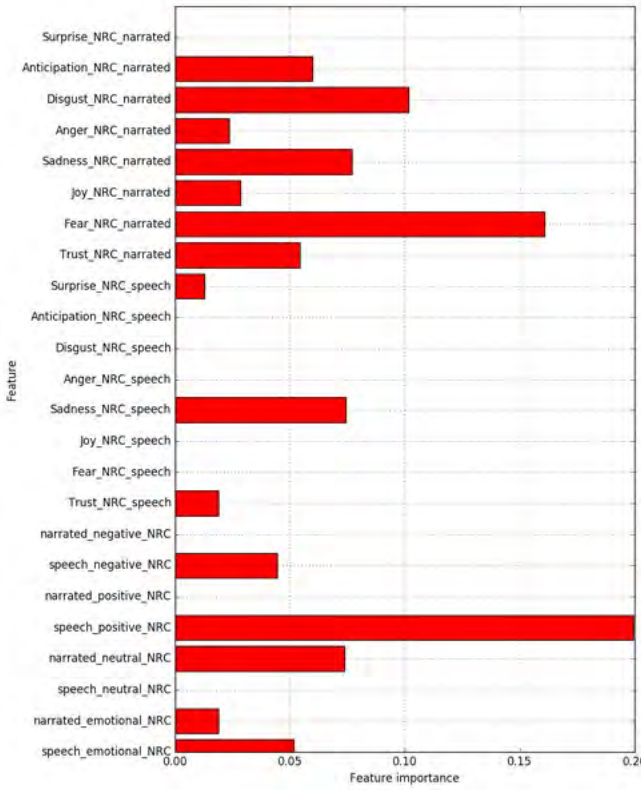
When comparing the results for the two different sentiment lexica, the NRC lexicon performs better than SentiWordNet, although the latter covers almost ten times as many words as the first one. A look into the feature importance shows that the eight basic emotions, which are only present in the NRC lexicon, are crucial (see Fig. 5 and 6).



Feature importance for a tree with depth 3, using NRC and speech vs. narrated text

9 Restricting the tree depth helps to prevent overfitting and usually leads to a better performance of the classifier on the test set.

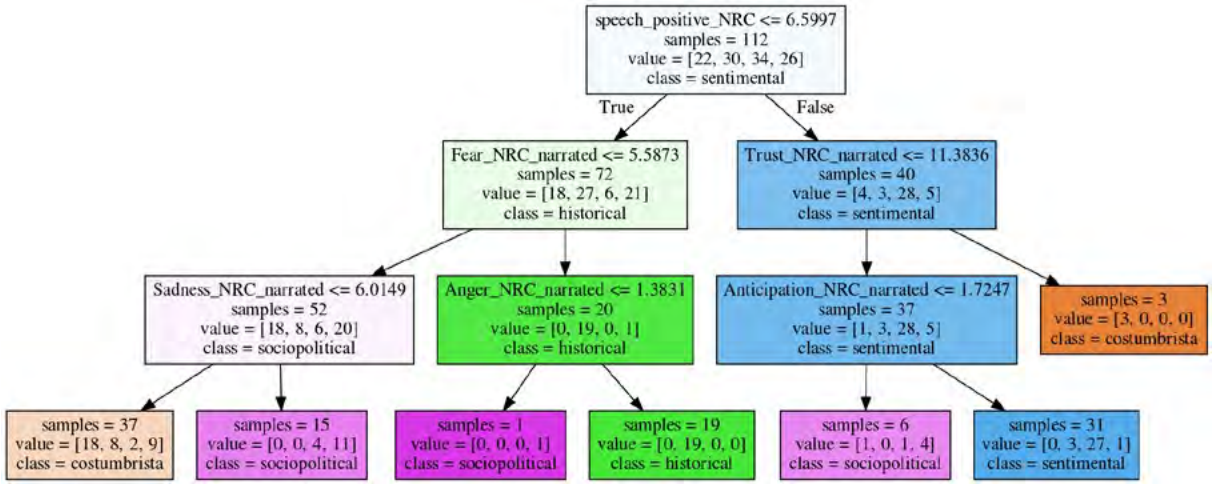
10 See Hettinger et al., 2016 for a discussion of various types of features (MFW, topics, networks) for subgenre classification, stating that genre classification in general works best with most frequent words, all words, and the like.



A Decision Tree for the classification of subgenres, based on the best parameters

Regarding the difference between speech and narration, Fig. 4 above also shows that the highest values for both WN and NRC are reached when the sentiment values are calculated separately for direct speech and narrated text. The best scores are obtained for the feature set "NRC speech". The most important feature in both example trees is positive speech, followed by narrated fear. Fig. 7 shows the Decision Tree corresponding to the feature importance in Fig. 5 above.

The tree shows that novels with higher values of positive speech are more likely to be sentimental novels. Other features that contribute to the distinction of sentimental novels are lower values of trust and higher values of anticipation in narrated text. The path for historical novels includes less positive speech and more fear and anger in narrated text. Costumbrista novels are characterized by less sadness in narrated text than sociopolitical novels and by more trust in narrated text than sentimental novels. Sociopolitical novels differ from historical novels in that they have a lower value of fear and anger in narrated text.[11]



Feature importance for a tree with depth 6, using NRC and speech vs. narrated text

11 The results of all experiments can be found at https://github.com/cligs/projects2018/tree/master/sentgenre-dh/.

Conclusion and Future Work

This exploration of sentiments in Spanish American Novels showed that Sentiment Analysis can be used as a basis for subgenre classification tasks. It has been shown that the distinction between emotions in direct speech and emotions in narrated improves the classification results considerably. Regarding the two sentiment lexica that were tested, the NRC Emotion Lexicon performs better than SentiWordNet.

The Decision Trees resulting from the classification give much insight into how sentiments in general, in direct speech and in narrated text are related to different types of novels. That the features can be interpreted easily contributes to a better understanding of what textual features are connected to the subgenres, but the classification results themselves can still be improved. Other classifiers, for example Random Forest trees or an SVM, might yield better results but will also be less interpretable. Another important next step is to increase the corpus size to make the results more stable.

References

Baccianella, S., Esuli, A. and Sebastiani, F. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. *Proceedings of LREC 2010.* Valletta, Malta: ELRA: 2200-2204. http://www.lrec-conf.org/proceedings/lrec2010/summaries/769.html (accessed April 27 2018).

García Ardeo, J. M. (2006). Eduardo Gutiérrez y sus dramas del terror. *Letras* 54: 77-94.

Henríquez Miranda, C. and Guzmán, J. (2017). A Review of Sentiment Analysis in Spanish. Una Revisión Sobre el Análisis de Sentimientos en Español. *TECCIENCIA* 12 (22): 35-48. doi: 10.18180/tecciencia.2017.22.5.

Hettinger, L., Jannidis, F., Reger, I. and Hotho, A. (2016). Classification of Literary Subgenres. *DHd2016.* Leipzig: Universität Leipzig: 154-158. http://dhd2016.de/boa.pdf (accessed April 27 2018).

Kim, E., Padó, S. and Klinger, R. (2017). Prototypical Emotion Developments in Literary Genres. *Digital Humanities 2017. Conference Abstracts.* Montréal: McGill University. https://dh2017.adho.org/abstracts/203/203.pdf (accessed April 27 2018).

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Cambridge, Mass.: The MIT Press.

Miller, G. A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM* 38 (11), 39-41.

Molina, H. B. (2011). *Como crecen los hongos. La novela argentina entre 1838 y 1872.* Buenos Aires: Teseo.

Padró, L. and Stanislovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. *Proceedings of the Language Resources and Evaluation Conference (LREC 2012).* Istanbul, Turkey: ELRA: 2473-2479. http://nlp.lsi.upc.edu/publications/papers/padro12.pdf (accessed April 27 2018).

Pang, B. and Lee, L. (2008). Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2 (1-2): 1-135.

Rivas, M. (1990). *Literatura y esclavitud en la novela cubana del siglo IXI.* Sevilla: Escuela de Estudios Hispano-Americanos.

Saif, M. and Turney, P. (2013). Crowdsourcing a Word-Emotion Association Lexicon. *Computational Intelligence* 29 (3), 436-465.

Zehe, A., Becker, M., Hettinger, L., Hotho, A., Reger, I., and Jannidis, F. (2016): Prediction of Happy Endings in German Novels based on Sentiment Information. *Proceedings of DMNLP, Workshop at ECML/PKDD.* Riva del Garda, Italy. http://ceur-ws.org/Vol-1646/paper2.pdf (accessed April 27 2018).

Zó, R. E. (2015). *Emociones escriturales. La novela sentimental latinoamericana.* Saarbrücken: Editorial Académica Española.

Winko, S. (2003). Über Regeln emotionaler Bedeutung in und von literarischen Texten. In Jannidis, F., Lauer, G., Martínez, M., Winko, S. (eds.), *Regeln der Bedeutung.* Berlin: de Gruyter, pp. 329-348.

Digitizing Paratexts

Kate Holterhoff
kate.holterhoff@gmail.com
Georgia Institute of Technology, United States of America

Digital archivists tend to disagree about the place of paratexts. Whereas *Google Books* often scans texts at such a low resolution that anything but printed words are difficult to discern, Andrew Stauffer's *Book Traces* project and Steven Olsen-Smith and Peter Norberg's The Melville Marginalia Project aims to identify individual copies of nineteenth- and early-twentieth-century books in libraries by highlighting their unique marginalia and inserts. Illustrations, advertisements, marginalia, boards, and decorative initials—the effluvium of the print form—does not digitize easily. Moreover, in terms of library and information science, paratexts resist standard means of categorization. Paratexts are problematic because they offer an exception rather than a type. To scholars, they often seem extraneous or even detrimental to the written texts they accompany. Marginalia, for instance, simultaneously defaces and compliments a text. Advertisements are a distracting and commercial accretion to an artwork. And yet, all paratexts provide necessary context for understanding the complexity and fullness of print history. The question I will address in this paper is how archivists ought broadly to understand paratexts, and how specifically should they treat nineteenth-century illustrations.

Numerous digital archives have taken on the task of scanning, categorizing, and tagging illustrations (*e.g.* the *William Blake Archive,* the *Cervantes Project*, Cardiff University's *Illustration Archive*), and yet the purpose and constraints of this task remain unfixed. In fact, Julia Tho-

mas notes in her recent *Nineteenth-Century Illustration and the Digital* (2016), that owing to the uniquely important role of context for these paratexts—usually the book or periodical—"the digital might appear an alien environment for historic illustrations." While the role of the digital image archive concerned with illustrations remains unsettled, recent scholars have used the affordances of the digital archive to open up new avenues for curation and exploration. Using as a case study a digital archive that I direct and edit titled *Visual Haggard*, a *NINES* indexed and peer reviewed archive that contextualizes and improves access to the illustrations of Victorian novelist H. Rider Haggard (1856 - 1925), I argue that digitizing illustrations must be inclusive.

I will consider the problem of inclusion and exclusion in digital archive curation. As paratexts, illustrations are lumped together with a number of visual objects that initially accompanied fictions. For this reason I explain the necessity of using metadata to differentiate illustration types. The large decorative initials which appear in many nineteenth-century texts, but originated in medieval manuscripts, are less illustrations of the text than embellishments. However, their ideological function is significant and multifold. Similarly, advertisements were often in conversation with serialized fictions—whether thematically or stylistically. In this paper I discuss strategies to enable digital image archivists committed to creating an authentic encounter with the history of print to avoid ignoring or marginalizing these types of unique and difficult paratexts.

A Corpus Approach to Manuscript Abbreviations (CAMA)

Alpo Honkapohja
alpo.honkapohja@ed.ac.uk
University of Edinburgh, United Kingdom

As anyone, who has worked with medieval manuscripts, will know, sometimes more than half of the words are abbreviated. For example, in a forthcoming paper on Middle English and Latin manuscripts of the *Polychronicon*, we found that in the most heavily abbreviated Latin sections as many as 59 percent of the words could be abbreviated, while the number for Middle English was 21 per cent (Honkapohja and Liira, in preparation). Studies comparining Latin and Romance have met with similar results (Hasenohr, 1997; Careri et al., 2011). Nevertheless, in digital scholarship, abbreviations are typically seen as something to get rid of rather than useful data to mine.

A major reason for lack of attention given to manuscript abbreviations can be found in editorial practices inherited from printed editions. It is a standard practice for editors to expand abbreviations as "a service to the reader" (cf. Driscoll, 2009). Twentieth-century editorial theory often treats abbreviations as scribal variation as "accidentals" (see e.g. W. W. Greg, 1950), not relevant for the authorial "work" contained in the manuscripts, as much scholarship focuses either directly on the work or uncovering the work under layers of scribal copying and errors. The outcome is an editorial tradition in which silently expanding abbreviations is very much the norm.

Digital approaches for making use of abbreviations as data are available, but are often not used. TEI P5 guidelines introduced the possibility of encoding both the abbreviations and their expansions using the <choice> elements with <abbr> and <expan> (cf. Driscoll, 2006, 2009; Honkapohja, 2013). Still, many digital resources continue the practice of silently expanding abbreviations. Reasons may range from considering encoding abbreviations to be too labour intensive to basing the digital resources on printed editions which expand the abbreviations (cf. Honkapohja et al., 2009). Moreover, text retrieval systems are typically unable to recognize different forms of the same word and the problem is usually solved by normalisation (cf. Kestemont, 2015: 160). Furthermore, some research questions, including investigations into syntax or stemmatology, also require normalisation. However, while normalisation may be necessary for some research questions, it also discards large amount of potentially useful data, which makes other types of research impossible.

The fairly few scholars who do work with abbreviations have identified a number of potentially interesting lines of enquiry. Abbreviations can be used, for example, for identifying change of scribe in the text (cf. Kestemont, 2015) or in historical dialectology for identifying regional characteristics in scribal language (see e.g. Smith, 2016), or studying the effect of right-margin justification on scribal spelling (Shute, 2017), or hiding endings in multilingual business writing (Wright, 2011). Consequently, the practice of expanding abbreviations is discussed and criticised by a number of scholars (Driscoll, 2006; Kytö et al., 2011; Rogos 2011, 2012; Stutzman, 2014, Lass, 2004).

Even though the problems related to the prevailing practice of silently expanding are well known, and some resources such as the *Medieval Nordic Text Archive* (MENOTA) do encode them, there have been relatively few studies which would have attempted to use them as data (e.g. Camps, 2016; Honkapohja, 2018; Kestemont, 2015; Rogos, 2012; Smith, 2016; Shute, 2017), especially in comparison to fields such as stemmatology and stylometry. My proposal for short paper presents project plan and early results for a project, called *Corpus Approaches to Manuscript Abbreviations* (CAMA), funded for September 2017- February 2020.

The current project focuses on applying methodologies developed for corpus linguistics on abbreviations in the spelling system of Early Middle English, 1150-1350. The period is of interest as it was a formative one for the writing systems of English. Linguistic situation in England changed dramatically after the Norman Conquest of 1066, which introduced a new ruling class and relegated Engli-

sh to a tertiary role after Latin and Anglo-Norman French. When Middle English texts become more numerous in the 13th century, we find a very diverse dialect landscape in which the lack of a prestigious vernacular has led to the proliferation of local varieties, with almost every text appearing to represent a separate linguistic system.

Within the Early Middle English period, my project focuses on four research questions:

- (Q1) Does each scribe have an individual scribal profile of abbreviations?
- (Q2) Are some abbreviation usages connected to certain geographic areas?
- (Q3) How are Latin and Old English abbreviations distributed in Germanic and Romance vocabulary?
- (Q4) What is the function of abbreviations in the spelling system(s) of Middle English?

The data comes from the *Linguistic Atlas of Early Middle English* (LAEME), a corpus of ca. 650,000 divided into scribal samples of localised Middle English. Each text in LAEME is based on a diplomatic transcription from manuscript facsimiles, not editions, and using a mark-up system that encodes the expansions of abbreviations, but in a way which makes identifying the abbreviation easy and workable (LAEME: 3.3.1). Consequently, it can be used to compile a dataset, which can be analysed quantitatively.

The methodology is based on corpus linguistics, statistical analysis and historical dialectology. I will use corpus enquiries to compile a dataset of the findings, then subject the dataset to statistical analysis using R and tried and tested techniques such as linear regression, linear correlation, principal component analysis, chi square test and cluster analysis which have yielded results in previous studies of abbreviations and spelling variation (cf. Kestemont, 2015; Smith, 2016).

Compiling the dataset consists of three steps:

1. Corpus enquiries, using the web interface and scripts of LAEME.
2. Corpus enquiries for unabbreviated forms of the abbreviated words found in stage 1 in each text, in which a particular abbreviation is used. These can be localised, using the lemmas tagged in the LAEME (see 2.3.2: E).
3. Compiling a dataset the results, which will include **a) results of the corpus enquiries**, i.e. the abbreviation type, the abbreviated word, non-abbreviated variant(s), frequencies, **b) information included in the LAEME metadata**, i.e. text, lemma, grammatical tag, manuscript, date, script, place, co-ordinates in the LAEME localisation grid, and **c) additional variables needed for research questions Q1 and Q3**, i.e. word origin: Germanic/Romance/Latin (12), content vs. function word (13).

The dataset will be subjected to further analysis, using:

A) The inbuilt mapping function in LAEME, which allows dynamically creating feature maps, based on the distribution of any form, its lemma, or grammatical tag.
B) Statistical analysis,
   a) linear correlation and linear regression, using the form of the abbreviation as the dependant variable, and the results encoded in the dataset (2.3.3: 3) as independent variables, calculating which of them interact with the type of the abbreviation in a certain specimen (cf. Smith, 2016),
   b) Principal component analysis common in stylometry (cf. Kestemont, 2015: 168-70).

As I am giving the presentation fairly early in the funding period, I hope to receive valuable feedback on the methodology and also to build a bridge between corpus linguistics and stylometry, creating discussion on the value and potential of scribal 'accidentals' as data.

References

Camps, J-B. (2016). *La `Chanson d'Otinel':* édition *complète du corpus manuscrit et prolégomènes à l'édition critique*. Paris-Sorbonne.<https://doi.org/10.5281/zenodo.1116735>

Careri, M., de Saint-Pol Ruby, C. and Short, I. (2011). *Livres et* écritures *en français et en occitan au XIIe siècle: catalogue illustré*, *Scritture e libri del Medioevo 8*, Viella.

Driscoll, M. (2006). Levels of transcription. In Burnard, L., O'Brien O'Keeffe, K. and Unsworth, J. (eds), *Electronic Textual Editing*, Modern Language Association, pp. 254–261.

Driscoll, M. (2009). Marking up abbreviations in Old Norse-Icelandic manuscripts. In M.G. Saibene, M. and Buzzoni, M. (eds), *Medieval Texts – Contemporary Media: The Art and Science of Editing in the Digital Age*. Ibis, pp. 13–34.

Greg, W. W. (1951/1951). The Rationale of Copy-Text. *Studies in Bibliography* Vol. 3, pp. 19-36.

Hasenohr, G. (1997). Écrire en latin, écrire en roman: réflexions sur la pratique des abréviations dans les manuscrits français des XIIe et XIIIe siècles. In Banniard, M. (ed.), *Langages et peuples d'Europe: cristallisation des identités romanes et germaniques (VIIe-XIe siècle*). Toulouse-Conques, pp. 79-110.

Honkapohja, A. (2018). "Latin in Recipes?" A corpus approach to scribal abbreviations in 15th-century medical manuscripts. In Pahta, P, Skaffari, J. and Wright, L. (eds), *Multilingual Practices in Language History: English and beyond*. De Gruyter, pp. 243-271.

Honkapohja, A. (2013). "Manuscript abbreviations in Latin and English: History, typologies and how to tackle them in encoding." *Studies in Variation, Contacts and Change in English Volume 14: Principles*

and Practices for the Digital Editing and Annotation of Diachronic Data. <http://www.helsinki.fi/varieng/series/volumes/index.html>

Honkapohja, A., Kaislaniemi, S. and Marttila, V. (2009). Digital Editions for Corpus Linguistics: Representing manuscript reality in electronic corpora. In Jucker, A., Schreier, D. and Hundt, M. (eds), *Corpora: Pragmatics and Discourse. Papers from the 29th International Conference on English Language Research on Computerized Corpora (ICAME 29)*. Ascona, Switzerland, 14–18 May 2008. Rodopi, pp. 451–474.

Honkapohja, A. and Liira, A. (in preparation). Abbreviations and Standardisation in the *Polychronicon*: Latin to English, and Manuscript to Print. In Wright, L (ed.), *The Multilingual Origins of Standard English (MOSTE)*. De Gruyter.

Kestemont, M. (2015). A Computational Analysis of the Scribal Profiles in Two of the Oldest Manuscripts of Hadewijch's Letters, *Scriptorium,* 69: 159-75.

Kytö, M., Grund, P. and Walker, T. (2011). *Testifying to Language and Life in Early Modern England: Including CD-ROM: An Electronic Text Edition of Depositions 1560-1760 (ETED)*. Benjamins.

LAEME = Laing, M. (2013). *A Linguistic Atlas of Early Middle English, 1150–1325*, Version 3.2. Edinburgh: © The University of Edinburgh.: <http://www.lel.ed.ac.uk/ihd/laeme1/laeme1.html>

Lass, R. (2004). Ut Custodiant Litteras: Editions, Corpora and Witnesshood. In Dossena, M. and Lass, R. (eds), *Methods and Data in English Historical Dialectology*. Peter Lang, pp. 21-48.

MELD = Stenroos, M., Thengs, K. and Bergstrøm, G. *A Corpus of Middle English Local Documents*, v. 2017.1. University of Stavanger. <http://www.uis.no/research/history-languages-and-literature/themest-programme/a-corpus-of-middle-english-local-documents-meld/>

MENOTA = *Medieval Nordic Text Archive*. <http://www.menota.org/EN_forside.xhtml>

Rogos, J. (2011). On the pitfalls of interpretation: Latin abbreviations in MSS of the Man of Law's Tale. In Fisiak, J. and Bator, M. (eds), *Foreign Influences on Medieval English*. Peter Lang, pp. 47–54.

Rogos, J. (2012). Isles of systemacity in the sea of prodigality? Non-alphabetic elements in manuscripts of Chaucer's 'Man of Law's Tale. <http://www.isle-linguistics.org/resources/rogos2012.pdf>

Shute, R. (2017). Pressed for Space: The Effects of Justification and the Printing Process on Fifteenth-Century Orthography. *English Studies* 98 (3): 262–82.

Smith, D. (2016). The predictability of {-S} abbreviation in Older Scots manuscripts according to stem-final littera. AMC Symposium. Conference paper.

Stutzmann, D. (2014). Conjuguer diplomatique, paléographie et édition électronique : les mutations du XIIe siècle et la datation des écritures par le profil scribal collectif. In Ambrosio, A., Barret, S. and Vogeler, G. (eds), *Digital Diplomatics. The computer as a tool for the diplomatist?*, Archiv für Diplomatik. Beiheft 14, 27190.

Wright, L. (2011). On Variation in Medieval Mixed-Language Business Writing. In Schendl, H. and Wright, L. (eds.), *Code-Switching in Early English*. De Gruyter, pp. 191-218.

On Natural Disasters In Chinese Standard Histories

Hong-Ting Su
r03944039@ntu.edu.tw
National Taiwan University, Taiwan

Jieh Hsiang
jhsiang@ntu.edu.tw
National Taiwan University, Taiwan

Nungyao Lin
nungyao@gmail.com
National Taiwan University, Taiwan

This paper describes a study which analyzes natural disasters described in the Chinese Standard Histories. We first define the scope and nature of disasters as presented in the Standard Histories. The records, in plain text but usually contain the dates, locations, type, and severity of the natural disasters, are then extracted. The extracted records are further annotated with metadata so as to meet the needs of the studies on the history of disasters. In order to ensure flexibility and extensibility, we have designed a markup language, WXML, to tag the information. A search/retrieval system with GIS is then developed to provide visualization of the distribution of time, space, and type of disasters of the search result.

We have made some preliminary observations. For instance, the number of disasters recorded during the Yuan Dynasty is significantly higher than the other dynasties (both in absolute number and on average). As another example, disasters seem to disproportionately concentrate around urban centers, in particular the capital of the time. This shows that the records in the Standard Histories may not accurately reflect the actual events, but rather how they were documented by the officials.

Natural Disasters described in the Chinese Standard Histories

Chinese Standard Histories (正史), 24 in total, are the official histories of the Chinese Dynasties. A Standard History is usually written during the succeeding dynasty, based on existing, often meticulously kept, records of the previous dynasty. These tomes start from *Shiji* (史記), written by Sima Qian (司馬遷) in the Han Dynasty around 90 BCE, and ends with *Ming Shi* (明史), the Standard His-

tory of Ming Dynasty (1368-1644). Together they cover about 2,500 years of China's written history. Fourteen of the standard Histories contain volumes of *Wuxingzhi* (Book of the Five Elements, 五行志), which record natural disasters and mysterious phenomena. Disasters are also documented in the *Benji* (Chronical of an Emperor, 本紀), another port of a Standard History. These records document the nature of disaster, time, location, and severity; thus serve as important source for modern studies of the history of disasters in China.

In this paper, we focus on the natural disasters recorded in the *Benji*'s and *Wuxingzhi*'s in the Standard Histories.[1] We exclude the human-caused and unexplainable phenomena described in the *Wuxingzhi*.[2] After analyzing the formation of the *wuxingzhi*'s and other studies of natural disasters, we classified the natural disasters into 14 categories: flood, rain, frost, hail, famine, drought, cold, snow, wind, locust, borer, plague, earthquake, and landslide.[3]

Processing the Records and Markup

We have designed an XML format (Wuxing Markup Language, or WXML) to tag the texts.

A *record* is a writing of natural disaster indicated in the text. A record contains the following elements: *event*, *time period*, *area*, *severity*, and *frequency*. A record may describe several *events*. For instance, a record of drought often also mentions famine. In this case, both events are tagged. *Time period* (written using dynasty, era, year, month, day) has three subtags: starting time, ending time, and duration. If only a date is indicated, that date is considered the starting date. If there's no mentioning of duration or ending date, then the ending date is the same as the starting date. If duration is vague (such as "it rained for some 30 days"), then the ending date tag will not be filled. The element *area* contains two subjects: location and range. Since one or several administrative regions, a river or a mountain range may be indicated in a disaster, the location tag may have multiple values. The range tag could also be an administrative region or a geographical entity. When a record describes the area as "capitol and its surrounding prefectures", the location will be the capitol of the time, and the range will be the "surrounding prefectures". *Severity* includes the effect, the damages, and the reactions that followed. For example, a flood may include the effect of the breaking of the embankment which results in flooding of the farms and houses (damages), which leads to the reduction in taxes in the following year (reaction). *Frequency* is less complicated, although not entirely trivial. A record may mention several earthquakes, without indicating the exact number. In this case, it will simply be tagged as "several".

Producing and Counting the events

We first use the 14 keywords of disasters to extract descriptions mentioning the disasters. The paragraphs are then parsed automatically to identify the records and their time, event, area, etc. We remark that each description may contain several events, several locations, or even several time periods. We then tag the events, time periods, and locations automatically from the descriptions. The dates are standardized using the Buddhist Studies Time Authority Databases developed at Dharma Drum College (http://authority.dila.edu.tw/time/). Geographic coordinates are provided using the Chinese Civilization in Time and Space developed at the Academia Sinica (http://ccts.ascc.net/). An expert is then asked to go through the result to correct manually.

Several ways have been used in the literature to count the number of events. A record may involve multiple locations, different years, and multiple disasters. The same disaster may also appear in different books. A simple way that counts only the appearance of a type of disaster was used in (Deng**,** 1973) (regardless of the frequency, locations and severity, it is counted as 1 if it appeared in China during that year at least once. Otherwise it is 0 for that year). This method was adopted later by other researchers (Luo, 2005). At the other extreme, each tuple of time, disaster, location is recorded as one event (Yuan**,** 2008). A third option is to specify a tuple of time and location as an event without consider the other attributes (Wang, 2005; Zhang, 2007). By using tags, our approach provides the flexibility of being able to adjust to any of these counting methods, without being forced to pre-select one, by simply turning on or off an attribute.

Using single time and type as the event unit (while counting multiple locations as one), we tabulated a total of 9,717 events of natural disasters mentioned in Chinese Standard Histories, after removing duplicates from 6,653 events mentioned in *Wuxingzhi* and 3,848 in *Benji*. (We also removed 489 duplicate events between *Yuanshi* and *New Yuanshi*, and 79 duplicate events between *Old Tangshu* and *New Tangshu*.) The time distribution is as follows:

1 We have also included the *Book of Signs* (靈徵志) of *Weishu* (魏書), which also contains a fair amount of natural disasters.
2 The name *Wuxingzhi* indicates a view of the world in which the five elements, metal, wood, water, fire, and earth interact with each other. Thus certain phenomena were interpreted as signals of the missing of balance. However, the portion of this type of writing diminished significantly after the 10th century (You**,** 2007).
3 *Fire* is not considered a natural disaster. Although some fire might be due to natural reasons such as forest fire caused by lightning, researchers of natural disasters usually regard fire, as a general category, a manmade disaster since it is often hard to identify the cause (Zhang, 2012)).

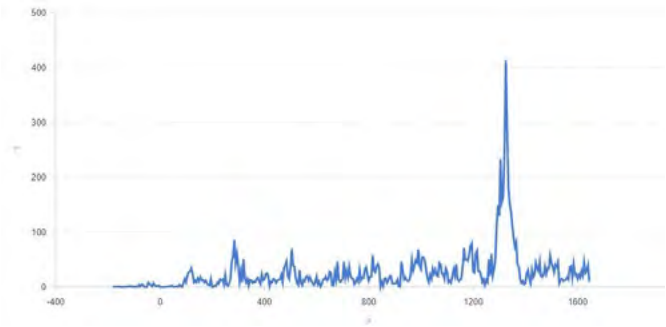


Figure 1 Distribution of natural disasters (X-axis: 5-year as a unit; Y-axis: frequency)

Note that the number of natural disasters recorded reached a peak during the Yuan Dynasty (1271-1368 BCE). (*Yuanshi*, 元史, only documented events occurred in China proper, not the Mongolian empire that ruled most part of the known world at the time.)

The system and some observations

We have built a system using the events of natural disasters mentioned above. Our interface allows one to specify one or several types of disasters, the era, and/or the areas and show the resulting data in number (or in graphs), on map, and also the texts of the events and their sources. The following is an example of disasters in the Guanzhong (關中) area.

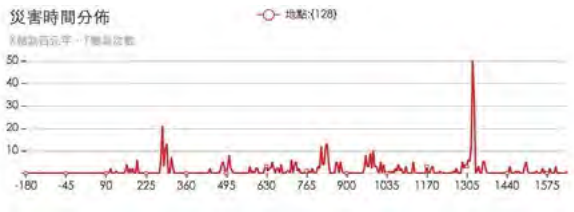


Figure 2 The number of disasters in Guanzhong area

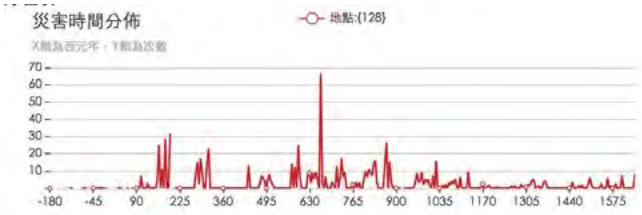


Figure 3 The percentage of disasters in Guanzhong area vs the country

The x-axis in both figures are years (in 5 years) in western calendar, while the y-axis of Figure 2 is the absolute number of disasters and the y-axis of Figure 3 is the *percentage* of *all* natural disasters recorded in the entire China during that time period. Note that although the number of disasters peaked around the year 1300, the percentage was dramatically high during the early Tang dynasty (618-907 BCE), when Changan (長安), a city in Guanzhong (關中), was the capital at the time. After the demise of Tang, the attention of latter empires gradually shifted to the northeast and south, and the percentage of disasters trailed off significantly, as Guanzhong gradually became irrelevant.

There are other interesting phenomena. For instance, there seemed to be more natural disasters during prosper periods. This may indicate that when the country was going through great turbulence such as foreign invasion or peasant revolt, the local officials simply did not bother to report natural disasters.

Concluding Remarks

In this paper we described a study on the natural disasters documented in the Chinese Standard Histories. We analyzed previous work on natural disasters and classified the events into 14 categories. We extracted texts of the records from *Wuxhingzhi* and *Benji*, and developed a markup language WXML to tag the events. We then build a system which is flexible in that one can use any of the measures mentioned above to show the results. Since the records are time-standardized and geo-referenced, our system also allows one to specify the type of disasters, time period, and locations and present the results either as charts or geographically. We are currently developing our system to allow full-text search to add flexibility.

We presented some preliminary observations. They seem to show that the natural disasters documented in the Standard Histories may not truthfully reflect the actual natural disasters that occurred. In other words, the records may reflect more on the circumstances under which the books were produced rather than the actual disasters that occurred. To more accurately capture natural disasters in Chinese history, one should at least also consult the local gazetteers (difangzhi, 地方志) (Chen, 2016). The WXML that we have designed is sufficiently flexible to incorporate those records as well.

References

You, Z. (2007). The Zheng (徵) and Ying (應) of Middle-age China. *Journal of Capital Normal University*, 2007.6: 10-16.

Zhang, G. (2012). *General Theory of Disasters*.

Deng, Y. (1973). *History of Disasters of China*. The Commercial Press.

Luo, C. (2005). Temporal-Spatial Distribution of East Han, MS Thesis, Zhengzhou Univ.

Yuan, Z. (2008). *Chinese Disaster History: Yuan Period*, Zhengzhoug University Press.

Wang, F. (2005). Disasters of two Jin, MS Thesis, Jiangxi Normal University.

Zhang, W. (2001). Preliminary Studies on the Natural Disasters of Han, PhD Thesis, Shaanxi Normal University.

Chen, S. (2016). Remapping Locust Temples of Historical China and the use of GIS, *Review of Religion and Chinese Society,* 149-163. Doi 10.1163/22143955-00302002.

REED London and the Promise of Critical Infrastructure

Diane Katherine Jakacki
diane.jakacki@bucknell.edu
Bucknell University, United States of America

Susan Irene Brown
sbrown@uguelph.ca
University of Guelph, Canada

James Cummings
james.cummings@newcastle.co.uk
Newcastle University, United Kingdom

Kimberly Martin
kmarti20@uguelph.ca
University of Guelph, Canada

Alan Liu has called upon digital humanists to think more critically about infrastructure - the "social cum technological milieu that at once enables the fulfillment of human experience and enforces constraints on that experience" (Liu, 2017). Liu's invitation comes at the moment when researchers involved in large-scale, long-term projects are shifting focus from remediation and the creation of digital incunabula to transmediation and the development of systems that support sustained discourse across ever-morphing digital networks, when we are recognizing the potential for "dynamism of the base or serialized form of the text—the state in which it is stored—as opposed to dynamic modes of presentation" (Brown, 2016: 288). REED London is one such project with a polyvalent dataset that spans over 500 years' worth of archival records, embracing from the start the need to establish a stable, responsive production and presentation environment primed for use by a wide range of scholarly audiences. Thus we find that we are immediately testing those infrastructural constraints. In this paper, members of the REED London project team will address the challenges we face as we develop and implement a framework that trains us to think about our collected data in relation to much larger networks of disparate resources and user needs.

REED London develops from a partnership between the Records of Early English Drama (REED) and the Canadian Writing Research Collaboratory (CWRC). Together we are establishing an openly accessible online scholarly and pedagogical resource of London-centric documentary, editorial, and bibliographic materials related to performance, theatre, and music spanning the period 1100-1642. With support from the Andrew W. Mellon Foundation and a CANARIE Research Software Program grant, a team of researchers in the digital humanities and performance history from the U.S., Canada, and the U.K. are building a stable, extensible editorial production and publication environment that will create new possibilities for scholarly presentation of archival materials gathered from legal, ecclesiastical, civic, political, and personal archival sources in and around London. The REED London project combines materials from three printed REED collections (*Inns of Court*, *Ecclesiastical London*, and *Civic London to 1558*), the prosopographical material from REED's *Patrons & Performances (P&P)*, the bibliographical materials of the *Early Modern LondonTheatres (EMLoT)* database, and in-progress and planned digital collections focusing on London area performance spaces, most notably the Globe, Rose, and Curtain theatres and Civic London 1559-1642.

REED is an internationally renowned scholarly project that has worked to locate, transcribe, and edit evidence of drama, secular music, and other communal entertainment in Britain from the Middle Ages until 1642. Since 1979 REED has published twenty-seven printed collections of transcribed records plus contextual materials. REED has long recognized the importance of online access to its resources,first with *P&P* and *EMLoT*, and more recently with the born-digital collection *Staffordshire*. REED has wrestled with the balance between what was once considered its "core" print publication activities and "adjunct" digital efforts, in the process migrating its data across a succession of programs and formats from Basic and dBASE to TEI P5 XML and MySQL (Hagen, MacLean, and Pasin, 2014). REED has developed its digital resources in ways that complicate integration (*P&P* exists in a Drupalinstance; *EMLoT* was built in a version of Django that is now out-of-date; *REED Staffordshire* was lightly tagged in TEI and relies on EATSML for entity management, an XML format used by the Entity Authority Tool Set (EATS) for serialisation of its data). The components of REED London must therefore first be made intra-operable before they can become interoperable (Jakacki, 2016). The partnership with CWRC supports broader adoption of standards for TEI text markup, RDF metadata specifications, and named entity aggregation, most immediately with the ingestion of *EMLoT* and the printed *Inns of Court* collection.

CWRC is an online infrastructure project designed to enable unprecedented avenues for studying the words that most move people in and about Canada. Built with funding from the Canada Foundation for Innovation, the CWRC platform supports best practices in the production of online collections, editions, born-digital essays, anthologies, collections, monographs, articles, or bibliographies, and supports the inclusion of visual, audio, and video sources (About CWRC/CSÉC). It supports collaboration through the use of interoperable data formats and

interlinking of materials, and for teams like REED London provides invaluable tools for communicating, tracking activity, and workflow. We envision that as the partnership develops and as REED London advances through production toward publication we will take full advantage of CWRC's functionality. From the start we have worked directly in CWRC's unique editor, CWRC-Writer, which allows us to edit REED London records, essays, and bibliographical material using more diplomatic and critical TEI P5 XML markup and at the same time creating semantic web annotations with RDF to identify, manage, and interlink entities contained within. The platform is also helping us to develop a better editorial workflow through management of access to data and editing by role, team communications, tracking and reporting of team activities.

To ensure REED London's stability and sustainability while extending its content and value to new generations of scholars the project is being built within the CWRC environment. The scope of REED London would not be possible without the sophisticated, integrated platform that CWRC provides. The focus of our first year is the design and construction of a collaborative online production and publication environment. Extending from CWRC's existing integrated content management and preservation system, the enhanced environment will accommodate the range of record texts, editorial and bibliographical content from the source materials, while a customized browser-based CWRC-Writer platform will support the team's goal of developing online editorial collaboration and review. The resulting streamlined production and publication environment will yield multi-faceted user-centered editions, meaning that agile component archival and editorial parts can cohere according to various criteria in response to scholars' research and teaching needs. In this way we are establishing a platform that produces new forms of "edition" that combine customized textual and contextual materials, exportable customized datasets and dynamic data visualizations. It also means that we will be able to realize the promise of extending the value of these materials to colleagues in fields beyond performance history, including political, religious, and cultural studies, and linguistics.

The partnership between CWRC and REED allows us to explore the potential for new research applications associated with prosopography, networks, and deep contextualization. REED London's wealth of references to very itinerant individuals across contemporaneous records means that we will be able to discern patterns through linking, analysis, and visualization. We will leverage REED's named entities for linking people, places, events, and organizations. Our team has healthy debates about the problematic present of linked data. Brown has stated that, "linking up with other data means connecting one ontology to another, and this brings with it a pressure toward generalization rather than specificity" (Brown, Simpson, et. al., 2015). Cummings has posited that "being able to seamlessly integrate highly complex and changing digital structures from a variety of heterogeneous sources through interoperable methods without either significant conditions or intermediary agents is a deluded fantasy" (Cummings 2014). Still, as a group we hope that by publishing our ontologies as a means of relating these entities as linked open data, we will be able to contribute to larger dialogues about class and society in Britain - certainly over the 500 years covered by REED London, but also about the development of Britain and Europe. CWRC content will be aggregated by the Advanced Research Consortium (ARC), and REED London will benefit from that aggregation, as we anticipate that people who figure in the REED London corpus, such as Elizabeth I, Francis Bacon, and Inigo Jones will be discoverable by scholars searching for these known figures across other linked resources. Perhaps more important, REED London records include extended references to thousands of Londoners who were in some way connected to performance, but who were not defined by that connection: civic officials, guild members, lawyers, clerks, priests, etc. The work of this project thus holds as yet unrealized value for a much broader understanding of British historical subjects.

Working within CWRC's platform and optimizing CWRC-Writer has allowed the core REED London team to move efficiently to an advanced planning phase. By the end of 2017 we will have designed templates for all record formats from *Inns of Court* and mapped database fields from *EMLoT* to align with the record parts from the print collections. We will have harvested a preliminary "white list" of named entities (people, places, organizations) from all three print collection indexes, P&P, and Staffordshire. Because of this efficient onramp we will be able to focus in the first half of 2018 on ingesting data, records, and contextual materials from Inns of Court and EMLoT. We will test the REED-specific entity list on ingested materials. We will also begin to user-test the editorial workflow system with the larger project team of REED editors and staff. By June 2018 we will have begun semantic tagging and experimentation with the CWRC HuViz semantic web visualization tool. At the DH 2018 conference we will report on further customization of the CWRC interface, our plans for data discovery and research collaboration, and present preliminary plans for user-responsive editions and data linkage.

References

Brown, S. (2016). Tensions and Tenets of Socialized Scholarship. *Digital Scholarship in the Humanities*, 31 (2): 283-300.

Brown, S., Simpson, J., CWRC Project Team, and Inke Project Team. (2015) An Entity By Any Other Name: Linked Open Data as a Basis for a Decentered, Dynamic Scholarly Publishing Ecology. *Scholarly and Research Communication* 6 (2). http://src-online.ca/index.php/src/article/view/212/409.

Canadian Writing Research Collaboratory project website. http://www.cwrc.ca/en/.
Cummings, J. (2014). The Compromises and Flexibility of TEI Customisation. In Mills, C., Pidd, M. and Ward, E. (eds), *Proceedings of the Digital Humanities Congress 2012*.
CWRC: About CWRC/CSÉC webpage. http://www.cwrc.ca/about/#whatis
CWRC Humanities Visualizer webpage. http://www.cwrc.ca/uncategorized/huviz-tool/
Early Modern London Theatres website. http://www.emlot.kcl.ac.uk
Entity Authority Tool Set (EATS) website. https://eats.readthedocs.io/en/latest/index.html
Hagen, T., MacLean, S., and Pasin, M. (2014). Moving Early Modern Theatre Online: the Records of Early English Drama introduces the Early Modern London Theatres. http://static.michelepasin.org/public_articles/2014-REED_McLean-Pasin.pdf
Jakacki, D. (2017) REED London: Humanistic Roots, Humanistic Futures. Paper given at MLA 2017. http://dx.doi.org/10.17613/M67794
Jakacki, D. (2016) REED and the Prospect of Networked Data. Paper given at the Conference of the Canadian Society for Renaissance Studies.http://dx.doi.org/10.17613/M6CK59
Liu, A. (2017) Toward Critical Infrastructure Studies", paper given at the University of Connecticut. https://www.youtube.com/watch?v=2ojrtVx7iCw
Records of Early English Drama project website. http://reed.utoronto.ca
REED Patrons and Performances website. https://reed.library.utoronto.ca
REED Staffordshire Collection website. https://ereed.library.utoronto.ca/collections/staff/

Large-Scale Accuracy Benchmark Results for Juola's Authorship Verification Protocols

Patrick Juola
juola@mathcs.duq.edu
Duquesne University, United States of America

Authorship attribution, the analysis of a document's contents to determine its author, is an important issue in the digital humanities. An accurate answer to this question is important, as not only do scholars rely on this type of analysis, but they are also used, for example, to help settle real disputes in the court system (Solan, 2012). It is thus important both to have analyses that are as accuracy, and to know what the expected accuracy levels are.

In keeping with good forensic practice, scholars such as Juola (2015) have proposed formal protocols for addressing authorship questions such as "were these two documents written by the same person?" Juola (2015) described a simple and understandable protocol based on a relatively small number of distractor authors, multiple independent analyses (e.g, separate analyses based on character n-grams, on word lengths, and on distributions of function words), and a data fusion step based on the assumption that the analyses were biased towards giving correct answers. Juola (2016) proposed minor revisions using Fisher's exact test to formalize the probability of a spurious match. The revised protocol has been formalized into a software-as-a-service product called Envelope to provide a standard (and low cost) authorship verification service.

We reimplemented Juola's (2016) protocol on a corpus of blog posts to determine whether, in fact, the protocol yields acceptable accuracy rates. Our reimplementation used the JGAAP open-source software package, an ad-hoc distractor set of ten authors (plus the author of interest), and the five analyses listed in Juola (2016): Vocabulary overlap, word lengths, character 4-grams, 50 MFW, and punctuation.

Blog data was taken from the Blog Authorship Corpus [Schler et al. (2006)] a collection of collected roughly 140 million words of blog text from 20,000 bloggers collected in August 2004. From this collection, we gathered 4000 examples of authors who had written 300 or more sentences. Ten of these authors were reserved, following Juola (2015;2016) as fixed distractor authors, while the others were randomly paired to create wrong-author test sets.

To test same-author accuracy, the first hundred sentences of each of the remaining 3990 blogs were used as "known documents" in the Envelope protocol, while the last hundred sentences of that author were used as "unknown documents." Perhaps obviously, the correct answer for these tests is that the documents should verify as the same author. To test different-author accuracy, the first hundred sentences of every author in the set was used as a "known document" and compared to the last hundred sentences of the other, paired, author. This procedure generated nearly four thousand test cases of both same and different authors. Each test case was analyzed five times and the rank sum of the known document within the eleven candidate authors calculated as an overall similarity measure from 5..55. This was converted to a *p*-value using Fisher's exact test.

Juola (2016) recommends a seven-point evaluative scale, as follows:

- $p < 0.05$ (Strong indications of same authorship)
- $p < 0.10$
- $p < 0.20$
- $p < 0.80$ (Inconclusive)
- $p < 0.90$
- $p < 0.95$
- $p >= 0.95$ (Strong indications of different authorship)

The results of these experiments are presented in table 1. The final column indicates the odds ratio; the likelihood that any particular finding at that level corresponds to an actual correct author.

p-value	Same Author	Different author	Odds
< 0.05	2948	748	3.941
< 0.10	246	359	0.686
< 0.20	195	396	0.492
< 0.80	409	1390	0.294
< 0.90	54	234	0.231
< 0.95	47	230	0.204
> 0.95	91	663	0.137

These results show that, in the same-author case, the proposed protocol is very good at identifying same-authors; roughly 3/4 of the actual same-author cases tested at the 0.05 level or better. Because of this, any result less stringent than "strong indications of same authorship" is actually evidence *against* same-authorship. The different-author case is more problematic; in theory, if there is no relationship between the known and questioned documents, the p-value should be uniformly distributed, representing a variety of chance relationships. However, the $0.20 < p < 0.80$ range ("inconclusive") contains 60% of the probability space, but only 1390/3990 = 35% of the different-author analyses. By contrast, the $0 < p < 0.05$ contains 19% of the analyses, while $0.95 < p < 1.00$ contains 17% of the different-author analyses. The observed distribution is thus highly weighted to the extremes of the probability space.

These results indicate that the underlying independence assumptions -- that (e.g.) similarity measured by analysis of word lengths is independent of similarity derived from the most common (function) words -- are not held generally. If a set of genuinely independent analyses could be found, the accuracy of this protocol would be greatly enhanced. Assuming the same distribution for the same author case, the odds ratio for the "strongly indications of same authorship" would be closer to 15:1 rather than 4:1.

Nevertheless, these results do show that, suitably interpreted, Juola's proposed protocol yields accurate results in a high proportion of test cases. We continue to work both on the development of a better analysis suite (with better independence properties) as well as continuing to replicate this experiment to obtain more accurate estimates.

References

Juola, Patrick. (2015). The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions. *Digital Scholarship in the Humanities. Volume* 30, Issue suppl_1, 1 December 2015, Pages i100–i113,https://doi.org/10.1093/llc/fqv040

Juola, Patrick. (2016). Did Aunt Prunella Really Write That Will? A Simple and Understandable Computational Assessment of Authorial Likelihood. *Workshop on Legal Text, Document, and Corpus Analytics - LTDCA 2016*, San Diego, California.

J. Schler, M. Koppel, S. Argamon and J. Pennebaker. (2006). Effects of Age and Gender on Blogging in- *Proceedings of 2006 AAAI Spring Symposium on Computational Approaches for Analyzing Weblogs*.

Solan, Lawrence M. "Intuition versus algorithm: The case of forensic authorship attribution."*JL & Pol'y* 21 (2012): 551.

Adapting a Spelling Normalization Tool Designed for English to 17th Century Dutch

Ivan Kisjes
i.kisjes@uva.nl
University of Amsterdam, The Netherlands

Wijckmans Tessa
tessa_wijckmans@hotmail.com
Huygens ING/Nederlab, The Netherlands

Context

One of the bigger problems in comparing historic Dutch texts is wildly differing spelling of the same word. Seventeenth century Dutch did not have standardized spelling. Many spelling variants of the same word coexisted, making it very difficult to use any language processing tools on such texts because they depend on the same word being spelled the same way. So, for example basic algorithms like named entity recognition to recognize place or personal names, or even just part-of-speech tagging to find the grammatical context of words to analyze, for example, changing meanings of words of phrases work less well on older texts. Other languages, of course, have the same problem.

The Dutch digital research platform *Nederlab* aims to provide researchers with as many current and historic Dutch text and a toolset to do research on them. As such, spelling normalization would be an important addition to their tools. This project is a collaboration between the CREATE-project of the University of Amsterdam and *Nederlab* to tackle that problem. To deal with the problem, rather than developing a tool from scratch, we chose to adapt an existing tool to this situation: VARD2.

VARD2

VARD2[4][5] (an acronym of VARiant Detector) is a Java tool developed by Alistair Baron. It uses two lists (a normalized word list and a variant list) to suggest or replace variant words with their normalized counterparts. The normalization suggestions using a combination of four different methods: 1. known variant replacements; 2. character edit distance; 3. letter rules and 4. phonetic distance. Not all of these were useful for Dutch: the phonetic matching algorithm for example is based on English phonemes and hence did not work on these texts, but the re-spelling rules and the known word replacements worked very well.

VARD2 was designed to normalize Early Modern English, but is modifiable for other languages with a custom configuration. To create a configuration we used the modifiable parts of VARD2: the letter rules, the variant list and the normalized word list.

Corpus

We used the 1657 edition of the Dutch translation of the bible as a training set. Not only because there was a modernized version of it available that stuck rather closely to the original word order, but also because it would make it possible to later include another edition of the same book printed in 1637 to easily find more spelling variants for the words we had manually respelled or checked in the 1637 edition. We were able to make a golden standard of modernized spelling for the books Genesis and Exodus.

Choices

We chose to only do orthographic respelling, in order to preserve grammatical relevant elements of the texts as those may be relevant to research using natural language processing. One problem were words that did not follow Dutch re-spelling rules or did not have a clear Dutch respelling: foreign words, particularly place names and personal names, We chose to ignore such words as they would taint re-spelling rules for Dutch.

Problems & solutions

The first problem we encountered was the lack of any usable existing word list of all possible conjugations in modern Dutch. To get as many possible conjugations of every Dutch word that occurs in the *Woordenboek der Nederlandse Taal*[6] *(WNT)* a two-pronged approach was necessary. A set of algorithms, one per word class provided possible conjugations for each word in the *WNT*. First approach: for some word classes we were able to check the conjugations manually, but the large numbers of nomina and verbs made that impossible to do in this project. Second approach: for those the resulting word lists were checked automatically against the occurrences of those words in the *Corpus of Spoken Dutch*[1], *Dutch Wikipedia*[2] and *Verbix*[3].

Another problem, there was no set of respelling rules available that was effective for respelling Early Modern Dutch - the rule sets available did correct some spellings but caused mistakes in others. Extracting re-spelling rules from patterns in our golden standard provided an effective set of rules, especially when we generalized the rules where possible to catch similar instances.

Third, VARD2 could not handle word variations where two words should be re-spelled to a single word. Our solution was to pre-process texts with a script to remove spaces from such words.

The fourth problem was that some homonyms had overlapping spelling variations but needed to be re-spelled to different spellings in modern Dutch. An example is the word 'nog': spelling variations 'nog' and 'noch' were used interchangeably, but in modern spelling those two spellings denote differences in meaning. The only way to determine the correct modernization is to take the grammatical context of the word into account, which VARD2 does not do. This necessitated a second pre-processing step: we were only able to run a few tests, but part of speech tagging the original text and (manually) selecting a few patterns that marked one meaning or the other seemed to provide enough information to deduce the correct re-spelling.

Results

All in all, with a few additions and modifications a tool like VARD2 can be successfully converted to work on a Early Modern Dutch. Tests on other types of texts (a treatise on mathematics from 1605, the description of a beached whale from 1599, a description of the New World from 1770, a poetry book from 1637 etc) show promising results, indicating that a little extra training can make this configuration work well for different genres. Automatic respelling of the entire 1657 bible at a 95% confidence level resulted in automatic re-spelling of 62% of 340,000 variants. For the earlier edition (1637), automatically correcting at 95% confidence corrects 60% of just short of 350.000 unknown words, at 75% confidence 84% of the variants were corrected. The paper will show the results of automatically re-spelling 17th century texts using a VARD2 trained on just the first two chapters of the bible.

4 http://ucrel.lancs.ac.uk/vard

5 Baron, A. and Rayson, P. (2008). VARD 2: *A tool for dealing with spelling variation in historical corpora. Proceedings of the Postgraduate Conference in Corpus Linguistics*, Aston University, Birmingham, UK, 22 May 2008.

6 http://wnt.inl.nl

Differential Reading by Image-based Change Detection and Prospect for Human-Machine Collaboration for Differential Transcription

Asanobu Kitamoto
kitamoto@nii.ac.jp
Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems; National Institute of Informatics, Japan

Hiroshi Horii
a-horii@amane-project.jp
AMANE LLC, Japan

Misato Horii
ymisachi@amane-project.jp
AMANE LLC, Japan

Chikahiko Suzuki
ch_suzuki@nii.ac.jp
Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Japan

Kazuaki Yamamoto
yamamoto.kazuaki@nijl.ac.jp
National Institute of Japanese Literature, Japan

Kumiko Fujizane
zanezane@post.ndsu.ac.jp
Notre Dame Seishin University, Japan

Introduction

Japanese books in the Edo period (1603-1868) were mainly published by woodblock print. Their caligraphic writing style using differnt characters prevents native Japanese people to read and understand the content, and the knowledge of the past has been buried in libraries. To change this situation, NIJL-NW project started a ten-year mass digitization program to create the open dataset of 300,000 old Japanese books [7]. To take advantage of emerging big data of Japanese culture, we are working on the development of "deep access" technology to make the content of books accessible by structuring the content by either manually or automatically.

This paper focuses on a series of old Japanse books called "Bukan" [6]. Bukan offers the directory of families of the state king (Daimyo) and bureaucrats of the central government (Bakufu) in the Edo period. Bukan has a unique history. It had been a best seller book for as long as 100 to 200 years, had been updated and published frequently with a peak frequency of a few times in a month, and had been the battle field of two commercial publishers competing each other to improve the quality of their own Bukan editions. Because of good coverage and quality of Bukan, the comprehensive analysis of Bukan is expected to improve our understanding on the political, administrative, and cultural structure in the Edo period.

Comprehensive analysis cannot be achieved, however, without a solution to the problem of multiple versions. Bukan had been published for a long period with high frequency, and it is not known how many versions had been published, or how to decide the proper ordering of existing versions. Moreover, the complete transcription of Bukan is not realistic due to a large amount of text across multipe versions. In short, two major problems, management of versions and reduction of transcription, need to be solved for comprehensive analysis of Bukan.

Method

We first propose the concept of "differential reading," which refers to the mode of reading books, such as close reading and distant reading. It is a reading focusing only on changes between different versions with support from digital tools. Algorithms to detect changes in different versions are two-fold; namely text-based and image-based approaches.

Text-based change detection is effective for manuscripts. Many tools, such as CollateX [2] and ViTA [9], have been developed for text comparison, or Versioning Machine [8], for structured text or TEI (Text Encoding Initiative). In the case of woodblock print, however, image-based change detection has a number of advantages. In the terminology of old Japanese bibliography, versions can be further classified into "publication" and "correction," where the former refers to the complete re-creation of the woodblock, while the latter refers to the application of small patches to the woodblock. Change detection on publication is an easy problem for image processing, and change detection on correction is also feasible by image matching because only a small part is corrected and other parts remain the same. Other advantages of image-based change detection include transcription-less change detection and non-textual change detection.

By taking advantage of image-based change detection, we formulate differential reading as a two-step process; namely machines work first to detect changes, and humans work next to read changes.

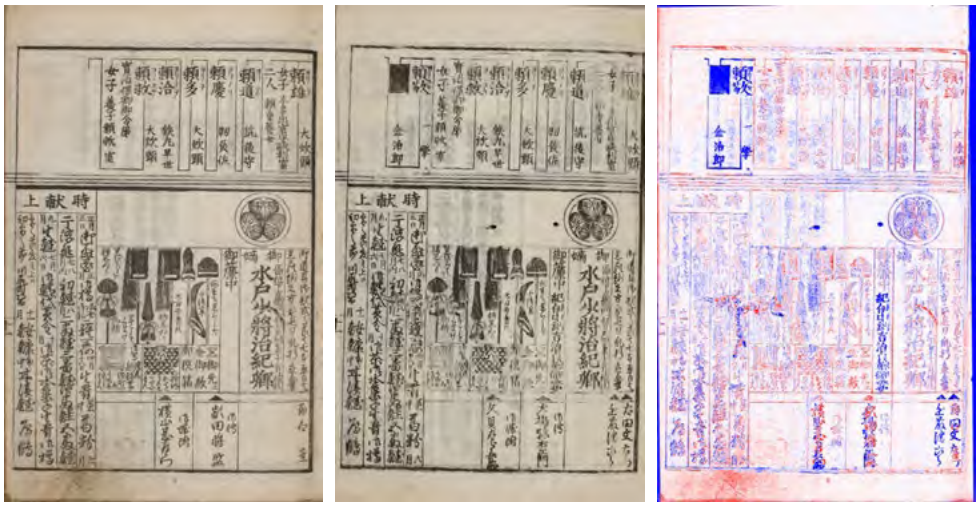


Figure 1: Comparison of two different versions of Bukan. Left: Kansei Bukan (1789); middle: Kansei Bukan (1791); right; the result of change detection, where red color represents regions present only on the 1789 version, and blue, the 1791 version.

Results

An image-based change detection algorithm was implemented on image processing library OpenCV 2.4 with a combination of algorithms such as FAST for feature detection, BRIEF for feature description, and Hamming distance for feature matching. In addition, RANSAC was used for estimating homography matrix for matching two images. Changes are then emphasized using a coloring scheme by assigning red and blue for large difference in pixel values and white for small difference in pixel values.

We compared two different versions of Bukan, Kansei Bukan (1789) [3] and Kansei Bukan (1791) [4] to check if the image-based change detection algorithm can identify changes between versions two years apart. Figure 1 shows the result of image-based change detection. It is clear that a part of the page, such as the genealogy of the family, has been changed from the 1789 version to the 1791 version. In the workflow of differential transcription, machine gernerated change infromation will be transferred to planned differential reading interface so that humans can focus only on a part of the image.

Differential transcription needs base transcription, on which transcription of subsequent versions depend. Initially the database of "Bukan Complete Collection" [1] uses Kansei Bukan (1789) as the base transcription. The database not only contains basic information about Daimyo, but also offers visualization about "Sankin Kotai," which is a required travel for Daimyo between their states and Edo city to meet Shogun (the national leader) in every two years or more often. Animated visualization in Figure 2 shows spatio-temporal and seasonal patterns of their trips coordinated by Bakufu. The database also offers the graphic design collection of Daimyo, such as family emblem, costumes, and tools they used for official activities.

We found one important missing element in creating the database; namely the standard ID system agreed within the community. Bukan is a collection of entities, such as people and political organization that changes over time. To uniquely identify entities appearing in different sources and to create a time-series database of linked entities, we need the standard ID system in the Edo period through collaboration with historians. With a proper ID system, this system may evolve into the information infrastructure of people and political entities for the historical studies of the Edo period.

Figure 2: Bukan Complete Collection website. Left: the list of Daimyo family emblems; right: animated visualization of spatio-temporal patterns of Daimyo trips. Only Japanese website is available at this moment

Discussion and Conclusion

The advantage of differential reading is two-fold. First, when reading two similar versions, differential reading has advantage over close reading by reducing the burden of human attention. A traditional approach of side-by-side comparison is error-prone, and machines can be optimized for pixel-level comparison without loss of attention by fatigue. For this type of task, human-machine collaboration should evolve into a combination that machines are in charge of low-level change detection while humans are in charge of high-level interpretation. Second, differential reading can be used as a component for differential transcription. The base transcription is required in any case, but the amount of transcription for subsequent versions is significantly reduced. A version management system may play an important role to optimize the transcription workflow, which is left for future work.

A proposed approach of differential transcription by human-machine collaboration is not only effective for Bukan, but also applicable to other woodblock print books with different versions. Our tools have been developed on IIIF (International Image Interoperability Framework), which allows us to apply our tools not only on NIJL-NW datasets but other datasets in the same manner. In the future, we plan to make a user interface on top of our IIIF Curation Viewer [5] and combine it with a workflow management tool to support efficient work of transcriptors.

References

Bukan Complete Collection, http://codh.rois.ac.jp/bukan/, (accessed April 27 2018).

CollateX, https://collatex.net/, (accessed April 27 2018).

Kansei Bukan (1789), https://doi.org/10.20730/200018823

Kansei Bukan (1791), https://doi.org/10.20730/200018825

IIIF Curation Viewer, http://codh.rois.ac.jp/software/iiif-curation-viewer/, (accessed April 27, 2018).

KITAMOTO, A., et.al. (2017) Structuring Time-Series Historical Sources by Human-Machine Specialization: Toward the Construction of Edo Information Platform Referring to "Bukan", *IPSJ SIG Computers and the Humanities Symposium 2017*, pp. 273-280 (in Japanese).

NIJL-NW project, http://www.nijl.ac.jp/pages/cijproject/index_e.html, (accessed April 27 2018).

Versioning Machine, http://v-machine.org/, (accessed April 27 2018).

ViTA (Visualization for Text Alignment), http://ovii.oerc.ox.ac.uk/vita/, (accessed April 27 2018).

The History and Context of the Digital Humanities in Russia

Inna Kizhner
inna.kizhner@gmail.com
Siberian Federal University, Russian Federation

Melissa Terras
m.terras@ed.ac.uk
University of Edinburgh, United Kingdom

Lev Manovich
manovich.lev@gmail.com
City University of New York, United States of America

Boris Orekhov
nevmenandr@gmail.com
National Research University Higher School of Economics, Russian Federation

Anastasia Bonch-Osmolovskaya
abonch@gmail.com
National Research University Higher School of Economics, Russian Federation

Maxim Rumyantsev
m-rumyantsev@yandex.ru
Siberian Federal University, Russian Federation

The history and context of the development of Digital Humanities in Russia as outlined in this paper shows that there are various influences at play which have led to the forming of the Russian DH field. We link the quantitative methods used to previous trends in scholarship, including mathematics, Russian editorial practices, and the development of museum computing in the country. By doing so we can consider the individual societal contexts which encourage a field to emerge, and although that field may look similar to outsiders, identify the lineage of intellectual approaches which still influence methods and cultures within the discipline.

The connection between Russian Formalism and the Digital Humanities (Allison et al., 2011; Moretti, 2013; Jockers, 2013; Stanford University, 2015) relates to the tradition that originated following the strengthening of Russian mathematics at the turn of the nineteenth century after the Moscow Mathematical Society was established in 1864. The influence of this school on literary studies can be traced through the twentieth century from Andrey Bely's experiments at the threshold of mathematics and poetry (Akimova, Shapir, 2006; Giansiracusa and Vasilieva, 2017) to the Moscow Linguistic Circle with Roman Jacobson as its chair (Akimova, Shapir, 2006; Pil'shchikov, 2015), to the Prague Linguistic Circle and further to the Tartu-Moscow School (Uspensky, 1998). Boris Jarkho's 'Research Methods for Literary Studies' written in 1936 anticipated the approach of Stanford Literary Lab not

only in its 'quantitative interpreting' (Underwood, 2017) but also in a skill of a scholar able to see wider contexts and make bridges across disciplines. The traditions are currently developed at the Centre for Digital Humanities at the Higher School of Economics in Moscow via digital tools (Skorinkin, 2017; Bonch-Osmolovskaya and Skorinkin, 2016; Orekhov and Tolstoy, 2017; Kuzmenko and Orekhov, 2016; Fischer et al, 2017).

Another tradition related to building the National Corpus of the Russian Language[1] can be traced back to Alexei Lyapunov (Sitchinawa, 2006), another famous Russian mathematician. The point here is not that mathematics sustained and influenced all the Russian humanities (Bakhtin's famous studies can provide an opposite example[2]) or that quantitative approach as a trendy international methodology was also present in this part of the world in the 1960s-1970s but that it provided the rigor and method to the field which was disconnected from the international research methods and standards. This disconnection resulted in a dramatic difference in academic cultures.

A recent paper (Underwood 2017) discusses distant reading as a part of the digital humanities project aimed at coping with confirmation bias. Underwood shows that (social) sciences provide the 'experimental structure' and help us build research design around hypothesis, samples and results. A specific Russian feature was that research methodology of this type was provided via mathematics, linguistics, and sciences. Social science and anthropology played a minor role in the interplay of influences (Gasparov, 2016).

A major part of the current Russian digital humanities project is connected to linguistics. However, linguistics did not only provide a set of formal features and a methodology to trace a formal technique in a literary work. It was an important initial influence, a novel method to do literary studies as a part of a new scientific perspective (Tynjanov, 1971; Jarkho, 2006) in the early twentieth century. The Moscow Linguistic Circle active from 1915 to 1924 held its meetings in Roman Jacobson's flat in Moscow and its members were over 60 linguists and scholars working in text analysis and literary studies[3]. Apart from its significant international influence, the society had an important impact on how Russian scholarship developed (Akimova, Shapir, 2006; Shapir, 1996; Pil'shchikov, 2015). Its traditions were continued in applying quantitative methods to studying poetry in the second half of the twentieth century (Akimova, Shapir, 2006; Bodrova, 2017). Its influence can be traced in a highly influential approach of applying structural linguistics to interdisciplinary cultural studies at Tartu University[4] also in the second half of the twentieth century (Gasparov, 2016).

A part of current projects in Russian digital humanities are connected to this tradition. The project of creating a semantic edition of Leo Tolstoy's complete works[5] (Bonch-Osmolovskaya, 2016) includes representative and interpretive components. The edition's interpretive part works with a humanistic data model of the characters' roles in *War and Peace* validated through the digital tools of natural language processing and extracting semantic roles (Bonch-Osmolovskaya and Skorinkin, 2016), this approach also includes a classification of characters using character networks (Skorinkin, 2017). The connection of digital approaches to the previous trends of scholarship (Russian Formalism and structural interdisciplinary studies initiated by scholars from Tartu and Moscow) is explicitly proposed and maintained through the Moscow-Tartu Summer School annually organized at the Higher School of Economics in Moscow.

Quantitative approaches to studying poetry has been a path traditionally pursued by Russian mathematicians or people related to mathematics. Andrey Bely who was closely related to Nikolai Bugaev[6], one of the first chairs of the Moscow Mathematical Society, developed a quantitative approach to studying poetic rhythm in 1910 and initiated a society where scholars were taught to use statistics to study poetry (Semyonov, 2009). Andrei Kolmogorov, a famous Russian mathematician, organized a seminar and published several papers in this field in the early 1960s (Semeyonov, 2009; Kolmogorov, 2015).

The tradition has been continued via digital tools where the authors show the limitations of digital analysis (Orekhov, 2014) or integrate mapping poetry in interdisciplinary cultural studies following the Tartu tradition (Kuzmenko and Orekhov, 2016).

Russian editorial practices in the second half of the twentieth century were focused on publishing complete works of the authors from the canon of the time. Thorough editorial work was limited by the editors' attempts to combine international standards of scholarly apparatus and the requirements of the moment. Twentieth century's attempts to create scholarly editions using interpretive practices of the time (Bonch-Osmolovskaya, 2016) resulted in a current need to build new epistemological foundations for contemporary scholarly editions. Digital methods and digital scholarly standards are probably the

1 The National Corpus of the Russian Language (http://www.ruscorpora.ru) includes over 600 million words. It was published online in 2004 and developed by the linguists from the Russian Academy of Sciences (Sitchinawa, 2006).

2 See, for example (Gasparov, 2002; Sedakova, 1992), for the discussion of the difference between Bakhtin and the Russian Formalism.

3 Tynaianov and Schklovsky, famous for their contribution to Russian Formalism, were members of the Moscow Linguistic Circle (Shapir, 1996).

4 Tartu University in Estonia, a part of Russia at that time, was home for the literary studies done in the tradition of the methodology looking at formal structural features.

5 A project that is currently developed at the Higher School of Economics and Leo Tolstoy museum. Apart from using a representational mark-up in TEI standards, the project includes experiments towards an interpretive component (Bonch-Osmolovskaya, 2016; Bonch-Osmolovskaya and Skorinkin, 2016).

6 Boris Bugaev's (Andrey Bely's) relations with his father and the influence of the academic environment on Bely's development have been widely discussed in literature (see, for example, Janecek, 2015 and Giansiracusa and Vasilieva, 2017).

best possible option to cope with epistemological difficulties in the field.

While editing textual materials was complicated by interpretive practices, visual editions in the 1970s, 1980s and early 1990s were introducing new standards of metadata and data models. Their editors made an important step towards digital practices and museum computing.

The editorial practices of printed visual editions of artworks related to the standards of publishing museum images (Kizhner et al, forthcoming), the quality of images and the scholarly apparatus accompanying visual editions in the 1970s and 1980s prepared the anticipations of standards for digital publishing and placing images in a wider context via digital tools (Polulyakh, 2009; Sher, 2006).

A specific Russian feature was looking for formal (structural) components to interpret a literary work, bringing a wide interdisciplinary context to interpretation. The tradition was sustained during the twentieth century before Russian scholars turned to digital humanities. The influence of social science, gender and race studies, enlarging or changing a canon did not leave significant traces even if (when) the ideas reached the community of scholars. A current exception are projects aimed at studying the nineteenth century literary canon and future developments seeking to compare it with contemporary canons (Vdovin and Leibov, 2013). The authors propose to build a canonical corpus and study the changes using a mark-up. The idea relates to Moretty's evolutionary theories (ibid) and the Russian traditions of observing the dynamics of a formal feature that can be traced back to Boris Jarcho's papers written in the 1930s.

The paper will demonstrate, using evidence from various sources that Russian traditions of quantitative interpreting, the influence of strong mathematics and a trend of placing cultural objects within a broader context were crucial for our understanding of how digital humanities, as a quantitative methodology, developed in the country, in a different way than it did elsewhere. Understanding these alternative histories will help us understand the range of activities taking place in Digital Humanities worldwide, by looking at the social, scholarly, and cultural contexts, helping the community to navigate and bridge differences.

References

Akimova, M. and Shapir, M. (2006) 'Boris Isaakovich Jarkho and the Strategy of Research Methods for Literary Studies', in Jarkho Boris, *Research Methods for Literary Studies (*ed. Maxim Shapir*),* Moscow: Phologica. In Russian.

Allison, S., Heuser R., Jockers, M., Moretti, F., Witmore, M. (2011) *Quantitative Formalism as Experiment,* Pamphlet 1. Stanford Literary Lab.

Bonch-Osmolovskaya, A., (2016) 'Digital Edition of Leo Tolstoy Works: Contributing to Advances in Russian Literary Scholarship'. *Journal of Siberian Federal University. Humanities and Social Sciences* 7 (9), pp. 1605-1614.

Bonch-Osmolovskaya, A. and Skorinkin D. (2017) 'Text Mining War and Piece: Automatic Extraction of Character Traits from Literary Pieces', *Digital Scholarship in the Humanities,* Vol. 32, Supplement 1.

Fischer, F., Orlova, T., Skorinkin, D., Palchikov, G., and Tyshkevich, N. 'Introducing RusDraCor - A TEI-Encoded Russian Drama Corpus for the Digital Literary Studies, in *International Conference Corpus Linguistics 2017: Book of Abstracts,* June 27-30, 2017, Saint Petersburg, pp. 28-32.

Gasparov, M., (2002) 'Michael Bakhtin in the Russian Culture of the 20th Century', in *Michael Bakhtin: Pro and Contra: Mikael Bakhtin's Heritage in the Context of the World Culture* (ed. Konstantin Isupov), Vol. 2, Saint Petersburg. In Russian.

Gasparov, B., (2016) 'Between Methodological Strictness and Moral Appeal: Questions of Language and Cultural Theory in Russia', History of Humanities, vol.1, number 2.

Giansiracusa, Noah and Anastasia Vasilyeva, 'Mathematical Symbolism in a Russian Literary Masterpiece', arXiv: 170902483v1 https://arxiv.org/pdf/1709.02483.pdf

Janecek, G., (2015) *Andrey Bely: A Critical Review,* Lexington: The University Press of Kentucky.

Jarkho B., (2006) *Research Methods for Literary Studies (*ed. Maxim Shapir*),* Moscow: Phologica. In Russian.

Jockers, M., 'Microanalysis: Digital Methods and Literary History. University of Illinois Press, Urbana, IL.

Kizhner, Inna, Melissa Terras, Maxim Rumyantsev and Kristina Sycheva, 'Accessing Russian Culture Online: The scope of digitization in museums across Russia', forthcoming.

Kolmogorov, A., (2015) *Studies in Poetry,* Moscow Centre for Mathematical Education Press, 2015. In Russian.

Kuzmenko, Elisaveta and Boris Orekhov, (2016) 'Geography of Russian Poetry: Countries and Cities Inside the Poetic World', in *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 830-832.

Moretti, F., (2013) *Distant Reading,* London: Verso.

Orekhov, Boris and Fekla Tolstoy, (2017) 'Textograf: A Web Application for Manuscript Digitization', in *Digital Humanities 2017: Conference Abstracts*. McGill University & Université de Montréal, Montréal.

Pil'shchikov, I., (2015) 'The Legacy of the Moscow Linguistic Circle and the Digital Humanities Today', in *Russian Formalism and the Digital Humanities: Abstracts*, Stanford University. https://digitalhumanities.stanford.edu/russian-formalism-digital-humanities-abstracts

Polulyakh, A., (2009), Photo capturing and digital technologies in museums: following traditions, In *Proceedings of 'ICT fro Regional Development' Conference*, 5-6 February 2009, Smolensk Regional Administration, Smolensk, pp. 217-222. In Russian

Sedakova, O., (1992) 'Michael Bakhtin: An Alternative Interpretation', http://www.olgasedakova.com/Moralia/267 In Russian.

Semyonov, V., (2009) 'Methods of Statistics for Studying Russian Poetry: Andrey Bely and Andrey Kolmogorov', *Journal of Moscow State University*, series 9, number 6.
Sher, J., (2006). 'Department of Museum Informatics at the Hermitage Museum (1975 - 1985), *Information Technology for Museums,* No 2, Saint Petersburg. http://kronk.spb.ru/library/sher-yaa-2006.htm In Russian.
Sitchinawa, D., (2005) 'The National Corpus of the Russian Language: a Brief Prehistory', in *The National Corpus of the Russian Language: 2003-2005,* Moscow: Indrik, pp. 21-30. In Russian.
Skorinkin, D., (2017) 'Extracting Character Networks to Explore Literary Plot Dynamics', in Proceedings of *Dialogue: Conference on Linguistic Computing*, Moscow, 31 May - 3 June, Issue 16 (23), Vol.1, pp. 257-270.
Shapir, M. (1996) 'An editorial note to Jacobson, Roman 'The Moscow Linguistic Circle', Philologica 3. In Russian.
Stanford University, (2015) *Russian Formalism and the Digital Humanities Conference: Book of Abstracts*, Stanford University. https://digitalhumanities.stanford.edu/russian-formalism-digital-humanities-abstracts
Tynjanov, J., (1971) 'On literary revolution', in *Readings in Russian Poetics (ed.* Ladislav Matejka and Kristina Pomorska), MIT Press.
Underwood, T., (2017) 'A Genealogy of Distant Reading', Digital Humanities Quarterly, Vol.11, No 2, http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html
Uspensky, B., (1998) 'On the origin of the Tartu-Moscow School', in 'The Tartu-Moscow School: History, Memoirs, Reviews' (ed. Sergey Neklyudov), Moscow: Languages of the Russian Culture, pp. 34-44. In Russian.
Vdovin, Alexey and Leibov, Roman (2013) 'Canonical texts: Russian poetry and teaching literature in the 19th century', in *Canonical Texts: Russian Pedagogical Practice in the 19th Century and Poetical Canon,* Tartu University Press. In Russian.

Urban Art in a Digital Context: A Computer-Based Evaluation of Street Art and Graffiti Writing

Sabine Lang
sabine.lang@iwr.uni-heidelberg.de
Heidelberg Collaboratory for Image Processing, Germany

Björn Ommer
ommer@uni-heidelberg.de
Heidelberg Collaboratory for Image Processing, Germany

Summary

The paper presents how digital photographs of street art and graffiti writing[1] are analyzed with computational methods by the Computer Vision group of Heidelberg University, where an interdisciplinary collaboration between art history and computer vision is embedded since 2009. The project on urban art started in November 2017 and has the following aims: It studies the effect of digital possibilities on street art and graffiti writing regarding access, dissemination and mobility. Per definition urban art is strongly attached to a street environment, which is canvas and frame at the same time. This resulting immobility of urban art is in contrast with traditional art, where the materiality simplifies a display at alternating locations. Eventually, the paper highlights why urban art can only endure within a digital context. An example of Bristol-born artist Banksy (*1974) illustrates this: In 2015, he put up a stencil on a wall in Calais, depicting *The Raft of the Medusa* (Fig.1); only two years later, workers painted over the wall and covered Banksy's work (Samuel, 2017). The project also establishes a data collection of urban art, consisting of reproductions from *Google Arts and Culture*, other image archives and a private collection by art historian and street art scholar Ulrich Blanché. Lastly, it demonstrates how computer-based tools are used to study images with regards to form and content. In this way, patterns over time and space or artistic networks are revealed and relations between artwork and urban environment can be evaluated. Therefore, the project team utilizes an interface, which was developed within the group and allows for a visual search based on multiple image regions in large image sets.

Evaluating street art and graffiti writing

In 2009, a collaboration between art history and computer vision was established within the Computer Vision group. Thus building a bridge between the two disciplines, which resulted in the realization of works, including the creation of an interface (Bell et al., 2014), reconstruction of drawing processes (Monroy et al., 2011) or the detection and analysis of gestures in medieval manuscripts (Yarlagadda et al., 2013), (Schlecht et al., 2011), (Yarlagadda et al., 2010). The group uses deep learning algorithms and unsupervised approaches to study visual similarities on image level (Bautista et al., 2017), (Bautista et al., 2016) and whole sequences (Milbich et al., 2017). The current project utilizes existing methods to study urban art. The presence of digital image collections of urban art and computational approaches enable both large-scale evaluations and detailed studies, which has not been done by scholars so far. Previous work mainly concentrated on terminology (Blanché, 2015), social aspects (Ross, 2016) or individual artists (Blanché, 2012), (Blanché, 2010), highlighted its mediality (Glaser, 2017) and generally justified its study in art history.

The presentation highlights the influence of digitization on urban art, describes the building of a suitable dataset and its evaluation through computational methods. (1) Digital possibilities have influenced all of humanities; for urban art, however, the effect is even more profound. Most traditional artworks are mobile; artists paint on canvas or paper, which allows for easy transportation and pu-

blic display at various places. In this way, styles, content, or individual motifs spread and art reveals itself to be less bound to a specific place. On the contrary, urban art is per definition tied to the street; its meaning only fully unfolds on site. The street not only provides a canvas, but also imposes form and additional meaning. As a result, urban art is greatly ephemeral: Works are being over-painted by authorities and artists (Samuel, 2017) – as the example of Banksy showed – or buildings are torn down. In reaction, works are increasingly documented and made available online. Since its start in November 2017, the project has studied the presence of urban art on the Internet. Its visibility on different websites has impacted the community and visuality of urban art: Communication between artists and fans has increased and is simpler, motifs are disseminated faster and wider, breaking national borders and indicating a tight network. It is only through digital possibilities that urban art can be preserved and disseminated – this distinguishes it from traditional art.

(2) In order to study form and content of images with computer-based tools, the project gathers a dataset of urban art, providing metadata if available. Images are taken from *Google Arts and Culture* or *Facebook's Global Street Art*. However, the project team also received a comprehensive set of photographs, capturing urban art in various cities worldwide between 2007 and 2017. All images were taken by art historian Ulrich Blanché; this unique data enables to address new questions regarding the capturing process: How did the photographic perspective and thematic focus change over time? Does it vary for different locations? Is there a correlation between alternating perspectives and Blanché's social role? Eventually, he captured urban art first as a simple admirer, then as a student and finally as a scholar – although the first role persisted throughout time. The final image collection, including metadata, will be published and can be used by other scholars. A large number of images contain large context regions and objects, including buildings or cars. To improve performance and detection, the data was pre-processed: Around 200 images from the Blanché-dataset were annotated with bounding boxes marking artwork or context.

(3) The project studies the visuality of urban art on the basis of this image collection using computer-based methods. It aims to find recurrences and variances of a motif, ultimately not only pointing to the same but different artists. On a smaller scale, the example of Cologne street artist 'kurznachzehn' illustrates this: She uses old family photographs to create paste-ups, which she attaches to walls in various German cities. Her most recognized motif is a young girl – the artist's mother as a child. The girl appears throughout her oeuvre in a similar pose but in varying scenarios: while picking up a dandelion (Fig.2), painting or feeding a little bird (Glaser, 2017), ('kurznachzehn', 2017). In order to study image collections, the project team utilizes unsupervised methods, which have been successfully applied on other tasks and do not rely on labeled data. This is valuable, since digital reproductions of urban art rarely have information regarding artist, title or creation date – this is mainly due to the anonymity of artists and legal reasons. Reproductions are evaluated on an interface, which not only allows to search for individual but also multiple regions and thus to consider geometrical relations between artworks and urban environment. The example of the dandelion-picking girl (Fig.2) by 'kurznachzehn' illustrates this: True to the nature of her gesture, she always appears close to the ground. Underlying algorithms use a SVM-classifier trained with one positive against many negative examples. While other retrieval systems require manual tagging, the algorithm purely operates on visual qualities. Currently state-of-the-art methods are being implemented, using CNN instead of HOG-features to train the classifier. Eventually, the interface not only detects identical motifs but also variations. First tests showed promising results; the project team studied images of artworks by Brazilian street artists OsGemeos. The user was interested in a figure seen from behind and a text region to its right; (Fig.3) shows the search results for the given queries after the second training round; the bottom row includes all correctly retrieved images as selected by the user. Results can now be analyzed regarding formal and semantic similarities or variances; also, it allows to evaluate the position of the motif in relation to the urban context. Future work should study the motif of the figure seen from behind also in the context of its general appearance in art history.

Applying computational methods to urban art data has emphasized chances and benefits, not only for art history but also for computer vision. Existing algorithms have been tested on challenging data and proofed their efficiency. However, working with urban art data has also highlighted some challenges: Collections are biased towards certain time periods, nationalities and dominated by works of popular artists. The new dataset, established within the project, is therefore extremely valuable. First tests, although overall successful, showed that algorithms are challenged by a dominating background, imaging mode (perspective) and the size of artworks. To remedy the latter, the project team decided to annotate part of the Blanché dataset with bounding boxes, which improved detection.

Conclusion

The presentation consists of two parts: A theoretical basis will be established in the first, discussing the influence of digital possibilities on aspects, such as mobility, access or dissemination, while the dataset will be introduced in the second half, which also includes presentations of search results on the interface. The project team aims to further establish urban art as a profound research topic in academia, point to new research questions and possible challenges when working with urban art data. Most importantly, the presentation emphasizes the chances offered by computer-based methods to study urban art in detail and on large-scale. (Words: 1485)

List of Illustrations

Fig.1: Banksy, *The Raft of the Medusa*, Calais, 2015

Fig.2: 'kurznachzehn', *Girl picking Dandelion*, Dusseldorf, 2013

Fig.3: Search results for figure and text to its right on interface; image collection of Brazilian street artists OsGemeos

References

Bautista, M., Sanakoyeu, A. and Ommer, B. (2017): Deep Unsupervised Similarity Learning Using Partially Ordered Sets. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. CVPR.

Bautista, M., Sanakoyeu, A., Sutter, E. and Ommer, B. (2016): CliqueCNN: Deep Unsupervised Exemplar Learning. *Proceedings of the Conference on Advances in Neural Information Processing Systems*. NIPS.

Bell, P., Ommer, B. and Takami, M. (2014): An Approach to Large Scale Interactive Retrieval of Cultural Heritage. *Eurographics Workshop on Graphics and Cultural Heritage*. The Eurographics Association.

Blanché, U. (2015): Street Art and related terms – discussion and attempt of a definition. *Street & Urban Creativity Scientific Journal. Methodologies for Research,* (1), pp. 32-40.

Blanché, U. (2012): *Konsumkunst. Kultur und Kommerz bei Banksy und Damien Hirst.* Bielefeld: Transcript Verlag.

Blanché, U. (2010): *Something to s(pr)ay: Der Street Artivist Banksy. Eine kunstwissenschaftliche Untersuchung.* Marburg: Tectum Verlag.

Glaser, K. (2017): *Street Art und neue Medien. Akteure, Praktiken,* Ästhetiken. Bielefeld: Transcript Verlag.

Milbich, T., Bautista, M., Sutter, E. and Ommer, B. (2017): Unsupervised Video Understanding by Reconciliation of Posture Similarities. *Proceedings of the IEEE International Conference on Computer Vision*. ICCV.

¿Metodologías en Crisis? Tesis 2.0 a través de la Etnografía de lo Digital

Domingo Manuel Lechón Gómez
domingo@sursiendo.com
Doctorado de Ecosur, Mexico; Sursiendo, Mexico)

La presente propuesta está basada en la experiencia y las reflexiones que han ido surgiendo en el transcurso del trabajo de investigación de doctorado todavía en proceso. Con el título "La disputa de Internet. Análisis de los marcos de acción colectiva del activismo tecnológico en México", este estudio busca describir, analizar e interpretar cuáles son los problemas sociotécnicos que diagnostican los propios actores sociales, quiénes son los causantes de esos problemas, quiénes componen un "nosotros" entre los movimientos ciberactivistas, quién es la audiencia a la que va dirigida la acción colectiva y cuál es la propuesta sociopolítica que plantean para resolver el problema. Todo ello desde la propuesta de los marcos de acción colectiva y la metodología basada en la etnografía de lo digital.

La participación en el Congreso tiene la intención de proponer preguntas para reflexionar e iniciar diálogos necesarios sobre los cambios epistemológicos y metodológicos que pueden darse con las investigaciones con/desde/en Internet, y las cuestiones éticas que subyacen en las ciencias sociales y las humanidades relacionadas con las redes digitales.

Partiendo de que Internet se inserta en un contexto histórico de profundos cambios sociales, a la vez que es uno de los dispositivos que potencia esas cambios en las sociedades actuales, cambios epistemológicos e incluso ontológicos. Como explicaba Priani (2012) desde hace años que se está dando un "desplazamiento del proyecto ilustrado", de la Modernidad, y con ello se ponen en cuestión el saber científico imperante y las formas de construir conocimiento adscritas a él. Esas transformaciones se vienen fraguando desde los años 60, y los llamados nuevos movimientos sociales dan cuenta de ello, impugnando al sistema desde el ecologismo, el feminismo, el antirracismo, el anticolonialismo, el antimilitarismo, etc. Los tecnoactivistas, que recogen enseñanzas de esos movimientos previos, de sus principios, acciones y propuestas, ahora actúan en el terreno de Internet incorporando cuestionamientos y evoluciones desde y hacia las ciencias.

Así, estos y otros cambios se están produciendo dentro mismo de las academias, como presentó Wallerstein (1996) en el Informe de la Comisión Gulbenkian, la hibridación de disciplinas es un hecho (necesario). Wallerstein y colaboradores invitan a explorar y dar palabra a lo que está ocurriendo en la actualidad en el campo de la ciencia y a idear las medidas institucionales que lo asienten y hagan operativo, para que las ciencias "sean más verdaderamente pluralistas y universales" (Wallerstein, 2006, 101). ¿Internet puede que haga más factibles esas transformaciones?

Uno de los aspectos que entran en debate ahí son las posiciones de objetividad y subjetividad, que por ejemplo Donna Haraway impugnó también por esas fechas con el "conocimiento situado" (1996). Con este concepto se pone en cuestión la construcción de conocimiento "desde afuera", problematiza aspectos tales como la influencia de la situación de "encuentro con el otro" en el investigador y los aspectos sensibles de la relación social que se plantea con los sujetos entrevistados u observados; y al abordar un hecho social prioriza la construcción conjunta de conocimiento entre el investigador y quienes devienen su objeto de estudio.

Con ello se puede vislumbrar que está en crisis el paradigma científico moderno, lo cual puede ser una buena oportunidad para debatir y trazar nuevos itinerarios.

En los estudios sociales sobre algún aspecto de las Tecnologías de la Información y la comunicación (TIC), o de Internet en concreto, también han ido cambiando los enfoques. Como por ejemplo es lo que Gálvez y colegas apuntaban: "El determinismo, ya sea tecnológico o social, ha marcado gran parte de las aproximaciones que se han hecho desde las ciencias sociales al estudio de la tecnología" (Gálvez y otros, 2003; p1). Ya cada vez más se mira desde una posición sociotécnica, tanto lo social como lo técnico se influyen mutuamente, y es necesario que cualquier investigación se aproxime desde ahí.

Por ello, entrando en temas metodológicos, por ejemplo, la etnografía de lo digital o virtual, que en un principio se asumió como el estudio de la práctica online, en la actualidad, lo que prevalece es un enfoque holístico en el que se superponen los campos online y offline (Hine, 2004). En definitiva, la etnografía virtual es un híbrido, en cuanto apunta a grupos en línea relacionados con situaciones fuera de línea.

Desde los movimientos conectados se da cuenta de otras formas de mirar Internet; por ejemplo, Carmona Jiménez (2011) apunta que junto a la noción de dispositivo sociotécnico que sume en cierta medida a Internet como un artefacto (socio-facto), el ciberespacio además permite considerarlo como un "lugar" en el que se gesta cultura (Hine, 2004) y proporciona una forma de "habitar", por lo que en verdad es un "espacio antropológico", pues hay una construcción simbólica del espacio. Estar en terreno exige que el investigador se convierta en usuario y su "observación participante" significa participar e interactuar (Carmona Jiménez, 2011).

Para Estalella y colaboradores, la etnografía de lo digital es "la adaptación de la metodología etnográfica a las propiedades de los fenómenos que se desarrollan a través de lo digital implica repensar muchos de sus conceptos básicos y planteamientos metodológicos" (Estalella y otros, 2006; p2).

Además, al tratarse de una investigación que se introduce en el mundo tecnoactivista en México, es impor-

tante considerar factores éticos, como el tratamiento de los anonimatos, el uso de programas de análisis de datos de código libre, las licencias de publicación, etc.

La participación en el Congreso puede aportar esas reflexiones que busquen nuevos itinerarios para iniciar diálogos sobre estas novedades, con sus dificultades y sus retos, para la reflexión sobre las ciencias sociales y humanísticas en la sociedad-red.

Referencias

Carmona Jiménez, J. (2011) Tensiones de la etnografía virtual: teoría, metodología y ética en el estudio de la comunicación mediada por computador. *Revista F@ro* No 13. Facultad de Ciencias Sociales, Universidad de Playa Ancha, Chile. En línea: http://web.upla.cl/revistafaro/n13/art03.htm

Estalella, A. (2007) *Etnografías de lo digital. borrador*. En línea: http://www.prototyping.es/wp-content/uploads/2014/05/Estalella_Etnografias-de-lo-Digital-borrador-parcial.pdf

Estalella, A.; Ardévol, E.; Domínguez, D.; y Gómez Cruz, E. (2006) Etnografías de lo digital, Actas del Grupo de trabajo, *III Congreso Online - Observatorio para la Cibersociedad* Del 20/11/2006 - 03/12/2006. En línea: http://mediaccions.net/wp-content/uploads/etnografias-digital-actas.pdf

Gálvez, A.M.; Ardèvol, E.; Nuñez, F. y González, I. (2003). "Los espacios de interacción virtual como dispositivos sociotécnicos". Comunicación presentada para el *VIII Congreso Nacional de Psicología Social*. Torremolinos, Málaga, Abril 2003.

Haraway, D. (1995) *Ciencia, cyborgs y mujeres. La reinvención de la naturaleza*. Madrid: Cátedra.

Hine, C.. (2000). Etnografía virtual. UOC, Barcelona.

Laraña, E. (1999). *La construcción de los movimientos sociales*. Madrid, Alianza, 1999.

Melucci, A. (1994) ¿Qué hay de nuevo en los nuevos movimientos sociales? En *Los nuevos movimientos sociales: de la ideología a la identidad* / coord. por Joseph Gusfield, Enrique Laraña Rodríguez-Cabello. págs. 119-150.

Mosquera Villegas, M. A. (2008) De la Etnografía antropológica a la Etnografía virtual. Estudio de las relaciones sociales mediadas por Internet. Fermentum. *Revista Venezolana de Sociología y Antropología*, vol. 18, núm. 53, septiembrediciembre, 2008, pp. 532-549 Universidad de los Andes Mérida, Venezuela.

Priani, E. (2012) Molinos o gigantes. Cambio y nuevas tecnologías en las humanidades. *Revista Virtualis* No. 5 27 de junio 2012. Publicado por Centro de Estudios sobre Internet y la Sociedad y el Tecnológico de Monterrey. p 9 a 12

Ruiz Méndez, M. R. y Aguirre Aguilar, G. (2015) Etnografía virtual, un acercamiento al método y a sus aplicaciones. En *Estudios sobre las Culturas Contemporáneas*. Época III. Vol. XXI. Número 41, Colima, verano 2015, pp. 67-96.

Tarrow, S. (1997). *El poder en movimiento. Los movimiento sociales, la acción colectiva y la política*. (H.b. Resines, Trad.) Madrid, España: Alianza.

Wallerstein, I. (ed.) (1996). *Abrir las ciencias sociales, Comisión Gulbenkian para la reestructuración de las ciencias sociales El Mundo del Siglo XXI*. México, ed. Siglo XXI.

Wallerstein, I. (2005). *Análisis de Sistema-Mundo Una introducción*. Madrid, España: Siglo XXI.

Hashtags contra el acoso: The dynamics of gender violence discourse on Twitter

Rhian Elizabeth Lewis
rhian.lewis@mail.mcgill.ca
McGill University, Canada

Introduction

The spring of 2016 has become known as the "Primavera Violeta" ("Purple Spring"), a period that saw the emergence of new digital activist networks tackling gendered and sexual violence in Latin America. Of the hashtags generated by these movements, few gained the public recognition and "celebrity status" of #MiPrimerAcoso ("My First Harassment" or "My First Abuse"), a hashtag that asked users to publically share their first experiences of sexual violence. On April 23, 2016, women in Mexico and across Latin America shared their stories via their personal Twitter accounts in response to a request tweeted by journalist Catalina Ruiz Navarro of the pop-feminism collective (e)stereotipas: "¿Cuándo y cómo fue tu primer acoso? Hoy a partir de las 2pmMX usando el hashtag #MiPrimerAcoso. Todas tenemos una historia, ¡levanta la voz!" *(When and how did your first acoso happen? Today from 2pm on, use the hashtag #MiPrimerAcoso. We all have a story, raise your voice!)*

Figure 1: A typical #MiPrimerAcoso tweet. In English: I was eleven years old and a man passed on a bicycle and grabbed my breast. A woman in the street blamed me for wearing that blouse.

After its initial launch, #MiPrimerAcoso spread rapidly throughout Mexico and quickly became a trending topic across Latin America. This analysis investigates the ways that Twitter users– activists, laypersons, public figures–

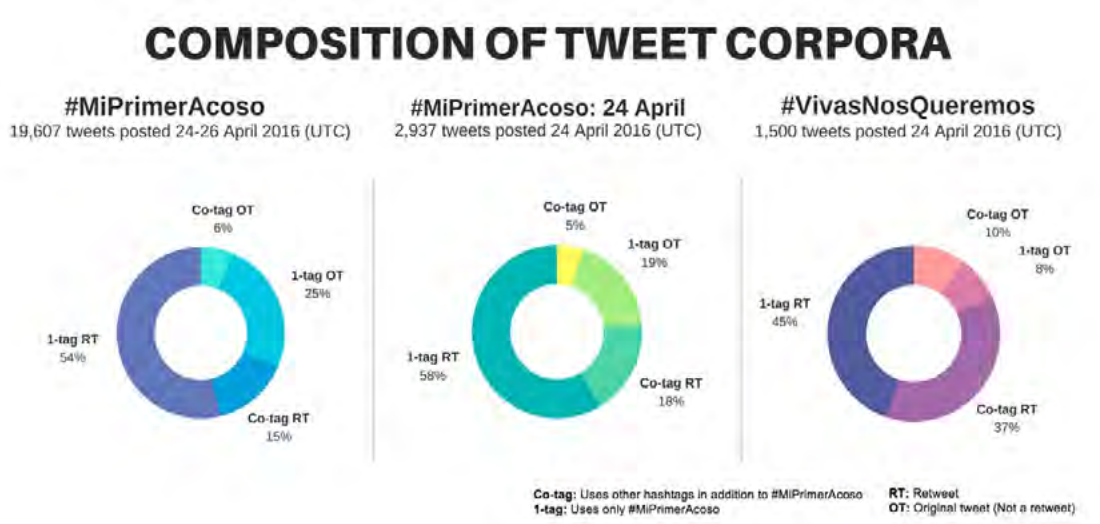
use hashtags to talk about trauma, paying special attention to the ways that quantifiable modes of Twitter engagement point to more complex affective experiences.

Methods

This project undertakes both qualitative and quantitative analyses of tweets posted using #MiPrimerAcoso in order to examine the key actors, contexts, and conditions that emerged from the hashtag's narrative premise. For the initial assessments, this analysis uses the #MiPrimerAcoso corpus collected and published by media company Lo Que Sigue. To provide a point of comparison, this project also analyzes a collection of tweets posted using another Primavera Violeta hashtag– #VivasNosQueremos ("We Want to Live")– whose corpus was collected and published by Lo Que Sigue at the same time as the #MiPrimerAcoso corpus.

Affective (and Effective) Tweeting

Hashtag dialogues serve to construct and re-construct bridges between different streams of dialogue within movements, between movement collaborators and stakeholders, and between activists, political powers, and the general public. To illustrate some of the preliminary findings of this exploration, I evaluate the prevalence of retweets and multiple-hashtag use (or "co-tagging") in the #MiPrimerAcoso corpus and another corpus published by #LoQueSigue of tweets posted using #VivasNosQueremos. Throughout this paper, I call upon Papacharissi's (2015) work on the affective properties of Twitter dialogues to further illustrate the forms of personal and political affect that drove the trans-national trajectory of #MiPrimerAcoso.



Although #MiPrimerAcoso is entangled with other Twitter dialogues on gender violence, it "stands alone" more often than one of its closest peers, and is less frequently retweeted and co-tagged. Here, I find that these concrete metrics summarize diverse modes of engagement: retweeting another user's personal story of violence is necessarily a different act than retweeting a popular news story about the hashtag. However, these metrics do demonstrate the ways in which use characteristics reflect the discursive mandate of a hashtag. Engagement with #MiPrimerAcoso might include reading, listening, creating original content, rebroadcasting, or responding to the content of other users within the affective public generated by the hashtag. This diverse set of practices allows Twitter users to "tune into an issue or a particular problem of the times but also to affectively attune with it, that is, to develop a sense for their own place within this particular structure of feeling" (Papacharissi 118). The Twitter users who tweeted their experiences of violence undertook a delegated task of content creation in response to the prompts posted by Ruíz Navarro. This guiding of the discussion allowed Twitter users to act and to *feel* using a pre-constructed response frame. By asking users to share the how and when of their first acoso, users tasked with personifying the political and *making it about themselves*. By focusing on a tweet structure that outlines an individually expressive personal action frame through the medium of shared experience, #MiPrimerAcoso allows its users to make "small and fitful contributions" (Bennett and Segerberg 2011) to a cause while feeling a profound sense of identification with the movement.

If we want to understand what it is people want from digital activism, #MiPrimerAcoso offers captivating insights regarding our need to see ourselves within online political movements. The secret of #MiPrimerAcoso's handling of collective and individual resonance lies in its personalization and generalizability: although the hashtag calls on a specific category of experience, it is sufficiently broad that many interpretations of *acoso* fit the bill, and many users were able to affiliate with the has-

htag without necessarily sharing a personal story of sexual violence. As Papacharissi (2015) notes, the use of hashtags as "open" signifiers allows various publics to affiliate with a movement and "fill in" the open hashtag with their own desired meanings. Women were able to link their own experiences of sexual violence to the individual narratives that had already been shared using the hashtag #MiPrimerAcoso. What, then, of those who did not contribute their original narratives to the library of primer acosos, but instead chose to respond or rebroadcast existing #MiPrimerAcoso content? In responding to a tweet, users may amplify, stifle, or otherwise alter the public life of the digital *acoso*. Although Papacharissi and others have linked the act of retweeting to the expression of solidarity with a movement, this conclusion may prove reductive in the context of #MiPrimerAcoso. However, solidarity does not adequately summarize the act of rebroadcasting another person's *acoso*: it is an expansion of the tweet's intangible audience of ethical witnesses to the tweeted acoso, a "re-telling" of scene of violence. Like any other hashtag, #MiPrimerAcoso needed to meet specific communicative and technical (in the case of Twitter) requirements in order to maximize its "reach" and extend beyond the core audience of (e)stereotipas. Referring to the act of retweeting, Papacharissi argues that refrains reinforce affect (Papacharissi 2015). By posting tweets tagged #MiPrimerAcoso, users spread the affective and contextual implications of the hashtag to their own Twitter audiences: those in digital "earshot" of their tweets. Similarly, the authors of original #MiPrimerAcoso tweets were also invited to act as amplifiers of the larger movement by adding their story to a collaborative, polyvocal narrative of lived violence.

Conclusions

In our study of digital movements, the use of the hashtag is the tip of the iceberg in comparison to the forms of knowledge, feeling, and understanding that emerge from these affective discourses. The results of this research have also suggested that conventional Twitter analysis methods may not adequately assess the affective clout of digital dialogues. For this reason, this analysis has strived to use the concrete metrics of the #MiPrimerAcoso data as guide to direct a "closer" reading of the narrative attributes of the tweets. When examining Twitter data, we must strive to expand the possibilities behind a simple, quantifiable act such as a retweet, and understand the hashtag as a point of contact between the user and digito-phenomenological processes of which we are largely unaware. Of course, there are key characteristics of the hashtag itself that are crucial to our understanding: its connectivity, for example, or its capacity to understand individual content as part of a larger dialogue. The hashtag is a departure point: an entity that gives rise to visible manifestations of trauma, digital acts of vulnerability and moments of personal catharsis, responses of support, condemnation, or indifference.

We should consider the tweet, then, as the execution of a series of digital actions, but also as the manifestation of a confluence of contacts between the ontological and phenomenological worlds of Twitter. To better assess these intangible qualities of Twitter data, we can listen to the testimonies of #MiPrimerAcoso authors, and pay attention to the strategies they employ to construct the *acoso* in relation to their present selves, the ways in which they reflect on the act of tweeting the *acoso* in front of an intangible digital audience. Here, I want to emphasize the diversity of experiences that users bring to the discursive space of Twitter, and the need to pay attention to the varied motivations that drive Twitter users to participate in social campaigns. These experiences do not easily reduce themselves to quantitative metrics, but we can search for their traces in the textual manifestations of our digital activity: the stories we tell, the words we use, the affective investments that we make as observers and participants.

References

Bennett, W. Lance, and Alexandra Segerberg. "The logic of connective action: Digital media and the personalization of contentious politics." *Information, Communication & Society* 15.5 (2012): 739-768.

Gerbaudo, P. (2014) The persistence of collectivity in digital protest,*Information, Communication & Society*, 17:2, 264-268, DOI: 10.1080/1369118X.2013.868504

Lo Que Sigue TV (2016). Tuits de #MiPrimerAcoso disponible en "table_5d787653". Database available on Carto. https://lqs.carto.com/tables/table_5d787653/public

Papacharissi, Z. (2015) *Affective publics: Sentiment, technology, and politics.* Oxford University Press.

Novas faces da arte política: ações coletivas e ativismos em realidade aumentada

Daniela Torres Lima
danielatorreslima@yahoo.com.br
Universidade Federal de Juiz de Fora, Brazil

Da modernidade até aquilo que Lipovetsky e Serroy (2010) nomearam como hipermodernidade, o sistema de construção de imagens e recepção delas passaram por diferentes fases e evoluções, transformando radicalmente as relações individuais e interações sociais. Para Pierre Lévy (2010), a partir da multiplicação de dispositivos móveis e suas funções comunicativas a nível global, as relações sociais deslocaram-se de contextos locais de interação e foram rearranjadas em extensões indefinidas, baseadas

em uma noção de espaço-tempo diferenciados. Essas novas conexões aglutinam indivíduos em afinidades de interesses e conhecimentos, propiciando um processo de cooperação e troca entre eles que independe de proximidades geográficas, mas que evidentemente constitui uma nova forma de organização social (Lévy, 2010:134).

Dessa nova interpretação de lugar, sem fronteiras geográficas e hiperconectado, emergem outras abordagens perceptivas sobre ser cidadão e de atuação sobre esses espaços e assuntos cada vez mais comuns. Percebemos, então, a emergência de formas de engajamento político que ultrapassam a prática do ativismo em partidos, sindicatos e movimentos sociais locais, assumindo também no ciberespaço e na utilização de novas mídias uma postura ativista, tomando tais sistemas como suporte para suas práticas. Na atualidade, notamos que os atos de ativismo e militância tem se apoiado em tecnologias cada vez mais avançadas, não apenas para usufruir dessa fácil e acessível forma de divulgação de informações, mas também explorando o potencial político de pressão e engajamento com a realidade pelo permeio de ambos espaços que mediam interações sociais e culturais atualmente: o virtual e o físico.

Nesse cenário, a tecnologia de realidade aumentada se destaca ao exigir a participação contínua de um interveniente, promovendo a interação entre objetos virtuais tridimensionais e usuários reais, interatuando em tempo-real no espaço. Para Gonçalves (2006), o uso de imagens atraentes e passíveis de interação e manipulação tem a função de mobilizar e despertar o interesse para esse gênero de iniciativas, criando um entusiasmo para o engajamento político, levantando questões e discutindo-as de forma crítica e lúdica ao mesmo tempo. Portanto, a arte tem papel fundamental nessas ações, uma vez que age como um fator de atração e reflexão para questões socioculturais relevantes (Gonçalves, 2006:12).

De acordo com Mark Skwarek (2017), artista multimídia que tem trabalhado na construção e articulação de atos de resistência em redes, a tecnologia de realidade aumentada ganhou notoriedade neste século na mediação de narrativas elaboradas permitindo que artistas usufruam da potência de visualização e comunicação digital para alcançar propósitos reflexivos e experiências estéticas contemporâneas. Segundo o autor, os primeiros ativistas a utilizarem-se desta tecnologia foram inspirados pelo trabalho de *culture jammers* e artistas de grafite dos anos 80, que apoiados em uma semiótica de guerrilha, colocavam em voga técnicas de anti-consumismo e anti-capitalismo a fim de romper ou subverter a cultura *mainstream*. Esses grupos criavam grosseiramente sobreposições e intervenções sem permissão de um estabelecimento, desafiando a noção do espaço público e privado ao serem aplicadas sobre muros, portas de instituições ou no logotipo de corporações (Skwarek, 2014: 17). Na realidade aumentada, entretanto, essas sobreposições ocorrem de forma virtual em ações interativas através da digitalização de um código *Quick Response* (QR) ou de reconhecimento de um objeto pré-codificado via câmera de celular, disparando elementos virtuais tridimensionais (frases, desenhos, vídeos e pichações) que aparecem sobrepostos ao mundo cotidiano através da tela do telefone. Enquanto visualiza elementos virtuais sendo sobrepostos à 'realidade', o usuário tem a possibilidade de registrar sua interação através da câmera fotográfica ou de um *print screen* disponibilizado em algumas aplicações. A etapa seguinte, apesar de não ser o único modo de conferir credibilidade a um movimento visto que a maioria dos aplicativos interativos computam o número de participantes e os identifica por localização, é a divulgação voluntária e em rede dos materiais visuais obtidos durante a interação. Talvez este seja o ponto crucial desses atos de ativismo coletivo uma vez que coloca maior agência e destaque nas mãos de pessoas comuns que se prontificam e se declararam como apoiadoras de um movimento coletivo, expondo suas identidades em uma espécie de fragmentação da cobertura midiática, além de ramificarem um ponto inicial de protesto para diferentes comunidades e pontos do globo de forma instantânea, o que de alguma forma aumentam o reconhecimento de lutas que tem se tornado cada vez mais universais.

Como um exemplo notável dessa nova forma de manifestar-se politicamente encontra-se o protesto *Occupy Wall Street*, organizado através de redes de internet em 2011 nos Estados Unidos, e que contou com a colaboração de artistas e programadores de todo o mundo para seu sucesso. Esse movimento teve estopim na cidade de Nova Iorque em um protesto que reivindicava o fim da desigualdade social e econômica, a corrupção e a grande influência de empresas sobre o governo, particularmente do setor de serviços e o financeiro. Os manifestantes não tiveram permissão para protestar em Wall Street, onde somente parte da calçada estava acessível ao público, sob constante vigilância da polícia. Foi a partir deste impedimento que ativistas de 82 países se organizaram para criar o movimento virtual sobre a hashtag #arOCCUPYWALLSTREET, visando que utilização de aplicativos de realidade aumentada levasse o protesto ao coração do distrito financeiro e desse voz aos manifestantes barrados (Fig.1)

Fig 1. *AR Occupy Wall Street*, em 2011. Imagens retiradas do site do evento.

Já num propósito de discussões de gênero e articulações feministas, *The Whole Story Project* propunha ocupações simbólicas dos espaços a partir da reformulação de imagens femininas, visando discutir a presença das mulheres na sociedade, inclusive na história, atuando num devir entre arte e ativismo. O projeto foi inspirado na campanha *Monumental Women Campaign*, que teve início em 2016 a partir de uma constatação que apenas 7,5% das 5193 estátuas espalhadas pela cidade de Nova York retratavam mulheres. Desde então este projeto passou a ser articulado por artistas com o objetivo de a arrecadar fundos para colocar as primeiras estátuas em homenagem à história das mulheres no Central Park de Nova York, que além de contribuírem para a maior representatividade feminina na sociedade, promoveria a conscientização sobre as contribuições das mulheres para a história compartilhada. Foi no início de 2017, durante as preparações para a maior marcha feminista já realizada no mundo, a *Women's March*, que artistas multimídias se apropriaram do projeto físico para criarem um aplicativo de realidade aumentada que colocasse diferentes mulheres em monumentos públicos, relatando suas histórias de luta e transformações sobre a cidade. O aplicativo *The Whole Story*, embora virtual, pretendia fazer ondas no mundo físico, chamando a atenção para que mulheres comuns se inspirem com a história de outras grandes realizadoras, empoderando-as politicamente em sua comunidade, além de permitir ao público recuperar a narrativa histórica das grandes cidades. (Fig.2)

Diante desses exemplos, e apoiados nos estudos de Ricardo Rosas (2003), podemos dizer que tais movimentos tratam-se de novas formas de ativismo uma vez que se articulam como modos de resistências temporárias e nômades, baseadas em ações coletivas de intervenção em espaços públicos que se fundamentam pelas redes virtuais ou no uso de mídias diversas (Rosas, 2003). De um modo pragmático, esses ativismos em Realidade Aumentada passam a abranger diversas províncias de significado e experimentar universos múltiplos em forma de manifestações, evidenciando que os métodos do passado podem estar se tornando menos efetivos e se faz necessário a reestruturação de um modelo que condiz com uma realidade hipermoderna.

Sobre antigos modelos de manifestação política e suas funcionalidades, o coletivo americano *Critical Art Ensemble* (1996) argumenta sobre a ocupação de espaços públicos em protesto da atualidade. Para o grupo, embora alguns dos monumentos do poder permaneçam fixos, ostensivamente presentes em locais estáveis, o poder já não reside nesses locais, já que a ordem e o controle agora se movem livremente. O que é proposto em seus manifestos é a apropriação de algo que tenha um valor comum na atualidade, tanto para as instituições de poder as quais se combate quanto para a sociedade de forma geral.

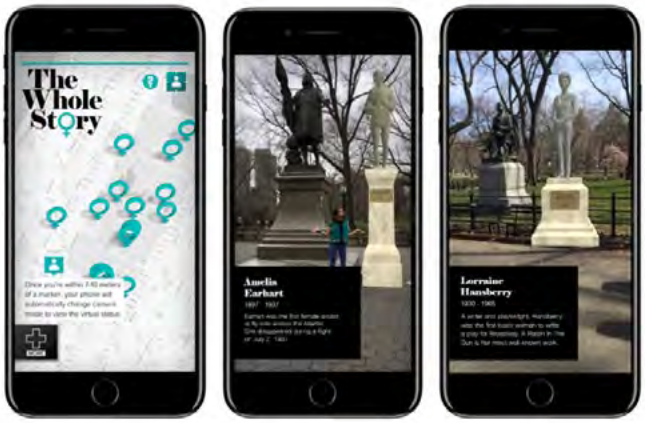


Fig. 2. RA em discursos de empoderamento feminino (2017). Imagem retirada do site do projeto.

Nesse sentido, a ocupação e uso das novas mídias e do espaço político que elas fornecem mostram-se como uma abordagem diferenciada sobre as competências da comunicação, flexibilizando seus usos comuns e trazendo a reflexão sobre as potencialidade de um trabalho colaborativo. Dentro do contexto cultural-midiático existente, isso seria uma forma de combater um sistema através de seus próprios meios, sendo, portanto, uma estratégia válida de enfrentamento, e mostrando um caminho possível para inversões temporárias no fluxo do poder.

Ao tentar compreender as novas formas de exercício político na contemporaneidade, esbarramos com uma grande gama de ações que têm sido realizadas através de tecnologia de Realidade Aumentada. A sobreposição de camadas virtuais e moldáveis às referências visuais do mundo físico tem sido usada como forma de atração e engajamento para atos de ativismos na tentativa de discutir sistemas políticos, econômicos e ambientais da forma mais atual possível. Essas novas formas de envolvimento político coletivo são reflexos de tempos onde as mídias móveis e a hiperconectividade reoganizaram nossas relações com o mundo, motivando indivíduos a ingressarem em novas experiências estéticas e reflexões singulares através da fusão entre arte, tecnologia e política.

References

Critical Art Ensemble (1996), "Electronic civil desobedience and other unpopular ideas". Disponível em: http://critical-art.net/books/ecd/ (acesso em 13 janeiro de 2018).

Gonçalvez, F. (2006). *Resistência nômade: arte, colaboração e novas formas de ativismo na Rede.* Rio de Janeiro: Revista Compós, p. 85-90.

Lemos, A. (2006). *Ciberespaço e tecnologias móveis: processos de territorialização e desterritorialização na cibercultura.*Porto Alegre: Sulina.

Lèvy, P. (2010). *Ciberespaço.*São Paulo: Editora*34*, p.133-134.

Lipovetsky, G., e Serroy, J. (2010). *O ecrã global: cultura midiática e cinema na era hipermoderna.*Lisboa: Edições 70.

Rosas, R. (2003). Que venha a mídia tática. *In: Rizoma.net.* Disponível em: http://www.rizoma.net/interna.php?id=174&secao=intervencao/ (Acesso 17 de dezembro de 2017).

Skwarek, M. (2017). Augmented reality activism. *InAugmented Reality Art.* Berlim: Springer, p. 3-29.

The Whole Story Project (2017). Website oficial. Disponível em: https://thewholestoryproject.com (Acesso em: 10 nov. 2017).

Modeling the Fragmented Archive: A Missing Data Case Study from Provenance Research

Matthew Lincoln
mlincoln@getty.edu
Getty Research Institute, United States of America

Sandra van Ginhoven
svanginhoven@getty.edu
Getty Research Institute, United States of America

Historians grapple with missing information constantly. While there are many statistical tools for gauging the impact of missing source data on quantitative results and conclusions, DH researchers have rarely deployed these tools in their work. This paper presents one implementation of data imputation used in the study of the New York City art dealer M. Knoedler & Co.Demonstrating the significant contribution imputation had on our study and its conclusions, this paper will discuss specific, practical rhetorical strategies, including static and interactive visualization, for explaining this methodology to an audience that does not specialize in quantitative methods.

Missing Data in the Digital Humanities

Miriam Posner has argued that both data structures and rhetorical conventions for computing with missing information, uncertainty, and highly subjective/viewpoint-contingent knowledge remains a key desideratum of DH scholarship. (Posner, 2015) Several attempts have been made by the information science community to express uncertainty in a structured format, ranging from generalized ontologies for reasoning in a networked world (Laskey et al., 2008), as well as more specific projects such as the *Topotime* library for reasoning about temporal uncertainty. (Grossner and Meeks, 2013)

However, many DH projects have sidestepped these approaches. Matthew Jockers, for example, has asserted that the availability of full text is becoming such that literary historians will no longer have to be concerned about drawing a representative sample. (Jockers, 2013: 7–8) More commonly, though, scholars have attempted to carefully constrain their conclusions based on what they know to be missing from their data. Theorizing and documenting the difference between one's data set and one's subject has become a genre of DH work unto itself. Katherine Bode has argued that such documented datasets should be understood as *the* object of DH inquiry. (Bode, 2017)

While statistical literature on the problem of missing-data imputation is quite mature (see Gelman and Hill, 2006 for a valuable review), few DH research projects have openly explored the use of statistical procedures for reckoning with missing data, nor have they grappled with how to theorize and present such imputation in the context of their home disciplines. (An important exception includes Brosens et al., 2016) Bode, for one, has explicitly rejected such approaches, arguing (without specific evidence) that quantitative error assessment cannot be usefully performed in historical analysis. (Bode, 2017: 101)

We argue that such methods should be *central* to data-based digital humanities practice. Simulation and imputation allow us to realize multiple, sometimes conflicting assumptions about the nature of missing data. In doing so, these affordances allow us to evaluate how certain assertions may propagate their assumptions through the transformations we perform on our sources.

Case Study: Modeling M. Knoedler & Co.'s Business from Sparse Stock Books

As part of a research initiative into data-based approaches to the study of the art market, we are investigating the changing strategies of the New York City art dealer M. Knoedler & Co., whose stock books have been encoded by the Getty Research Institute (http://www.getty.edu/research/tools/provenance/search.html). Based on these transaction data, we have built a predictive model that classifies whether a given artwork would result in a profit or a loss, using a host of variables such as how much money the work of art originally cost, the genre and size of the work, their prior relationships with buyers and sellers, and the time the work remained in stock before it was sold, to name but a few. Predictive modeling illuminates complex relationships between these variables and highlights unusual sales for further archival research.

As informative as these stock books are, however, many of their notations are partial: Knoedler's staff may have neglected to record the date of sale; there may be a listed purchase without a description of the type of work (i.e.portrait, landscape, etc.); the identity of the buyer, and whether they were a first-time customer or a well-known shopper, may also have gone unrecorded. Because our random-forest-based model (Liaw and Wiener, 2002) does not allow missing values, we must either discard incomplete records (and thus eliminate nearly half of the records from consideration), or we must find ways to impute values for our predictor variables.

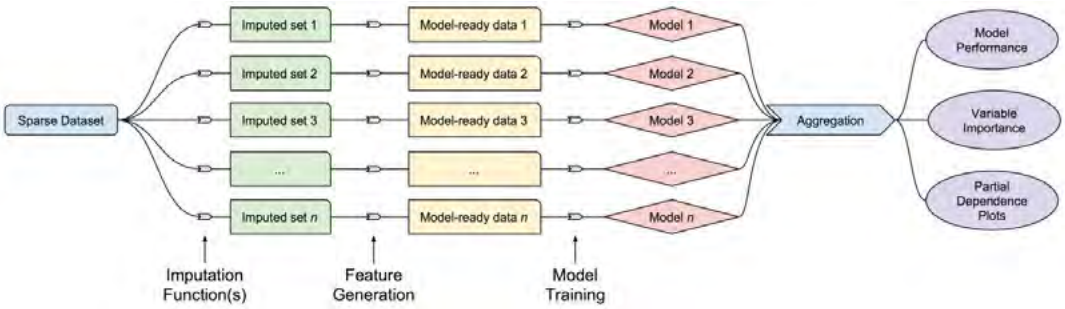


Figure 1 Schematic workflow for imputing missing data, producing derivative features, building models, and then aggregating statistics from the multiple models produced.

While it is impossible to perfectly reconstruct these missing records, it *is* possible to operationalize educated guesses about their possible values. (Figure 1) Purchase and sale dates for artworks, for example, can be predicted with some accuracy based on their location in the roughly-chronological series of stock books. Likewise, unknown genres can also be imputed as a function of the existing distribution of genres across stock books, with, e.g.abstract paintings being far less common in the pre-20th c. books than in the later ones. By defining an informed range of possibilities for these missing data, and then sampling from that range, we can produce ensemble models and results that provide a more nuanced representation.

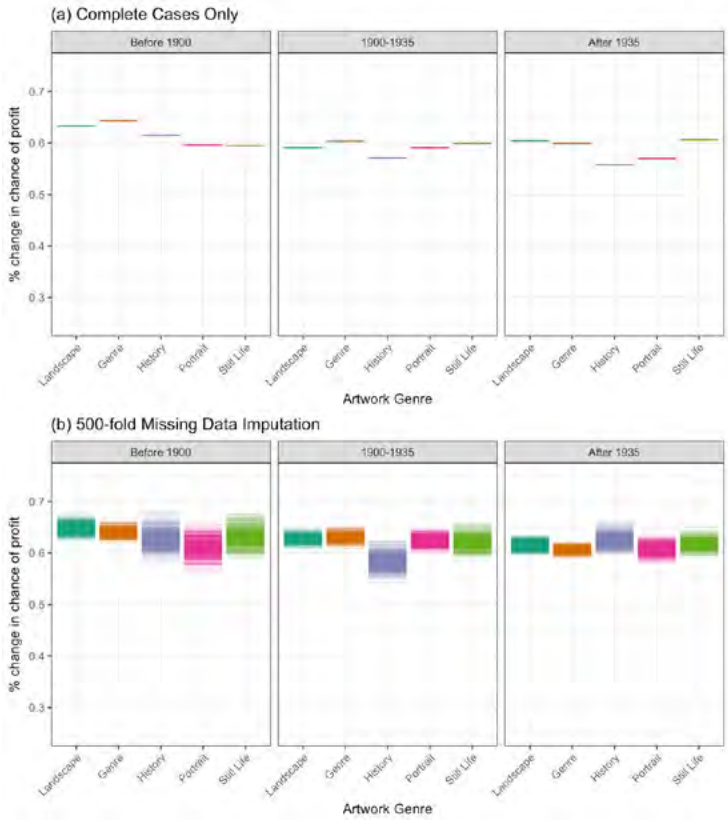


Figure 2 Partial dependence plots illustrating the marginal effect of artwork genre on Knoedler's chance at profitability.

Figure 2(a) shows the marginal effect of artwork genre on Knoedler's chance of turning a profit across three periods of their business, only considering around 20,000 "complete" cases from the Knoedler transaction records (approximately 60% of the known transactions they made.) A first glance suggests that history paintings were markedly less profitable after 1935, while still lifes became comparatively more profitable after 1935.

However, 2(b) shows the results not from 1 model, but from 500 models, each trained on a slightly different set of stochastically-imputed data. By visualizing one bar for each model, this plot drives home the effect of increased uncertainty on these measurements, while visually foregrounding the crucial methodological decision - 500 models instead of 1 - in a way that a box plot or other summary visualization method does not (at least, not in the eyes of a reader unused to reading such idioms.) The apparent advantage of still life in Knoedler's post-1935 business has evaporated, although the notably-lower value of history paintings between 1900-1935 may have withstood this simulation of uncertainty. While this model affirms that genre is largely an anachronistic construct that has little effect on prices, these results complicate a simplistic reading by indicating that, in some cases, there *is* a significant relationship that must be reckoned with.

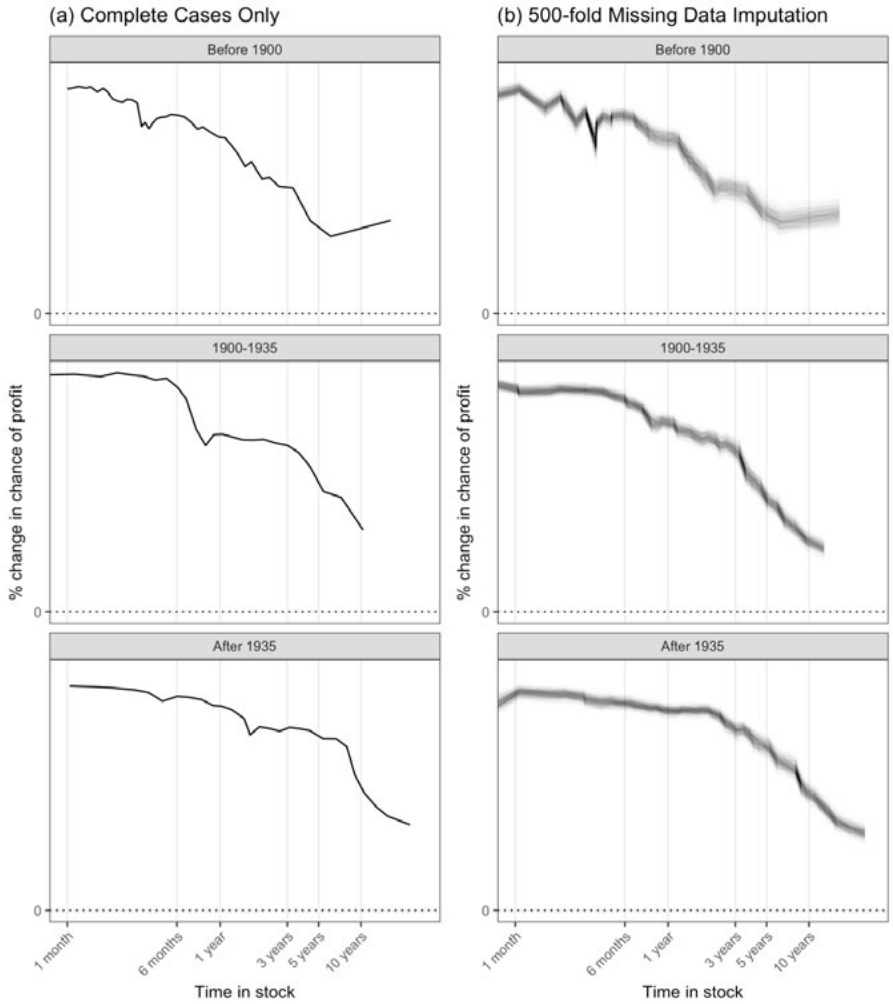


Figure 3 Partial dependence plots illustrating the marginal effect of time in stock on Knoedler's chance at profitability.

Figure 3 shows a similar comparison of complete case vs.imputed data for a continuous variable: the time a painting spent in stock. Both 3(a) and 3(b) support the conclusion that not only did a longer time in stock contribute to lower chances of turning a profit, but that Knoedler's window for making a profitable sale grew throughout the lifetime of the firm, from around 2 years before 1900, to more than 5 years after 1935. The increased uncertainty added by the multiplicity of models in 3(b) discourages the kind of over-interpretation that the seeming-precision of 3(a) allows. However, it also demonstrates that, even in the face of so many missing or imprecise dates in the Knoedler stock books, we can still recover meaningful quantitative conclusions.

Figure 4 Screenshot of an interactive application allowing users to modify imputation assumptions and see the effect on modeling and analysis results.

These static visualizations are easily enhanced through animation that shows the buildup of individual model characteristics into aggregate confidence intervals. (Lincoln, 2015) We have also experimented with interactive applications (Figure 4) that allow the user to specify different imputation assumptions, and then immediately see the downstream results on our predictive models, reinforcing the close relationship between starting assumptions and modeled conclusions. (An early demo of this work: https://mdlincoln.shinyapps.io/missingness/)

Computationally, these imputations are simple, perhaps even simplistic. More complex approaches, such as iteratively modeling every missing variable (Buuren and Groothuis-Oudshoorn, 2011), might lead to more accurate modeling. However, these less parsimonious methods are more opaque to humanities scholars. Operationalizing the historian's habit of educated guessing and thoughtful assumptions, and visualizing those operations straightforwardly, may allow missing data imputation to work its way into the accepted suite of DH methodologies.

References

Bode, K. (2017). The Equivalence of 'Close' and 'Distant' Reading; or, Toward a New Object for Data-Rich Literary History. *Modern Language Quarterly*, 78(1): 77–106 doi:10.1215/00267929-3699787.

Brosens, K., Alen, K., Slegten, A. and Truyen, F. (2016). MapTap and Cornelia: Slow Digital Art History and Formal Art Historical Social Network Research. *Zeitschrift Für Kunstgeschichte*, 79: 1–14.

Buuren, S. van and Groothuis-Oudshoorn, K. (2011). Mice: Multivariate Imputation by Chained Equations in *R*. *Journal of Statistical Software*, 45(3) doi:10.18637/jss.v045.i03.

Gelman, A. and Hill, J. (2006). Missing-data Imputation. In, *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Oxford: Cambridge University Press, pp. 529–43.

Grossner, K. and Meeks, E. (2013). *Temporal Geometry in Topotime*. Stanford University Libraries http://dh.stanford.edu/topotime/docs/TemporalGeometry.pdf.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.

Laskey, K. J., Laskey, K. B., Costa, P. C. G., Kokar, M. M., Martin, T. and Lukasiewicz, T. (2008). *Uncertainty Reasoning for the World Wide Web*. W3C Incubator Group Report World Wide Web Consortium (W3C) https://www.w3.org/2005/Incubator/urw3/XGR-urw3-20080331/ (accessed 26 November 2017).

Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R News*, 2(3): 18–22 https://www.stat.berkeley.edu/~breiman/RandomForests/.

Lincoln, M. D. (2015). DataGIFs: Animate Your Visualizations for Fun and Clarity *Matthew Lincoln, PhD* https://matthewlincoln.net/2015/12/18/datagifs-animate-your-visualizations-for-fun-and-clarity.html.

Posner, M. (2015). What's Next: The Radical, Unrealized Potential of Digital Humanities *Miriam Posner's Blog* http://miriamposner.com/blog/whats-next-the-radical-unrealized-potential-of-digital-humanities/ (accessed 20 April 2017).

Critical Data Literacy in the Humanities Classroom

Brandon T. Locke
blocke@msu.edu
Michigan State University, United States of America

Humanities data and data in our daily lives

As our world becomes increasingly data-driven, data skills and literacies (including the ability to assess data gaps and coverage, misleading visualizations, and the ethics surrounding data collection, usage, and sharing) are becoming crucial tools to our lives, both inside and outside of higher education. Scholarship across disciples is moving towards more data-intensive work, and scholars are increasingly expected to include open access to the data collected and used. At the same time, devices and software we use, the platforms we use to communicate, and the places we shop are increasingly enabled by the collection of data about our purchasing habits, web history, and contents of our email inboxes. Governments at all levels are increasingly collecting and using data to alter policies and direct day-to-day activities, ranging from transportation infrastructure to policing.

While much (though certainly not all) data-driven scholarship may seem significantly different from third-party data collection and data-driven policing, the former provides an opportunity to prepare students to understand, critique and improve the latter. Learning about the accurate and ethical and collection and usage of data and algorithms is a crucial part of liberal education that can help students better understand the processes around them, and better prepare them to apply those ethics and practices in the workplace and civic realm after graduation.

While many may think of data literacy as being the work of Computer Science departments, or perhaps library workshops targeted at researchers, the author argues that teaching these skills in the humanities classroom is fruitful for both the development of disciplinary knowledge and for developing crucial skills for use outside of the humanities classroom. Humanities data provides an excellent space to think critically about how people, ideas, and culture can and cannot be captured and analyzed through data. Comparing the data structures of colonial record keeping with the structures communities develop to document themselves provides clear lessons in the power of determining who and what gets documented, in the values that each community holds, and in privacy, ethics, and consent. Text mining novels, government records, or newspapers facilitates critical thinking about the value of metadata, the ability (or lack thereof) to derive meaning from large collections of text, and the use of different algorithms and approaches to ask different questions.

At the same time, the ability to think about humanities sources as data, and to properly curate and analyze them as such, provides a productive way to engage more with the way we conceptualize the sources, the way disciplinary knowledge is constructed and practiced, and the affordances provided by digitized and born-digital resources.

Data Challenges in Higher Education

The process of gathering, "cleaning," and organizing data can be incredibly time-consuming and difficult to prepare for. It can be tempting (and in many cases, required), to provide students with pre-prepared data to for analysis. Allotting time, either as in-class instruction or independent, project-based work, can take up weeks of time and can be a grueling disincentive for engagement. However, working critically with data rather than working with pre-packaged, pre-prepared datasets also aides us in the integration of digital humanities methods into the classroom, and better enables us to teach students emerging research methods through the full course of humanities research. Students can get a glimpse of the intellectual labor that goes into data collection, organization, and curation; not just in the final analysis.

There are several data literacy models that have shown success in other contexts. Data curation training often occurs in university-wide workshops or seminars, and are often brief and necessarily divorced from content and community practices (Carlson and Johnston 2015 p. 2-3). The Data Information Literacy (DIL) initiative, led by Jake Carlson and Lisa R. Johnson, is an extension of the ACRL Information Literacy Framework that focuses on

both the creation and consumption of data (ibid.). DIL is designed to be integrated into courses and research labs in the context of subject-specific data and domain-based community practices, but is primarily intended for faculty, staff, and graduate students working on peer-reviewed publication (ibid., p. 2-3). The Library-Led DH Pedagogy: Modeling Paths Toward Information and Data Literacy symposium facilitated productive conversations about the topic of data and information literacy in the digital humanities, but has not produced significant scholarship, models, or frameworks (Padilla et al. 2015).

In addition to making the case for teaching critical data literacy in the digital humanities classroom, the author will discuss both practical and theoretical approaches to data literacy in the undergraduate classroom that speak to the impetus behind teaching data literacy in the humanities: for greater disciplinary knowledge and understanding, to better facilitate digital scholarship and knowledge production, and to prepare students to better grasp, interrogate, and work with data in the public and private sector as citizens, employees, and employers.

References

Carlson, J. and Johnston, L. eds. (2015). *Data Information Literacy: Librarians, Data, and the Education of a New Generation of Researchers. Purdue Information Literacy Handbooks*. West Lafayette, Indiana: Purdue University Press.

Padilla, T., Smiley, B., Miller, S., and Mooney, H. (2015). "Modeling Approaches to Library–Led DH Pedagogy," *DH 2015 Global Digital Humanities Conference Abstract*. http://dh2015.org/abstracts/xml/PADILLA_Thomas_George_Modeling_Approaches_to_Libr/PADILLA_Thomas_George_Modeling_Approaches_to_Library_le.html (accessed 15 August 2017).

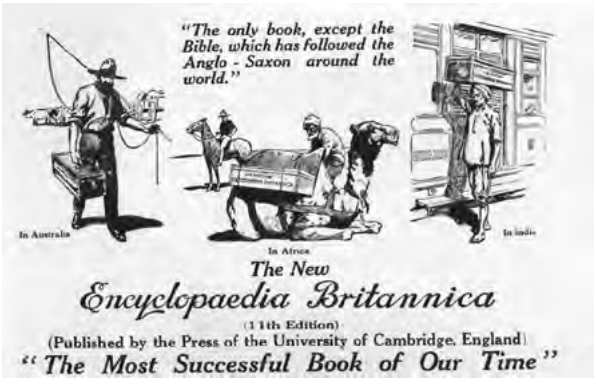
Ontological Challenges in Editing Historic Editions of the Encyclopedia Britannica

Peter M Logan
peter.logan@temple.edu
Temple University, United States of America

First published in 1771, *Encyclopedia Britannica* continues in publication today and is the only encyclopedia in any language to survive that 250-year period. Historical editions of the Encyclopedia offer scholars a unique means of examining the evolution of ideas and beliefs about sensitive cultural topics – such as suicide, race, and hysteria – by studying their treatment in different editions. But what can this curated dataset as a whole can tell us about larger patterns in the social construction of knowledge in the nineteenth-century English-speaking world?

We are creating a data set of all text from these historic editions for use in text mining. The corpus will include over 100,000 entries, all of which need to be tagged with essential metadata fields. How do we identify the different subject areas in this body of knowledge? This article briefly discusses the use of an automatic-metadata-generating algorithm, HIVE, created by the Metadata Research Center at Drexel University. But the central issue it addresses is the theoretical problem encountered in defining a subject vocabulary for this corpus.

The *Encyclopedia* claims to represent the "Sum of Human Knowledge," and while we can dispute this claim, it nonetheless represents the existence of older knowledge taxonomies used in its creation. How do we construct a subject vocabulary without distorting this older organizational scheme for subject categories? Those older vocabularies were clearly biased. For example, the decision to include or exclude entries, as well as the size assigned to entries, were all based on assumptions about what mattered as "legitimate" knowledge. Many of these are assumptions we no longer share; the editors excluded forms of knowledge rooted in folk and tribal cultures, and female authors were wholly absent until 1889. Racism and the perspective of British Imperialism are evident in many entries. These prejudices reflect the social beliefs of the writers and editors, of course, and as such, they illustrate the degree to which knowledge in the nineteenth century was clearly socially constructed. And the invented nature of that taxonomy needs to be captured accurately. The value of the curated content of Britannica to researchers today is that is the most comprehensive representation we have of that older knowledge system in its totality, and so it makes it possible to study that system as a structure and to observe how it changed over time.



The problems in tagging this biased dataset take three forms. First is the danger of historical anachronism. Applying a C21-century ontology, like Library of Congress Subject Areas, to C18 and C19 editions makes it accessible to modern researchers, but it also misrepresents the older system of knowledge. For example, the entries on

"History" from the important 3rd (1797) and 7th (1842) editions present authoritative accounts of human prehistory. While we might tag them under "anthropology," that field of knowledge was not recognized by the Royal Academy of Sciences until the 1880s (as a subset of Biology) and does not appear in the Encyclopedia itself until 1889. In fact, the older references cite the Book of Genesis as their authority, and a tag on applications of scripture to the interpretation of external reality might better represent the entry than an anachronistic "Anthropology, history of" tag could do.

The second difficulty is encountered when trying to reconstruct the older ontology used by the Encyclopedia, because it was a moving target. Subject categories changed over the first 150 years, with new categories added, others (like human prehistory) moving from one field to another, and still others disappearing. While we might construct a stable ontology for one edition, any historically-accurate ontology will have to become a system of multiple ontologies, whose relationships with one another need to be explained at the very least.

Third is the question of how to treat the built-in biases within the corpus. Older ontologies of knowledge are rife with bias, often through omission. Historically-accurate subject terms duplicate that bias. Information on attitudes toward women and national minorities, for example, exists within multiple entries, but there are no subject terms for minorities and no entries for women as such, making that data largely invisible without some form of intervention.

We are in the process of creating this new dataset and by summer of 2018 we will be completing preliminary tests on tagging systems, so the final paper will share preliminary results.

Distinctions between Conceptual Domains in the Bilingual Poetry of Pablo Picasso

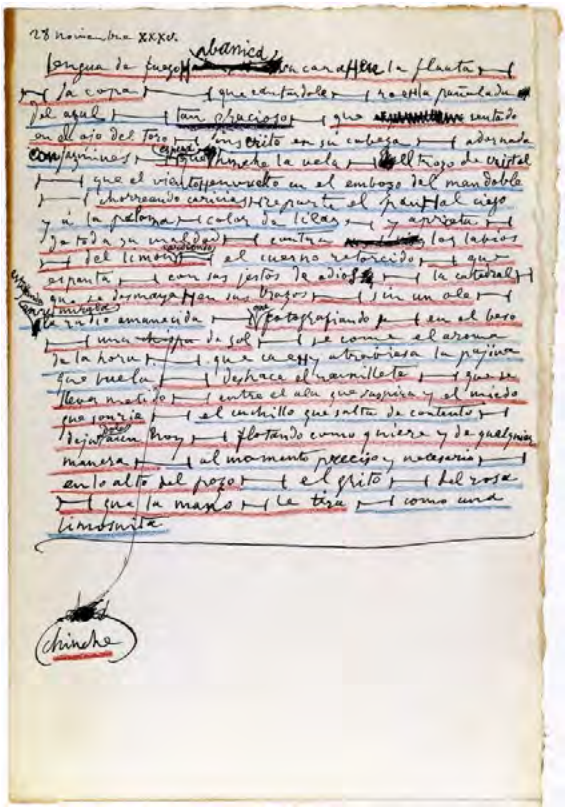
Enrique Mallen
mallen.shsu@gmail.com
Electronic Textual Cultures Lab, University of Victoria, Canada

Luis Meneses
ldmm@uvic.ca
Sam Houston State University, United States of America

Introduction

Picasso started writing poetry in April, 1935 during a period of personal crisis. Many have cited, among other possible causes the political turmoil in Europe in the period between the two wars. These views are predicated on an assumed irreducible conflict between visual composition and verbal expression. However, even before this he had been fascinated by linguistic structure and alternative methods of expression during his cubist period.

His poetry is not only fascinating as a form of communication from someone who is primarily known for his plastic output, it is also puzzling for anyone researching the interconnection between language and writing, i.e. verbal and graphic signs. His poetry is an attempt to expand the expressive power of language, as he adjoins words in unordered strings, following a technique very similar to cubist collage. Figures 1 and 2 show examples of this technique. The relation between words remains open and bidirectional, so that the reader is free to establish multiple semantic relations. And yet, while we see a close correlation between his poems and his artworks (Elizabeth, 2002) (Picasso and Baldassari, 2005), one cannot deny that his texts are primarily verbal; and this is precisely what makes them fascinating, as they provide a window into Picasso's mind that is separate from his own artistic creations –although they share with his artworks a predominance of ambiguity (Rubin et al., 1992) and the presence of unresolved conceptual oppositions.



Example of multiple additions and deletions in Picasso's poetry. P. Picasso, "lengua de fuego abanica ... (7)", Musée Picasso, Paris, 1935.

Example of the visual components in Picasso's poetry. P. Picasso, "si yo fuera afuera ... (2)", Claude Ruiz-Picasso Collection, 1935.

For some time now, our research has taken us through different approaches to analyze Picasso's artistic legacy (Meneses et al., 2008a), his poetry (Meneses et al., 2008b) and its semantic domains (Meneses and Mallen, 2017). In this paper, we propose to investigate how Picasso explored subtle differences between words within specific concepts in French and Spanish –as he composed his poems in two languages. This new perspective allows us to identify how the two languages offered Picasso a wide range of semantic domains to choose from when establishing subtle contrasts.

We have determined that, in Picasso's poems, certain semantic domains are predominant in each of the two languages. For instance, Picasso is more inclined to refer to food items and everyday objects in his Spanish poems. On the other hand, given the influence French Surrealist writers exerted on him, his French poems concentrate on more abstract concepts involving politics, religion and sexuality. Why did he choose to use these languages in the way he did? Daix (Daix and Emmet, 1993) has pointed out that "Picasso did not believe in spontaneous poetry – or painting". Our research will address the question of why did Picasso choose to write in a given language about a specific semantic domain.

Methodology

We have already classified the semantic interconnections between the concepts that Picasso explored (Meneses and Mallen, 2017). For this purpose, we used a taxonomy-based approach to identify the semantic domains in Picasso's poetry. We created a set of database tables that allowed us to specify concepts and then map them to their related terms in Picasso's poetry. It is important to note that these concepts are not bound to a given language per se: we were able to overcome the language barrier by linking concepts using the English translation of relevant terms –a language that Picasso didn't use in his poems.

We observed that some of the existing semantic categories are linked to a higher number of concepts than others. For example, we find a high number of nominal artifacts in his poems. Some are related to art, such as engraving, impression, ornament, paint and palette; others related to war, such as armor, axe, blade, bomb, bow, bullet, camouflage and rapier. These may appear antagonistic, but in Picasso's world there is a close relation between destruction in war and creation in art. Not surprisingly for a painter and writer, nominal communication is another frequent semantic category, with such concepts as advice, agreement, alphabet, fable, language, news, outcry and parable.

Given that we had a refined taxonomy, we decided to approach this problem from a purely computational perspective and expand on our previous efforts based on statistical models and algorithms (Meneses et al., 2016). More specifically, we propose to address our research questions by analyzing Picasso's semantic domains using Latent Dirchlet Allocation (Blei et al., 2003) and Term frequency-Inverse document frequency. We will do this by linking sets of words with their corresponding semantic concepts. Our analysis has shown that these techniques are capable of highlighting patterns and trends in Picasso's poetry that escape other forms of traditional analysis.

Conclusions

Our study is an attempt to further understand the semantic domains that Picasso operated with. Again, we know that Picasso's style, both in his visual and his verbal compositions, was very much inspired by collage. What makes them interesting is that those elements he placed together belonged to a restricted set, so that their interconnection, while not obvious to the viewer/reader, must have been somewhat determined in Picasso's "view" of reality. It is that determined interconnection which Picasso saw that we propose to explore with this study. In other words, we want to get closer to Picasso's "vision" of the world through his poems in order to investigate how that "vision" may differ from what he depicted in his graphic works.

To summarize, in this paper we propose to analyze why Picasso chose to write in a given language about concepts in a specific semantic domain. More so, throu-

gh the use of statistical models we propose to identify and pinpoint representative themes and correlations across different concepts and languages. Picasso once said: "Computers Are Useless. They Can Only Give You Answers". In this case, through our use of computational methods we are attempting to do just that: help us as researchers to get a better understanding of Picasso's poetry and artworks –and consequently, from the artist that created them as well.

References

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, **3**: 993–1022.

Daix, P. and Emmet, O. (1993). *Picasso: Life and Art*. Thames and Hudson.

Elizabeth, C. (2002). Picasso: style and meaning.

Meneses, L., Estill, L. and Furuta, R. (2016). This was my speech, and I will speak it again": Topic Modeling in Shakespeare's Plays.

Meneses, L., Furuta, R. and Mallen, E. (2008a). Exploring the Biography and Artworks of Picasso with Interactive Calendars and Timelines. pp. 160–62.

Meneses, L. and Mallen, E. (2017). Semantic Domains in Picasso's Poetry Paper presented at the Digital Humanities 2017, Montreal, Canada.

Meneses, L., Monroy, C., Mallen, E. and Furuta, R. (2008b). Picasso's Poetry: The Case of a Bilingual Concordance. pp. 157–59.

Baldassari, A. (2005). *The Surrealist Picasso*. Flammarion.

Rubin, W., Varnedoe, K., Reff, T., Cottington, D., Fry, E. F., Poggi, C., Krauss, R. and Bois, Y. A. (1992). *Picasso and Braque: A Symposium*. Museum of Modern Art.

A formação de professores/ pesquisadores de História no contexto da Cibercultura: História Digital, Humanidades Digitais e as novas perspectivas de ensino no Brasil.

Patrícia Marcondes de Barros
patriciamarcondesdebarros@gmail.com
UNESPAR, Brazil

A presente comunicação (resultante de pesquisa em fase inicial) tem como objetivo geral, a análise e a reflexão, sob uma perspectiva metodológica qualitativa, acerca da formação de professores/pesquisadores de História no contexto da cibercultura e na sua esteira, da chamada História Digital (*Digital History*) e Humanidades Digitais (*Digital Humanities*). Os contributos metodológicos e práticos que as diversas tecnologias podem oferecer aos profissionais de História, as competências técnicas necessárias ao usufruto das mesmas e o entendimento das novas subjetividades erigidas, são de suma importância para a análise das mudanças paradigmáticas contemporâneas que abarcam o ensino, a pesquisa e, portanto, a formação docente neste devir.

Desde os anos 60 e 70 do século XX, observam-se mudanças culturais relacionadas aos meios comunicacionais, estudadas por grandes pesquisadores das mais diversas áreas de saber, a exemplo de Marshall McLuhan, filósofo e teórico da comunicação, que postulou a ideia de que a interdependência eletrônica recriaria o mundo numa aldeia global resultando no neotribalismo, erigindo assim, uma nova cultura.

Seus aforismos como "o meio é a mensagem", "os meios como extensões do homem" e "O homem cria a ferramenta. A ferramenta recria o homem", permanecem atuais na análise do mundo contemporâneo com suas múltiplas conexões, dotado da dimensão de universalidade (e assim sendo, "extenso, interconectado e interativo") e, portanto, menos totalizável (LÈVY, 1999, p.120) e de difícil apreensão.

Mcluhan aponta para uma sensibilidade na qual o meio traz consequências sociais e pessoais resultantes do estalão introduzido em nossas vidas por uma nova tecnologia, que é a extensão de nós mesmos. A máquina, por exemplo, independente do tipo de produção que faz, constitui a mensagem e transforma as relações. O autor pretende assim postular que o meio, geralmente pensado como um simples canal de passagem do conteúdo comunicativo, é um elemento determinante da própria comunicação: "o meio é a mensagem" (MCLUHAN, 2006).

A ideia postulada por Mcluhan de "aldeia global", de "ser planetário" relaciona-se aos movimentos de contracultura dos anos 60, que junto à instantaneidade dos meios comunicacionais eletrônicos, construiu uma subjetividade diferenciada (não-linear), denominada por muitos como pós-moderna e que foi a gênese da cibercultura que eclodiu em 1989 (BOLESINA; GERVASONI, 2015, p.08). Surge assim um novo mundo, relacionados à tecnociência e **às** tecnologias de Informação e de Comunicação (TICs).

É no universo educacional, o *locus* de grande visibilidade das mudanças sociais e culturais, tendo em vista a construção das identidades e apreensão da alteridade cultural através dos processos de aprendizagem e socialização. Com a pós-modernidade este universo passa por ressignificações e buscam metodologias que se integrem **às** novas tecnologias da informação, a interdisciplinariedade - entendida como os saberes comuns a uma ou mais matrizes de conhecimento-, e principalmente, a Antropologia, esfera privilegiada que aborda a cultura como dimensão fundadora da sociedade e permite o entendimento da alteridade, importante valor de reconhecimento das diversas culturas que permeiam o ambiente escolar.

A complexidade e a diversidade cultural observada reflete o espaço sem fronteiras, desterritorializado da cultura engendrada no ciberespaço, denominada como cibercultura.

A cibercultura representa um conjunto de técnicas, modos de pensamento e valores que se instituíram no ciberespaço (LÈVY,1999) que especifica não apenas a infraestrutura material da comunicação digital, mas também o universo de informação que ela abriga, assim como os seres humanos que navegam e alimentam esse universo. Pode ser entendido como a união de redes e recursos de comunicação formada pela interconexão global dos computadores pelo qual passou a ser possível o acesso à distância aos recursos de um computador, a exemplo da troca de arquivos digitais de forma simplificada, o envio de mensagens de forma sincrônica ou assincrônica, conferências eletrônicas em tempo real e transmissão de vídeo/som, entre horizonte de outras possibilidades. O conjunto dessas novas práticas, suportadas pelas tecnologias digitais e que foram apropriadas pela sociedade contemporânea transformaram os saberes e as práticas educacionais.

Vale ressaltar que tais transformações culturais se dão não somente com o aparato tecnológico, mas principalmente pelos tipos de signos que circulam nesses novos meios engendrando mensagens e processos de comunicação (SANTAELLA, 2003, p.24).

Com o advento da cultura digital e sua universalização, as interações sociais e a produção de conhecimento são amplamente transformadas através da virtualidade, reverberando na chamada História Digital e Humanidades Digitais.

Dentro do contexto educacional brasileiro esta nova configuração tecnológica não se enquadra ainda a realidade escolar e não apenas devido à questão estrutural de precarização de escolas e universidades públicas, como se observa atualmente. Há também resistência dos profissionais no campo das licenciaturas, especificamente nas das Ciências Humanas, como a História, em relação à inserção e discussão acerca de metodologias do ensino relacionadas ao novo processo comunicacional, o que nos coloca em situação de atraso frente a outros países.

Hansen (2015) assinala, segundo dados da centerNet, que existem 196 centros de pesquisa sobre Humanidades Digitais, sendo 88 na América do Norte, 75 na Europa e os 12 restantes pelo resto do mundo. A História Digital (*Digital History*) e o campo das Humanidades Digitais (*Digital Humanities*) são termos ainda recentes no léxico acadêmico brasileiro e não há consenso sobre seus significados.

A presente pesquisa analisará os processos comunicativos contemporâneos e suas interfaces com a Educação e a História (tanto no ensino, quanto na pesquisa), assim como as tecnologias, interações e convergência (História Digital e Humanidades Digitais) e a produção de linguagens e produção de sentidos no contexto da cibercultura.

Trata-se de forma geral, de repensar sob a **égide** das mudanças paradigmáticas postuladas pela cibercultura, os novos sentidos para a educação e a pesquisa que seja consonante **às** tecnologias, mas também humanista e aberta à complexidade e diversidade que os novos "meios e mensagens" nos trazem na contemporaneidade. Como afirma Hansen (2015), preparar futuros historiadores para o uso de outras mídias, que não as convencionalmente usadas, significa equipá-los com ferramentas que permitam explorar criativamente diferentes formas de apresentação do conhecimento histórico, e também avaliar criticamente produções e recursos disponíveis.

References

HANSEN, P.S. *Digital History* e formação de historiadores: sugestões para um debate.In BUENO,A.;ESTACHESKI, D.; CREMA (organizadores). *Tecendo amanhãs: O ensino de História na atualidade*. Rio de Janeiro/União da Vitória: edição especial *Sobre Ontens*, 2015.

LEVY, P. *O que é o virtual*. São Paulo: Ed. 34, 1996.

_____. *Cibercultura*. São Paulo: Ed. 34, 1999.

MCLUHAN, Marshall. Os meios de comunicação como extensão do homem (understanding media:The Extensions of Man). Editora Cultrix, São Paulo, 2006.

SANTAELLA, L. *Culturas e artes do pós-humano: da cultura das mídias à cibercultura*. São Paulo: Paulos, 2003.

Presentation Of Web Site On The Banking And Financial History Of Spain And Latin America

Carlos Marichal
cmari@colmex.mx
El Colegio de Mexico, Mexico

The purpose of this ten-minute presentation is to present this thematic site which we have constructed in 2016 (and is bilingual) and is of use for professors, students, and general public. The object this academic web page (the first of its kind in this specific field) is to provide a large amount of documents (including links to over 200 working papers) historical statistics (over 600 Excel charts and graphs), bibliographies, guides to archival sources and short historical summaries of the banking histories of many countries in Latin America as well as Spain. Both El Colegio de México and the University of Cantabria (Spain) have collaborated in this project under the direction of Dr. Carlos Marichal. The site will soon be transferred to the electronic resources of the Libraries of both academic institutions.

I argue that this thematic webpage corresponds to an increasing trend in contemporary academics to *combine*

concrete and deep research results in *subdisciplines* with complementary resources of a varied nature, including historical statistical series, reference texts, images (photos and engravings), timelines, and resources for teaching.

Such resources are especially useful for consultation on line by professors and students in local universities, many of which – in Mexico and Latina America- do not have really rich library/digital resources In addition I might remark that there is a demand from schools and universities for advanced online courses in humanities and social sciences that can be especially useful for updating university professors, especially in the provinces, where there is urgent need for support to achieve a substantial improvement in teaching and research in humanities and social sciences in Mexico or other countries in the region.

To accomplish this, a multidisciplinary working group of academics has been set up to gather pertinent information from the various humanities and social sciences in the field of banking and financial history of Latin America and Spain, which explains the international consortium engaged. The interest of the project lies in the pioneering projects in this field in the humanities and social sciences both in academia in Mexico and elsewhere. The site can be consulted at http://codexvirtual.com/hbancaria/

Spatial Disaggregation of Historical Census Data Leveraging Multiple Sources of Ancillary Data

João Miguel Monteiro
joao.miguel.monteiro@tecnico.ulisboa.pt
University of Lisbon, IST and INESC-ID, Portugal

Bruno Emanuel Martins
bruno.g.martins@ist.utl.pt
University of Lisbon, IST and INESC-ID, Portugal

Patricia Murrieta-Flores
p.murrietaflop.a.murrieta-flores@lancaster.ac.uk
University of Lancaster, United Kingdom

João Moura Pires
jmp@fct.unl.pt
Universidade NOVA de Lisboa, FCT / NOVA LINCS, Portugal

Accurate information about the human population distribution is essential for formulating informed hypothesis in the context of several social, economic, and environmental issues. Government instigated national censuses are authoritative sources of population data, subdividing space into discrete areas (e.g., fixed administrative units) and providing multiple snapshots of society at regular intervals, typically every 10 years. Many research institutions or national statistical offices have developed historical Geographical Information Systems (GIS), containing statistical data from previous censuses together with the administrative boundaries (i.e., records of administrative boundary changes) used to publish them over long periods of time. However, using these data can still be quite challenging, particularly when looking at changes over time.

There are multiple reasons why population data aggregated to administrative units is not an ideal form of information about population counts and/or density. First, these representations suffer from the modifiable areal unit problem (Lloyd, 2014), which states that the results of an analysis that is based on data aggregated by administrative units may depend on the shape and arrangement of the units, rather than capturing the theoretically continuous variation in the underlying population. Second, the spatial detail of aggregated data is variable and usually low, particularly in the context of historical data. In a highly aggregated form these data are useful for broad-scale assessments, but using aggregated data has the danger of masking important local hotspots, and overall tends to smooth out spatial variations. Third, there is often a spatial mismatch between census areal units and the user-desired units required for particular types of analysis. Finally, the boundaries of census aggregation units may change over time from one census to another, making the analysis of population change, in the context of longitudinal studies dealing with high spatial resolutions, difficult.

Given the aforementioned limitations, high-resolution population grids (i.e., geographically referenced lattices of square cells, with each cell carrying a population count or the value of population density at its location) are often used as an alternative format to deliver population data. All cells in a population grid have the same size and the cells are stable in time. There is no spatial mismatch problem as any partition of a given study area can be rasterized to be co- registered with a population grid.

Population grids can be built from census data through the application of spatial disaggregation methods (Monteiro et al., 2014), which range in complexity from simple mass- preserving areal weighting, to intelligent dasymetric weighting schemes that leverage regression analysis to combine multiple sources of ancillary data.

Nowadays, there are for instance many well-known gridded datasets that describe the modern population distribution, created using a variety of disaggregation techniques (e.g., the Gridded Population of the World (Doxsey-Whitfield et al., 2015) or the WorldPop databases (Tatem, 2017)). However, despite the rapid progress in terms of disaggregation techniques, population grids have not been widely adopted in the context of historical data. We argue that the availability of high-resolution population grids within historical GIS has the potential to improve the analysis of long-term geographical population changes. Perhaps more importantly, this can also facilitate the combination of population data with other

GIS layers to perform analyses on a wide range of topics, such as the development of the transport network, historical epidemiology, the formation of urban agglomerations, or climate changes.

This work reports on experiments with a hybrid disaggregation technique that combines the ideas of dasymetric mapping and pycnophylactic interpolation (Monteiro et al., 2014), using machine learning methods (e.g., linear regression models, ensembles of decision trees, or deep learning approaches based on convolutional neural networks, which previously have only seldom been used for spatial disaggregation (Robinson et al., 2017)) to combine different types of ancillary data (e.g., historical land-coverage data from the HILDA project (Fuchs et al., 2015), together with modern information that we argue can correlate with historical population), in order to disaggregate historical census data into a 200 meter resolution grid. Apart from few exceptions related to the use of areal interpolation for integrating historical census data, most previous related studies have focused on modern datasets.

We specifically report on experiments related to the disaggregation of historical population counts from three different national censuses which took place around 1900, respectively in Great Britain, Belgium, and the Netherlands. All three statistical datasets, together with the corresponding boundaries for the regions at which the data were collected (i.e., parishes or municipalities), are presently available in digital formats within national historical GIS projects. The obtained results indicate that the proposed method is indeed accurate, outperforming simpler schemes based on mass-preserving areal weighting or pycnophylactic interpolation. Moreover, the obtained results also show that modern data, particularly pre-existing gridded datasets that describe the modern population distribution (i.e., data from the Gridded Population of the World (Doxsey-Whitfield et al., 2015) project), are particularly useful as features for supporting the disaggregation of historical population counts. The best results were obtained with regression models leveraging multiple features (i.e., different models attained the best results in each of the three national territories that were considered), although a simple dasymetric technlque, leveraging the modern population gridded data to define the disaggregation weights, achieved very competitive results.

Acknowledgements

This research was partially supported by the Trans-Atlantic Platform for the Social Sciences and Humanities, through the Digging into Data project with reference HJ-253525. The researchers from INESC-ID also had financial support from Fundação para a Ciência e Tecnologia (FCT), through the project grants with references PTDC/EEI-SCR/1743/2014 (Saturn) and CMUPERI/TIC/0046/2014 (GoLocal), as well as through the INESC-ID multi-annual funding from the PIDDAC program, which has the reference UID/CEC/50021/2013.

References

Doxsey-Whitfield, E., MacManus, K., Adamo, S. B., Pistolesi, L., Squires, J., Borkovska, O. and Baptista, S. R. (2015). Taking advantage of the improved availability of census data: a first look at the gridded population of the world, version 4. *Papers in Applied Geography*, 1(3).

Fuchs, R., Herold, M., Verburg, P. H., Clevers, J. G. and Eberle, J. (2015). Gross changes in reconstructions of historic land cover/use for Europe between 1900 and 2010. *Global change biology*, 21(1).

Lloyd, C. D. (2014). *The Modifiable Areal Unit Problem. Exploring Spatial Scale in Geography*. John Wiley & Sons.

Monteiro, J., Martins, B. and Pires, J. M. (2017). A hybrid approach for the spatial disaggregation of socio-economic indicators. *International Journal of Data Science and Analytics*, 5(2-3), pp 189–211.

Robinson, C., Hohman, F., and Dilkina, B. (2017). *A Deep Learning Approach for Population Estimation from Satellite Imagery. Proceedings of the ACM SIGSPATIAL Workshop on Geospatial Humanities*. New York: ACM Press.

Tatem, A. J. (2017). WorldPop, open data for spatial demography. *Scientific Data*, 4, 170004.

The Poetry Of The Lancashire Cotton Famine (1861-65): Tracing Poetic Responses To Economic Disaster

Ruth Mather
r.m.mather@exeter.ac.uk
University of Exeter, United Kingdom

Our project will make freely available a searchable webapp, built in eXist-db (http://exist-db.org/exist/apps/homepage/index.html), containing a database of poems responding to Lancashire Cotton Famine of 1861-65, along with audio recitations and musical performances drawing directly on these poems (Rennie, 2017). This poetic response is important in that it often represents labouring-class voices from the mid-nineteenth century, which, in spite of renewed academic interest in such material, remain underappreciated (Goodridge et al, 2012 provides a useful introduction and bibliography). The study of this material and its digital publication will significantly enrich literary scholarship and historical perspectives of this economic crisis, and provides the opportunity to draw public attention to an episode of history that is little known beyond the scholarly sphere. The project seeks to establish a much more detailed understanding of

the nature of Lancashire Cotton Famine poetry: its extent, its intents, and its functions. To date, there is no critical literature specifically addressing the poetry written and published during the Cotton Famine, though the period is touched upon in Brian Hollingworth's anthology *Songs of the People: Lancashire dialect poetry of the industrial revolution* (1972: 98-113).

The project draws predominantly on the local newspaper collections of Lancashire's various civic archives and local studies centres for material. Though some of these newspapers have been digitised in collections such as the British Newspaper Archive and Gale Historical Newspapers, the majority are in hard-copy or microfilm format, so that users attempting to access the poetry encounter practical obstacles relating to both geography and the preservation of materials. Additionally, as the archives and local studies collections in Lancashire are significantly under-resourced, there are long-term concerns about maintenance of the equipment and staffing levels required to ensure that access to these significant historical collections is sustainable. The recovery element of our project therefore aims to ensure long-term, free-of-charge access to a near-complete repository of Cotton Famine Poetry without the requirement to visit multiple archives or local studies centres. As the recovered poems have been transcribed by hand, we are also avoiding replicating the Optimal Character Recognition errors which have been incurred by some of the existing newspaper databases (Joulain-Jay, 2016).

Alongside the vital recovery and collation of this material, the experience of the investigators in the field of labouring-class poetry enables a simultaneous critical analysis of the poetry as it emerges, focussing on local, regional, national, and international fields of interest. We encounter a wide range of poetic styles, written both in Lancashire dialect and standard English, which demonstrate the sophisticated literary engagements of their authors. In terms of subject matter, the poems describe not only the direct, local experience of the Cotton Famine, but also offer more abstract reflections on issues including work, poverty, war, slavery and abolition. We want to determine the extent to which political dissent is present in the poetry, and to what degree opposing discourses relating to slavery and the American Civil War were articulated through literature of this type. We are already beginning to establish that a significant proportion of Cotton Famine poetry represents a labouring-class address to a regional and national middle-class readership, and part of our analysis will involve mapping a transatlantic discourse between the Lancastrian labouring classes and writers on both sides of the American conflict. The popular narrative of the Cotton Famine has Lancashire textile workers staunchly supporting the North in spite of the deprivation caused by the war, because of their strong support for abolition. Early analysis of poetic responses suggests a more complicated engagement with the American conflict, including some elements of support for the south (see also Ellison, 1972). We hope that by making the poetry freely available online, we encourage its further use as an important historical and literary resource for understanding some of these complex themes.

The digitisation of a large and varied body of work such as this presents both challenges and possibilities, and this paper will reflect upon the difference that digital methods make to our interpretation of the material and the ways in which it can be used. The process of marking up text for the database enables us to make our editorial decisions in presenting this body of work transparent, and to group similar poems and themes for the reader's ease of analysis. Nonetheless, in so doing, we impose our own interpretations upon what was a fluid form - often orally transmitted, and published in different versions across different media. A key concern in presenting this material has been the desire to ensure its usefulness for scholars who might have different approaches and questions to our own. In forcing us to grapple with these challenges, the use of digital methods has encouraged us to make explicit our own methodologies and thought processes, and enabled the creation of a resource that could be considered more intellectually 'open' than the traditional analogue anthology.

The design of our webapp therefore reflects our desire for a flexibility which in turn offers better representation of a literature which was originally available in multiple, sometimes changing forms. Some poems were written to be read or sung aloud, while others endeavoured to capture in writing the transient forms of local dialect, and different variations of the same poem appear across the newspapers. The use of XML enables us to continuously add layers of interpretation as they occur in the data for macro analysis, while marking up at word-by-word level enables a careful close reading in which we are forced to be conscious of the decisions we are making about the presentation of material.

An important part of the project is its public-facing element, including the involvement of school groups in finding and transcribing the poetry. We also welcome submissions of potential Cotton Famine poetry from members of the public, local historical societies, and educational projects with an interest in this material. Managing the upload and editing of these submissions, and ensuring that appropriate credit is given, is one of the tasks that the project team has taken on, and it is likely to present its own set of challenges: gathering this data provides an opportunity to involve the public in undertaking research and giving them insights into this process, but we also need to ensure that the results are useful and worthwhile for the project's outputs. At present we are not planning to train people beyond the team in how to encode in TEI, but giving contributors a 'behind the scenes' tour of the database and offering an introduction to how we create and structure our digital materials (and why)

has the potential to enable further discussions and may encourage contributors to get involved with digital humanities activity beyond our immediate project. We feel that this is an important step in ensuring that contributors gain an understanding of what happens to their data once they submit it, and how it is transformed into what they see in the final digital publication. We will discuss these challenges and how we intend to maintain interest and engagement amongst our contributors.

At this stage of the project, in which we are making key decisions about how to manage our own data and that from our external contributors, we would welcome discussion and comments from the wider digital humanities community on how we can ensure that our resource is an effective tool for both research and teaching. We hope, too, that the challenges that we are encountering and some of our proposed solutions might prove helpful for others working with comparable datasets or audiences.

References

Ellison, M. (1972) *Support for Secession: Lancashire and the American Civil War.* Chicago & London: University of Chicago Press.

http://exist-db.org/exist/apps/homepage/index.html

Goodridge, J. et al (2012) 'Introduction', *Labouring-Class Poets Online.* https://lcpoets.wordpress.com/introtobibliography/

Hollingworth, B. (1972) *Songs of the People: Lancashire dialect poetry of the industrial revolution*. Manchester: Manchester University Press.

Joulain-Jay, A. (2016) 'Dealing with Optimal Character Recognition errors in Victorian Newspapers', *British Library Digital Scholarship Blog*, July 20th 2016. http://blogs.bl.uk/digital-scholarship/2016/07/dealing-with-optical-character-recognition-errors-in-victorian-newspapers.html

Rennie, S. (2017) *The Poetry of the Lancashire Cotton Famine (1861-65)* http://cottonfaminepoetry.exeter.ac.uk

READ Workbench – Corpus Collaboration and TextBase Avatars

Ian McCrabb
ian@prakas.org
University Of Sydney, Australia

The Research Environment for Ancient Documents (READ) project commenced in 2013 with development support from a consortium of institutions (University of Washington in Seattle, Ludwig Maximillian University in Munich, University of Lausanne, University of Sydney and Prakaś Foundation) involved in the study and publication of ancient Buddhist documents preserved in Gāndhārī.

READ has been developed as a comprehensive multi-user platform for the transcription, translation and analysis of ancient Sanskrit, Gāndhārī, Pali and other Prakrit texts: manuscripts, inscriptions, coins and other documents. It is based on open source software (Postgres, PHP and JQuery), supports the TEI standard and provides an API for integration with related systems. READ is positioned as a research environment, complementary to existing textual repositories and integrated with existing dictionaries. Existing transcriptions can be imported, elaborated upon, analyzed, and then published as research output in standards-based formats. The defining innovation of READ is atomization to a semantically linked network of objects; a paradigm shift in data structure from strings of marked-up text to sequences of linked objects.

The underlying design and entity model was presented at both the Digital Humanities conference in 2015 and the 2016 Australian Digital Humanities Conference. READ has been publicly released and is supporting a wide range of corpus development projects. Whilst this presentation will follow on to briefly precis the range of research projects currently supported by READ, the focus will be on a related platform. READ Workbench (Workbench) is a server portal hosted at the University of Sydney since 2016 to 'harness' READ. Developed using the same technology stack as READ, it is a comprehensive management framework to support the integration of people and processes in the collaborative development, maintenance and publishing of textual corpora.

The design of Workbench evolved organically as the implementation requirements of READ expanded from a single researcher working on a single text, to the capacity to support multiple projects, each with a team of researchers collaborating on the development of an integrated corpora, with widely divergent research objectives. The fundamental objective was to design a supporting framework with which manage the balance between autonomy and collaboration in large scale projects. The approach adopted was to implement strategies, models and workflow patterns consistent with those applied in the IT consulting industry to digital content design, development and migration projects.

Workbench is a software as a service (SaaS) platform managing multiple READ installations, each with project and language specific configurations, supporting researcher collaborations across multiple institutions. It provides a self-service portal for researchers to develop, maintain, manage and publish texts without requiring technical support or the mediation of a database administrator; critical to the longer-term sustainability of corpora projects. Workbench's three facets (configuration management, database management and corpus workflows) provide a scalable implementation architecture for READ and instantiate a comprehensive corpus development methodology.

Whilst the configuration management services might be bracketed as conventional for a SaaS platform, database management is predicated upon an architectural innovation that enables researchers to build, share, manage, maintain and publish their own texts. The adoption of a single text/single database (TextBase) as the fundamental object of development, collaboration and portability is quite a departure from conventional models where a centralized administrator manages a single monolithic corpus database.

This TextBase architecture underpins a corpus development, analysis and publishing methodology that provides significant flexibility in terms of the iteration and synchronization of three fundamental workflows: text alignment, analysis registration and text aggregation.

The text alignment process integrates image, text and model configuration data to automatically generate a database. This approach allows for the distribution of responsibility to specialists and integration of their research output to align the image and the transcription at their most atomized to generate a 'substrate'. Rather than requiring researchers to command exacting markup schemas, substrate databases can be automatically generated from Word processing and Spreadsheet inputs. Workbench enables each of the specialist roles to work independently and their contributions be separately managed and integrated, ameliorating project risk by minimizing dependencies and bottlenecks.

The analysis registration process synchronizes analysis data with an existing substrate. This approach allows a researcher to work independently and externally to READ in developing their own analysis 'strata' and then register that strata on a substrate. Grammatical analysis, translation, semantic, syntactic and structural analysis can all be independently developed and iteratively registered. Researchers from other disciplines can develop and register their own analysis (archaeological, historical etc.). Each stratum is registered on a particular substrate (an edition) of the text within a TextBase, is separately owned and attributed, and its visibility is controlled by the researcher registering it.

The text aggregation process allows individual researchers to work and innovate in private to the point where they elect to participate in research collaborations or their text is ready for publishing. A TextBase might be aggregated with others to form a 'sequenced' collection, a 'mapped' collection or a 'merged' corpus; a continuum expressing an increasing degree of synthesis and harmonization of analysis ontologies and methodological standards. Researchers may contribute their TextBase to any number of aggregates. This approach allows a researcher to align a TextBase with the analysis standards of an established corpus as a predicate to participation as a constituent of that merged aggregate. In parallel, that same TextBase might be mapped to the analysis ontology of an entirely different collection. The potential exists for the same TextBase substrate to manifest as a constituent of separate aggregates, with alternative configurations of registered analysis strata, supporting widely divergent (aggregate specific) research objectives; the emanation of multiple TextBase avatars.

The strategy adopted with Workbench was to design a solution architecture within which to reframe some of the ubiquitous issues in the conventional corpus development model; ownership, control, confidentiality, innovation, standardization, portability, resourcing and support. The critical innovation in maximizing development flexibility and in balancing autonomy and collaboration across the range of individual, collection and corpora development projects is the TextBase; the target of text alignment, the substrate for registration of analysis and the object aggregated.

Preserving and Visualizing Queer Representation in Video Games

Cody Jay Mejeur
cmejeur@gmail.com
Michigan State University, United States of America

The nascent field of queer game studies has expanded exponentially in recent years thanks to the work of scholars such as Adrienne Shaw, Bonnie Ruberg, and Edmond Chang. This growth in scholarship has paralleled a significant rise in LGBTQ representation in games, including games such as *Gone Home*, *The Vanishing of Ethan Carter*, *Dream Daddy*, and others. Yet, despite growing representation and scholarly attention to queer characters and players, queer game studies continues to face the multivalent marginalizations of queer folks and their experiences in gaming. A prime example of this marginalization is the difficulty of preserving queer gaming cultures: queer representations and gaming communities are recorded largely in ephemeral, unofficial digital forms such as wikis, blogs, and fan-made websites due to a lack of access to mainstream platforms that often minimize and reject queer perspectives and desires. There is some advantage to these forms in that they allow queer gamers to create online communities as "counterpublics," which are communities defined against normative rules and expectations, but this means that queer gaming cultures are also in constant danger of being ignored, becoming outdated, or disappearing suddenly due to lack of resources (Warner, 2002).

A case in point is GayGamer.net, a website dedicated to game news, commentary, and community for LGBTQ gamers that went dark without notice in May 2016. GayGamer.net was a valuable resource for documenting LGBTQ game characters and communities, and while parts of it were captured by the Internet Archive, much of the site is no longer accessible outside of an old Facebook

page (GayGamer.net). While many digital objects face similar issues of compatibility and archiving, queer game artifacts and documentation are especially endangered because of the marginalized status of queer gamers and characters in gaming culture. With fewer individuals (almost all volunteers) and institutional resources to support them, these sources must be actively preserved now before they—and crucial LGBTQ cultural heritage with them—are lost.

The LGBTQ Video Game Archive, founded by Adrienne Shaw at Temple University, was created to address these issues by collecting LGBTQ representations from the 1980s to the present in order to "offer a record of how characters are explicitly coded, what creators have said about these characters, as well as how fans have interpreted these characters" (Shaw). The archive aims to allow easy, comprehensive access to queer gaming sources for queer game scholars, queer gamers, and the interested public, and further to ensure that these sources remain available in the future. To this end, one of the archive's current projects is an ongoing preservation effort in association with the Strong National Museum of Play that seeks to save copies of the many online media artifacts that document queer gaming cultures. The archive's preservation project demonstrates how archiving can be used to further social justice projects in digital spaces by safeguarding the cultural productions (including personal blogs, community forums, wikis, etc.) of marginalized peoples. As part of this presentation, I will share the process I developed for collecting and storing the websites, images, and videos referenced in the archive, and then transferring these materials to the Strong for permanent storage and public access. Using a combination of browser plugins and command line tools such as wkhtmltopdf and youtube-dl, the sources are saved as common file types that are entered into an Omeka database. The database allows the Strong to make the files publicly accessible, and the common file types allow for easier maintenance and curation of the collection. This process could be of use to other scholars and activists working to collect, curate, and sustain digital cultural resources, especially those significant to marginalized communities.

By preserving this cultural heritage, the LGBTQ Video Game Archive allows for new analyses of queer gaming culture and representation that highlight ongoing issues and emerging possibilities in games. For example, Utsch et al. used the archive to create data visualizations of queer representation throughout video game history, and revealed several trends such as a predominance of gay men in LGBTQ representation and an exponential growth in overall number of representations (Utsch et al., 2017: 7). To date, however, an intersectional analysis of the archive that addresses sexuality alongside identity categories of race, class, or disability has not been attempted, and this paper presentation will address these intersections using new interactive data visualizations. Completing these visualizations required revisiting each representation in the archive and recording additional data about the character's identity. The visualizations are interactive in order to make them more fluid and dynamic: in other words, to make them better representations of identity than the static categorizations that intersectionality has sometimes been accused of (Puar, 2005: 125). This intersectional analysis of the archive is only the beginning of the archive's potential, and it has a number of limitations. For example, it only includes games currently in the archive, and only what is observable and documented about each representation. Future work will add more games to the analysis, and provide more granular analysis of particular genres, developers, and intersectional identities in games. Together, preservation and critical analysis are essential tools for developing archival practices that support social justice in digital humanities, and both are much needed forms of public, academic, community-oriented activism.

In sharing the LGBTQ Video Game Archive's ongoing efforts to preserve and visualize queer representation in games, this paper presentation calls for increased attention in digital humanities to the needs of marginalized groups such as queer gaming communities. Concepts and design practices such as imagining a QueerOS can help guide the field's attempts at better inclusion, but we as digital humanities scholars can and must do more (Barnett et al., 2016). As we build and make with our digital tools, we must constantly confront the question of who we are building and making for. I argue that digital humanities should be the digital theories and practices of social justice, and it should do the crucial work of engaging with communities and supporting their efforts to make and shape themselves. Representation in queer games and queer gaming communities provides some practical methods for doing so, and contributes to ongoing discourse of what digital humanities can be.

References

Barnett, F., Blas, Z., Cárdenas, M., Gaboury, J., Johnson, J. M. and Rhee, M. (2016). QueerOS: A User's Manual. In Gold, M. K. and Klein, L. F. (eds), *Debates in the Digital Humanities: 2016*. University of Minnesota Press.

Chang, E. (2017). Queergaming. In Shaw, A. and Ruberg, B. (eds), *Queer Game Studies*. University of Minnesota Press, pp. 15–24.

Condis, M. (2015). No homosexuals in Star Wars? BioWare, 'gamer' identity, and the politics of privilege in a convergence culture. *Convergence*, **21**(2): 198–212 doi:10.1177/1354856514527205.

GayGamer.net. Social Media. *Facebook*. https://www.facebook.com/gaygamernet/?ref=br_rs. (accessed 28 Nov. 2017).

Gold, M. K. and Klein, L. F. (eds). (2016). *Debates in the Digital Humanities: 2016*. Minneapolis London: University of Minnesota Press.

Malkowski, J. and Russworm, T. M. (eds). (2017). *Gaming Representation: Race, Gender, and Sexuality in Video Games*. (Digital Game Studies). Bloomington: Indiana University Press.
Puar, J. K. (2005). Queer Times, Queer Assemblages. *Social Text*, **23**(3-4-85): 121–39 doi:10.1215/01642472-23-3-4_84-85-121.
Shaw, A. (2014). *Gaming at the Edge: Sexuality and Gender at the Margins of Gamer Culture*. Minneapolis: University of Minnesota Press.
Shaw, A. *LGBTQ Video Game Archive*. https://lgbtqgamearchive.com. (accessed 28 November 2017.
Shaw, A. and Friesem, E. (2016). Where is the queerness in games?: Types of lesbian, gay, bisexual, transgender, and queer content in digital games. *International Journal of Communication*, **10**: 13.
Utsch, S., Braganca, L. C., Ramos, P., Caldeira, P. and Tenorio, J. Queer Identities in Video Games: Data Visualization for a Quantitative Analysis of Representation.
Warner, M. (2002). *Publics and Counterpublics*. New York: Zone Books.
Warner, M. and Social Text Collective (eds). (1993). *Fear of a Queer Planet: Queer Politics and Social Theory*. (Cultural Politics v. 6). Minneapolis: University of Minnesota Press.

Segmentación, modelado y visualización de fuentes históricas para el estudio del perdón en el Nuevo Reino de Granada del siglo XVIII

Jairo Antonio Melo Flórez
jairom@colmich.edu.mx
El Colegio de Michoacán, Mexico

Introducción

Una de las características de la cultura jurídica del antiguo régimen era la evidente polisemia de sus conceptos (Hespanha, 2002). Términos que actualmente pertenecen al plano teológico, moral, histórico y literario; tenían un valor normativo dentro del lenguaje jurídico que afectaba la práctica del gobierno y la justicia. Como explica Alejandro Agüero, la perspectiva crítica develó una lógica del orden natural del mundo y del poder político en la que el derecho escrito si bien no es irrelevante tampoco cumple el papel protagónico en la organización de las ciudades y los reinos (Agüero Nazar, 2012: 84).

El giro metodológico, de la exégesis a la auscultación del pensamiento jurídico a través de los conceptos centrales del discurso normativo desde la edad media europea, ha implicado cruzar las fronteras de la historia del derecho como campo especializado del estudio del derecho para entrar a discutir con la disciplina histórica. La hermenéutica jurídica trasciende por lo tanto el ejercicio de la interpretación del texto para remitirse al "sentido" (*sinn*) del derecho en un enfoque cercano a la filosofía hermenéutica (Costa, 1972: 46), cuya derivación más conocida por los historiadores la representan Gadamer y Koselleck (1997).

Con este trabajo pretendemos explorar las posibilidades que brindan los métodos de análisis computacional para la historia de los conceptos y en general del lenguaje jurídico-político anterior al siglo XIX, en el entendido que el perdón permite analizar un elemento fundamental del poder político de la Edad Moderna (Foucault, 1975: 56–57), así como el proceso de secularización y de pretendida tecnificación del indulto en el marco del proyecto de modernización legislativa decimonónico (Prodi, 2000).

La construcción del corpus: modelado básico de la información.

El corpus textual objeto de esta muestra está compuesto de documentos seleccionados de fondos de justicia y gobierno de archivos españoles y colombianos. Al no existir una serie documental consistente para el problema del perdón, fue necesario realizar una exploración y recolección de documentación en diversos repositorios que permitiera representar el universo de la clemencia en el ámbito del virreinato del Nuevo Reino de Granada.

Con el propósito de estructurar la información, se aprovechó el entorno Omeka para facilitar tanto la transcripción como la asignación de metadatos al contenido y a los elementos (Melo Flórez, 2016). Se identificaron distintos tipos documentales que se agruparon en cédulas, peticiones y concesiones de indulto, legislación, doctrina, prisiones, perdones particulares, expedientes judiciales y biografías. Para estos dos últimos elementos se construyó un tipo de elemento (ítem-type) con lo cual se puede recuperar información específica como suplicaciones, vistas fiscales, sentencias, alegatos de defensores, testimonios.

La transcripción de los documentos se realizó de manera tradicional intentando conservar el valor fonético o literal de las fuentes, por lo cual el texto digitalizado no fue modernizado en su ortografía ni en la acentuación. Un problema consistió en el desarrollo de abreviaturas, las cuales por lo general se indican haciendo uso de corchetes, por ejemplo, N^{vo} R^{no} de Granada se desarrolla como N[ue]vo R[ei]no de Granada. Por el momento se ha optado por imitar la etiqueta <expan> del modelo TEI del modo [expan = Nuevo Reino de Granada] con lo cual se adelanta la identificación de algunos elementos semánticos y por otra parte solventa la lectura automática del texto. La misma operación se realiza con etiquetas como <abbr>,

<gloss>, <note>, <corr>, <sic>, <placeName>, <geo>, <textLang mainLang> y <name type>.

Finalmente, la información se recuperó mediante la función *metadata* de Omeka que permite seleccionar entre diferentes tipos de metadatos para luego exportarlos en HTML y convertirlos a texto plano (Turkel and Crymble, 2012). Con el propósito de visualizar el cambio de significado el corpus se segmentó en seis grupos temporales: 1739-1775, 1776-1789, 1790-1807, 1808-1818, 1819-1829, 1830-1842.

Segmentación

Antes del análisis textual, el texto requiere ser *tokenizado*, es decir, segmentado en unidades lingüísticas con la intención de conocer las métricas de las fuentes (Mikheev, 2005). Este proceso tiene el propósito de agrupar caracteres alfanuméricos en palabras, diferenciar tipos (número de palabras diferentes en un corpus), la frecuencia de cada palabra representada como *tokens*, y aplicar el proceso de *stemming* (reducir las palabras a su raíz) y la lematización (formas flexionadas de una palabra). Por lo tanto, en esta etapa, el análisis se reduce a la estructura básica del texto, su construcción y la medición del peso de sus elementos (Jockers, 2013: 4). El resultado se presenta en la tabla 1, aunque el primer segmento (1700-1775) revela la disparidad temporal respecto a las demás divisiones, por lo cual se comprende deberá corregir esta discrepancia en un futuro ejercicio. Los periodos con mayor cantidad de tokens están representados por aquellas etapas más convulsas del periodo: la rebelión de los comunes de 1781 y el proceso de revolución e independencia desde 1808 hasta 1830.

Corpus	Tokens	Types	Lemmas
1700-1775	112136	15514	15514
1808-1818	93301	12198	12538
1776-1790	80015	11438	11714
1819-1830	51548	8510	8605
1830-1842	32968	6578	6736
1790-1808	25885	5932	6043

Tabla 1. Resultados del proceso de segmentación del corpus por segmentos

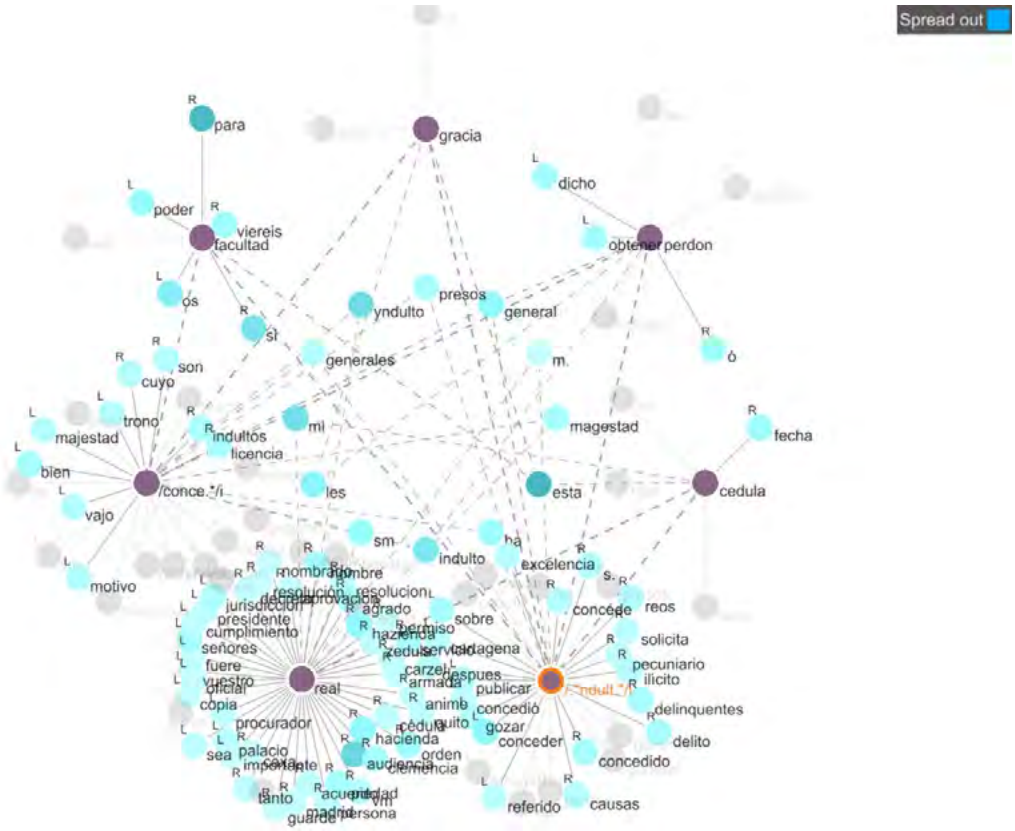
La abundancia de tipos y lemas se deriva de la cantidad de variaciones que el software no tiene la capacidad de deducir formas de una misma palabra, por ejemplo, indulto e yndulto es leído como dos vocablos separadas. La manera más simple de solucionar este inconveniente consiste en modernizar la ortografía de las expresiones arcaicas, sin embargo, esto disociaría el corpus de las fuentes doctrinales y legales impresas, cuya información se recupera por técnicas de OCR. En este caso nos remitimos nuevamente a la representación de grafemas, tarea propia de la paleografía, y su uso por parte de escribanos en la Edad Moderna, así como las posibilidades de semi-automatización y estandarización de esta tarea.

Análisis y visualización

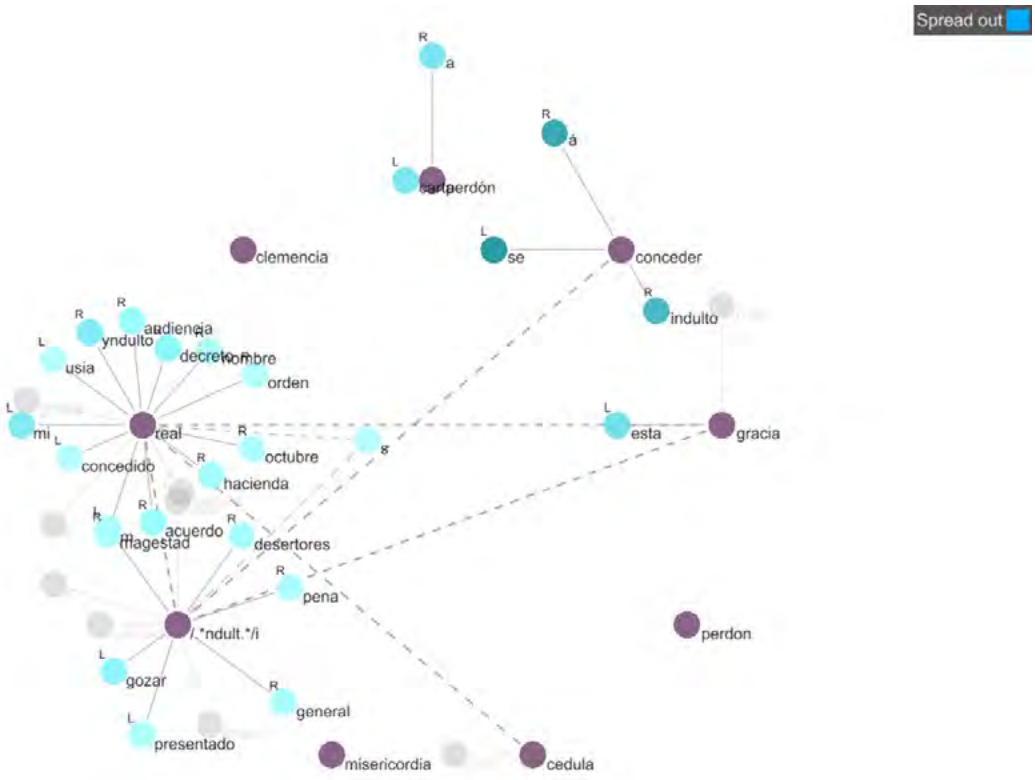
La colocación es un fenómeno léxico que da cuenta de las unidades fraseológicas más allá de las locuciones con significado propio. Su interpretación se fundamenta en la frecuencia estadística con la cual ciertos vocablos se relacionan entre sí y cuál es la relevancia de dicha asociación (Alonso Ramos, 1995: 9–28). En este sentido, consideramos esta es una de las estrategias de la lingüística que mejor podemos aprovechar para percibir un posible cambio semántico (Pazos-Bretaña, 2016). Para realizar el análisis de colocación nos servimos de la herramienta informática *LancsBox* (Brezina et al., 2015), así como de las propuestas metodológicas del lingüista Paul Baker (2016: 142–48). Cada segmento del corpus fue interpretado con la opción "GraphColl" del mencionado programa, la cual se configuró con una estrategia estadística MI (*mutual information statistic*) que favorece las relaciones léxicas entre palabras evitando al mismo tiempo artículos de uso frecuente como "el", "la" o "de". Se utilizó la extensión de análisis estándar de cinco palabras hacia la izquierda y la derecha del término.

El resultado de cada segmento se asemeja al presentado en la gráfica 1, en el cual se despliegan los valores más significativos de la colocación. En este ejemplo, el término indulto (representado con caracteres comodín para solventar los problemas de *semmatization* y lematización) se despliega en una red que comprende en un primer nivel los términos real, facultad, gracia, perdón, cédula, delito, reos, presos, concesión, entre otros. Todos estos son términos que coinciden con el discurso tradicional del perdón real que dominó durante la Edad Moderna castellana (Rodríguez Flores, 1971; Sandoval Parra, 2014).

Si se compara con la gráfica 2, el vocablo perdón es reemplazado por el nombre gracia, separándose completamente los términos perdón, clemencia y misericordia del nombre indulto, aunque siguen formando parte de las impetraciones y de las cartas de perdón de parte. Esto parece indicar que en el lenguaje de la práctica jurídico-política hubo un fenómeno de metonimia en el cual la gracia reemplazó al perdón, en este caso, gracia era equivalente a perdón pero no a indulto (por ello se añadiría la conjunción copulativa "indulto y gracia").



Gráfica 1. Red semántica centrada en el concepto indulto (1739-1775). R5-L5, MI(5)



Gráfica 2. Red semántica centrada en el concepto indulto (1790-1807). R5-L5, MI(5)

Este ejercicio abarca varios procesos relevantes para el análisis de corpus de información histórica no estructurada. Con el avance de la digitalización de fuentes documentales en archivos y bibliotecas en ambos lados del Atlántico la tarea de recuperación y macroanálisis de la información se hace más compleja, por lo cual es necesario ya no sólo introducir metodologías computacionales utilizadas en contextos anglosajones, sino construir estrategias propias que permitan lidiar con una tradición paleográfica y archivística particular.

Las tareas inmediatas que se plantean para este proyecto incluyen el resolver la transcripción de documentos para lo cual se está explorando la plataforma Omeka S, así como la posible exportación de elementos y modelarlos en XML-TEI. Del mismo modo, se pretende mejorar el modelo de segmentación construyendo una estrategia para resolver las disparidades grafológicas. Se espera que estos ejercicios en un futuro pueden ser relevantes para los proyectos de digitalización actuales en Latinoamérica.

Referencias

Agüero Nazar, A. (2012). Historia política e Historia crítica del derecho: convergencias y divergencias. *PolHis*, 5(10): 81–88.

Alonso Ramos, M. (1995). Hacia una definición del concepto de colocación: de J. R. Firth a I. A. Mel'ĉuk. *Revista de Lexicografía*, 1: 9–28.

Baker, P. (2016). The shapes of collocation. *International Journal of Corpus Linguistics*, 21(2): 139–64 doi:10.1075/ijcl.21.2.01bak.

Brezina, V., McEnery, T. and Wattam, S. (2015). Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics*, 20(2): 139–73 doi:10.1075/ijcl.20.2.01bre.

Costa, P. (1972). Semantica e storia del pensiero giuridico. *Quaderni fiorentini per la storia del pensiero giuridico moderno*, 1(1): 45–87.

Foucault, M. (1975). *Surveiller et punir: naissance de la prison*. Paris: Gallimard.

Gadamer, H.-G. and Koselleck, R. (1997). *Historia y hermenéutica*. (Ed.) Villacañas, J. L. & Oncina Coves, F. Barcelona: Paidós.

Hespanha, A. M. (2002). *Cultura jurídica europea: síntesis de un milenio*. (Trans.) Soler, I. and C. Valera Madrid: Tecnos.

Jockers, M. L. (2013). *Macroanalysis: Digital Methods and Literary History*. (Topics in the Digital Humanities). Urbana: University of Illinois Press.

Melo Flórez, J. A. (2016). Metadatos *Cibercliografía* http://ciberclíografia.org/manuales/crear-un-fichero-de-investigacion-con-omeka/metadatos/ (accessed 28 April 2018).

Mikheev, A. (2005). Text Segmentation. In Mitkov, R. (ed), *The Oxford Handbook of Computational Linguistics*. Oxford: Oxford University Press http://www.oxfordhandbooks.com/view/10.1093/oxfordhb/9780199276349.001.0001/oxfordhb-9780199276349-e-10 (accessed 25 November 2017).

Pazos-Bretaña, J.-M. (2016). El efecto de la historia sobre el cambio semántico en el español peninsular. *Itinerarios: revista de estudios lingüisticos, literarios, históricos y antropológicos*(23): 123–39.

Prodi, P. (2000). *Una Storia Della Giustizia: Dal Pluralismo Dei Fori Al Moderno Dualismo Tra Coscienza e Diritto*. (Collezione Di Testi e Di Studi). Bologna: Il mulino.

Rodríguez Flores, M. I. (1971). *El perdón real en Castilla (siglos XIII-XVIII)*. Salamanca: Universidad de Salamanca.

Sandoval Parra, V. (2014). *Manera de Galardón: Merced Pecuniaria y Extranjería En El Siglo XVII*. (Sección de Obras de Historia). Madrid: Fondo de Cultura Económica : Red Columnaria.

Turkel, W. J. and Crymble, A. (2012). From HTML to List of Words (part 2). *Programming Historian* https://programminghistorian.org/lessons/from-html-to-list-of-words-2 (accessed 28 April 2018).

Part Deux: Exploring the Signs of Abandonment of Online Digital Humanities Projects

Luis Meneses
ldmm@uvic.ca
Electronic Textual Cultures Laboratory - University of Victoria, Canada

Jonathan Martin
jonathan.d.martin@kcl.ac.uk
King's College London, United Kingdom

Richard Furuta
furuta@cse.tamu.edu
Center for the Study of Digital Libraries, Texas A&M University, United States of America

Ray Siemens
siemens@uvic.ca
Electronic Textual Cultures Laboratory - University of Victoria, Canada

Introduction

Building online research components for projects in the digital humanities is a common practice. However, not many researchers have a plan for these online components once the project halts or comes to an end. Consequently, many of these projects become abandoned and

slowly degrade over time –some more gracefully than others. Additionally, there is a certain inherent fragility associated with software and our online research tools. In turn, this fragility threatens the completeness and the sustainability of our work over time.

Previous studies have attempted to harness and manage the fragility of online resources. Studies have been carried out to address their potential reconstruction (Klein et al., 2011), the overall decay of websites (Bar-Yossef et al., 2004) and the decomposition of their shared resources (SalahEldeen and Nelson, 2012). Recently, our research has been focusing on analyzing the perceptions of change in distributed collections (Meneses et al., 2016)NY, USA","page":"273–278","source":"ACM Digital Library","event-place":"New York, NY, USA","abstract":"It is not unusual for documents on the Web to degrade and suffer from problems associated with unexpected change. In an analysis of the Association for Computing Machinery conference list, we found that categorizing the degree of change affecting digital documents over time is a difficult task. More specifically, we found that categorizing this degree of change is not a binary problem where documents are either unchanged or they have changed so dramatically that they do not fit within the scope of the collection. It is in part, a characterization of the intent of the change. In this paper, we present a case study that compares change detection methods based on machine learning algorithms against the assessment made by human subjects in a user study. Consequently, this paper will focus on two research questions. First, how can we categorize the various degrees of change that documents endure? And second, how did our automatic detection methods fare against the human assessment of change in the ACM conference list?","URL":"http://doi.acm.org/10.1145/2914586.2914628","DOI":"10.1145/2914586.2914628","ISBN":"978-1-4503-4247-6","author":[{"family":"Meneses","given":"Luis"},{"family":"Jayarathna","given":"Sampath"},{"family":"Furuta","given":"Richard"},{"family":"Shipman","given":"Frank"}],"issued":{"date-parts":[["2016"]]},"accessed":{"date-parts":[["2017",4,12]]}}], "schema":"https://github.com/citation-style-language/schema/raw/master/csl-citation.json"} . However, we believe that the inherent characteristics of online digital humanities projects present an interesting (and unique) area for inquiry for two reasons. First, the research aspect of digital humanities projects hinders previous approaches –as the methods for identifying change in the Web do not fully apply. And second, digital humanities projects have a limited useful life –which is accompanied by research from primary investigator, which may or may not be indicated by updates in the project's content and tools.

We presented a paper in Digital Humanities 2017 that explored the abandonment and the average lifespan of online projects in the digital humanities (Meneses and Furuta, 2017). However, we believe that managing and characterizing the online degradation of digital humanities projects is a complex problem that demands further analysis. In this abstract, we propose to explore further the distinctive signs of abandonment of online digital humanities projects. For this second instalment of our study we took a different direction: we departed from strictly using retrieved HTTP response codes and incorporated additional metrics such as number of redirects, DNS metadata and a detailed analysis of content features.

This study aims to answer four questions. First, can we identify abandoned projects using computational methods? Second, can the degree of abandonment be quantified? Third, what features are more relevant than others when identifying instances of abandonment? Our final question is philosophical: can an abandoned project still be considered a digital humanities project?

Methodology

A complete listing of research projects in the Digital Humanities does not exist. However, the Alliance of Digital Humanities Organizations publishes a Book of Abstracts after each Digital Humanities conference as a PDF. Each one of these volumes can be treated as a compendium of the research that is carried out in the field. To create a dataset, we downloaded the Books of Abstracts corresponding from 2006 to 2016 –except for 2015 which was not available for download. We must thank and acknowledge Dr. Jason Ensor from Western Sidney University for providing us the abstracts for the 2015 Digital Humanities conference –which completes our dataset of conference abstracts. We obtained these abstracts after we had carried out our preliminary analysis. Therefore, we will present our findings using the complete dataset of abstracts in the presentation of our paper.

Then we proceeded to extract the text from these documents using Apache Tika and parse the 5845 unique URLs that we found using regular expressions. Then we used Python's Requests Library to retrieve the HTTP response codes and headers corresponding to the URLs, which we used to classify the websites into two groups depending on their correctness: valid (correct) and decayed (showing signs of degradation). Figure 1 shows the distribution of decay for each year. Based on our preliminary findings we approximate the average lifespan of a research project to 5 years, which aligns with reports from previous work (Goh and Ng, 2007). The average time in years since the last modification of the websites in the study is shown in figure 2.

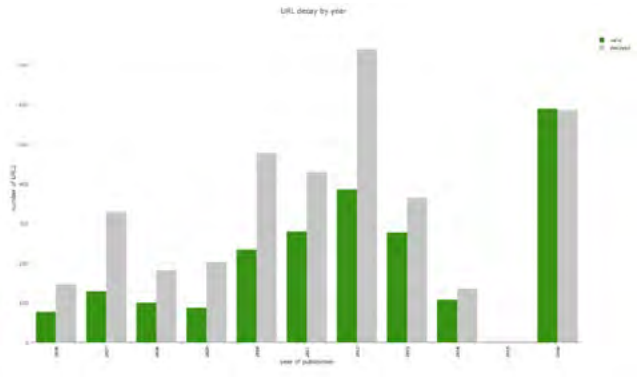


Figure 1: URL decay by year.

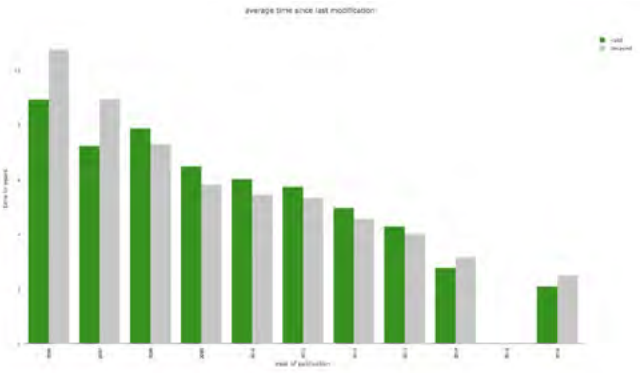


Figure 2: Average time in years since last modification.

Developing classifiers

To develop classifiers for the degradation identified in the previous section, we considered features computed based on DNS metadata, the initial HTTP request, number of redirects, and the contents and links returned by traversing the base node. The features we included are divided into topology, content-type, anchor-text and child node features. These features stem from concepts we used in our previous work (Meneses et al., 2016)NY, USA","page":"273–278","source":"ACM Digital Library","event-place":"New York, NY, USA","abstract":"It is not unusual for documents on the Web to degrade and suffer from problems associated with unexpected change. In an analysis of the Association for Computing Machinery conference list, we found that categorizing the degree of change affecting digital documents over time is a difficult task. More specifically, we found that categorizing this degree of change is not a binary problem where documents are either unchanged or they have changed so dramatically that they do not fit within the scope of the collection. It is in part, a characterization of the intent of the change. In this paper, we present a case study that compares change detection methods based on machine learning algorithms against the assessment made by human subjects in a user study. Consequently, this paper will focus on two research questions. First, how can we categorize the various degrees of change that documents endure? And second, how did our automatic detection methods fare against the human assessment of change in the ACM conference list?","URL":"http://doi.acm.org/10.1145/2914586.2914628","DOI":"10.1145/2914586.2914628","ISBN":"978-1-4503-4247-6","author":[{“family":"Meneses","given":"Luis"},{“family":"Jayarathna","given":"Sampath"},{“family":"Furuta","given":"Richard"},{“family":"Shipman","given":"Frank"}],"issued":{“date-parts":[[“2016"]]},"accessed":{“date-parts":[[“2017",4,12]]}}],"schema":"https://github.com/citation-style-language/schema/raw/master/csl-citation.json"} .

The text associated with resources is the most obvious feature for determining the topics. Given that we are dealing with a very specialized domain, we developed a domain-oriented expectation model. In particular, we generated topic and term frequency models to examine the similarity among the documents in a given project (the contents of the base node and the metadata and the contents of the child nodes). We used Latent Dirichlet Allocation to model the content of the text (Blei et al., 2003) and a simple Tf-Idf ranking function to measure and compare them. This ranking function is based on adding the Tf-Idf values for the documents, which were calculated using the terms from the topic modelling as a vocabulary. We will present a detailed version of our results on the longer version of our paper.

Discussion

This study is an attempt to categorize change in a very specific domain. More so, this study constitutes one step towards addressing potential strategies for the archival and the long-term preservation of abandoned digital projects. It is important to highlight that not all projects are equal and thus require different approaches towards long-term preservation. In the case of dynamically generated projects, a common approach nowadays is to produce a static set of HTML files which are easier to store. However, this approach assumes the backwards compatibility of Web browsers over time –something that has not always been the case.

To summarize, in this study we aim to computationally identify the indicators of the abandonment of digital humanities projects –as well as quantify their degrees of neglect. To address our philosophical question, we believe that an abandoned project can still be considered a valid digital humanities project depending on its audience. However, this has several nuances that should be considered. Digital online projects in the humanities have unique characteristics that make them impervious to the metrics that used in the Web as a whole –which make them worthy of study. In the end, we intend this study to be a step forward towards better preservation strategies and for the planned obsolesce of digital humanities projects.

References

Bar-Yossef, Z., Broder, A. Z., Kumar, R. and Tomkins, A. (2004). Sic transit gloria telae: towards an understanding of the web's decay. *ACM* doi:10.1145/988672.988716.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3: 993–1022.

Goh, D. H. and Ng, P. K. (2007). Link decay in leading information science journals. *Journal of the American Society for Information Science and Technology*, 58(1): 15–24.

Klein, M., Ware, J. and Nelson, M. L. (2011). Rediscovering missing web pages using link neighborhood lexical signatures. ACM doi:10.1145/1998076.1998101.

Meneses, L. and Furuta, R. (2017). Shelf life: Identifying the Abandonment of Online Digital Humanities Projects Paper presented at the *Digital Humanities 2017*, Montreal, Canada.

Meneses, L., Jayarathna, S., Furuta, R. and Shipman, F. (2016). Analyzing the Perceptions of Change in a Distributed Collection of Web Documents. *Proceedings of the 27th ACM Conference on Hypertext and Social Media*. (HT '16). New York, NY, USA: ACM, pp. 273–278 doi:10.1145/2914586.2914628. http://doi.acm.org/10.1145/2914586.2914628 (accessed 12 April 2017).

SalahEldeen, H. M. and Nelson, M. L. (2012). *Losing My Revolution: How Many Resources Shared on Social Media Have Been Lost?*

A People's History? Developing Digital Humanities Projects with the Public

Susan Michelle Merriam
merriam@bard.edu
Bard College, United States of America

In this short paper I will explore some of the problems—particularly those having to do with power and access—inherent in collaboratively produced community digital history projects. I will focus on two projects currently in development in which I (working from within an academic institution and digital media lab) have partnered with people from marginalized populations located in geographic areas that have been given relatively little attention. One of my goals in initiating these projects has been to explore how to use institutional resources, including grants and IT support, to work outside of institutional structures. In each instance, my community partners and I have created projects centered on individual, personal narratives as they relate to place. Our objective has been to develop a kind of "people's history," giving voice to those who have traditionally been excluded from historical research and writing. In the course of conceptualizing and beginning to make these projects, however, we've encountered a number of thorny questions about notions of community, access, and narrative form and content.

Conceptually, these projects have been fundamentally enabled by digital technologies that allow new makers to produce historical narratives. Indeed, digital media has fed an emerging industry in small-scale, creative historical projects, many of which academic historians would term "micro histories." Explored perhaps most famously by Carlo Ginzburg, micro histories can be viewed as correctives to "great man" theories of history or macro narratives that are easily undermined when challenged by specific circumstances. Focusing on seemingly "small" events—a day in an individual's life, for example—micro histories often make transparent the point of view of the researcher, thus destabilizing hegemonic forms of historical writing. Micro histories can also bring attention to, or use, lacunae in the historical record, as well as offer narrative forms for "regular" people to engage in the construction of history.

Working with digital tools and a loose concept of micro history, last spring I founded Bard College's "Mobile History Van," which operates under the umbrella of Bard's Digital History Lab (http://eh.bard.edu/dhl/). Both are funded by a major grant from the Mellon Foundation. The Mobile History Van uses digital technology to record and publish local history, and has worked closely with a local library and museum on digitizing archival materials and recording community history. While these projects have excavated important aspects of the historical record, they were executed with institutional partners—a college and historical society—and are thus still inscribed in easily recognized power structures.

In pursuit of developing projects outside of institutional structures, I approached students from Bard's Clemente program (https://www.clementecourse.org), a college credit granting year-long course for people who earn below a certain income level. Many of these students are struggling with substance abuse issues, criminal records, and post-traumatic stress disorders, but they wish to find ways to engage in the world. My intention: to work together with them and develop digital projects from the ground up, including working with other people who might not consider themselves part of the community described by the local historical society.

The first part of this talk will briefly introduce the audience to the genesis of the projects, and to their current state. By the time of the conference, both projects will be near completion, one as a series of personal narratives mapped onto the city of Kingston NY, the other in the form of a podcast about storytelling.

The second half of this talk will examine a series of thorny questions we have encountered in the process of our work: Who controls the projects? What role does the institution play in supporting the projects, and how does

this institutional affiliation shape the outcome in each case? How do we develop the projects for maximum input from collaborative partners? As we are working outside of an institutional structure, how do we define "community"? When should we, or is it necessary to, protect the storyteller? And finally, what does this type of project reveal about access to digital media?

Peer Learning and Collaborative Networks: On the Use of Loop Pedals by Women Vocal Artists in Mexico

Aurelio Meza
meza.aurelio@gmail.com
Concordia University, Canada

PoéticaSonora is a research group interested in the study of sound, listening, and legibility at the intersection of art and literary studies. The group has worked together for two years, organizing several events and projects that operate under two main axes, activation and preservation. The most important project on the preservation axis is dedicated to the design, creation, and development of a digital audio repository (DAR) for sound art and sound poetry in Latin America. During the first phase of the project to gather audio files for the repository, we have conducted fieldwork and archival research in different public centers and private collections, mostly in Mexico City. The DAR prototype was designed and developed by graduate students at Concordia University, in Montreal, with an initial sample of 317 audio tracks, performed or composed by around 180 artists, most of them from Mexico but also Argentina, Brazil, and the US. These tracks have previously been classified as sound art, sound poetry, radioart, experimental music, spoken word, poetry slam, performance, hip hop, indigenous language poetry, among many other terms.

The interest of PoéticaSonora members has focused on studying audio recordings of poetry readings, as well as organizing and curating sound art and experimental music events. This presentation, however, studies how musical instruments and other sound-generating devices accompany, even modify the human voice, and how the DAR contributes to understand the understanding of these works and the context where they are developed. Texts by scholars who have taught, studied, and/or conducted fieldwork in Montreal, such as Mark J. Butler, Jeremy Wade Morris, Tara Rodgers, and Jonathan Sterne will serve as a theoretical framework to discuss some artistic practices by Mexican women vocal artists who participate in collaborative creative networks (sometimes called "bands," "collectives," "jams," among other labels) and use sound-generating devices as a fundamental element in their performance. As a case study, I will focus on the path of two such artists, Edmée García and Leika Mochán, who combine spoken word and singing with the use of loop pedals. For several years they collaborated on the LP *Frágil*, along with jazz songwriter Iraida Noriega, and have ever since worked in different creative projects, both solo and with other artists. Some of the pieces where they use loop pedals are also analyzed here, such as García's *Chilanga habla*, described by herself as a "piece for poet and Line6," and Mochán's "Kaleidocycle," consisting of an amplifier and a Line6 DL4 attached to a customized bicycle.

The work of García and Mochán contributes to the discussion about what is intuitive and what is not in the use and adaptation of digital devices to produce sound with artistic or aesthetic purposes. These artists generate their own learning networks, transmitting to each other the empirical knowledge they acquire from free experimentation with a device. García calls Mochán "the loop pedal guru," and learned from her and Noriega how to use it during the creation of *Frágil*. This experience completely shaped the way García would perform her next poetry book, *Chilanga habla*, up to the point of deciding not to publish the text-based version, as it did not portray the project's whole scope and shape. As for the "Kaleidocycle," it allows Mochán to interact with the audience in a direct way, and posits questions about the false distinction between liveness and recording, particularly at the moment of performance. The different paths followed by García and Mochán after *Frágil* are a good example of how knowledge is not always prefigurative (from an elder to a youngling), but also configurative (among peers) and sometimes even postfigurative (from a youngling to an elder). This presentation sheds light on how Mexican artists face a device's intended use and how their actual uses diverge and become mainstreamed within certain collaborative networks.

Frágil and some works by García have already been integrated to PoéticaSonora's DAR. The presentation will start with a brief showcase of how their collaborative networks are illustrated in the prototype, as well as the roles and instruments each participant plays in a particular composition. It will then discuss how to integrate new works by García and Mochán, how to possibly solve some of the prototype's limitations, and reflect upon the next steps in the project, considering the implications it may have on the prototype's data schema. As it stands, does the DAR help us visualize these collaborative networks? Is it necessary to have an entity for groups and collectives, or can it be inferred from other categories in the database? It will finally discuss a brief genealogy of loop pedals to understand how such a marginal guitar effects unit (like the Boss RC 20 or 30, Line6 DL4) evolved into a device for singers (Boss RC 500, but more specifically the TC Helicon series), and in so doing re-purposed this device. With these case studies I will explain how the functions delegated to the loop pedal allow these artists to overco-

me the fact of not being a "musician," even though both have a strong musical background, and to perform "solo" despite holding a creative relation with the loop pedal.

Next Generation Digital Humanities: A Response To The Need For Empowering Undergraduate Researchers

Taylor Elyse Mills
mills@hope.edu
Michigan State University, United States of America

Introduction

Integrating the digital humanities (DH) into undergraduate level higher education programs has often been a difficult and ambiguous process. Faculty sometimes struggle to create syllabi that incorporate technologies but that do not require constant redesign as technologies evolve. Institutions may lack systems to connect students with faculty and staff who are interested in collaborative research, and collaboration beyond one's own institution can be complicated or inaccessible for students. These are real challenges; as institutions increasingly develop DH courses and degrees, the impact on undergraduate students is diverse, ranging in minimal involvement, to career-altering. So, what should the role of the undergraduate in DH be, and how can we address these challenges? For the past three years I have explored these questions. This exploration has led to helping redesign and teach the foundational seminar for Hope College's Mellon Scholars DH Program, as well as co-founding and chairing the Undergraduate Network for Research in the Humanities (UNRH), an undergraduate-led organization with the mission of reimagining the undergraduate role in DH through the establishment of a network of digital humanists who present research, collaborate, and share ideas. On the basis of these experiences as an alumna of Hope's DH Program and UNRH Chair, I have been considering the ways in which faculty, staff, and institutions might support undergraduate DH researchers. My work has culminated in a series of models, programs, and initiatives that address the need for fostering the next generation of digital humanists in the classroom, at the institution, and beyond.

Method

Classroom

The first challenge I consistently identified in DH courses was an incohesive structure that treated the digital and the humanities as separate units rather than an interconnected academic space. Secondly, seminar themes grounded in particular technologies had to be redesigned frequently as these technologies evolved or became outdated. This was the case for the year-long introductory seminar for Hope's DH Program. Each year students felt that the seminar was two unrelated courses, one focusing on a particular area in the humanities, the other, teaching technologies like GitHub and data analysis. The course was a noble attempt but ultimately inconsistent, incohesive, and not a truly interdisciplinary approach to DH. I set about designing a seminar model that was adaptable to new technologies yet still focused on an intersectional theme. I consulted with educators at conferences and researched seminar formats at other institutions, but unsurprisingly there was a wide range of approaches that seldom emphasized independent research quite like Hope's program. Thus, I grounded the seminar model in that very aspect: a chronological approach to independent research in the humanities. Over course of four units students engage with the evolution of humanities-based research and with the research process from beginning to end. During the first unit, students work in the archives, practice cataloging primary sources with tools like Zotero, develop strong but focused research questions, and discuss literature to answer the ever-present question "What is DH?" The second unit follows the progression in humanities-based research, moving from sources like libraries and datasets into the first examples of DH: text analysis. Students curate their own text-based datasets, analyze and visualize them, present them with Omeka, and discuss research project methodologies of source compilation and argumentation. The third unit it titled: CCP-Collaboration, Communication, & Presentation. It involves group research collaboration and finalizing research projects through effective communication and presentation. Students complete writing workshops in which they must adapt a piece of writing for different audiences and styles, from conference abstracts to blogs and tweets; they also practice oral and web presentation skills. The final unit addresses advanced topics and tools which require students to focus on race, gender, sexuality, politics, and socioeconomic status. Students learn that equity and accessibility are paramount when creating public scholarship, digital or otherwise, and they are exposed to a survey of technologies in efforts to broaden their concept of what form research can take. The outcome of this course should be a comprehensive and diverse approach to humanities-based research projects through the chronological progression that research in the humanities has followed.

Institution

For collaborative research, students and faculty alike find it challenging to make necessary connections with one another in the four short years that students have on campus.

My solution is Bin(d)r: the Baccalaureate Interdisciplinary Network for (Digital) Research. Stemming from an initial idea of a physical binder with pages featuring the profiles of faculty, staff, and students interested in collaborative research, Bin(d)r: is ideally implemented as a searchable database of anyone on campus with research interests and skills. It is like Tinder for academics. All faculty and staff interested in collaborating simply create a profile on a site with tools like WordPress's "Ultimate Member" Plugin. Students are invited to create profiles if they are interested in research. By including specific research interests and skills, faculty and students can get "matched" in a timely manner. Bin(d)r: has parentheses around "digital" because this tool does not have to be exclusively for digital projects, but it would provide an extra level of support for digital projects, connecting computer science students with humanities faculty, for example. Bin(d)r: is capable of being entirely free, low maintenance, highly interdisciplinary, and ultimately a tool for encouraging undergraduate research. Furthermore, if the digital Bin(d)r: takes off at numerous institutions, searching others' databases would foster cross-institutional collaboration.

While considering the institutional level, I would also argue that institutions must make space to hear the voices of their students. I propose that institutions establish a quarterly forum for undergraduates, faculty, and administrators to gather and discuss how the institution can better support students. Academic institutions are designed first and foremost to educate their students, so I assert that students have the right to tell institutions how they can improve, and institutions have the responsibility to listen. Simply creating space for dialogue is empowering.

Beyond

I also argue that empowering undergraduate researchers means providing agency, accreditation, and opportunities to join a community. Because DH is emerging at different rates across the globe, many students never meet other students engaging in their work. Furthermore, exposure to different methodologies, technologies, and project ideas has a profound impact. Faculty and staff gain this exposure at academic conferences and within their departments. UNRH aims to give this space and community to students, too.

Our method of creating UNRH relied heavily upon initial organization, forming a Steering Committee, review system, and website. The format of our conference was meticulously designed. We created a "speed-dating" session for rapid introductions and elevator pitch practice, a formal project presentation session, informal poster-style presentation sessions, a keynote address, and workshop sessions. These workshops include technology tutorials, panel discussions about different students' roles and experiences at their institutions, and design-thinking sessions to address the needs and concerns of students striving to develop DH projects.

Beyond the conference we have been developing an online network space in which students create profiles and can share project updates, articles, conference opportunities, and requests for peer review. In essence, each of our decisions was an effort to create space and flexibility for students to answer for themselves the question of what the undergraduate role in DH can be.

Results

Classroom

The feedback from my students who experienced my seminar model have been positive. The survey results indicate that the seminar has largely met the learning outcome goals, and students indicated increases in confidence and preparedness in conducting independent research (approximately 30% average increase) and using new technologies (approximately 37% average increase) according to a seven-point scale. Those who indicated having less prior experience (1-4) had an average increase of about 33% in independent research and about 39% in technology use. I plan to track program retention rates in the coming years to hopefully see improvements as the sophomore students navigate from the structured seminar into the independent research spaces of their junior and senior years.

Institution

Bin(d)r: has not yet been implemented but is in development for implementation at Hope College in the coming year.

Beyond

The results of our efforts exceeded expectations. Since our first conference in 2015, we have accepted over 50 projects, involving over 80 undergraduates from 31 institutions all across the United States, Canada, Nigeria, and Pakistan. According to in-person comments and our post-conference evaluations, students have felt empowered, encouraged, and independent in their research. Moreover, students were amazed at what they learned and accomplished by interacting with undergraduates from other institutions.

Through our initial design and modifications over the years, we feel confident in the model for an organization and conference that grants agency to undergraduates, and space to understand their own roles. Now in my third year as Project Manager/Chair, when I consider again the undergraduate role in DH, I think of students as connected learners and independent researchers pursuing their own interests while learning from peers and mentors ali-

ke. Within and beyond this space, each student must determine her role for herself.

Instructors, institutions, and organizations, invest in these students, for they are the next generation of digital humanists.

La creación del Repositorio Digital del Patrimonio Cultural de México

Ernesto Miranda
mirandatrigueros@gmail.com
Secretaría de Cultura, Mexico

Vania Ramírez
vaniasa.ramirez@gmail.com
Secretaría de Cultura, Mexico

Introducción

La creación de la Secretaría de Cultura Federal del Gobierno de México, trajo consigo la creación de la Dirección General de Tecnologías de la Información y Comunicaciones, y con esta, el mandato de construir la Agenda Digital de Cultura. Dentro de las atribuciones que tiene la Dirección se encuentra la interoperabilidad de las colecciones digitales albergadas y administradas por la Secretaría de Cultura. Para poder responder a este mandato se ha puesto en marcha el desarrollo del "Repositorio Digital de Patrimonio Cultural de México (RDPCM)", primer esfuerzo de largo alcance y con visión integral desde el gobierno de México para integrar los acervos digitales de museos, bibliotecas, televisoras, radiodifusoras y diferentes instituciones culturales que son coordinadas por la Secretaría de Cultura.

En el presente artículo, se describirán los diferentes módulos de trabajo que se han planteado para sentar las bases de este Repositorio y el grado de avance que cuenta al día de hoy. Asimismo se plantearán los retos técnicos, económicos y de gestión que ha implicado e implicará un proyecto de esta envergadura.

Contexto y antecedentes

Uno de los retos prioritarios para la Agenda Digital de Cultura, es la integración de los acervos culturales y ofrecerlos a los mexicanos para su divulgación, difusión y a través de una herramienta digital, convertirlos en una poderosa herramienta educativa de apoyo para la formación de la población.

El desafio es enorme, ya que actualmente los acervos en su gran mayoría no se encuentran normalizados bajo ningún tipos de esquema de datos, los contenidos descriptivos carecen de información y los objetos digitales son precarios o inexistentes, lo cual implica no únicamente una difusión carente de estructura, libre y abierta al público, sino que también, no se cuenta con ningún programa de preservación a largo plazo.

Objetivos

Algunos de los objetivos que el desarrollo del RDPCM tiene contempledos son:

- Generar una base sólida tecnológica, interoperable, libre y sustentable para la institución.
- Estandarizar los acervos bajo un mismo modelo de datos, para ser utilizado no sólo en la SC sino que sea extensivo en todo el país.
- Preservar el enorme y valioso patrimonio cultural de México de forma digital.
- Generar una plataforma web que permita a las audiencias acceder al vasto patrimonio cultural mexicano de manera enriquecida, sencilla y atractiva.
- Definición de derechos de los objetos digitales para la difusión y divulgación. Cabe destacar que actualmente en México no se cuentan con buenas prácticas en este tema.

Módulos de trabajo

Para el desarrollo del RDPCM, se plantearon los siguientes módulos que generaron productos específicos para la preservación y esquematización del patrimonio cultural digital:

1. Modelo de Datos México: creación de un modelo de datos único que permita normalizar e interoperar los metadatos de las instituciones, para coadyuvar a la mejor gestión de información producida desde la administración pública.
2. Normalización de registros: heterogeneidad de los registros y vocabularios, así como enriquecimiento editorial.
3. Tesauro Regional Mexicano:creación de un tesauro que incluirá terminologíamexicana especializada disponible a través del modelo LOD (Linked Open Data) e incluirá autores, obras, términos y relaciones de distintas instituciones.
4. Desarrollo: creación de sistema que cumplan con la visión de la Web Semántica, que permita exponer en formatos estándar todo la información. Contará con un meta-modelo ontológico, una herramienta de extracción y normalización, un cosechador-Indexador, buscador, CMS, API EndPoint y un módulo de preservación.
5. Sistemas gestores de colecciones de museos mexicanos: definidos en sistemas transparentes para el intercambio de datos con el RDPCM, pero que cuentan con total independencia al Repositorio.
6. Declaratorias de derechos: análisis del caso mexicano vs. el panorama internacional para la definición de derechos, según las leyes mexicanas.

Figura 1. Primera maquetación del RDPCM

Retos y responsabilidades

En la primera etapa de conceptualización y desarrollo del RDPCM, contendrá los acervos de 14 instituciones de la Secretaría de Cultura, que ascienden a más de 600,000 objetos digitales que represetan los acervos arqueológicos, históricos, artísticos, videográficos y sonoros de México. Uno de los retos más importantes en esta etapa es seguir incrementando proveedores de datos institucionales y continuar aumentandolos registros y objetos digitales de los que ya se encuentran en el Repositorio, además de seguir desarrollando contenidos de alto nivel.

También ser desarrollarán contenidoscurados por expertos, que permitan a los usuarios finales entender mejor las colecciones digitales y conectarse virtualmente con el patrimonio.

Prospecciones

- Sustentabilidad técnica y financiera: crear un sistema sólido y escalable en módulos, que permita adaptarse a las necesidades futuras y permitir reducir los costos de desarrollo, para poder ser aplicados en recursos humanos que administre las colecciones digitales.
- Aumento de audiencias: creación de una estretegia en medios y vinculación ciudadana a través de Hackatones y activaciones en redes sociales.
- Investigación y creación de contenidos: reconocer el valor de los investigadores y creadores de contenidos en las intituciones mexicanas, vinculado sus conocimientos para hacerlos participes en la creación y edición de contenidos.
- Proveedores de datos y agregadores: incrementar y exponer el mayor material disponible libre de derechos de autor en el RDPCM y generar salidasinnovadoras para las las nuevas audiencias.
- Creación del programa de digitalizción permanente: que permita incrementar el acervo digital para su posterior integración al repositorio de difusión y preservación.
- Profesionalización de los recursos humanos: generar redes para compartir conocimiento y solventar procesos a través de la creación de estrategias para el mejoramiento del sector cultura mexicano en el ambito de catalogación y preservación digital.

Conclusiones finales

México es un país con un enorme y valioso patrimonio cultural, la creación del RDPCM es una medida prioritaria y necesaria para el estado mexicano, que permitirá coadyuvar al acceso universal al patrimonio cultural mexicano para beneficiar a más audiencias educando, compartiendo conocimiento y transformando el mundo a través de la cultura.

Esta es una oportunidad para enriquecer la web con acciones positivas que refirmen la cultura en el mundo digital.La estrategía del RDPCM a 10 años, es proyectar el mayor número de objetos con una excelente calidad y a través de colaboraciones creativas e innovadoras, ofrecer un gran número de colecciones curadas en línea de acceso público para la investigación, el aprendizaje y la sociedad.

Referencias

Organización de las Naciones Unidas (2003). "Carta sobre la preservación del patrimonio digital". París.

Scholz, Henning, Devarenne,Cécile, Freire,Nuno, Kyrou,Panagiotis,Pekel, Joris (2017). "Europeana Content Strategy. Getting the right content to the right user at the right time".https://pro.europeana.eu/post/europeana-content-strategy

Digital Public Library of America (2015). Strategic Plan 2015-2017. Boston: DLPA.https://dp.la/info/wp-content/uploads/2015/01/DPLA-StrategicPlan_2015-2017-Jan7.pdf

Towards Linked Data of Bible Quotations in Jewish Texts

Oren Mishali
oren.mishali@gmail.com
Technion, Israel Institute of Technology, Israel

Benny Kimelfeld
bennyk@cs.technion.ac.il
Technion, Israel Institute of Technology, Israel

Introduction

The Hebrew Bible (the Tanakh) is the most ancient and sacred collection of Jewish texts. Throughout the history, additional religious Jewish texts have been written such as the Mishna, the Babylonian Talmud, and many more. These additional texts are often related to (or inspired by)

the Bible. As such, many of them quote verses[1] from the Bible (as in Figure 1). Depending mostly on their frequency and location within the text, the quotations may indicate a weak or strong semantic relation between a given text and a specific portion of the Bible. Knowing these semantic relations may be beneficial for those interested in studying or investigating the Bible.

Nowadays, a variety of Jewish texts are publicly available over the Internet, yet the identification of Bible quotations within them is often sparse and sometimes entirely absent. Moreover, the existing identification lacks a rigorous representation, which makes it difficult to automatically infer semantic correspondence and to develop supporting software applications.

We report an ongoing project that aims to establish the machinery for the automatic detection and rigorous representation of quotations of Bible verses within Jewish texts. The project consists of three interleaving components. In the first component, an algorithm for identifying Bible quotations in text is developed. In the second, the results of executing the algorithm on a large and open text corpus are represented as a Linked Data graph (RDF dataset). In the third component, we develop a web frontend for making the dataset accessible to end users. Exposing the data to end users may also engage their participation in data-driven crowdsourcing (Ched et al, 2015), and hence, will serve to collectively help in improving the dataset quality. In what follows, we elaborate on each of the project components.

Algorithm

Quotation detection is gaining popularity in fields such as copyright enforcing and political analysis, and within ancient texts (Ernst-Gerlach andCrane, 2008; Gesche et al, 2016). The algorithms in use share common characteristics, yet each domain brings its own specificities and challenges. Given an input text, our algorithm first matches maximal n-grams[2] in the text to candidate Bible verses. For example, the green bigram (2-gram) in the first line of Figure 1 will have one matching verse, since its text ('לך לך') appears in exactly one Bible verse. This matching is maximal, since the words that appear before and after the bigram are not part of the quoted verse.

```
PREFIX jbo: <http://jbs.technion.ac.il/ontology/>
PREFIX jbr: <http://jbs.technion.ac.il/resource/>

SELECT ?uri ?text FROM <http://jbs.technion.ac.il> WHERE {
    ?uri a jbo:Text; jbo:text ?text.
    ?uri jbo:quotes jbr:text-tanach-1-1-1.
}
```

A portion of ancient Jewish Text (from Midrash Raba), that quotes two Bibles verses. Quotations to the same verse are marked in a similar color. Note that each quotation refers only to a part of the verse (1-4 words of it).

A first challenge that we face is related to variations found between the quoting text and the original Bible text, mostly related to the omission (or inclusion) of Hebrew vowel letters. As an example, consider the red quotation in the second line of the figure, that contains the word 'ומה הירו', where in the original Bible source the 'ו' (vav) vowel is omitted. We have implemented two alternative solutions, one is based on *fuzzy search* (Levenshtein distance), and the other on *exact search* performed simultaneously on two versions of the Bible, with and without vowels.

Not all verse candidates are valid quotations of Bible verses in the text. For instance, the phrase 'ויבא תיב' in the third line of the figure (underlined) is a common phrase that appears in eleven different Bible verses. Nevertheless, the phrase is mentioned in a different context, which is not related to any of them. False candidates occur mostly in bigrams and trigrams (3-grams), and the algorithm makes an effort to filter them out. One approach is to keep a candidate if a matching candidate appears in a larger n-gram in the same text. For instance, the green bigrams and trigram shown in the figure are reported as valid quotations since there is a 4-gram that quotes the same verse in the text ('וארא רשא קראה לא', line 3). We are considering additional filtering approaches related to statistical data inference and machine learning. We are also creating collections of labeled data for a systematic evaluation of the algorithm.

Linked Data

The detected quotations are represented as RDF Linked Data, making them accessible to machines for standard consumption and integration. We use a lightweight ontology that we have defined, augmented with standard properties taken from known ontologies such as RDF, RDFS, and Dublin Core (DC). We are working on the integration of additional ontologies such as CIDOC-CRM, FRBR, and SPAR. Key ontology classes are *Book*, *Section*, *Text*, and *Quotation*. A *Book* is composed of *Sections*, that may be composed of other *Sections*, and eventually of *Text* elements. Each Bible verse is a node of type *Text* in the RDF graph. To date, our graph contains a total of 23,206 *Text* nodes of Bibles verses. Additional 355,181 *Text* nodes represent text elements within other Jewish books (where quotations are searched for). An edge from a *Text* node of the latter kind to one of the former kind indicates a 'quotes' relationship. Nodes of class *Quotation* hold additional details such as the exact location wherein a quotation appears in the text.

A SPARQL query that retrieves all text elements quoting the first verse of the Bible.

1 The Bible is divided into basic text elements called *verses*.
2 An *n-gram* is a contiguous sequence of n words from a text.

A Linked Data graph may be accessed by expert users using the SPARQL query language. An example SPARQL query is shown in Figure 2. To make our data widely accessible, we have implemented a graphical web frontend that acts like a search engine for Bible verses. A user selects a set of verses from the Bible, and then being presented with all text elements that quote one or more verses from the set. (The elements are retrieved from the RDF graph.) The results are sorted by significance, and may be filtered using predefined categories. We plan to enhance the web interface with data-driven crowdsourcing support, where the crowd will help in improving the accuracy of the algorithm by marking false negatives (places in the text that the algorithm has missed), as well as false positives (incorrect detections). The web tool, as well as the detection algorithm and related artifacts, are accessible via our main GitHub repository.

References

Ched, L. and Lee, D. and Milo, T. (2015). *Data-driven Crowdsourcing: Management, Mining, and Applications*. International Conference on Data Engineering (ICDE), Tutorial.

Ernst-Gerlach, A. and Crane, G. (2008). *Identifying Quotations in Reference Works and Primary Materials* . Research and Advanced Technology for Digital Libraries, 78-87.

Gesche, S. and Egyed-Zsigmond, E. and Calabretto , S. (2016). *Was it better before? Automated Quotation Detection in Ancient Texts*. CORIA-CIFED, 167-182.

Towards a Metric for Paraphrastic Modification

Maria Moritz
mmoritz@gcdh.de
University of Goettingen, Germany

Johannes Hellrich
johannes.hellrich@uni-jena.de
Graduate School "The Romantic Model", Friedrich-Schiller-Universität Jena, Germany

Sven Buechel
sven.buechel@uni-jena.de
JULIE Lab, Friedrich-Schiller-Universität Jena, Germany

Introduction

Clarifying the genesis of a passed down text is of outmost importance for many scholarly disciplines within the humanities such as history, literary studies, and Bible studies. Often, historical text sources have been copied over and over for hundreds or even thousands of years, thus being subjected to paraphrasing and other kinds of modifications, repeatedly. Despite the significance of source criticism for the humanities as a whole, algorithmic support in this matter is still limited. While current approaches are able to tell **if** and **how frequent** a text has been modified—to the best of our knowledge—there has been no work on determining the **degree** of paraphrastic modification. To a human reader, the introduction of, say, spelling variations is indubitably a minor modification compared to substituting entire words. Yet, how can the different "degrees" of alterations, which are intuitively clear to scholars, be captured in an algorithmic way?

To this end, we present a first approach for designing a metric for paraphrastic modification in text (henceforth paraphrasticality). Based on an English Bible corpus (three literal Hebrew and Greek translations and three standard translations) we measure the frequency of different classes of textual variations between each pair of Bibles. We then use the probability of these variations in a machine learning experiment to derive weights for these classes of modifications. Ultimately, this allows us to define a metric for paraphrasticality which we validated with promising results.

Related work

Measuring the **similarity** or **distance** between two spans of text is relevant to many areas in and related to natural language processing (NLP). One of the earliest approaches is Levenshtein's (Jurafsky and Martin, 2009) edit distance which is based on character-level removal, insertion, and replacement operations. BLEU (Papineni, 2002) is the most common evaluation metric in machine translation, capturing the difference between gold and automatic translations based on (word-level) n-gram overlap. In **stylometry**, different kinds of delta metrics are used to compute the difference between the writing style of authors or texts (Jannidis et al., 2015). These are typically based on the frequency distribution of the most frequent words. These first three approaches have in common that they rely on surface features (token and character-level) alone and do not incorporate semantic proximity. In contrast to that, computing the **semantic similarity** between two sentences is a popular task within NLP (Xu et al., 2015). However, approaches in this field are typically not indented for manual inspection and are thus less suited for applications in the humanities. Lastly, Moritz et al. (2016) quantify modification operations on a parallel Bible corpus yet do not present a way to aggregate these counts into a distance metric. In contrast to these related contribution, here, we aim to develop a metric which is both semantically informed as well as human interpretable.

Data

We use a parallel corpus of the Old Testaments of six English Bible translations[3] from the 19th century, half of them being literal translations that closely follow the primary source texts' language and syntax while the other half are standard translations (see Table 1).

name	abbr.	publication	source	translation
The Webster Bible	WBT	1833	bst	standard
Brenton's English Septuagint	LXXE	1851	mys	literal
Young's Literal Translation	YLT	1862	bst	literal
Smith's Literal Translation	SLT	1876	mys	literal
English Revised Version	ERV	1881-1894	mys	standard
Darby Bible	DBY	1890	ptp	standard

Table 1: Bible editions used for investigation. Sources: bst: https://www.biblestudytools.com/; mys: https://www.mysword.info; ptp: Parallel Text Project (Mayer and Cysouw, 2014)

Literal translations: Robert Young, the translator of **YLT,** created a highly literal translation of the original Hebrew and Greek texts. Thus, Young tried to be as consistent as possible in representing Greek tenses with English ones, e.g., he used present tense where other translations used past tense (see Young, *1898a; Young, 1898b*) as in: 'In the beginning of God's preparing the heavens and the earth —' (Genesis 1:1). **SLT**: Upon publication, Julia Smith's Bible translation was considered the only one directly translating the historical source texts to contemporary English. She aimed at complete literalness and tried to translate each original word with the same English word, consistently, and tended to translate the Hebrew imperfect to English future tense (Malone, 2010). **LXXE** by Sir Lancelot Charles Lee Brenton is an English translation from the Codex Vaticanus version of the Greek Old Testament, which itself is a translation of the Hebrew Old Testament (Roger, 1958).

Standard translations: WBT by Noah Webster is a revision of the King James Bible mainly eliminating archaic words and simplifying Grammar (Marlowe, *2005*). **ERV** is today's only officially authorized revised version of the King James Bible in Britain (no author, 1989). The most recent edition in our study is **DBY**, Darby's translation of the Bible. The Old Testament was published by his students in 1890 and is based on Darby's German and French versions (Marlowe, 2017).

Methods

Preprocessing and alignment: We use MorphAdorner (Burns, 2013), a specialized toolkit for early modern and modern English, to tokenize and lemmatize the Bibles. After removing punctuation and verse identifiers, we pair up our six Bibles in every possible combination (15 in total). Since the different Bible versions are inherently aligned on the verse-level (by their verse identifier), our next step builds up a statistical alignment at the token level for each pair of bibles using the Berkeley Word Aligner (De Nero and Klein, 2007), a tool originally designed for machine translation.

Counting modification operations: Building on these word-aligned pairs of Bibles, we can describe the divergence between a pair of verses in terms of the **modification operations**—such as replacing a word by its synonym—which would be necessary to convert one version into another. We automatically apply and count the modification classes introduced by Moritz et al. (2016) for each verse and Bible pair (see Table 2). Synonyms, hypernyms, hyponyms and co-hyponyms, are identified based on BabelNet (Navigli and Ponzetto, 2012).

abbr.	operation	estimated coefficient θ_{relative}
lower	case-folding matches	0.060
lem	lemmatizing matches	0.195
low_editdist	writing variant	0.068
syn	synonyms match	0.190
hyper	source word is hypernym of target word	0.117
hypo	source word is hyponym of target word	0.170
co-hypo	co-hyponyms match	0.122
fallback	other	0.078

Table 2: Operations used as features together with normalized estimated weights (coefficients) of the fitted model

Weight identification: By counting modification operations, we gain a fine-grained description of the *exact* differences between two spans of text. However, to construct a metric, we had to find a way to condense these modification frequencies down to a single number. For that we exploit the fact that we deal with two classes of Bible translations, literal and standard ones. Thus, to estimate a human judgment of deviation, we assume that standard translations are more homogenous to each other than literal translations (since the latter demand for more creative language use; see Section 3). Hence, we can train a classifier to distinguish whether a pair of Bible verses is from the same class (both Bibles being standard or literal translations, respectively) or from different classes. For this task, we train a maximum entropy classifier[4] where we use the relative frequencies of the modification operations as features. Now, the key part of our contribution is that we can exploit the coefficients of our fitted model as the first ever presented empirical estimate of the relative importance of these modification operations for paraphrasticality.

3 Note that our approach is not limited to applications on historical text and that our choice of textual material is based on technical reasons only. In fact, any paraphrastic, parallel corpus would work equally well for our proposed method.

4 Using the scikit-learn.org implementation. Training for this binary classification task was done using 10-fold cross-validation achieving an accuracy of .68.

Results

Feature weights: Table 2 lists the final, normalized (summing up to 1) feature weights of our fitted model. Lemmatization, hyponym and synonym relations turn out to be especially important for the classification task.

Metric: Based on these coefficients, we define the paraphrasticality metric *par* between two word-aligned text spans *a* and *b* as

$$par(a,b) = \sum_{i=0}^{n} \theta_i x_i^{a,b}$$

where n is the total number of features (or classes of operations), θ_i is the absolute weight for feature i determined via the classifcation experiment and $X_i^{a,b}$ is the relative frequency of the respective operation. In order to gain face validity for this newly defined metric, we compute the paraphrasticality score for each one of the 15 Bible pairs in our corpus (as average of their verse paraphrasticality). The results are presented in Table 3.

	DBY	ERV	WBT	LXXE	YLT	SLT
DBY	-	0.13	0.13	0.29	0.31	0.29
ERV	-	-	0.09	0.3	0.32	0.31
WBT	-	-	-	0.28	0.33	0.29
LXXE	-	-	-	-	0.42	0.37
YLT	-	-	-	-	-	0.31
SLT	-	-	-	-	-	-

Table 3: Deviation between each pair of Bibles in terms of our newly developed paraphrasticality metric; higher values indicate higher distance

Bible pair	operation 1	operation 2	operation 3	classes
DBY-ERV	lem (1.6%)	syn (1.1%)	cohypo (.9%)	standard
DBY-WBT	lem (1.6%)	syn (1.1%)	cohypo (.9%)	standard
ERV-WBT	lem (1.6%)	syn (.7%)	cohypo (.6%)	standard
DBY-LXXE	lem (3.1%)	syn (2%)	cohypo (1.9%)	standard/literal
DBY-YLT	lem (6.6%)	low (4.7%)	syn (2.6%)	standard/literal
DBY-SLT	lem (5.9%)	syn (2.6%)	cohypo (2.2%)	standard/literal
ERV-LXXE	lem (3.5%)	low (2.1%)	syn (1.9%)	standard/literal
ERV-YLT	lem (6.6%)	low (4.7%)	syn (2.5%)	standard/literal
ERV-SLT	lem (5.9%)	syn (2.6%)	cohypo (2.2%)	standard/literal
WBT-LXXE	lem (3.4%)	low (2.2%)	syn (1.9%)	standard/literal
WBT-YLT	lem (6.8%)	low (4.8%)	syn (2.7%)	standard/literal
WBT-SLT	lem (5.8%)	syn (2.6%)	cohypo (2.2%)	standard/literal
LXXE-YLT	lem (7.%)	low (4.4%)	syn (2.6%)	literal
LXXE-SLT	lem (5.8%)	cohypo (2.6%)	syn (2.6%)	literal
YLT-SLT	lem (5.4%)	low (4.8%)	syn (2.5%)	literal

Table 4: Top 3 most frequent operations (without fallback) per Bible pair

Qualitative validation: We can identify three regions in the plot. The upper left triangle shows that our standard translations do not differ much from each other (as expected), especially since WBT and ERV are revisions of the same Bible. The 3x3 rectangle in the upper right corner represents pairs of one literal and one standard translation, respectively. We can see that the distance between those is about 0.3 thus displaying increasing paraphrasticality compared to pairs of *only* standard translations. The highest deviation however is between the literal translations by Smith (SLT) and Young (YLT) compared to the English Septuagint (LXXE). This can be explained by the choice of vocabulary by each translator and by the purpose they follow in their translations. For example, SLT and LXXE use "firmament" when YLT uses "expanse", SLT and YLT use "rule" when LXXE uses "regulating". We thus conclude that our metric yields valid and—perhaps even more important for applications in the humanities—interpretable results.

Our approach also enables to judge distance on a fine-grained level based on pure operation counts. In Table 4 we show the top 3 operations for each Bible pair. As we can see, most of the top 3 operations per Bible pair relate to semantic relations between the aligned word pairs (matching lemma, synonymy, or co-hyponomy) underscoring the advantage that our metric has as opposed to more surface feature-dependent approaches (to textual similarity) such as Levenshteindistance or delta measures.

Conclusion

We presented the first study on designing a metric for paraphrasticality. Different from existing approaches on measuring distance or similarity between texts, we describe paraphrasticality as frequency of specific modification operations for which we tried to find empirically adequate weights via a machine learning experiment. As demonstrated, our approach is specifically useful for applications in the humanities as operation frequencies, and feature weights, as well as paraphrasticality scores are open to manual inspection. A more comprehensive comparison against existing similarity metrics and a human judgment is left for future work.

References

Burns, P. R. (2013). Morphadorner v2: A java library for the morphological adornment of English language texts. *Northwestern University, Evanston,* IL, no page numbers.

De Nero, J. and Klein, D. (2007). Tailoring word alignments to syntactic machine translation. *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*. Prague, pp. 17–24.

Jannidis, F., Pielström, S., Schöch, C. Vitt, T. (2015). Improving Burrows' Delta—An empirical evaluation of text distance measures. *Digital Humanities Conference 2015*. Sydney, no page numbers.

Jurafsky, D. and Martin, J. H. (2009). *Speech and language processing: An introduction to natural language processing, speech recognition, and computational linguistics.* Englewood Cliffs: Prentice-Hall.
Malone, D. (2010). Julia Smith bible translation (1876), https://recollections.wheaton.edu/2010/12/julia-smith-bible-translation-1876/ *(accessed November 2017).*
Marlowe, M. *(2005).* Webster's Revision of the KJV (1833), http://www.bible-researcher.com/webster.html*(accessed November 2017).*
Marlowe, M. (2017). John Nelson Darby's Version, http://www.bible-researcher.com/darby.html*(accessed November 2017).*
Mayer, T. and Cysouw, M. (2014). Creating a massively parallel bible corpus. *Proceedings of the Ninth International Conference on Language Resources and Evaluation.* Reykjavik, pp. 3158–61.
Moritz, M., Wiederhold, A., Pavlek, B., Bizzoni, Y. and Büchler, M. (2016). Non-literal text reuse in historical texts: An approach to identify reuse transformations and its application to bible reuse. *Empirical Methods in Natural Language Processing.* Austin, pp. 1849–59.
Navigli, R. and Ponzetto, S. P. (2012). Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193(2012): 217–50.
Papineni, K., Roukos, S., Ward, T. and Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, pp. 311–18.
Roger, N. (1958, 1959). New Testament Use of the Old Testament. In Henry, C. F.H. (ed.), *Revelation and the Bible. Contemporary Evangelical Thought.* Grand Rapids: Baker, 1958. London: The Tyndale Press, 1959, pp. 137–51.
Xu, W., Callison-Burch, C. and Dolan, B. (2015). SemEval-2015 Task 1: Paraphrase and Semantic Similarity in Twitter (PIT). *SemEval@ NAACL-HLT.* Denver, pp. 1–11.
Young, R. *(1898a). Young's Translation: Publisher's Note and Preface,* http://www.ccel.org/bible/ylt/ylt.htm*(accessed November 2017).*
Young, R. *(1898b). Young's Literal Translation*, http://www.bible-researcher.com/young.html (accessed November 2017).
No Author. (1989). *The Holy Bible. Revised Version.* London: Cambridge University Press. Synopsis.

Temporal Entity Random Indexing

Annalina Caputo
annalina.caputo@adaptcentre.ie
Adapt Centre, Trinity College Dublin, Ireland

Gary Munnelly
gary.munnelly@adaptcentre.ie
Adapt Centre, Trinity College Dublin, Ireland

Seamus Lawless
seamus.lawless@adaptcentre.ie
Adapt Centre, Trinity College Dublin, Ireland

Introduction

In this exploratory research, we sought to investigate how we might identify and quantify the contextual shift surrounding significant entities in news based corpora. For example, might we be able to see changing public opinion such as that experienced by George W. Bush Jr. after the events of 9/11 and thus note how a population can rally behind their leader in the face of cultural trauma?

Our method of identifying these changes has its roots in the field of distributional semantics and the measurement of semantic shift. A typical approach to solving this problem involves building multiple word models across subsets of the sample corpus which are organized by date. By comparing the outputs of the different models we can see how the definitions of words have evolved. We adopt Temporal Random Indexing (TRI) (Basile et al., 2014) as our method of measuring semantic shift over time as it allows for a direct comparison between word representations on the basis of simple cosine similarities.

Method

In order to apply our method of measuring contextual shift in relation to entities we require a consistent representation of each entity that will span the entire collection e.g. the algorithm will need to know that "President Bush", "G.W." and "Dubyah" all refer to the same individual. In order to achieve this, an Entity Disambiguation process is applied to the source text prior to building the semantic space. This step substitutes mentions of each entity with a URI obtained from DBpedia, allowing the algorithm to track an individual through the collection irrespective of how they are referenced. We use CogComp NER[5] (Ratinov and Roth, 2009) for entity recognition and AGDISTIS[6] (Usbeck et al., 2014) for disambiguation.

Given the output from the disambiguation tools, a different semantic space for each year in the collection's timespan is built using the TRI implementation by Basile[7]

5 https://github.com/CogComp/cogcomp-nlp/tree/master/ner
6 https://github.com/dice-group/AGDISTIS
7 https://github.com/pippokill/tri

(Basile et al., 2014). Each space provides a semantic representation of words and Named Entites (NE) in terms of their proximity in space, which reflects their semantic relatedness. A time series for each NE is extracted by computing the cosine similarity between two consecutive semantic spaces (e.g. 2001 and 2002). Finally, candidate dates for the shift in meaning are extracted using the Change Point Detection algorithm as implemented by Kulkarni[8] (Kulkarni et al., 2015).

Evaluation

For test data we utilized the New York Times collection curated by LDC[9] (Sandhaus, 2008) which spans 20 years of American news from 1987 to 2007. While methods which measure semantic shift in word sense typically require collections which span hundreds of years, because circumstances evolve more quickly than language, we believe that a 20 year span is more than enough to produce interesting results when the same methods are applied to the examination of entities.

The collection was preprocessed and analysed using the method described in Section 2. This yielded a series of 20 language models which provided semantic representations for each entity identified and linked by CogComp NER and AGDISTIS. We computed the temporal shift for all the entities in the corpus and ranked them by the magnitude of this shift (p-value from the Change Point Detection algorithm). We selected the top 100 entities from this ranking (i.e. those with the greatest semantic shift) and selected the largest group of entities which underwent a semantic shift in the same year from within that group.

Results

The evaluation methodology described in Section 3 yielded a shortlist of 12 entities which undergo a sizeable semantic shift in 2001: Federal_Bureau_of_Investigation, Pentagon, White_House, New_York, Congress, Department_of_Justice, George_H._W._Bush, Texas, West, Saddam_Hussein, Republican_Party_(United_States), and American_Motors. Almost all of them are related to politics and have strong connections with the happenings of 9/11. Notably, while a member of the Bush family is connected with these events and does indeed undergo a shift in semantic representation, it is the wrong individual - the father rather than the son. This assignment of a semantic shift to George_H._W._Bush in 2001 is certainly due to the disambiguation process.

While we believe the inclusion of the entity disambiguation step is an interesting contribution of this work, we observed a number of problems with the process.

The contents of the knowledge base, which informs the disambiguation software, has a dramatic impact on the quality of the results obtained. So too does the nature of the entities being disambiguated. One notable example of this was our results with regards to mentions of "the Internet". Our method showed a dramatic increase in discourse surrounding the Internet from the mid 90s up into the second millennium. However, while the representation was consistent, the referent chosen by the disambiguation software was an American band known as "The Internet", rather than the network of computers we use today.

While the error with the Internet is obvious, more challenging was distinguishing between mentions of George W. Bush Jr. and George H. W. Bush Sr. The former's role in the events post 9/11 (reports of which were included in our corpus) made him an important entity for the disambiguation software to correctly annotate. However, in many cases this proved to be extremely difficult. This is understandable given the similarity in context surrounding both Bush Jr. and Bush Sr., We can work with an incorrect annotation provided it is consistently incorrect. However the unpredictability surrounding the name "Bush" presents a difficult problem when this information is used as part of the Random Indexing process.

Conclusion

We have presented a preliminary case study, which although not robust enough to infer any conclusions, highlights the potential of this type of analysis. We conducted our preliminary investigation guided by a major cultural trauma that occurred between 1987 and 2007, and which caused a sudden reaction and change in the public discourse. It is clear that a weakness in the method is the disambiguation process. Future work will focus on improving the quality of disambiguation as well as investigating the possibility of building time series models over shorter spans of time e.g. months or weeks.

References

Basile, P., Caputo, A. and Semeraro, A. (2014). Analysing word meaning over time by exploiting temporal Random Indexing. *Proceedings of the First Italian Conference on Computational Linguistics CLiCit 2014 and of the Fourth International Workshop EVALITA 2014 911 December 2014 Pisa* doi:10.12871/CLICIT201418. http://www.pisauniversitypress.it (accessed 25 April 2018).

Kulkarni, V., Al-Rfou, R., Perozzi, B. and Skiena, S. (2015). Statistically Significant Detection of Linguistic Change. ACM Press, pp. 625–35 doi:10.1145/2736277.2741627. http://dl.acm.org/citation.cfm?doid=2736277.2741627 (accessed 25 April 2018).

Ratinov, L. and Roth, D. (2009). Design challenges and misconceptions in named entity recognition. Association for Computational Linguistics, p. 147 doi:10.3115/1596374.1596399. http://portal.acm.

8 https://github.com/viveksck/langchangetrack
9 https://catalog.ldc.upenn.edu/ldc2008t19

org/citation.cfm?doid=1596374.1596399 (accessed 25 April 2018).
Sandhaus, E. (2008). The new york times annotated corpus. *Linguistic Data Consortium, Philadelphia*, 6(12): e26752.
Usbeck, R., Ngomo, A.-C. N., Röder, M., Gerber, D., Coelho, S. A., Auer, S. and Both, A. (2014). AGDISTIS-graph-based disambiguation of named entities using linked data. *International Semantic Web Conference*. Springer, pp. 457–471 http://link.springer.com/chapter/10.1007/978-3-319-11964-9_29 (accessed 12 February 2017).

IncipitSearch - Interlinking Musicological Repositories

Anna Neovesky
anna.neovesky@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Germany

Frederic von Vlahovits
frederic.vonvlahovits@adwmainz.de
Akademie der Wissenschaften und der Literatur Mainz, Germany

Open research data is facilitating broader ways of using, reusing, enriching, and linking research results. Many services use metadata to bring the information of different repositories together. Europeana, for example, links material from various thematic focal points with diverse origins and makes a wide range of collections, archives and source objects searchable. Other platforms interlink and aggregate material for one distinct discipline or thematic interest.

To connect musicological collections and repositories, we created a metasearch that builds up on annotated music. IncipitSearch is a tool and a service specifically tailored for research on music incipits, the initial sequences of notes that characterise a work. It is simultaneously a centralised data endpoint, where multiple aggregated catalogues can be accessed and searched by their music incipits, as well as a decentralised software and data cluster.

Open Data and Meta Search Engines: Perspectives for Digital Musicology?

Open research data is facilitating broader ways of using, reusing, enriching, and linking research results. Many services use metadata to bring the information of different repositories together. Europeana (https://europeana.eu), for example, links material from various thematic focal points with diverse origins and makes a wide range of collections, archives and source objects searchable. Other platforms interlink and aggregate material for one distinct thematic interest such as Ariadne (http://ariadne-infrastructure.eu), which makes manifold archaeological contents accessible, or correspSearch (http://correspsearch.net), which enables to search through collections of editions of letters.

Meanwhile, musicological projects do not only often have digital components, too. Ambitious global catalogue projects like the Répertoire International des Sources Musicales (RISM, https://opac.rism.info) or national library services such as the catalogues of the Italy's Servizio Bibliotecario Nazionale (SBN, http://opac.sbn.it) or the Deutsche Nationalbibliothek, (DNB, https://portal.dnb.de) substantially rely more and more on the digital representation of their data. In addition, overall demand of digital research platforms has lead to born digital editorial projects, e.g. Freischütz Digital, a genuinely digital edition of Carl Maria von Weber's Freischütz (http://freischuetz-digital.de) exploring the possibilities of multimedial digital musicological work editions, or the digital thematic work catalogue of the complete edition of Gluck's works (http://gluck-gesamtausgabe.de). The researcher's stronger trust and belief in the benefits of open and accessible research data has led to a stronger emergence of open data policies in musicological projects. In order to interlink existing data repositories and encourage new proposals, a digital data hub is needed. But how can musicological data collections be connected and linked together? In our approach, we concentrated on musical incipits, the initial sequences of notes, that function as identifier for works, and created IncipitSearch, a metasearch for musical incipits.

Encoding Music Incipits

One of the main goals of musicological catalogues is making musical works findable and researchable. The main problem that often occurs, especially for music composed before 1800, is that it originally was composed for a singular religious or secular cultural event, e.g. at an aristocratic court to be performed only once or just a few times. Music was additionally bound to formalised genre standards and therefore unambiguous titles were not required. But how to search for a Sonata in D of a composer who has composed 20 sonatas in D? As early as the 1960s, Music librarians introduced the idea to generate a human and machine readable standardised format to identify music by its melodic beginning. For that purpose, Barrey S. Brook and Murray Gold developed the Plaine & Easie Code that allows the transcription of the beginning notes of a musical piece into a combination of numbers and letters. What Brook and Gould pointed out in 1964 was already a distinct definition of and guide to the Plaine & Easie code system. They introduced it as "an accurate shorthand for musical notation, especially useful for incipits and excerpts." With some foresight they also stated that "it must be so devised to be readily transferable to electronic data-processing equipment for key transposi-

tion, fact-finding, tabulaturing and other research purposes." (Brook and Gold, 1964)

Plaine & Easie Code is a simple to parse plaintext format and therefore suitable to deliver important metadata for manifold musicological interests. IncipitSearch adopts this standard and at the moment is purely concentrated on Plaine & Easie. However, the future goal is to be capable of reading incipits notated in other formats as well, e.g. MEI (http://music-encoding.org) or abc notation (http://abcnotation.com).

Searching Music Incipits

Musc information retrieval systems either build up on audio or symbolic music notation. In digital musicology, that deals with notation and critical digital edition of works, the search in notated music is widely used (Typke et. al. 2005).

RISM is undoubtedly the most established repository for musical data. It contains over one million records of historic music materials and over 1,7 million musical incipits (for manuscripts only), which can be accessed using an incipit search (RISM search). Further incipit search engines build up on the RISM datasets. For example, Utrecht University has developed an extended and experimental search approach offering extended functionalities for user input as well as using sophisticated matching and ranking methods (Van Nuss et. al. 2017).

But other musical incipits exist which cannot be accessed via RISM because they either have not been implemented as data yet or because they are not a type of resource the RISM collection is focusing on and will not be added to the catalogue, such as work catalogues.

IncipitSearch

Scope and Functionalities

The efforts to implement incipits in the digital work catalogue of the complete edition of Gluck's works and to make them searchable have led to the idea of connecting this research data with other repositories and creating even easier ways to instantiate new machine readable incipit repositories. Both digital and analogue catalogues, editions and collections which provide their data in a standardised format can be interlinked with IncipitSearch.

IncipitSearch addresses music that can be displayed in common western music notation. Its main focus lies on music composed prior to 1800. Nevertheless, through its openness it can be furthermore used as a platform to explore challenges in researching culturally and historically different forms of musical notation.

IncipitSearch is a tool and a service specifically tailored for research on music incipits. It is simultaneously a centralised data endpoint where multiple aggregated catalogues of incipits can be accessed as well as a decentralised open source software that can be integrated as stand-alone search in other platforms. A microservice based software architecture allows high flexibility in usage and extension of individual components (Haft et. al. 2015).

IncipitSearch enables users to enter search queries in the search field by playing them on a virtual piano keyboard while Plaine & Easie Code can also be directly entered into the search field. Search with transposition or with exact matching can be selected (https://incipitsearch.adwmainz.net). Next to the found concordant incipts, the result list displays backlinks to the entry in the respective catalogue.

Screenshot of the search interface of IncipitSearch.

Metadata Schema

To enable a standard suitable for the different types of musicological repositories such as digital and analogue catalogues, editions and collections and to provide an output of the collected data, we have developed an easy to understand RDF schema using the schema.org vocabulary. Besides being recommended by the W3C, cross-linking possibilities for data and the possibility to rely on various vocabularies for specific topics, the interoperability and the multiple serialisation formats for RDF are advantageous.

Schema.org provides a vocabulary for the description of web pages. The initiative of several major search engine companies aims to develop a simple vocabulary to add semantic information to webpages. These vocabularies were designed in collaboration with domain experts. For the markup of music information, the data type MusicComposition (http://schema.org/MusicComposition) supplies most elements to describe a work and its parts. To add the possibility of describing music incipits, we have expanded the vocabulary with further elements. The format can be used directly for data interchange - a feature request for the extension of schema.org with incipit declaration is planned.

The metadata format functions as an acquisition format for the repositories. It can be used to add information to the catalogue by adding music incipits to existing re-

source as well as a schema for the annotation and digital publication of analogue catalogues. Moreover, it will provide the aggregated data in a standardised format to enable further usage.

Conclusion

At the moment, IncipitSearch aggregates the incipit data of the catalogue of Gluck's works, the SBN OPAC, the RISM OPAC and includes a sample data set of the thematic Breitkopf Catalogo delle Sinfonie 1762.

IncipitSearch builds on the potential of open musical data and provides the possibility to interlink musicological repositories of various types. This is accomplished by concentrating on musical incipits and using a standardised data interface, a straightforward metadata schema and encapsulated software components.

Through consistent usage of authority control and metadata standards, IncipitSearch is an open source tool and service warranting sustainability, transparency, and accessibility of research data.

External Links

- Europeana: https://europeana.eu
- correspSearch: http://correspsearch.net
- Deutsche Nationalbibliothek (DNB): https://portal.dnb.de
- Freischütz Digital: http://freischuetz-digital.de
- IncipitSearch: https://incipitsearch.adwmainz.net
- Répertoire International des Sources Musicales: https://opac.rism.info
- schema.org: http://schema.org
- Servizio Bibliotecario Nazionale (SBN): http://opac.sbn.it
- Work catalogue of the complete edition of Gluck's works (GluckWV): http://gluck-gesamtausgabe.de

References

Brook, B.S., Gold, M. (1964). Notating Music with Ordinary Typewriter Characters (A Plaine and Easie Code System for Musicke). *Fontes Artis Musicae,* vol. 11, no. 3, 1964, pp. 142–159. www.jstor.org/stable/23504533.

Haft, M., Neovesky, A. and Reimers, G (2016). Digitale Nachhaltigkeit von Forschungsdaten durch Microservices. FORGE 2016. Forschungsdaten in den Geisteswissenschaften: Conference Abstract, pp. 23 -24. https://www.fdm.uni-hamburg.de/ueber-uns/a-nachrichten/aktivitaeten/forge16/praesentationen/programmheft.pdf#page=23.

Typke, R., Wiering, F. and Veltkamp, R.C. (2005). A survey of music information retrieval systems. Proceedings of the 6th International Conference on Music Information Retrieval, ISMIR 2005, pp. 153 - 160. http://ismir2005.ismir.net/proceedings/1020.pdf.

Van Nuss, J., Giezeman, G.-J., Wiering, F. (2017). Melody retrieval and composer attribution using sequence alignment on RISM incipits. Proceedings of the International Conference on Technologies for Music Notation and Representation. Universidade da Coruña, pp. 79–90. http://doi.org/10.5281/zenodo.924135

OCR'ing and classifying Jean Desmet's business archive: methodological implications and new directions for media historical research

Christian Gosvig Olesen
c.g.olesen@uva.nl
University of Amsterdam, The Netherlands

Ivan Kisjes
I.kisjes@uva.nl
University of Amsterdam, The Netherlands

This paper discusses the endeavours of the research project *MIMEHIST: Annotating EYE's Jean Desmet Collection* (2017-2018) - funded by the Netherlands Scientific Research Organisation - to do optical character recognition (OCR) and apply various computer vision techniques on the business archive of film distributor and exhibitor Jean Desmet (1875-1956).

The Desmet collection consists of approximately 950 films produced between 1907 and 1916, a business archive containing around 127.000 documents, some 1050 posters and around 1500 photos. The Collection is unique because of its large amount of rare films from the transitional years of silent cinema, and because of the richness of its business archive which holds extensive documentation of early film exhibition and distribution practices in the 1910s. These features contribute to its immense historical value which was one of the main reasons why it was inscribed on UNESCO's Memory of the World Register in 2011.

By OCRing and classifying Jean Desmet's business archive, MIMEHIST will allow scholars to browse and annotate its documents - all scanned in high resolution - in the new 'Media Suite' of the Dutch national research infrastructure (CLARIAH). The results will be integrated in a search interface enabling media historians to identify word frequencies and topics as a basis for research on early film distribution and exhibition and, the paper argues, open for media historical research which productively builds on and expands the collection's use in previous scholarship.

Throughout the past decades, Desmet's business documents have offered a rich source for socio-economic

cinema history. Media historians such as Karel Dibbets and Rixt Jonkman have studied parts of the collection's (related) data by manually transcribing and organising it into databases (Jonkman, 2007; Dibbets, 2010). This work produced an empirical, quantitative foundation for network analysis of Dutch film distribution and exhibition in cinema's earliest years. However, this research also made evident that the archive is too large and diverse to organise and transcribe manually. A particular challenge is that collection contains many different kinds of documents: personal letters, business letters, records of film rentals, postcards, newspaper clippings, telegrams, scraps of paper with notes, photographs etc. Furthermore, some documents are typewritten, others handwritten.

To allow scholars to research and annotate larger amounts of the archival documents' data in CLARIAH's Media Suite, automated information extraction from the documents seemed challenging yet promising. MIMEHIST took up this challenge by trying OCR, document classification, topic modelling, named entity recognition and other visual and linguistic tools on the set of scans in order to extract as much data and metadata from the individual documents as possible. Different document types required different treatment. For instance, we quickly determined that it did not make much sense to do OCR on a tiny handwritten note, while handwriting detection on the other hand would be possible and could yield productive results on such an item.

Experiments were conducted in visual document classification, visual document analysis and distant reading. Visual document classification was performed by clustering a combination of color and texture histograms derived from the scans. This step was taken mostly because the existing index of the archive is incomplete: it has information on the folders in the archive, which contain the documents, but not the documents themselves. The Media Suite works with individual documents, not folders, so it became necessary to, for instance, discern sub-folder covers from the documents inside.

A second reason to do classification was that each type of document needs a different kind of processing - typed letters can be OCR-ed, but not photos, while handwritten letters could be classified by comparing handwriting styles. By separating different document types it became possible to employ the most effective information extraction tools on them. This procedure also allowed for finding visually similar documents, making it possible for researchers to look for similarities in for instance texture or color.

The typewritten documents were OCR-ed, then classified on the basis of the recognized text in order to differentiate e.g. personal letters from business correspondence. Named entity recognition on the texts provided us with a network of people and places, with links to the letters. Attempts at handwriting recognition on the basis of 'image texture' histogram comparisons provided mixed results, - for the instances where larger samples of a single person's handwriting were available it worked reasonably well, but for handwriting types occurring only a few times the confidence of the classifier was too low and such documents were classified as one of the more frequently occurring types. The results of these steps, in combination with the existing index's metadata, provided a rich enough metadata structure for the use of individual documents in the tool.

In addition to a discussion of these steps, our paper reflects on the results' epistemological implications for future research, by discussing them in relation to previous quantitative approaches to the Desmet Collection. From this vantage point, our paper argues that while previous quantitative studies of Desmet's business documents were premised in the coding and transcription procedures of Cliometrics and *Annales* historiography, MIMEHIST's results nurture exploratory and qualitative research coupled with serendipitous search and annotation procedures focusing also on visual features. Consequently, the paper argues, researchers may to a greater degree than hitherto highlight data contingencies and multiplicity of viewpoints in the Desmet business archive.

The 91st Volume — How the Digitised Index for the Collected Works of Leo Tolstoy Adds A New Angle for Research

Boris V. Orekhov
nevmenandr@gmail.com
National Research University Higher School of Economics, Russian Federation

Frank Fischer
ffischer@hse.ru
National Research University Higher School of Economics, Russian Federation

Introduction

The collected works of Leo Tolstoy were printed and published in 90 volumes of some 46,000 pages between 1928 and 1958. The visibility and usability of these volumes were increased by the project "Tolstoy Digital", a TEI-encoded version of this vast resource (Skorinkin & Mozhaev 2016).

This talk, however, is not about the 90 volumes themselves, but about the 91st volume of this edition, a supplement volume containing indexes of works and proper names, from both the fictional works and the many volumes containing Tolstoy's letters.

"The 91st Volume" is a web application based on the digitised index of proper names for the 90-volume collection of Tolstoy's collected works (http://index.tolstoy.ru/). The digitised data features additional properties,

which can be explored by the enthusiast as well as the specialist.

This talk tries not just to present a new tool for literary scholars, but tries to generalise how this kind of resources can be used to gain new insights into larger text collections.

Level 1: Enhanced Searches

First and foremost, the index retains its original functionality, which is to map names to volumes and pages. Collected works of a canonical writer are not primarily meant to be read one by one, line by line. A 90-volume collection of books does not only contain entertaining narratives, but it can also be viewed as a set of facts, dates, names, mentions, etc. An index is the key to this data, and it was the only means to gain some orientation in the pre-digital age.

In the web app version of the "91st Volume", the index is even more convenient to use than in the paper version, as it allows "fuzzy" searches. By entering "ava" it will list among the results terms like "Poltava", "Bavariâ", or "Abdulla-al'-Mamun Zuravardi". The higher the frequency of a name within the whole collection, the higher up it will be displayed in the results. These types of searches are already an enhancement over the traditional index search.

If we cannot define in advance what we are looking for, we still have the lists of all names in the index (which sum up to more than 16,000 entries). Once we've found what we were looking for, we don't need to remove any book from its shelf and open the right page, but can jump directly to the corresponding page.

A graphical word-cloud representation is also featured and conveys a first idea about the most frequent words in the corpus.

Level 2: Studying Life and Works of Leo Tolstoy by Means of Network Analysis

Turning an index of names into a network is a new approach to facilitate the study of contexts. The co-occurrence of names in the same environment (on the same page, in the same chapter, etc.) reveals similarities and relations between different entities, which on the scale of 90 volumes, helps us to understand larger contexts.

"The 91st Volume" unfolds a rather unconventional social network of Leo Tolstoy. It shows not only Tolstoy's connections with other people (e.g., his pen pals), but also the connections of people from the point of view of Tolstoy.

The co-occurrence of proper names on the same page within the 90 volumes establishes an edge of the emerging network as it creates a link between two entities. For example, the Hindu scripture "Bhagavat-gita" can be found five times on the pages of the Complete Works, and it shares these five pages with a total of 43 other names mentioned. The proximity of these mentionings is not accidental, of course, in our example they form some kind of "Indian cluster" containing works like "Gitopadeša", "Dhammapada", "Vamana Purana", or names like Ramakrišna Šri Paramagamza.

For Tolstoy, the mentioned texts are part of a set of carriers of philosophical knowledge, and are associated with names like Xenophon, Montaigne, Montesquieu, Pascal, Skovoroda, Socrates. These networks provide great opportunities for understanding the whole range of Tolstoy's interests and ideas. It presents a panoramic picture revealing general trends and larger thematic clusters. For each individual name there is also a small graph showing the most significant names associated with it.

Another new kind of access to the 90 volumes is a heat map that shows the density of proper names used in each of them (the more names mentioned, the warmer the colouring).

In the first volume of the collection containing youth experiments, a red splash suddenly appears in the middle of a rather calm blue background on page 269. You can view this page and will find that it contains a list of European cities: Rome, Naples, Dresden, Berlin.

Level 3: Editorial Evolution of the "Complete Works"

The index also allows scholars to study the coming into life of the "Complete Works of Leo Tolstoy", i.e., the difficulties that had to be overcome when working on this edition (as they are laid out in Osterman 2002). The "91st Volume" allows us to understand how editorial principles have changed over time, especially as regards the depth of commenting.

For example, the 13th volume, with draft editions of "War and Peace", has a weak commentary, and the 47th volume (diaries and notebooks) features such detailed comments that it is the most detailed in the entire 90-volume edition. Quantifications like this allow us to draw conclusions to the process of editing the Complete Works over three decades.

Like mentioned above, the web app retains all the capabilities of the traditional index, and at the same time extends its potential through computer-based information management, a multi-purpose search engine and different kinds of visualisations. The app is to be understood as a suggestion to apply the newly developed methods to the Collected Works of other authors.

References

Osterman L. (2002): *The Battle for Tolstoy: History of the Publication of Tolstoy's Complete Works*. [Srazhenie za Tolstogo. Istorija izdanija Polnogo sobranija sochinenij Tolstogo.

Skorinkin D., Mozhaev E. (2016). TEI markup for the 90-volume edition of Leo Tolstoy's complete works. In: *TEI Conference and Members' Meeting 2016. Book of Abstracts*. Vienna: Austrian Centre for Digital Humanities, pp.107–109.

Adjusting LERA For The Comparison Of Arabic Manuscripts Of _Kalīla wa-Dimna_

Beatrice Gründler
beatrice.gruendler@fu-berlin.de
Freie Universität Berlin, Germany

Marcus Pöckelmann
marcus.poeckelmann@informatik.uni-halle.de
Martin Luther University Halle-Wittenberg, Germany

Introduction

In this paper, we present the collaboration between the pilot project *Kalīla wa-Dimna – Wisdom encoded*[1] and LERA.[2] In the project's first phase, devoted to experimenting with existing tools and identifying necessary adjustments, we adopted and generalized LERA for the classical Arabic language. This modification worked well and thus will become a cornerstone for future research within the ERC-Advanced Grant Project "AnonymClassic" (Gruendler 2017).[3] The benefit is already apparent, yielding first observations of the text's development, and the tool was successfully applied in an undergraduate academic course at the Seminar for Semitic and Arabic Studies of FU Berlin.

Kalīla wa-Dimna – Wisdom encoded

Using Sanscrit sources, *Kalīla wa-Dimna* was composed in Middle Persian (version lost) in the 6th century CE and ultimately translated into in forty languages worldwide. Its key phase is the Arabic translation from the 8th century, from which all later translations derive. But this version has proliferated in many variants, which prevents a conventional critical edition by stemma (Gruendler 2013). This project seeks to assess via a (partial) synoptic critical edition the range of variation between selected Arabic manuscripts of this work. These derive from the 13th to the 19th century CE.

LERA – Locate, Explore, Retrace and Apprehend complex text variants

LERA is an interactive, digital tool for analyzing variations between multiple versions of a text in a synoptic manner with several differences to other well known collation tools (Schütz and Pöckelmann 2016). It was first developed for printed texts of the French Enlightenment (Bremer et al. 2015) within the SaDA-project[4] at Martin Luther University Halle-Wittenberg and since then adopted to other texts and languages.

The tool follows three major steps for generating the synopsis. The first is a segmentation of the given manuscripts. In the case of *Kalīla wa-Dimna,* the text-units are narrative steps. Second, an automatic alignment of these segments is done by the built-in algorithm. The researcher can intervene afterwards by moving, cutting or merging the segments if necessary. The final step is the detailed comparison of the aligned segments' words by the system. The identified variants are highlighted with colors depending on the kind of difference. Besides, a variety of filters are available, e.g., hiding all changes that purely concern punctuation. On this basis, a comprehensive and easily readable apparatus is generated. The proto-edition thus produced can be downloaded in several formats.

Moreover, LERA provides further assistance for the analysis of the variants. Most prominent is CATview (Pöckelmann et al. 2015), a graphical representation of the alignment that facilitates overviewing and navigating within the synopsis.[5] It is also associated with the word clouds and search functions of LERA.

Modification of LERA for Kalīla wa-Dimna

In this project, LERA made its debut in classical Arabic, which has required some language-specific adoptions. Processing the Arabic alphabet was rather uncomplicated, because the system already uses Unicode. Regarding the backend, the processes for tokenizing, indexing (for search) and language recognition were extended. On the other hand, modifications for the frontend included adding a font for the alphabet, displaying the correct writing direction, and revising the download function.

More important, some specific needs for the *Kalīla wa-Dimna* project have already been implemented. LERA now allows the manual alignment by experts using unique segment identifiers, which are encoded within the manuscripts' XML/TEI files. On this basis, we also added identifiers for the segments that can be edited and displayed in the synoptic view. On major improvement is the visualization of transposed segments. They occur if the order of the segments within one manuscript was changed or when similar segments appear somewhat distant to each other, but aligning them is blocked due to other aligned segments. We included an option to display copies of them in the synopsis that will be shown grayed and are linked to their actual position. They will also appear in CATview.

1 E-Learning/E-Research project, located at and funded by Freie Universität Berlin, homepage http://www.geschkult.fu-berlin.de/e/kalila-wa-dimna

2 Information and a demonstrator can be found at https://lera.uzi.uni-halle.de

3 No. 742635, "The Arabic Anonymous in a World Classic," PI Beatrice Gruendler, Freie Universität Berlin, see http://www.geschkult.fu-berlin.de/forschung/erc/anonym_classic/index.html

4 See the project's homepage: https://sada.uzi.uni-halle.de

5 Further information on CATview is also available at https://catview.uzi.uni-halle.de

In respect to the project's goal to investigate the interrelation of the manuscripts of *Kalīla wa-Dimna*, we developed two new modes for coloring the variants. The first one only highlights passages that are unique to one manuscript, which points to some independence of the copyist-redactor regarding the other manuscripts. The second mode only highlights those passages that appear in exactly two manuscripts. Finding such pairs is important evidence that suggests that both manuscripts are related to each other.

Another helpful extension is the so called 80%-filter, which leads to treating words as identical if they share at least 80% of their letters. This approximating similarity measure is grounded in the property of the Arabic language that words with identical roots tend belong to one semantic field.

Benefits

LERA could be adjusted for the first phase of this interdisciplinary collaboration. Based on this, we already made interesting observations: against our assumption, the first analysis shows that there are no distinct groups of manuscripts. Instead, variations fluctuate, forming continua in which some manuscripts cumulatively assemble reformulations that appear scattered among others.

Furthermore, in the summer semester of 2017 the project was integrated into an undergraduate academic course on *Kalīla wa-Dimna* at Freie Universität Berlin. The students used the synopsis to explore the variants of five aligned manuscripts in class and wrote papers applying this method individually.

Conclusion

With the work presented here, we established a foundation for a comprehensive analysis of *Kalīla wa-Dimna*. Owed to the text's complex history and manifold variants, this ambitious project is planned for a timespan of ten years. With the ongoing research, more features of analysis will be needed. This includes an advanced utilization of language specific information for the comparison, e.g., a root extraction for Arabic words. Moreover, the comparison and visualization of manuscripts of *Kalīla wa-Dimna* in other language is being considered. Finally, the functionality to comment on the identified variants is crucial for their scientific investigation.

References

Bremer, T., Molitor, P., Pöckelmann, M., Ritter, J. and Schütz, S. (2015). "Zum Einsatz digitaler Methoden bei der Erstellung und Nutzung genetischer Editionen gedruckter Texte mit verschiedenen Fassungen - Das Fallbeispiel der *Histoire philosphique des deux Indes* von Guillaume Thomas Raynal." In v. Nutt-Kofoth, R., Plachta, B. and Woesler, W. (eds), *Editio,* 29(1), pp. 29–51.

Gruendler, B. (2013). "Les versions de *Kalīla wa-Dimna*: une transmission et une circulation mouvantes." In Ortola, M. (eds), *Énoncés sapientiels et littérature exemplaire: une intertextualité complexe*. Nancy, pp. 385-416.

Gruendler, B. (2017). "The Arabic Anonymous in a World Classic (Acronym: AnonymClassic). Presentation of a Research Project." *Geschichte der Germanistik,* 51/52, pp. 156-57.

Pöckelmann, M., Medek (⁎Gießler), A., Molitor, P. and Ritter, J. (2015) "CATview - Supporting the Investigation of Text Genesis of Large Manuscripts by an Overall Interactive Visualization Tool Digital Humanities." *Digital Humanities 2015: Conference Abstracts.* Sydney: UWS. http://dh2015.org/abstracts/xml/POCKELMANN_Marcus__CATview____Supporting_The_Inve [accessed 11/27/2017]

Schütz, S. and Pöckelmann, M. (2016) "LERA - Explorative Analyse komplexer Textvarianten in Editionsphilologie und Diskursanalyse." In: *Book of abstracts of the third annual conference of Digital Humanities for German-speaking regions (DHd 2016)*, pp. 249-253.

Afterlives of Digitization

Lily Cho
lilycho@yorku.ca
York University, Canada

Julienne Pascoe
julienne.pascoe@gmail.com
Library and Archives Canada; Canadiana.org, Canada

This paper is based on our commitment to the possibilities of re-thinking the processes of digitization such that digitization does not end with the uploading the scanned object and archivally-mandated metadata. Rather, that point is merely the beginning of the life of any particular digital collection. The ways that any collection is used by academic researchers, community groups, and members of the public should contribute to the processes of digitization. These collections live when they are used and these uses should be reflected in the collection so that other researchers can see and build on this work. Every collection that has been digitized has an afterlife. But how can we use new technologies – in particular, the International Image Interoperability Framework (IIIF) and linked data – in order to make these afterlives visible and usable? How can we develop infrastructure and protocols so that the metadata *lives*?

Our project focuses on building a platform for annotations based around a specific collection of images: the Chinese Immigration records, initially captured by Library and Archives Canada (LAC), and subsequently digitized, preserved and made accessible online by Canadiana.

There are approximately 41,000 images in this particular set of archival images that we work with. Canadiana has recently completed the digitization of this collection. Because it is comprised of a nearly complete set of immigration certificates for individual Chinese migrants collected between 1910 and 1953, the collection is particularly rich for researchers working in the area of race, immigration history, and citizenship.

In working with these materials, Lily Cho's research team has identified several layers of annotations that would be pertinent to this material. For example, the research team has transcribed names of each immigrant on the record. Each image contains two names: the anglicized Chinese name written down by an immigration agent, and a name in Chinese script written by the migrants themselves. In our transcriptions of the first several hundred images, there is no correspondence between the name written by the immigration agent and the name in Chinese script. Because these records were used to identify individual immigrants for the purposes of allowing them to exit and enter Canada (and thus functioning much like a passport for Chinese immigrants who were, during this period, denied the rights of citizenship), this finding radically changes our understanding of how Chinese immigrants navigated racist immigration controls during this historical period. However, there is currently no way for her research team to contribute to the metadata already attached to this collection.

Such contributions to the metadata already in place function as annotations in this project. Working in partnership, Cho and Julienne Pascoe, who has been the Lead Metadata Architect for Canadiana and is now serving as a Digital Archivist at LAC, are developing a platform for supporting annotations for this archive using the Web Annotations standard, the IIIF, and linked data. Canadiana is currently in the process of implementing IIIF as well as the initial stages of developing a data model that would provide the foundation for such a partnership. In short, this project uses IIIF as a framework for enabling open standards for annotations that can then be reused as linked data - all three areas coming together to support the linking, sharing and re-use of metadata.

This paper reports on the progress we have made in developing this platform, and will also briefly outline the possibilities for the use of this platform beyond this specific collection of images. Although museums and archives are under enormous pressure to digitize their collections, and are rapidly in the process of doing so, these digitization initiatives are rarely undertaken in conversation with some of the primary users of these digitized texts and objects: academic researchers. For example, metadata that meets archiving standards is not necessarily useful for researchers, and is often based on hegemonic archival practices that reinforce colonial structures and narratives. At the same time, academic researchers often have resources to contribute to, and enrich, the digitization that has been accomplished as well as facilitate postcolonial interpretations of the archive. This project brings academics and digital archivists together in order to develop protocols so that digitized collections can be dynamically connected to the communities using them. Once digitized, a collection does not need to remain static. It can respond to, and include, the findings of researchers in the community; and these findings could and should be made available to other users of the collection. However, protocols for curating, organizing, and disseminating this information must be developed. This project will use one specific collection, an archive of approximately 40, 000 head tax certificates held by LAC and digitized by Canadiana, as a test case for developing precisely the kinds of protocols that would allow a digitized collection of materials to leverage the findings emerging from people using these materials.

Rapid Bricolage Implementing Digital Humanities

William Dudley Pascoe
bill.pascoe@newcastle.edu.au
University of Newcastle, Australia

This paper presents a practical approach to building digital humanities (DH) at a university, across disciplines with diverse requirements, starting without institutional support, with scarce staff on a low budget. Examples are provided from the Centre For 21st Century Humanities (C21CH) at University of Newcastle, Australia (UON).

Digital humanities (DH) requires expertise that crosses many fields from specific humanities disciplines to software development and production management. DH has a broad range in scale – from a scholar learning basic programming to hack a Python script, to multi-institutional collaborations on neural network learning. Few people are experts in all these fields meaning DH is often a collaboration. The requirements for any individual DH project can differ greatly also requiring IT skill sets that may not be easy to find in any one individual. This makes it difficult for university humanities departments with no spare cash, and often reluctant to invest heavily in IT, to successfully support DH, yet DH projects present problems beyond standard service offerings and provisioning and different to STEM. The Digital Lab of C21CH at UON has evolved an approach, here called 'rapid bricolage', that has successfully delivered a range of sustained internationally recognised DH projects influencing national debate. Some comparison will be made with other approaches, and while not necessarily suiting all circumstances, 'rapid bricolage' has proved an effective approach catering to characteristic issues in DH research, drawing from but differing to established IT practice.

This 'rapid bricolage' approach draws on 'rapid' software development and 'bricolage', both common practice in software development and humanities, but modifies them to meet the unique needs of Digital Humanities. These modifications are epistemic, structural, methodological and a matter of degree. It has also crucially involved consultative processes to identify and Pareto prioritise inter-disciplinary interests and achievable, feasible, high impact projects. The success of these feeds back to build interest and support for DH towards funding and growth, and results in project driven infrastructure, bridging the gap between projects without infrastructure and infrastructure without projects by beginning with demonstrable utility and developing with shared human and technical resources.

C21CH projects include (http://c21ch.newcastle.edu.au):

- Colonial Frontier Massacres (v1.2) – map of massacres in Australia.
- EMWRN archive (v2) – innovative archive of material cultures of early modern women's writing.
- Intelligent Archive (v3beta) – stylometry software.
- ELDTA site (v1alpha) – linguistics web player for media with tiered glosses, translations etc.
- Text To Map (prototype) – online automatic recognition and mapping of places in texts, linking to and from the text and the map.
- Scriptopict (v1) – annotations for images eg: *Battle of Kurukshetra Mural* and *Mixtec Glyph.*

Rapid

Rapid application development is an established methodology for software development focusing on getting a prototype working as early as possible followed by regular review with clients and incremental feature additions and bug fixes. For humanities departments this approach ensures that at least some software exists as an outcome of initial spending when the budget is tenuous and provides an encouraging proof of concept. For the cost of a meeting with several professors or executives a working prototype can be developed, making it worth simply trying it out rather than lengthy discussions about the value of proceeding. A rapid approach also helps greatly when the client is unclear of what is needed or has little understanding of IT. An early prototype establishes confidence and commitment early. Gaps in desired functionality immediately become clear through interaction. In particular because research is heuristic and highly changeable it allows for speculative requirements to change as the project progresses. Because of this, an even more rapid than usual approach is suited to humanities research because, as research, not all requirements cannot be known in advance. The process necessarily involves taking some action with ongoing revision, addition and enhancement. Not all aspects of humanities research activity, such as thorough rumination on a nuanced argument on a complicated problem, fit this 'rapid' model, but:

- the speculative, heuristic activities necessary to research are enhanced;
- some slower methodical activities essential for rigour and completeness can be sped up, sometimes making research possible that otherwise would not have been, or improved through the need for clear structures and definitions;
- the 'slow' process of rumination, of considering complex problems and developing arguments, while irreplaceable, can be augmented.

Bricolage

Bricolage is a well-established approach in software development. Software is typically put together from pre-existing libraries, frameworks and cut and pasted code that is modified and added to, to produce something that works in ways that conventional intellectual property and copyright are not practically applicable to. This approach is in sympathy with developments in critical theory in the late 20th century and after, with 'bricolage' and the problematization of authorship being major themes in describing postmodernity and in contemporary humanities methodology. Just as a very rapid approach suits humanities research so too is bricolage especially suitable for DH.

As research, DH typically requires constant and regular modification and adjustment, rather than delivery of a working system according to contracted specification. Much software is developed for a STEM or commercial purpose, or has a STEM like approach to problem solving. STEM and the commercial sector have larger budgets and devote larger budgets to software. This means that humanists are often in a pragmatic situation of re-using software from different disciplines despite having divergent requirements. Humanities often focus on complexity, exceptions, structural change and highly contingent historical (not repeated) events, while STEM and commerce focus on systemisation, normalisation, ceteris paribus and repeatability, for example. If humanities researchers are to avoid fitting research to the software limitations this means constantly adapting systems to their own different epistemic, ontological and methodological paradigm, ie: bricolage.

The DH research need for these two approaches, rapid application development and bricolage, combined in extremis presents challenges to established IT practice. These challenges can be met with appropriate staffing, strategy and a 'rapid bricolage' approach to build DH at a University despite diverse demands and resourcing adversity.

References

Craig, H., Pascoe, W. (2018). *Intelligent Archive v3.0* Newcastle: Centre For 21 Century Humanities

Ryan, L., Debenham, J., Brown, M., Pascoe, W. (2017). *Colonial Frontier Massacres* Newcastle: Centre For 21 Century Humanities http://hdl.handle.net/1959.13/1340762

Smith, R., Pender, P., Pascoe, W. (2017). *Early Modern Women's Research Network Digital Archive* Newcastle: Centre For 21 Century Humanities http://hdl.handle.net/1959.13/1326860

The Time-Us project. Creating gold data to understand the gender gap in the French textile trades (17th–20th century)

Eric de La Clergerie
eric.de_la_clergerie@inria.fr
ALMAnaCH, Inria, France

Manuela Martini
manuela.martini@univ-lyon2.fr
LARHRA, Université Lyon 2, France

Marie Puren
marie.puren@inria.fr
ALMAnaCH, Inria, France

Charles Riondet
charles.riondet@inria.fr
ALMAnaCH, Inria, France

Alix Chagué
alix.chague@enc-sorbonne.fr
ALMAnaCH, Inria, France

The role of women in industrial development is now largely recognized in sociological and economic studies on developing countries during the first industrial revolution in Europe. Yet data on their remuneration, schedules and domestic work, and that of men working in the same sectors, remains deficient for many regions, especially for France. The Time-Us project aims to reconstruct the remuneration and time budgets of women and men working in the textile trades in four French industrial regions (Lille, Paris, Lyon, Marseille) in a long-term perspective, by bringing together a multidisciplinary team of historians, natural language processing (NLP) experts and sociologists. It will create comparable series on the remuneration and time allocation of employed men and women (i)through classical sources and company and trade association archives, and (ii)by piecing together a series of qualitative sources identifying words and actions associated with work in both domestic and non-domestic activities. The project will provide keys to understanding the gender gap by analyzing changes in work and time uses during the first industrialization process.

The Time-Us team works on a heterogenous corpus of French handwritten and printed sources spanning from the seventeenth to the twentieth century. These documents are mainly preserved in French local archives, from the four industrial regions that have been mentioned above (for instance, Archives municipales de Lyon, Bibliothèque municipale de Lyon, or Bibliothèque nationale de France in Paris, etc.). The analyzed corpus brings together numerous historical sources, and includes court decisions, petitions, police reports and files, and sociological surveys on living conditions of the working class (especially *Les monographies de famille de l'École de Le Play* or Le Play's families' budgets (Hincker, 2011)). Many of these documents are manuscripts, written by various hands over long periods of time (more than a hundred years for the "Registre de contraventions aux règlements des métiers" that begins in 1670 and ends in 1781 (Lyon, Archives municipales).

This unpublished set of documents constitutes an important corpus of historical sources that is well-suited for applying computational analysis. In this paper, we will present the approach adopted by the Time-Us team to analyze this corpus. We will also discuss the prospects opened up by this project for historical research in terms of digital research workflow.

Our goal consists in applying NLP methods to heterogeneous historical documents, in order to identify and analyze the relevant semantic or syntactic patterns that describe work, remuneration and time budgets. The application of such methods, mainly parsing, will facilitate the analysis of the corpus by creating series of comparable quantitative and qualitative data:

Quantitative data on remunerations, household budgets and time spent for domestic (or unpaid) and non-domestic (or paid) work by women and men.

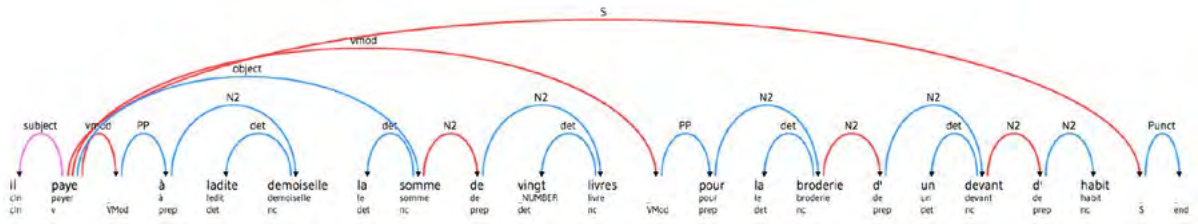
Qualitative data on paid and unpaid tasks realized by women at home and at work, namely information on the type of the task, its description, its duration and its results. Computational methods will also be used to extract statements describing the women performing these tasks (occupation, social status, age, marital status, family composition), and the relationships between the actors involved in these tasks, especially between men and women (family relationships such as husband and wife, brothers and sisters mothers and sons, or working relationships such as employers and employees).

The analyzed sources can take a number of varied forms. Thus, we chose to work closely with economic and labour historians in the data modelling process. As the corpus gathers together diverse historical sources, the definition of a light and flexible annotation schema, bringing together the history and language processing experts, is a major step to create "gold data" to train parsing models. This gold data take the form of annotated texts encoded in TEI (*Text Encoding Initiative*). TEI can be seen as

a bridge representation for historians and NLP experts: in this approach, historians annotate a first set of documents in TEI, in order to create training data that can be easily processed and analyzed by NLP experts. Besides, the choice of the TEI allows for the creation of sustainable data, that can be reused in the long term by other projects and researchers. Our aim consists also in creating a flexible TEI data model that will be relevant to modelize differents types of data, and that will enable NLP experts to extract comparable information such as quantitative data (amounts of money, period of time…). In this way, this model could be reused by other research projects especially, but not only, projects of economic and labor history.

A first step is the transcription of the manuscripts into a simple TEI representation, covering the text and a set of metadata. This task is nothing but trivial, due to the diversity of sources mentioned above, but it is not the scope of this paper. Then, the representation is enriched by annotation layers. The first annotation layer is the recognition of tasks and occupations, linked to their associated amounts of money, and the actors of the transaction. The extraction of Named Entities such as person and place names is also necessary in order to properly analyse how gender and localization influence remuneration.

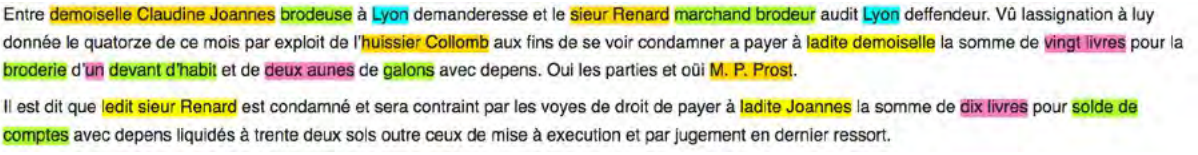
The annotation process will start as a collaborative effort, in order to get a first dataset that could possibly be used to train/configure NLP tools, but also to help designing a precise annotation guide between the NLP people and historians. At a later stage, we will progressively deploy more automatic NLP tools to create these annotations. In this regard, we plan to identify the elements of vocabulary (tasks, products) and the interesting phrases (e.g."*someone was paid (this amount) for (this product) for a (given amount of time)*"), using knowledge acquisition techniques based on the distributional hypothesis and syntactic analysis of the corpus. The knowledge of the domain will allow us to define syntactic extraction pattern to be applied on the corpus to detect and annotate specific instances of tasks, products, money, people and relationships between these pieces of information. Some human validation will still be needed to filter the vocabulary, refine the patterns, and propose missing elements (vocabulary and patterns). Language processing will be conducted with the French processing chain developed by the INRIA Almanach team, and in particular with the FRMG parser (Morardo and de La Clergerie 2014). Parsing produces dependencies between words, allowing us to identify who does what, when, how for some event. The processing chain has already by used several times for knowledge acquisition over specific domains (legal, medical). In our case, specific issues may arise because of the quality of the transcriptions and the peculiarities of the language used, which contains archaic constructions, whereas our parser was designed for contemporary French.



Example of a parse for one sentence of the corpus

The annotation task is therefore mainly collaborative, so the need for a shared framework has emerged. Several digital projects have already taken into account the specific needs of historians in terms of image visualization, transcription and collaboration. For instance, the *Transkribus* interface enables Humanities scholars to transcribe handwritten and printed historical sources, and offers a very powerful Handwritten Text Recognition engine. The project *Transcribe Bentham* takes account the collaborative dimension in transcribing historical documents. The *Old Bailey* transcription project uses a combination of hand encoding an automatic recognition and extraction systems. Nevertheless, they do not address all the requirements of Humanities scholars working on primary sources, and the need of comprehensive Digital Humanities-based publishing systems is emerging. We have chosen to setup a specific digital workflow enabling historians and NLP experts to work together. We will present the solution that has been put in place, and especially a customized wiki with:

the Transcribe Bentham transcription desk, adapted to our needs,
and a TEI toolbar, specifically customized for tagging named entities and measures.


Entre demoiselle Claudine Joannes brodeuse à Lyon demanderesse et le sieur Renard marchand brodeur audit Lyon deffendeur. Vû lassignation à luy donnée le quatorze de ce mois par exploit de l'huissier Collomb aux fins de se voir condamner a payer à ladite demoiselle la somme de vingt livres pour la broderie d'un devant d'habit et de deux aunes de galons avec depens. Oui les parties et oüi M. P. Prost.

Il est dit que ledit sieur Renard est condamné et sera contraint par les voyes de droit de payer à ladite Joannes la somme de dix livres pour solde de comptes avec depens liquidés à trente deux sols outre ceux de mise à execution et par jugement en dernier ressort.

Customization of the TEI toolbar

References

Clergerie, É. D. L., Sagot, B., Stern, R., Denis, P., Recourcé, G. and Mignot, V.
(2009). Extracting and Visualizing Quotations from News Wires. vol. 6562. Springer, pp. 522–32 doi:10.1007/978-3-642-20095-3_48. https://hal.inria.fr/inria-00607463/document (accessed 24 April 2018).

Hincker, L. (2001). Les monographies de famille de l'École de Le Play. Les Études sociales, n 131-132, 1er et 2e semestres 2000. *Revue d'histoire du XIXe siècle. Société d'histoire de la révolution de 1848 et des révolutions du XIXe siècle*(23): 274–76.

Morardo, M. and Clergerie, É. V. de L. (2014). Towards an environment for the production and the validation of lexical semantic resources. https://hal.inria.fr/hal-01005464/document (accessed 24 April 2018).

Seaward, L. and Kallio, M. (2017). Transkribus: Handwritten Text Recognition technology for historical documents. Montréal https://dh2017.adho.org/abstracts/649/649.pdf (accessed 24 April 2018).

Thomasset, F. and Clergerie, É. D. L. (2005). Comment obtenir plus des Méta-Grammaires. *Proceedings of TALN'05*. Dourdan, France.

University College London UCL Transcribe Bentham http://blogs.ucl.ac.uk/transcribe-bentham/ (accessed 24 April 2018).

Old Bailey Online - The Proceedings of the Old Bailey, 1674-1913 - Central Criminal Court https://www.oldbaileyonline.org/ (accessed 26 April 2018).

Modeling Linked Cultural Events: Design and Application

Kaspar Beelen
k.beelen@uva.nl
University of Amsterdam, The Netherlands

Ivan Kisjes
i.kisjes@uva.nl
University of Amsterdam, Netherlands, The

Julia Noordegraaf
j.j.noordegraaf@uva.nl
University of Amsterdam, The Netherlands

Harm Nijboer
harm.nijboer@huygens.knaw.nl
Huygens Institute for the History of the Netherlands, The Netherlands

Thunnis van Oort
t.vanoort@uva.nl
University of Amsterdam, The Netherlands

Claartje Rasterhoff
c.rasterhoff@uva.nl
University of Amsterdam, The Netherlands0

Introduction

This paper discusses the promises and pitfalls of linking historical data on cultural events. Quite a few datasets on historical European music, theatre and film are now publicly available online (Baptist 2017). The ones that contain programming information are, at least to some extent, already event-based. However, they are highly heterogeneous in scale and scope, and they generally do not use the same definitions for, for example, venues, events, or companies. Conceptualizing and embedding cultural events such as concerts or theatrical performances in a linked data framework helps to overcome such issues without forcing an overarching ontology, and it enables researchers to acknowledge the performative and interactive nature of cultural expressions within their (local) societal context (Nijboer and Rasterhoff 2018).

By linking event data internally as well as to external knowledge bases such as DBpedia and Wikidata by means of shared vocabularies, researchers are invited to systematically analyse cultural life cross-sectorally (i.e. theatre, music), internationally (European comparisons and connections), and contextually (in relation to local social, economic, political and cultural features) (cf. EPAD: European Performing Arts Dataverse). In this paper we discuss the conceptual and practical requirements for such a linked-data approach on the basis of a series of research projects on historical cinema, musical, and theatrical events in the research program Creative Amsterdam: An E-Humanities Perspective (CREATE).

Cultural events

Events play a key role in historical scholarship, and have gained even more urgency with the increasing importance of digital resources in humanities research. Many projects on historical events, however, employ them as devices to structure data collections and do not explicitly aim to develop analytical frameworks in relation to event data collection and data modeling (De Boer et al. 2015; Van Hage et al. 2011; Shaw 2013. An exception can be found in a statistical method known as event history analysis, which treats events as dependent variables, seeking to statistically describe, explain, or predict their occurrence (Allison 2004). Most research on (urban) arts and culture, however, does not try to statistically identify variables that predict or explain an event, for example the staging of the opera Norma or the screening of the movie Casablanca. Rather, historians may seek to identify (series of) events that have contributed to, for example, the canonical status of specific expressions or genres, to the shaping of local and international cultural taste cultures, or to the emergence of some places as particularly creative and cultural.

We therefore emphasize that (networks and series of) events should also be considered as independent variables that can help us identify and disentangle processes

of cultural change and continuity. Central in this view is the assumption that 1) events can be seen as units of analysis with structural properties (notably, a time and place) with links to, for example, actors, institutions, other events, and local properties, and 2) that these interlinkages are key to analysing their role in shaping, for instance, local cultural or social life (Tilly 2002). Turning individual event datasets into linked data versions would provide instantaneous insight into how much performing arts datasets overlaps, ontologically, with any of the others. This provides a roadmap for integrating these still scattered data and studying them in conjunction. A systematic analysis of cultural events therefore requires a data structure which allows for querying connections.

Linking cultural event data

A first analysis of performing arts datasets demonstrated that normalizing even the most basic data across datasets is tricky and that trying to completely harmonize and link all the relevant datasets is futile (Baptist 2017). Fortunately, the structure of linked data provides a way to transparently query heterogeneous data, without enforcing an overarching ontology. Breaking events down into variables such as 'people', 'venues', 'place', and 'time', for instance, circumvents the issue of formally defining a 'performance'. Linked data also allows researchers to test various different link-ups of two data sets so they can evaluate the results when they adjust their queries. In the case of cinemas, for example, one of the problems is that the typology of cinemas differs across countries and periods. In the Netherlands cinemas are divided into types 'A' and 'B' according to frequency of screenings; in Flanders the cinemas are classified according to how soon they tend to new films after their premiere. If the data was put into a relational database it would be necessary to 'reconstruct' either of the classifications for the other dataset. But linked data, because of its model of loose connections, allows querying both datasets, defining a classification only during the query.

For the datasets on cultural events such as historical musical and theatrical performances we build on a rigorous relational data model by Karel Dibbets et al. for the Cinema Context database (Van Vliet et al. 2009). All movies (often circulating under various titles), persons and companies in in this dataset have been identified and aligned to a master record, and where possible linked to the well known and well maintained Internet Movie Database (IMDb). We develop this approach for other datasets and by linking data on cultural events to other datasets and to other knowledge bases using shared vocabularies such as schema.org and Vocabulary of a Friend (VOAF). In this paper we illustrate research potential, but also practical issues by discussing a recent project on the establishment of movie theatres in the city of Amsterdam in the early twentieth century. By linking data on the history of cinema and movie-going to local contextual data (e.g. census data, municipal election data), we assess how linked data might be used to analyse how specific local historical characteristics shaped form and function of urban cultural life.

References

Allison, P. (2004). Event History Analysis. In Hardy, M. and Bryman, A. (eds.), *Handbook of Data Analysis*. Sage Research Methods, pp. 369-385

Baptist, V. (2017). Mapping European Performing Arts Databases. Presentation at the symposium *European Performing Arts Dataverse*, 9 November 2017, Amsterdam. http://www.create.humanities.uva.nl/epad

Cinema Context. http://www.cinemacontext.nl

De Boer., V., Oomen, J., Inel, O., Aroyo, L., Van Staveren, E., Helmich, W., De Beurs, D. (2015). DIVE into the event-based browsing of linked historical media. *Journal of Web Semantics*, 35(3), 152-158

European Performing Arts Dataverse (EPAD). http://www.create.humanities.uva.nl/epad

Nijboer, H. and Rasterhoff, C. (2018). Linked cultural events: Digitizing past events and its implications for analyzing and theorizing the creative city. In Münster, S., Friedrichs, K., Niebling, F. and Seidel-Grzesińska, A. (eds.), *Digital Research and Education in Architectural Heritage. 5th Conference DECH 2017 and First workshop UHDL 2017*, Dresden, Germany, 30-31 March 2017, Springer CCIS series, pp. 22-33

Tilly, C. (2002). Event Catalogs as theories. *Sociological Theory* 20(2), 248-254

Shaw, R. (2013). A Semantic Tool for Historical Events. In *Proceedings of the The 1st Workshop on EVENTS: Definition, Detection, Coreference, and Representation*. Atlanta, Georgia, 14 June 2013, pp. 38–46

Van Hage, W.R., Malaisé, V., Segers, R., Hollink, L., Schreiber, G. (2011). Design and use of the Simple Event Model (SEM). *Web Semantics*, 9(2), 128-136

Van Vliet, H., Dibbets., K., Gras, H. (2009). Culture in Context: Contextualization of Cultural Events. In Ross, M., Grauer, M., Freisleben, B. (eds.), *Digital Tools in Media Studies: analysis and research*.Transcript Verlag: Bielefeld, pp. 27-42

Bridging Divides for Conservation in the Amazon: Digital Technologies & The Calha Norte Portal

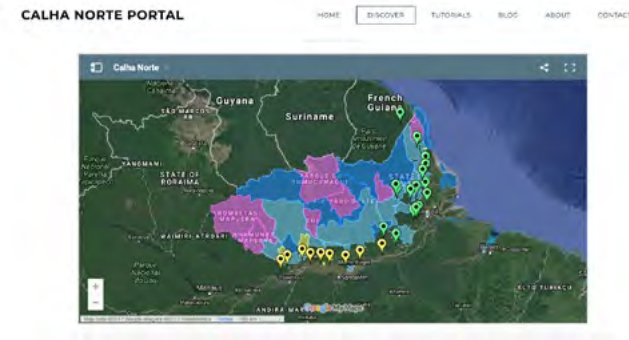
Hannah Mabel Reardon
hannahmreardon@gmail.com
McGill University, Canada

Calha Norte is the northernmost region of the Brazilian Amazon, and constitutes the largest mosaic of protected

areas in the world, encompassing nearly 14 million hectares of land. It stretches from the Amazon river in the south to the frontier regions between Brazil and the Guianese and Surinamese borders in the north, encompassing half of Pará and the state of Amapá in its entirety. Due to the vastness of the area, enforcement of protected areas can be poor, and far too little is done to involve local communities in the decision-making processes that inform conservation policy. Increasingly, digital technologies are helping to overcome some of the challenges to conservation.

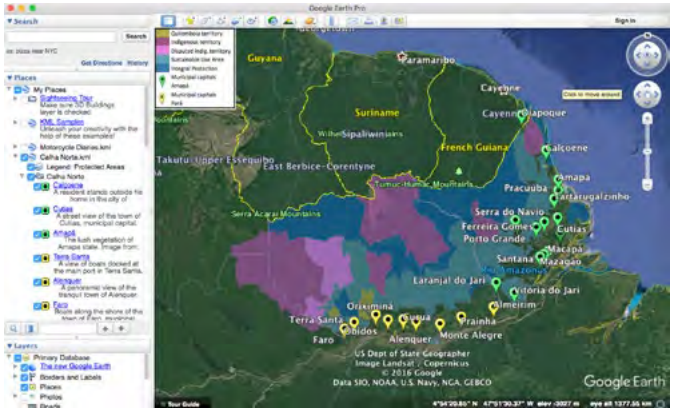
In the past ten years the Amazonian Institute for Man and the Environment (Imazon), an environmental NGO based in Belém, has made a name for itself by developing a "logging feasibility map" (Souza et al., 2010). The feasibility project uses GIS mapping software to chart data on transportation networks (including existing roads and navigable rivers), topography, biodiversity figures, deforestation areas, conservation zones and timber processing facilities. The combination of this data allows for reliable predictions of at-risk areas for deforestation, which has led to better zoning of different areas for community use, conservation and even tourism. By charting the data in a format easily interpreted by stakeholders, the feasibility maps produced by Imazon have become invaluable tools for raising public awareness and building consensus between enforcement officials, regulatory agencies, loggers, conservation groups, and local communities.

Given the immensity of the Calha Norte region and the isolation of many of its inhabitants, communication and monitoring are pressing challenges for authorities charged with the enforcement and protection of conservation zones. Initiatives like Imazon's feasibility map offer innovative solutions at a relatively low cost. Digital tools such as GIS mapping, data analytics software and tele-communications technologies are helping to bridge gaps in the knowledge and understanding of the ecological, political, and social realities of the region, thereby improving predictors of at-risk zones and response times to threats. This short paper presentation will discuss some of the ways in which digital technologies are helping environmentalists overcome the challenges of conservation

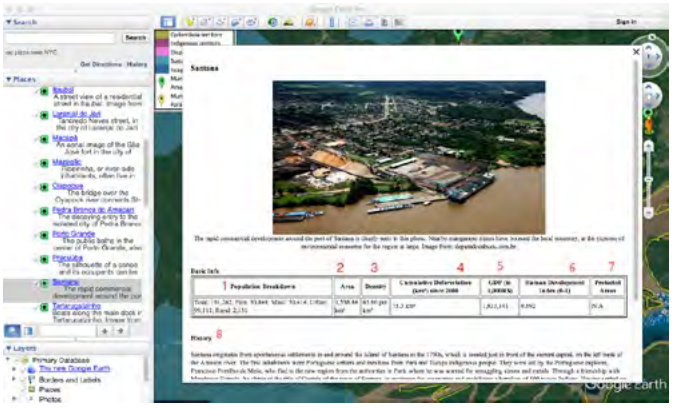


Ex. 1 & 2: Screenshots from CalhaNortePortal.com. Top: The online version of the map application, located on the "Discover" page. Bottom: The CalhaNortePortal.com homepage

The data used for Calha Norte Portal was gathered during my work with the Social Policy department at Imazon. In accordance with the department's focus on communities in the Calha Norte region, I compiled data from various sources about the region's history, cultural diversity, transportation networks, governing bodies, development indices, demographics, economic activities, protected area implementation, and accessibility. This data was then plotted in the creation of the Calha Norte Portal.



Ex. 3: Calha Norte in the Google Earth map. This application allows the user to navigate through the region by clicking on protected areas, indigenous territories, maroon communities and municipal capitals to access historical information, demographic statistics, economic and political data, photos, deforestation figures and an implementation index for protected areas. Users can also look back in time at satellite images from 1960 to the present and visualize patterns of deforestation, and urban sprawl over time.



Ex. 4: Example of the pop-up window for the municipality of Santana, Amapá.

The Calha Norte map focuses mostly on political, economic, historical, cultural and social data for populations in protected areas and municipalities. As an anthropologist, I am particularly interested in dispelling the myth of Amazonia as an uninhabited biological entity. Advocating for conservation policy which involves the participation of local communities has been the focal point of recent critical reports from the International Union for the Conservation of Nature (IUCN) (Cisneros and Orellana, 2017; Orellana, 2017). These reports reflect a general shift in attitudes on the management of protected areas away from traditional approaches which sought to isolate designated conservation zones and often neglected the history of interactions between local populations and the land in question.

My motivation in developing the Calha Norte map was to bridge some of the divides between disciplines in conservation studies. Inspired by Susanna Hecht's political ecology approach to the study of the Amazon (Hecht and Cockburn, 2010; Hecht, 2013), I wanted to create an interdisciplinary platform which would incorporate elements from the fields of anthropology, political science, history, sociology and geography. By breaking down the data I collected on the Calha Norte region in a visual, interactive format I hoped to facilitate an engagement with the region's political, social and cultural characteristics, and emphasize the importance of tailoring environmental policy to the realities of the region's inhabitants. The stand-alone map in Google Earth is meant to be played with, manipulated and explored, in ways that dismantle the linear narrative format of most textual information on the area and incorporate elements of critical cartography studies. Crampton and Krygier state that critical cartography demonstrates its political nature by "linking geographic knowledge with power" (2005: 1), suggesting that a democratization of mapping tools through digital technologies can also result in new avenues for democratizing political power. My project on the Calha Norte region aims to engage with this idea by reformulating knowledge of the region with a focus on mapping its social, cultural and historical characteristics. By advancing a more holistic vision of human interactions with the environment, I hope to make an argument for conservation policy which advocates for greater local-level management of natural resources by the inhabitants of designated protected areas.

Beyond this engagement with the critical cartography literature, my short paper presentation will also raise questions on open source access to information and digital technologies related to environmental issues in the Amazon. Particularly, my paper focuses its commentary on the complexities of conservation in the region, and the importance of greater transparency in the creation and management of protected areas, indigenous territories and traditional community lands. In relation to the over-arching theme of the conference, I hope that my presentation will demonstrate the immense potential for digital technologies to bridge divides of communication and understanding between institutional bodies, environmentalists, policymakers and the communities they serve, as well as bridging gaps in knowledge for scholars and researchers of the Amazon region.

References

Coronel Cisneros, M. y. Solórzano Orellana. J. (2017). *Comunidades locales y pueblos indígenas: Su rol en la conservación, mantenimiento y creación de áreas protegidas,* Quito: Iniciativa Visión Amazónica, REDPARQUES, WWF, FAO, UICN, ONU Medio Ambiente.

Crampton, J. and Krygier J. (2005). An Introduction to Critical Cartography. *ACME: An International Journal for Critical Geographies,* 4(1), pp. 11-33.

Hecht, S. (2013). *The Scramble for the Amazon and the "Lost Paradise" of Euclides da Cunha.* Chicago: University of Chicago Press.

Hecht, S. and Cockburn, A. (2010). *The Fate of the Forest: Developers, Destroyers and Defenders of the Amazon, Updated Edition.* Chicago: University of Chicago Press.

Reardon, H. (2018). *Calha Norte Portal.* [Online] Available at: calhanorteportal.com

Souza, C. J., Brandão, A. J. and Lentini, M. (2010). The feasibility of logging in the Pará Calha Norte region of the Brazilian Amazon. In: *Mapping Forestry.* Redlands(California): ESRI Press, pp. 1-5.

Solórzano Orellana, J. (2017). *El aprovechamiento de los bienes comunes en los bosques amazónicos: Impactos económicos, sociales y culturales de la creación y funcionamiento de áreas protegidas en dos paisajes amazónicos fronterizos,* Quito: Iniciativa Visión Amazónica, REDPARQUES, WWF, FAO, UICN, ONU Medio Ambiente.

Measured Unrest In The Poetry Of The Black Arts Movement

Ethan Reed
ecr6nd@virginia.edu
University of Virginia, United States of America

Introduction

"Anger is loaded with information and energy," says Audre Lorde ina 1981 speechon its political uses—but the nature of this affective information, sparked by a given political present, becomes highly vexed when articulated through literary objects (Lorde, 1997: 280). On the one hand, the cool detachment of aesthetic mediation keeps the politics of experimental works from being seen as mere propaganda, but runs the risk of appearing elitist or self-indulgent. On the other hand, the red-hot political outrage of a protest poem by Amiri Baraka or Sonia Sanchez grounds itself in the present, but may be attacked for subordinating aesthetic sophistication to political agendas.

Building on recent scholarship (like the work of Lauren Berlant and Sianne Ngai) suggesting that feeling gives structure to cultural formations, my research investigates the provocation and articulation of emotions like frustration, anger, and discontentment within recent US literary history as they relate to systemic injustice. An agitprop play that ends with shouts for workers to unite in class revolution; a poetic broadside that vents frustrations against white supremacy in America; a novel that indulges in a revenge fantasy against America's colonial history. Unlike plays, poems, or novels that seem to obscure, submerge, or confound their own political dimensions, these works wear their hearts on their sleeves: they are frustrated, fed up with how things are, and unafraid to speak truth to power in a direct, seemingly "un-literary" way.

"Measured Unrest in the Poetry of the Black Arts Movement" offers a proof-of-concept for performing sentiment analysis on some of the most politically and affectively charged poetry of the 20th century in America, that of the Black Arts Movement of the 1960s and 1970s. The BAM first took shape at the height of the Black Power Movement with the foundation of the Revolutionary Theatre by Amiri Baraka in 1965. As Larry Neal—one of BAM's principal theorists—says in a 1969 manifesto, the "Black Arts movement seeks to link, in a highly conscious manner, art and politics" toward "the liberation of Black people" (Neal, 1969: 54). Moreover, what Neal calls the movement's "black esthetic" is famous for its affective dimensions, often exploring the limits and political uses of anger, frustration, and militant poetic rage. But while BAM writers sought to link art and politics through explicitly racial terms, many—though by no means all—were marked by a failure to attend to the intersections of gender with racial injustice.

In this project I ask two questions in particular: first, how are the feelings associated with injustice in this corpus coded in terms of race and gender? And second, what can natural language processing techniques like sentiment analysis show us about the relations between different dimensions of poetry—like affect and gender—given that poetry is highly figurative and notoriously difficult to quantify in terms of sentiment or opinion?

Method

In addressing both these questions, this project uses a small corpus of poetry—currently 26 books—from prominent BAM authors. I employ both close reading as well as machine reading techniques, combining the powerful scale of sentiment analysis with the granularity of traditional literary analysis in an effort to explore the intersections of feeling, gender, race, and injustice in the radical poetry of this period. My goal in this project is not to develop a sentiment classifier that works on experimental poetry in English. Rather, it is to see what existing classifiers can show us about a specific corpus of poetry.

In this sense, I use pre-existing sentiment classifiers like VADER and Pattern (via TextBlob) to perform a kind of exploratory computational analysis on my corpus (Hutto and Gilbert, 2014; De Smedt, and Daelemans, 2012). Rather than use these tools to make general claims about this incredibly diverse body of poetry, I test, experiment, and make targeted use of sentiment analysis techniques to pursue research questions already present in existing scholarly conversations—for example, how poets might tie heightened affects to an explicitly political quest for racial justice in America. The insights I draw from my computational analyses, then, go hand in hand with more traditional literary practices. Moreover, my methodology aims to acknowledge the fact this poetry was written in the shadow of government surveillance programs, active FBI counterintelligence operations, and a larger culture fearful of radical thought. Because of this, my project explores the fraught methodological implications of using distanced, potentially decontextualizing computational text analysis techniques to think through BAM poetry, and how these methods might best be used to pursue questions, problems, and lines of inquiry centered around black thought and experience.

The already vibrant conversations on sentiment analysis and natural language processing more generally have been illuminating in forming these thoughts and questions. The discussion between Matthew Jockers and Annie Swafford on the*Syuzhet*package and "archetypal plot shapes" has helped me not only to consider the currentpossibilities and limitationsof sentiment analysis as applied to literary corpora, but also to think through the kinds of results we expect from digital projects and how we verify those results as an academic community (Swafford, 2015; Jockers, 2015). With regards to poetry and NLP more specifically, Lisa Rhody'stopic modeling of highly figurative ekphrastic poetryis a great model for

how unexpected failures in textual analysis can also be productive, prompting us towards new questions as well as new understandings of familiar methods like close reading (Rhody, 2012).

Results

I have implemented NLP techniques with NLTK and TextBlob, a text-processing Python library, on my collection of 26 books of poetry. I have also used two sentiment classifiers—Pattern (via TextBlob) and VADER—to evaluate my corpus for sentiment and interpret my results. While this work is ongoing, so far my work comprises explorations and experiments in the smaller-scale uses of sentiment analysis in the study of poetry and affect.

For example, Pattern considers Quincy Troupe's "Come Sing a Song"—from his 1972 collection *Embryo Poems, 1967-1971*—to be the most negative poem in my entire corpus. In a corpus of poetry containing direct attacks, extreme invective, and explicit takedowns of individuals, groups, and institutions, I did not find this poem to contain an exceptional amount of negative sentiment. On the contrary, I found "Come Sing a Song" to be positive and celebratory with regards to black life and black artistic expression. VADER, meanwhile, considers Nikki Giovanni's "The True Import of the Present Dialogue, Black vs. Negro"—from her 1968 *Black Feeling, Black Talk*—to have the most negative sentiment in the corpus. These results are very much in keeping with other human readers of this poem: critics consider it to be one of the most significant and famous examples of a certain type of angry, militant, even aggressive poem. Where Pattern and I disagree strongly over the feel of Troupe's "Come Sing a Song," critics and VADER seem to agree that Giovanni's "The True Import" has, on the surface, an exceptional amount of negative sentiment compared with its contemporaries.

Among other things, my project analyzes discrepancies and correspondences such as those described above. Already, my findings have revealed an interpretive disjoint between the denotative affective impact of words—what might be called their surface sentiment—and their more nuanced affective import as shaped by poetic, literary, social, and political contexts. A sentiment classifier like VADER, for example, highlights the intensity of negative sentiment in a poem according to the words and phrases it contains without the literary and historical context of their use. This kind of surface reading, attuned specifically to words' immediate affective impact, anticipates the space between a surface anger that can spark feelings regardless of context and a poetic form that, in the case of Giovanni's "The True Import," leverages negative sentiment to address meaningful social issues in a productive, ultimately positive way. By investigating these poems through conventional literary methods (i.e., historical contextualization, close reading, consideration of relevant scholarship) and computational methods (in this case Pattern and VADER), while also investigating the histories, intended use contexts, and potential biases of the chosen computational methods, this project provides an opportunity to examine what it is, exactly, that provides a book, poem, or poetic line with its emotional charge.

References

De Smedt, T. and Daelemans, W. (2012). "Pattern for Python." *Journal of Machine Learning Research* 13: 2063–67.

Hutto, C. J. and Gilbert, E. (2014). "VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text." *Eighth International Conference on Weblogs and Social Media*. Ann Arbor, MI, June 2014.

Jockers, M. (2015). "Revealing Sentiment and Plot Arcs with the Syuzhet Package," February 2, 2015. http://www.matthewjockers.net/2015/02/02/syuzhet/ (accessed 27 February 2018).

Lorde, A. (1997). "The Uses of Anger." *Women's Studies Quarterly* 25, no. 1/2: 278–85.

Neal, L. (1969). "Any Day Now: Black Art and Black Liberation." *Ebony* 24, no. 10: 54–62.

Rhody, L. (2012). "Topic Modeling and Figurative Language." *Journal of Digital Humanities* 2, no. 1. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-figurative-language-by-lisa-m-rhody/ (accessed 27 February 2018).

Swafford, A. (2015). "Why Syuzhet Doesn't Work and How We Know," March 30, 2015. https://annieswafford.wordpress.com/2015/03/30/why-syuzhet-doesnt-work-and-how-we-know/ (accessed 27 February 2018).

Does "Late Style" Exist? New Stylometric Approaches to Variation in Single-Author Corpora

Jonathan Pearce Reeve
jon.reeve@gmail.com
Columbia University, United States of America

Does "late style" exist? That is, do novelists exhibit a well-defined and distinctive stylistic shift as they reach old age, artistic maturity, or both? Edward Said's *On Late Style: Music and Literature Against the Grain* argues not only that such a style does exist, but that it has well-defined characteristics. Said describes late style as, somewhat paradoxically, involving "a nonharmonious, nonserene tension, and above all, a sort of deliberately unproductive productiveness going *against*" (Said 22). The term "late style," derived from Thodor Adorno's concept of Beethoven's *Spästil*, is one which Adorno conceives of as "catastrophic" (Adorno 567). As Adorno puts it, "the maturity

of the late works of significant artists does not resemble the kind one finds in fruit. They are, for the most part, not round, but furrowed, even ravaged. Devoid of sweetness, bitter and spiny, they do not surrender themselves to mere delectation" (564). To determine whether this claim is more than just anecdotally true, it deserves to be experimentally tested. Using new techniques of computational stylometric analysis, I test whether a writer's late works are statistically dissimilar to the rest of their corpus. I find that late style is not a statistically quantifiable phenomenon. Instead, the opposite is true: the novelists tested exhibit very distinctive early styles.

Twelve single-author corpora were prepared for this study. These include three novelists Said cites at length: Marcel Proust, Thomas Mann, and Jean Genet, as well as nine novelists from the 19th and 20th centuries, chosen for their prolificacy and electronic availability: Charles Dickens, Joseph Conrad, Ernest Hemingway, Henry James, Walter Scott, George Meredith, Willa Cather, Arnold Bennett, and Mary Augusta Ward. Two samples were taken from each novel in these corpora, so that the internal stylistic similarity of the samples serve as a metric check for the validity of the method. These samples were randomly chosen, to ensure that no text is longer than the shortest text in each corpus, and that that the analysis will compare equal amounts of text.

Each of these samples was then vectorized to 500-dimensional vectors, according to their top 500 word frequencies. These samples were then reduced to five dimensions using principal component analysis (PCA). Five dimensions were used here, instead of the usual two, since a cross-validated grid search in a previous study determined this value to be the most effective at clustering documents according to voice and style. This study also introduces two new metrics for stylistic difference. First, the "distinctiveness score" of a novel sample is calculated by determining the distance of the vector from the mean in five-dimensional space, using the Pythogorean theorem. A late novel that shows a high distinctiveness score, therefore, could correctly be called an instance of "late style."

Second, I introduce a metric representing the "periodicity" of the writer's style. This is calculated by first inferring prior category labels of early, middle, and late using publication years. Then, the novel's reduced vectors are clustered using a Baysian Gaussian mixture model, which probabilistically infers three or fewer clusters. These assignments are finally compared using a mutual information score, which calculates the similarity of these clusters with the prior inferences, regardless of label. A high periodicity score indicates that a novelist exhibits distinct stylistic periods, whereas a low score indicates that a novelist has a relatively unchanging or unpredictable stylistic progression.

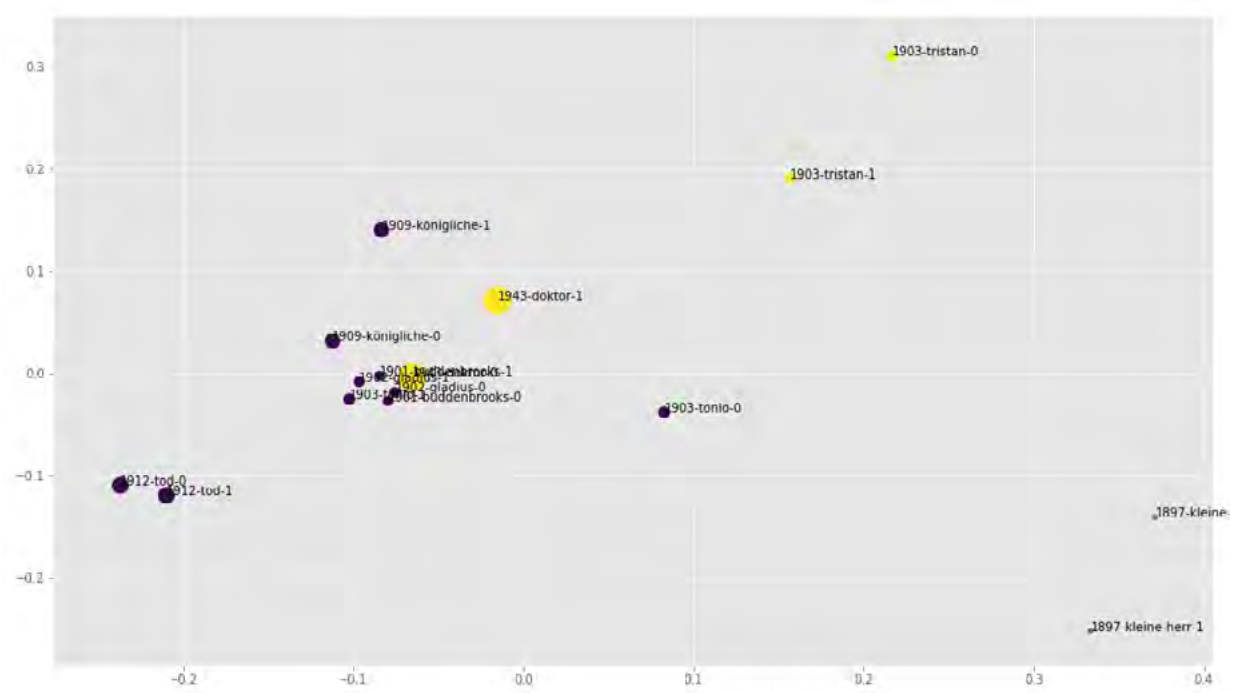


Figure 1: Thomas Mann

Figure 1 shows a projection of the first two dimensions of the vectors generated from Thomas Mann novels. The sizes of the points represent their relative publication years: small circles are early works, and large circles are late works. The colors represent the clusters predicted from the Bayesian Gaussian mixture model. The samples with the highest distinctiveness scores are from his first work *Der Kleine Herr*

Friedemann and his early work *Tristan*. The samples showing the least distinctiveness, are from *Doktor Faustus*, the very work Said cites as an example of a distinctive late style.

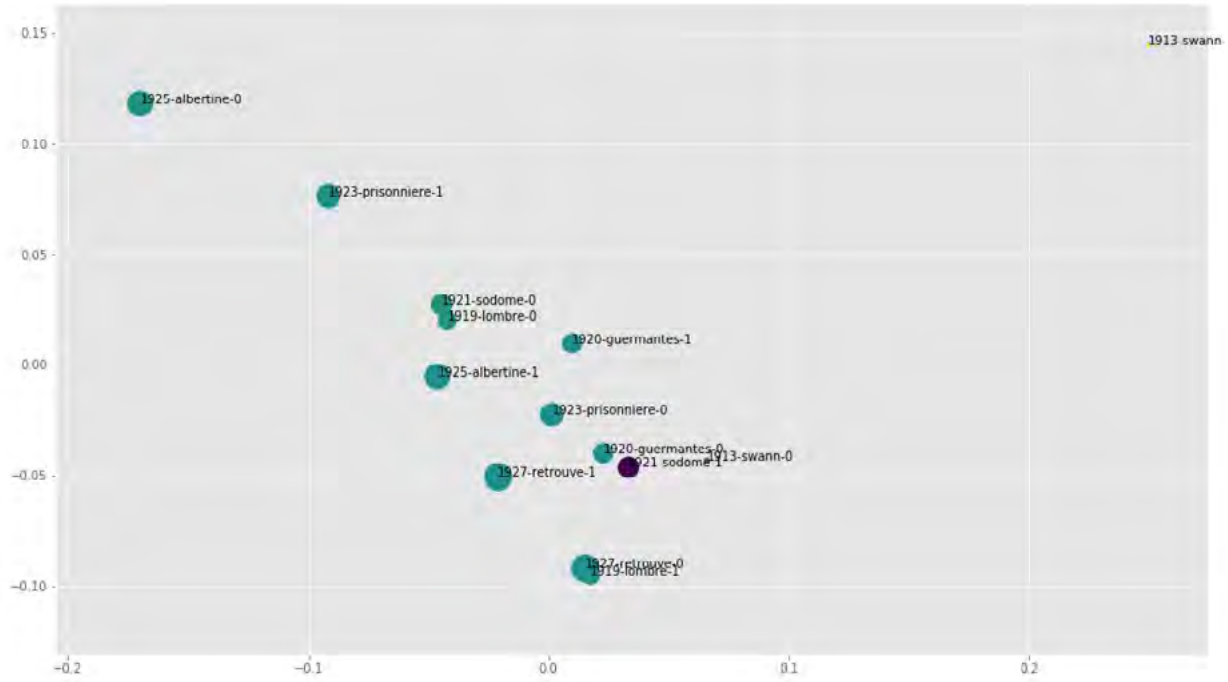


Figure 2: Marcel Proust

Figure 2 shows the same projection for samples from the works of Marcel Proust. Proust's first work, *Du côté du chez Swann*, is the most distinctive. Proust's last published work, *Le temps retrouvé*, which Said cites as an example of late style, is in fact very non-distinctive. Proust's middle works, however, *La prisonnière* and *Albertine disparue*, are only intermediary with respect to publication dates, since they were the final novels he wrote. Here, Said is somewhat correct that Proust has a late style, but misidentifies the works that exemplify it. Again, however, Proust's early style shows a stronger signal than his late.

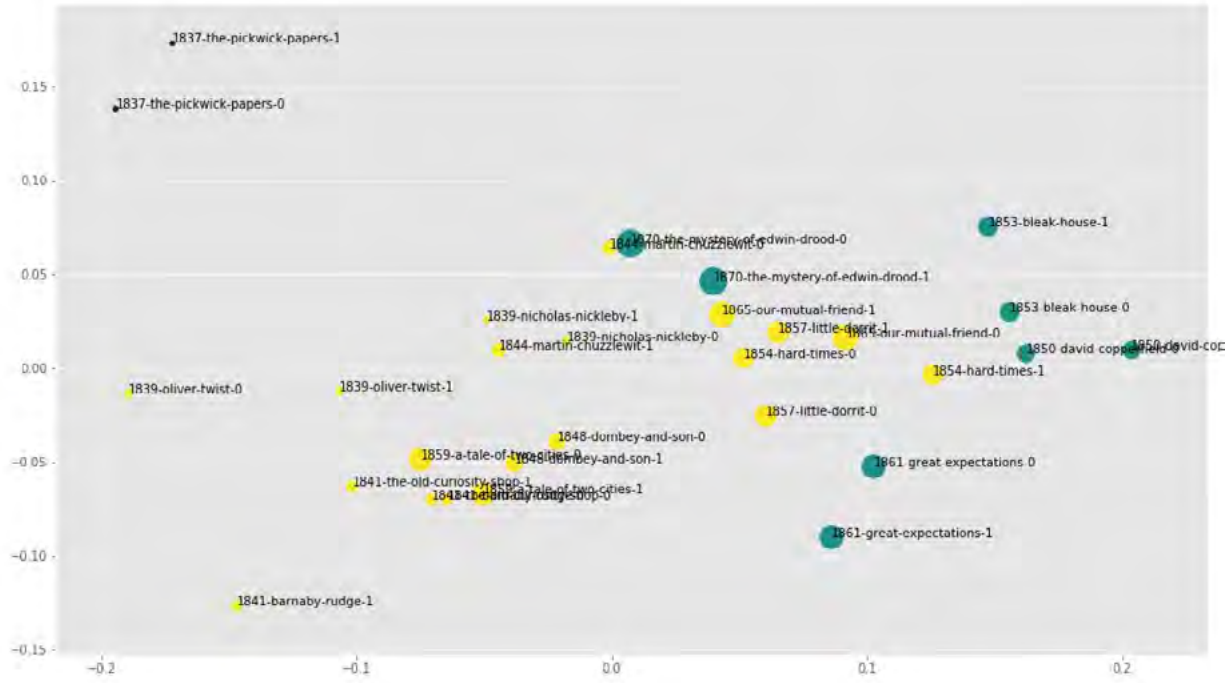


Figure 3: Charles Dickens

Figure 3 shows vectors generated from Charles Dickens novels. Here again, the early work *The Pickwick Papers* has the highest distinctiveness score, followed by *David Copperfield*. Late works like *Our Mutual Friend* are among the least distinctive. As the alignment of the point colors and sizes here suggests, Dickens shows a strong periodicity. At 0.469, his is the second-highest periodicity score.

Author	Periodicity Score
Proust	0.023
Meredith	0.028
Ward	0.166
Cather	0.177
Conrad	0.177
Bennett	0.220
Hemingway	0.326
Scott	0.360
Mann	0.367
Genet	0.457
Dickens	0.469
James	0.472

Table 1

Table 1 shows the periodicity scores of all the novelists studied here. Those novelists with well-known early and late styles, such as James and Dickens, have high periodicity scores. Writers like Proust, on the other hand, whose novels all form part of the series *À la recherche du temps perdu*, and were all published within about a decade, show the lowest periodicity scores.

This study, beyond simply testing and ultimately disproving the claims of Said and Adorno, provides a framework for stylometric analysis of textual difference, one which could be used to enhance authorship detection techniques and the techniques of forensic text analysis more generally. More experiments are needed, of course, to test the validity of these techniques beyond the domain of literature.

References

Adorno, T. (2002). Late Style in Beethoven. In: *Essays on Music*. Berkeley: University of California Press, 2002. pp. 564–568.

Said, E. (2006). *On Late Style: Music and Literature Against the Grain*. New York: Pantheon Books.

Keeping 3D data alive: Developments in the MayaCityBuilder Project

Heather Richards-Rissetto
richards-rissetto@unl.edu
University of Nebraska-Lincoln, United States of America

Rachel Optiz
rachel.opitz@glasgow.ac.uk
University of Glasgow, United Kingdom

Fabrizio Galeazzi
fabrizio.galeazzi@york.ac.uk
University of York, United Kingdom

Digital data preservation is complex and multi-layered. The digital humanities brings unique challenges and opportunities to „keeping data alive" that are leading to innovative cross- disciplinary solutions. Data preservation involves standards, guidelines, open-source vs. proprietary software, accessibility, and much more. While establishing best practices, cultivating a community of experts, and developing infrastructure for 3D data used in cultural heritage has been the focus of several coordinated efforts in Europe over the past decade (Campana and Remondino 2014; Fresa and Prandoni 2015; Vecchio et al. 2015), efforts have been less systematic in the United States. Recently, however, digital humanities practitioners have spearheaded 3D data preservation and sharing in the United States.

While scholars working with 3D data must deal with management and sustainability issues (Galeazzi 2016; Richards-Rissetto and von Schwerin 2017), endeavors are typically tailored to individual projects. To broaden and coordinate efforts, the Community Standards for 3D Data Preservation (CS3DP) project is bringing together librarians, curators, technical specialists, and scholars to begin the process of developing standards for preservation and sharing of digital 3D data. While long-term archival of these data, for example, in a dark archive, is integral to our research (Koller et al. 2010), the MayaCityBuilder project is contributing to "keeping data alive" by developing workflows to supporting reuse and repurposing of procedurally-generated 3D data in the humanities.

While many types of 3D models are being used in humanities scholarship, the case study focuses on 3D models of ancient Maya architecture generated from multiple data sources including architectural drawings, excavation reports, Geographic Information Systems (GIS), and airborne LiDAR. To contribute to 3D data preservation efforts, while maintaining realistic goals, the MayaCityBuilder Project focuses on procedural modeling—rapid prototyping of 3D models from a set of rules. Procedural modeling is ideally suited for the development of 3D modeling standards that promote data interoperability, dissemination, and reuse because they bring with them the underlying metadata, paradata (information about modeling choices) (Bentkowska-Kafel et al. 2016), and descriptive data (e.g., data sources, textures, building type).

Within these circumstances, the two objectives of the "keeping data alive" component of the MayaCityBuilder Project, supported by a Tier I Research and Development Grant from the Division of Preservation and Access of the National Endowment for the Humanities (NEH), are to de-

velop **workflows**: (1) to generate, store, and make accessible 3D models of ancient architecture in open-source and proprietary software to foster data (re)use and (2) to host, deliver, and visualize these 3D models, linked to metadata, paradata, and descriptive data, in 3D visualization environments. These objectives are part of a larger goal to contribute to *innovative methods of materials analysis and new modes of discourse using interactive 3D web visualizations.* To achieve this goal requires not only data accessibility but also data compatibility—scholars must also be able to combine and recombine data for reuse and repurposing.

Building on previous research and development and lessons learned from the MayaArch3D Project (von Schwerin et al. 2013), Gabii Goes Digital (Opitz et al. 2016), and the Archaeological Data Service (ADS) (Galeazzi et al. 2016), we present technical workflows to dynamically host, deliver, and visualize 3D models that are linked to metadata, paradata, and descriptive data in two 3D environments: (1) an open source 3D web-based environment based on 3DHOP (3D Heritage Online Presenter—an open-source software package for hosting interactive, high-resolution 3D models on the web that uses HTML, JavaScript, and WebGL (Web Graphics Library) (2) Unity—a proprietary and widely-used gaming engine that offers free access to many of its powerful tools. We present an overview of the workflows we have developed explaining how the steps serve our objective of data reuse and more broadly access and preservation of 3D data. Additionally, we discuss how these workflows relate to the next phase of the project, i.e., prototype development. The prototype will take advantage of recent developments in web technology, namely the adoption of WebGL that renders interactive 2D and 3D computer graphics in browsers without plugins.

The ability to efficiently generate, store, deliver, and visualize models in an interactive 3D web- based environment will help keep data alive by fostering collaborative and comparative humanities research. We focus on procedural models because they can be quickly generated and are directly linked to metadata and paradata. 3D models allow scholars to test architectural reconstructions and situate them within landscapes to investigate spatial relationships at multiple scales while providing a sense of embodiment (Barcelo et al. 2000; Dylla et al. 2010; Frischer and Dakouri-Hild 2008; Richards-Rissetto and Plessing 2015; Saldana 2015). However, the diversity of 3D data types, tools, and technologies in combination with a lack of standards requires workflows to promote reuse and repurposing of 3D data to contribute to long-term access and preservation of 3D data.

References

3D Heritage Online Presenter (3DHOP). http://vcg.isti.cnr.it/3dhop/index.php; last accessed on 04/24/18

Barcelo, J., M. Forte, and D. Sanders. (2000). Virtual Reality in Archaeology. *BAR Int. Series* 843.

Bentkowska-Kafel, A., Denard, H., Baker, D. (Eds.), (2016). *Paradata and Transparency in Virtual Heritage – Digital Research in the Arts and Humanities Series*. Routledge Taylor & Francis, London.

Campana, S., & Remondino, F. (2014). 3D modelling in archaeology and cultural heritage: theory and best practice. *BAR Int. Series* 2598.

Dylla, K., B. Frischer, P. Mueller, A. Ulmer, and S. Haegler. (2009). Rome Reborn 2.0: A Case Study of Virtual City Reconstruction Using Procedural Modeling Techniques. In *Making History Interactive*, pp. 62-66. Oxford: Archaeopress.

Fresa, A., Justrell, B., & Prandoni, C. (2015). Digital curation and quality standards for memory institutions: PREFORMA research project. *Archival Science*, 15(2), 191-216.

Frischer, B. and A. Dakouri-Hild (eds). (2008). *Beyond illustration: 2d and 3d digital technologies as tools for discover in archaeology*. Oxford: Archaeopress.

Galeazzi, F, M. Callieri, M. Dellepiane, M. Charno, J. Richards, R. Scopigno. (2016). Web-based visualization for 3D data in archaeology: The ADS 3D viewer. *Journal of Archaeological Science*: Reports 9: 1-11.

Galeazzi, F. (2016). Towards the definition of best 3D practices in archaeology: Assessing 3D documentation techniques for intra-site data recording. *Journal of Cultural heritage* 17: 159-169.

Koller, D., Frischer, B. and G. Humphreys. (2010). Research challenges for digital archives of 3D cultural heritage models. *Journal on Computing and Cultural Heritage* 2(3):7:1-7:17.

Opitz, R., Marcello Mogetta, and Nicola Terrenato. (2016). *A Mid-Republican House from Gabii.* Ann Arbor: University of Michigan Press.

Richards-Rissetto, H. and R. Plessing. (2015). "Procedural Modeling for Ancient Maya Cityscapes: Initial Methodological Challenges and Solutions." *2015 Digital Heritage International Congress*, Volume 2: 85-88. IEEE Conference Publications.

Richards-Rissetto, H. and J. von Schwerin. (2017). A Catch 22 of 3D Data Sustainability: Lessons in 3D Archaeological Data Management & Accessibility. *Journal of Digital Applications in Archaeology and Cultural Heritage*. 6: 38-48.

Saldana, M. (2015). An Integrated Approach to the Procedural Modeling of Ancient Cities and Buildings. *Digital Scholarship in the Humanities*, Volume 30, Issue suppl_1, 1 December 2015, Pages i148–i163,

Vecchio, P., Mele, F., De Paolis, L. T., Epicoco, I., Mancini, M., & Aloisio, G. (2015). Cloud Computing and Augmented Reality for Cultural Heritage. In *Augmented and Virtual Reality* (pp. 51-60). Springer International Publishing.

von Schwerin, J. , H. Richards-Rissetto, F. Remondino, and G. Agugiaro. (2013). "The MayaArch3D Project: A 3D WebGIS for Analyzing Ancient Maya Architecture and Landscapes at Copan, Honduras." *Literary and Linguistic Computing* 28(4):736-753.

Finding Data in a Literary Corpus: A Curatorial Approach

Brad Rittenhouse
bcrittenhouse@gatech.edu
Georgia Institute of Technology, United States of America

Sudeep Agarwal
hello@sudeep.co
Georgia Institute of Technology, United States of America

PI and Presenter: Brad Rittenhouse
Others involved: Taha Merghani, Sudeep Agarwal, Vidya Iyer, Madison McRoy, Sidharth Potdar, Nate Knauf, and Kevin Kusuma

In this short paper, I will discuss an ongoing text analysis project, which applies NLP, topic modeling, mapping, and other methodologies to the Wright American Fiction corpus. From a theoretical standpoint, the project is an extension of my qualitative work, which tracks a notable historical shift in literary data management strategies through the works of two canonical American writers: Herman Melville and Walt Whitman. Both wrote in New York as it grew from a small market town of around 60,000 residents to a global metropolis of nearly 1,000,000 and had to imagine strategies of data management to integrate newly urban, consumerist surroundings into their writings in an effective, efficient manner. Translating increasingly crowded material realities—populated by people, products, and print—into literary data, these writers illustrate an important ontological shift from the positivist data strategies of the Enlightenment to digital logics of aggregation, organization, and metonymic indexing that increasingly address the impossible scale of modern infospheres.

As relatively privileged subjects, however, these writers' very ability to integrate and innovate with this information was largely based upon a free access to information (and indeed information overload) that many contemporaries did not enjoy. In short, critics have historically apportioned literary status upon hegemonic standards of information, with prestigious genres like "encyclopedic writing" preferring masculinist topics and knowledge bases such as ballistics (Pynchon), cetology (Melville), violence (Bolaño) over spheres of knowledge historically more accessible and immediate to women and people of color.

My quantitative work looks to sidestep these biases, using an assortment of natural language processing techniques to recover works from the archive that may be performing similarly impressive literary acts of aggregation, but which critics may have overlooked because the works exist in alternative thematic and affective registers. By measuring the accretion of material information across the corpus, and identifying areas of relative density, my process points to writing which humans readers have overlooked but which machines are able to see as substantially similar to canonical encyclopedic works.

We intentionally made a very broad measurement of the text to identify a broader range of artistic expression. The process itself involves chunking all the texts into 500-word segments, performing a parts-of-speech tag with OpenNLP, then rendering these tags in "baby binary": a "0" for all non-nouns, a "1" for all nouns. We then summed the segments and divided by the total length of each to obtain a noun density measurement, which generally indicates an aggregation of material information. Though it is possible to use more specific grammatical measures (subjects, objects, etc.), we used nouns at-large so as to capture a fuller spectrum of thought, sentiment, and other immaterial objects that accompany the human masses of urbanization.

We also assembled a fair amount of demographic metadata for the corpus, which has allowed us retrieve relatively forgotten works from the archive. After identifying the densest chunks of text, we attempted to identify author gender with the use of the machine learning platform SexMachine. We cross-referenced these results with those derived from the noun-density analysis to pinpoint female authors of interest. To conduct this analysis, we first performed exploratory data analysis to understand the underlying distribution of noun ratios across the corpus, which appeared to be normally distributed, although with a slight right skew. Then we compared this distribution with that of the noun ratios identified for authors of each gender. The distributions seem to be largely similar. This naturally led to an outlier analysis within each gender, which identifies outliers as works with a noun ratio 1.5 interquartile ranges either above or below the median, yielding 71 outliers for male authors and 47 outliers for female authors (43 and 26 on the high-end, respectively). We then performed additional analyses on these outliers to get a better understanding of what differentiated them from the rest of the corpus.

One case study I will present from among these outliers is that of Emma Wellmont, a nineteenth-century temperance writer who the academy has largely ignored, I suspect because of the emotional, sensationalist overtones of her chosen genre. Nonetheless, her work is quantitatively similar to Walt Whitman's, with many extracts in the highest quadrant of noun density across the corpus and packed with what the latter evocatively refers to as "stuff." Unlike Whitman, however, her densest passages are often emotional, pathetic scenes of death and suffering. Critics, if they read Wellmont's work (and most do not), would likely label it sensationalistic or melodramatic, and therefore, unserious, writing. My methodology, on the other hand, makes an argument for her as an important encyclopedist, albeit of canonically unlikely subject matter. I will present the case study through a prototype interactive visualization that allows users to explore the corpus at-large, all the way down to significant passages within individual works (Figure 1).

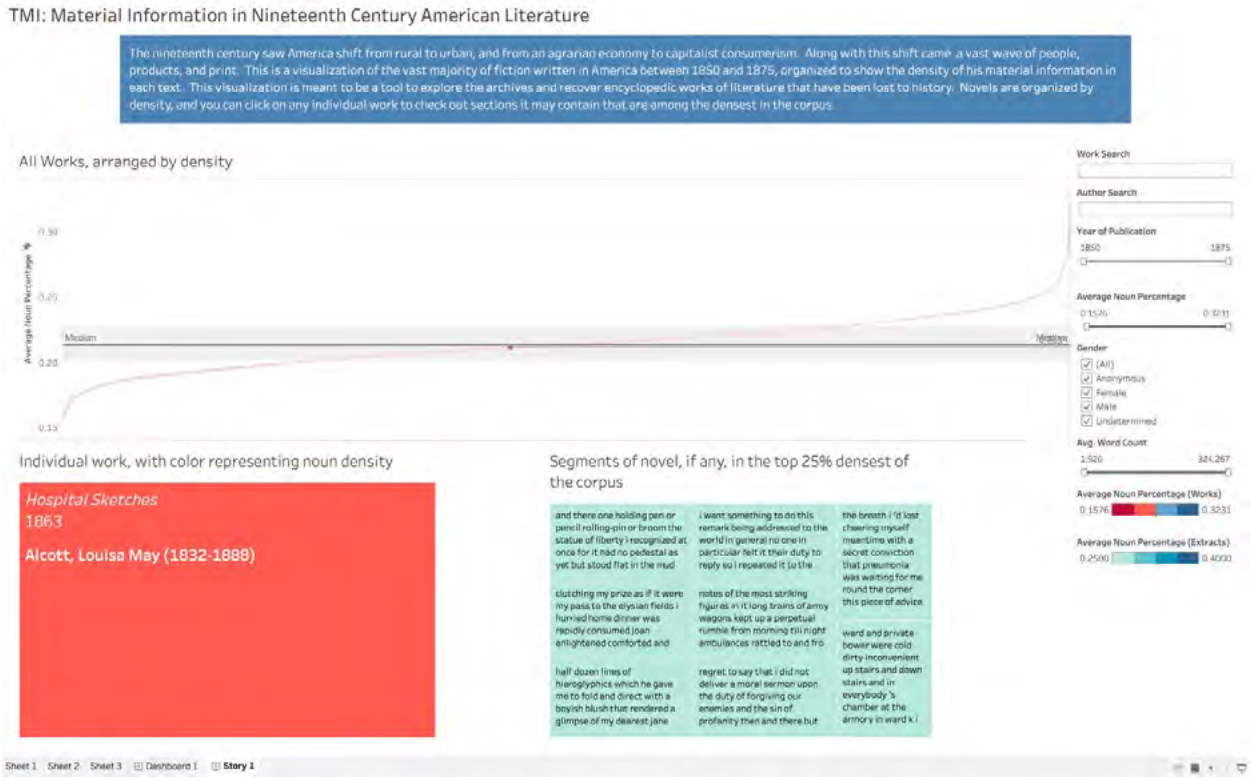


Figure 1

This curatorial process builds upon the methods described by Long and So in their recent article "Literary Pattern Recognition: Modernism between Close Reading and Machine Learning," using high-powered computing and statistical analysis on a corpus scale to identify information-dense passages for later close reading and analysis. Reading literature as information, the methodology is flexible in not only illuminating macro-scale trends, but also identifying human-readable works and passages for literary critics who also value critical reading practices. The project also runs in parallel to Dennis Yi Tenen's recent work in its "articulation of 'effect spaces' via material density," though it pulls from a broader range of quantitative, grammatical measures in its attempt to broaden the generic construct of encyclopedic writing.

References

Long, H. and So, R. (2016). Literary pattern recognition: modernism between close reading and machine learning. *Critical Inquiry*, 42(2): 235-267.

Yi Tenen, D. (2018). Toward a computational archaeology of fictional space. *New Literary History*, 49(1): 119-147.

Mapping And Making Community: Collaborative DH Approaches, Experiential Learning, And Citizens' Media In Cali, Colombia

Katey Roden
rodenk@gonzaga.edu
Gonzaga University, United States of America

Pavel Shlossberg
shlossberg@gonzaga.edu
Gonzaga University, United States of America

Engaging the "bridges/puentes" theme central to the conference, this paper presents first-hand knowledge and practical insights garnered from a collaborative digital mapping project between North/South academics, students, and community activists engaged in community-based social justice activism in Cali, Colombia. A foundational goal of this Digital Humanities project is thus to create intercultural and communicative bridges between not only the academic communities of Gonzaga University and Pontifica Universidad Javeriana, but also to provide a platform by which Colombian community organizers shape their presence in local as well as digital communities.

The paper discusses our goals and methods, and also the roadblocks we encountered, in establishing collabo-

rative pathways to embed Digital Humanities mapping tools as central elements within a field-based Communication and Community Development course. The Digital Humanities project at the heart of this course aligns with pedagogy as well as practical fieldwork in the area of Development Communication, which holds that communication processes and projects that support or foster the growth of grassroots civil society are essential elements of community development and empowerment. In this vein, Digital Humanities perspectives and methodologies that privilege the bottom-up democratization of access and information inform course content assigned to students from the Global North, who come to Cali, Colombia as part of an intensive immersion.

As such, the course invites students from Cali and the United States to engage, accompany, and shadow community-based organization that work in areas such as citizens' radio; street theatre and community-based performances; and grassroots documentary production. The work undertaken by the community-based organizations seeks to displace hegemonic media and dominant culture imaginaries, which routinely render these resource-deprived communities as being inherently abject, dangerous, chaotic, and pathological. The community organizations with whom we partner engage community problems by creating and claiming spaces for public expression that amplify popular voices within their own communities and beyond. The Digital Humanities mapping project developed for this course responds to these community initiatives, in that it serves as a community-academy collaborative space. The digital map produced collaboratively, provides a platform that presents, promotes, captures, and renders visible popular or grassroots media, communication activities, and products through which "citizens can learn to manipulate their own languages, codes, signs, and symbols, empowering them to name the world in their own terms" (Rodriguez 2011). The paper documents a central element of our work, the mutual efforts engaged in creating active and equitable roles for each party involved in the digital product's production, whether those groups come from the academy, the community, the Global North or the Global South.

Beyond discussing the compatibility and fit of Digital Humanities tools with the articulation of community-based approaches to citizens' media and popular communication, this paper also discusses the significant ways in which Digital Humanities mapping tools can be mobilized to foster or promote community-based experiential learning experiences in intercultural contexts bridging the global North and South. Experiential learning emphasizes the acquisition of knowledge through interactive processes of action and reflection; where students can take an active part in the creation of knowledge (Hale 1999). We contend that producing Digital Humanities projects in the contexts of an international immersion and hand-in-hand with local partners whose voices, perspectives, and needs drive project conceptualization and the mapping process, presents a unique opportunity for experiential learning that extends well beyond the classroom and into the lived realities of all the parties involved. In this vein, the experiential learning opportunities developed in such an environment embrace the broader humanistic agenda of Digital Humanities as a field, where people come together through and with technology to "produce a collaborative, connected, and relational knowledge production, of making and learning and learning through making" (Goldberg 2015). Accordingly, our project seeks to facilitate an experiential learning opportunity for our students, but in doing so we also seek to diminish the sometimes too rigid boundaries that privilege academic institutions as the sole purveyors and producers of knowledge. By collaboratively creating a digital map with and for local community members while in their communities, our project aims to decentralize knowledge production and encourage our students to become conscious of diverse forms of knowledge and authority.

Furthermore, our experience also suggests that with effective planning and development, community-based Digital Humanities mapping projects can productively alleviate issues and problems that commonly arise in the context of experiential- or service-learning courses taking place in intercultural contexts across the North/South boundary. It is well known that "service-learning can reinforce stereotypes and paternalism among students. Some scholars argue that many applications of service learning do little to question the role of students as providers of resources…" (Chupp & Joseph 2010). Additionally, service- or experiential-learning is "often implemented with a sole focus on the potential beneficial impact on the student, with little or no emphasis on the possible longer-term beneficial impact on those served by the activity and their broader community" (Chupp & Joseph 2010). The collaborative mapping project we have developed engages Digital Humanities approaches within an embedded community context, with the explicit intention of addressing potential problems linked to the implementation of experiential service learning project in partnership between the North and South.

In sum, the Digital Humanities mapping project nested within this Communication and Community Development course remains an experimental and open collaboration. Well-established and emergent issues and challenges continue to exist. With that caveat in mind, experience and evidence also suggests that digital technology mapping tools provide a set of ready enhancements to experiential learning, study abroad, and Communication and Community Development courses. These features begin to realize the promise and purpose of Digital Humanities by creating bridges that foster global collaboration, create open access platforms, and generate academy-community, North/South collaborations that equalize access to the generation and circulation of knowledge locally and globally.

KEYWORDS: Global South/North, Experiential Learning, Mapping, Community Development, Citizens' Media

References

Chupp, Mark G., and Mark L. Joseph. "Getting the most out of service learning: Maximizing student, university and community impact." *Journal of Community Practice* 18.2-3 (2010): 190-212.

Goldberg, David Theo. "Deprovincializing Digital Humanities." In *Between Humanities and the Digital*. Eds. Patrik Svensson and David Theo Goldberg. MIT Press, 2015. 163-71.

Hale, Aileen. "Service-learning and Spanish: A missing link." *Construyendo Puentes (Building Bridges): Concepts and Models for Service-Learning in Spanish*. Ed. Josef Hellebrandt and Lucía T. Varona. Washington, DC: American Association for Higher Education (1999): 9-31.

Rodríguez, Clemencia. *Citizens' media against armed conflict: Disrupting violence in Colombia*. U of Minnesota Press, 2011.

The Diachronic Spanish Sonnet Corpus (DISCO): TEI and Linked Open Data Encoding, Data Distribution and Metrical Findings

Pablo Ruiz Fabo
pablo.ruiz@linhd.uned.es
Universidad Nacional de Educación a Distancia, Spain

Helena Bermúdez Sabel
helena.bermudez@linhd.uned.es
Universidad Nacional de Educación a Distancia, Spain

Clara Martínez Cantón
cimartinez@flog.uned.es
Universidad Nacional de Educación a Distancia, Spain

Elena González-Blanco
elenagbg@gmail.com
Universidad Nacional de Educación a Distancia, Spain

Borja Navarro Colorado
borja@dlsi.ua.es
Universidad de Alicante, Spain

Introduction

Digital resources in poetry in Spanish are scarce, particularly for certain periods. This poses difficulties for Digital Humanities studies in Spanish.

Some digital editions of medieval poetry exist, e.g. BiDTEA (Gago Jover et al, 2015), ADMYTE (Marcos Marin and Faulhaber, 1992), PoeMetCa (Escribano et al, 2016), besides resources containing partial editions like ReMetCa (González-Blanco and Rodríguez, 2014). For the Golden Age, Navarro-Colorado et al. (2015) presented the *Corpus of Spanish Golden-Age Sonnets*. For later periods, we are not aware of poetry collections, although other genres are covered in Textbox (Schöch et al, 2017), BETTE (Santa María Fernández et al, 2017), Aracne (Álvarez and Martín, 2015) or Revistas Culturales 2.0 (Ehrlicher and Rißler-Pipka, 2015).

This paper describes the DISCO corpus and how it complements available digital materials for poetry in Spanish in several respects: First, the author and period range. Second, metadata concerning the authors and their works expressed in TEI-RDFa, given the importance of interoperability between literary datasets and the advantages of Linked Open Data as a paradigm. Finally, example findings that can be obtained with our corpus are provided, regarding metrical patterns diachronically.

The corpus is available on GitHub[1] and Zenodo.[2]

Corpus description

The corpus contains 4087 sonnets in Spanish by 1204 authors (15th to 19th century),[3] extracted from HTML sources at Biblioteca Virtual Cervantes (García, 2005, 2006a, 2006b) and Wikisource. Sonnets were chosen given the form's importance in European poetry, where it is even considered as its own genre. The form's clear restrictions make it easily amenable to computational treatment, facilitating meaningful comparison across poems. Several computational linguistics studies on the sonnet exist (Navarro-Colorado et al., 2015, 2016, 2017a, 2017b; Agirrezabal, 2017). A new sonnet corpus complements earlier work on both traditional and computational poetry analyses.

We focused on canonical and non-canonical authors, from different Spanish-speaking countries (Figure 1).

Period	Nbr of Sonnets	Nbr of Authors			Sources
Golden Age (15th-17th)	1088	477	Female	31	García (2006a)
			Male	446	
			America	12	
			Europe	458 (+7)	
18th century	323	42	Female	1	García (2005) Wikisource
			Male	41	
			America	6	
			Europe	36	
19th century	2676	685	Female	46	García (2006b)
			Male	637	
			America	334	
			Europe	348 (+3)	

Sonnet and author distribution per period, including the number of female and male authors, and the continent where they developed

1 https://github.com/postdataproject/disco/
2 https://doi.org/10.5281/zenodo.1012567
3 About 125 sonnets by approx. 20 authors whose production took place in the early 20th century (with date of death prior to 1936) are also included in the corpus; the documentation on the GitHub repository (footnote 1 above) provides more details.

their literary activity. Numbers in parentheses indicate authors which were probably active in Europe.

Encoding Paradigms: TEI and Linked Open Data

The poems are encoded in XML-TEI P5. A plain-text version is also provided. Together with the TEI-semantics, this corpus provides a layer of Linked Open Data (LOD) expressed in RDFa (Herman et al., 2015). To our knowledge, no out-of-the-box tools exist for publishing literary TEI corpora as LOD.[4] In this context, the enrichment of TEI with RDFa attributes is a solid approach to translate TEI semantics to the web (see precedents like Jewell, 2010) and benefit from the wide range of possibilities of the Semantic Web: First, we enrich our dataset by linking to third-party ones (as DBpedia), providing additional resources to complement the corpus. Second, we publish our data openly using standard schemas, thus supplying semantic interoperability that allows third-party applications to automatically use our data.

Author metadata

Author metadata were extracted or inferred from unstructured source content, and specified in the teiHeader: Year, place of birth and death, and gender. Two versions of the texts are available: one collecting every sonnet per author, the other with a single sonnet per file.

For the current corpus release we augmented the TEI annotation with URIs and class/property information, expressing them in RDFa. The most straightforward information concerns authors and their works, and the DCMI Metadata Terms (DCMI Usage Board, 2012) provides an appropriate scheme. Most features regarding authors' biographical data were formalised with the FOAF vocabulary (Brickley and Miller, 2014). Links to other resources were supplied. For instance, authors were assigned *Virtual International Authority File* (VIAF) identifiers, by querying VIAF's API supplemented with manual validation. Since the corpus includes non-canonical authors, LOD is an important asset to share their work thanks to the enhanced display of this type of data implemented by search engines.

Our documentation[1] provides further details.

Metrical encoding and enjambment

Using the **met** attribute, each line was annotated for scansion (strong and weak syllables) with the ADSO tool[5] (Navarro-Colorado, 2017), which specializes in Spanish fixed-meter forms, attaining a performance of 0.95 F1. A heuristic was used to automatically annotate the quatrains' rhyme-scheme, i.e. enclosed (ABBA) or alternate (ABAB).

Using an **enjamb** attribute, lines were annotated for enjambment[6] with the ANJA tool[7] (Ruiz-Fabo et al., 2017). The tool's performance at detecting enjambment is above 0.8 F1, and its efficacy at classifying enjambment types varies across periods and types. A *cert* attribute specifies the expected certitude for each enjambment type annotated.

The corpus documentation[1] provides more details.

How's this corpus different?

The metadata mentioned in 2.2. were unavailable in structured, machine-readable format in the corpus sources, or in other sonnet collections, like Sonnet-Archiv (Elf Edition). Regarding coverage, the corpus complements Navarro-Colorado et al's (2015) Golden Age Sonnet corpus, by including minor Golden Age authors. For later periods, we cover more poems and authors than existing digital corpora, up to the 19th century. Our corpus integrates RDFa annotations, which in a second version will be fully compliant with the POSTDATA model.[8] This is a pioneering model that will provide means to publish European poetry materials as Linked Open Data. Finally, combining the annotation of metrical patterns, stanza types and enjambment is not offered by prior corpora.

Some metrical findings

Corpus data on stress patterns (Figure 2) agree with existing descriptions[9] of the Spanish hendecasyllable based on small-sample analyses: A maiori patterns (with 6th-syllable stress) predominate, and a minori patterns (with 4th-syllable stress) follow. However, our data show an increase of a minori patterns in the 19th century, which might suggest an interest in metrical variety in that period.

Regarding diachronic data on the number of stressed positions (Figure 2), patterns with three stresses are

4 Whereas the publication of literary corpora in Linked Open Data formats is not widespread, inspiration could be drawn from the linguistics community, which has been especially successful in building the means to convert resources with linguistic annotations to the Resource Description Framework model (see McCrae et al., 2011; Chiarcos and Ejavec, 2011). In addition, more general projects, not limited to linguistic analysis, are being developed as well: see work on building a TEI ontology in Ciotti et al (2016).

5 https://github.com/bncolorado/adsoScansionSystem

6 The tool detects different types of enjambment (i.e. a mismatch between syntactic and metrical structure) as characterized by Quilis (1964). The tool also detect's Spang's (1983) concept of *enlace*, which takes place when a subject or direct object occur in a line adjacent to their governing verb's line, and which triggers a less noticeable effect than the enjambment types defined by Quilis

7 See https://sites.google.com/site/spanishenjambment/ for details

8 See Bermudez-Sabel et al. (2017). Version 0.2 of the POSTDATA model is available at https://doi.org/10.5281/zenodo.832906

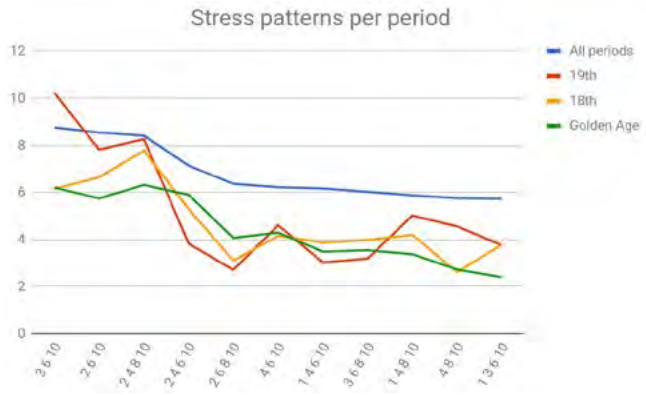
9 See Domínguez Caparrós (2014: 143) or Henríquez Ureña (1919: 132) for details on *a maiori* and *a minori* patterns. The main *a maiori* variants as described in previous literature are 2 6 10 and 3 6 10; this is confirmed in our data. Patterns are formalized as a series of numbers indicating stressed syllables, e.g. 2 6 10 for the second, sixth and tenth syllables. Note that 10th-syllable stress is mandatory in all patterns.

highly used across periods. However, most a maiori patterns with four stresses decrease in the 19th century. This might indicate a 19th-century preference for "lighter" patterns, with stresses further apart from each other.

Whereas the predominant meter for sonnets is naturally the hendecasyllable, alexandrines[10] are attested, mostly in the 19th century, preferentially used by American authors. The alexandrine sonnet uses an alternate rhyme scheme (ABAB) more often than the usual enclosed scheme (ABBA). See Figure 4.

Pattern	Pattern Class	Stress Count	Percentage of lines			
			All periods	19th	18th	Golden Age
3 6 10	mai	3	8.76	10.23	6.16	5.21
2 6 10	mai	3	8.55	7.82	6.65	5.75
2 4 8 10	*min*	*4*	*8.41*	*8.26*	*7.77*	*6.32*
2 4 6 10	mai	4	7.14	3.83	5.30	5.90
2 6 8 10	mai	4	6.37	2.71	3.10	4.07
4 6 10	mai	3	6.23	4.61	4.16	4.30
1 4 6 10	mai	4	6.17	3.03	3.87	3.49
3 6 8 10	mai	4	6.03	3.19	3.98	3.55
1 4 8 10	*min*	*4*	*5.88*	*5.02*	*4.2*	*3.38*
4 8 10	*min*	*3*	*5.76*	*4.56*	*2.62*	*2.73*
1 3 6 10	mai	4	5.73	3.76	3.79	2.40

Distribution of stress patterns per period (percentage of lines for each pattern) for the 10 most frequent patterns in the corpus, sorted by decreasing percentage of occurrence in the complete corpus. Pattern classes are also provided (*mai*: *a maiori*, i.e. stress on 6th syllable, *min*: *a minori*, i.e. stress on 4th and 8th syllable). Rows for *a minori* patterns are in italics. *Stress count* refers to the number of stresses in the pattern. Patterns with three stresses are widely used in any period. Most *a maiori* patterns with 4 stresses decrease in the 19th century, whereas *a minori* patterns increase in that century.



Distribution of stress patterns per period (percentage of lines for each pattern) for the 11 most frequent patterns in the corpus.

10 In Spanish, the alexandrine has 14 metrical syllables. In sonnets, the hendecasyllable predominates almost exclusively. However, particularly since the 19th century, alexandrine sonnets have been written.

Meter Length	Quatrain Rhyme	Sonnet Count		
		total	American	European
hendecasyllable	Enclosed	2269	1218	1051
	Alternate	122	96	26
	Total	**2391**	**1314**	**1077**
Alexandrine	Enclosed	122	98	24
	Alternate	145	121	24
	Total	**267**	**219**	**48**

Count of hendecasyllable vs. alexandrine sonnets according to the authors' continent of production, in the **19th century** (alexandrine sonnets are very rare before). The type of rhyme scheme in the quatrains (enclosed or alternate) is also specified. The alexandrine sonnet is preferentially used by American authors, and there's a preference for alternate rhyme for this meter length.

Acknowledgements

Supported by the project 'Poetry Standardization and Linked Open Data: POSTDATA' (ERC-2015-STG-679528), funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme, and led as a Principal Investigator by Dr. Elena González-Blanco, LINHD-UNED (http://postdata.linhd.es/).

References

Agenjo, X. (2015). Las bibliotecas virtuales españolas y el tratamiento textual de los recursos bibliográficos. Ínsula: revista de letras y ciencias humanas 822: pp. 12–15.

Agirrezabal, M. (2017). *Automatic Scansion of Poetry. PhD Thesis*. University of the Basque Country.

Álvarez Mellado, E. and Martín-Fuertes, L. (2015). *Aracne Project,* http://www.fundeu.es/aracne/ (Accessed 22 Sep. 2017).

Bermúdez-Sabel, H., Curado Malta, M. and González-Blanco, E. (2017). Towards Interoperability in the European Poetry Community: The Standardization of Philological Concepts, in Jorge Gracia et al. (ed.) *Proceedings of Language, Data, and Knowledge: First International Conference (LDK 2017)*: pp. 156–65. Springer International Publishing doi:10.1007/978-3-319-59888-8_14.

Biblioteca Virtual Miguel de Cervantes (1999). *Biblioteca Virtual Miguel de Cervantes,* http://www.cervantesvirtual.com/ (Accessed 22 Sep. 2017).

Biblioteca Virtual Miguel de Cervantes (2007). *Biblioteca del Soneto [Sonnet Library]*, http://www.cervantes-

virtual.com/bib/portal/bibliotecasoneto/ (Accessed 22 Sep. 2017).
Brickley, D. and Miller, L. (2014). FOAF Vocabulary Specification, http://xmlns.com/foaf/spec/ (Accessed 22 Nov. 2017).
Chiarcos, C. and Erjavec, T. (2011). Owl/dl formalization of the multext-east morphosyntactic specifications. *Proceedings of the 5th Linguistic Annotation Workshop, LAW V* '11, pp. 11–20 Stroudsburg: PA, USA,
Ciotti, F., Peroni, S., Tomasi, F., and Vitali, F. (2016). An OWL 2 Formal Ontology for the Text Encoding Initiative. *Digital Humanities 2016: Conference Abstracts,* pp. 151–153
DCMI Usage Board (2012). DCMI Metadata Terms, http://dublincore.org/documents/dcmi-terms (Accessed 22 Nov. 2017).
Domínguez Caparrós, J. (2009). *El moderno endecasílabo dactílico, anapéstico o de gaita gallega*. Sevilla: Padilla Libros.
Domínguez Caparrós, J. (2014). *Métrica española*. Madrid: UNED.
Ehrlicher, H., and Rißler-Pipka, N. (2015). *Revistas Culturales 2.0*, https://www.revistas-culturales.de/es. (Accessed 22 Sep. 2017).
Elf Edition: *Sonett-Archiv,* http://sonett-archiv.com. (Accessed 22 Sep. 2017).
Escribano, J., González-Blanco, E. and Río Riande, G. del (2016). *PoeMetCa—Recursos digitales para el estudio de la Poesía Medieval Castellana,* http://poemteca.linhd.es (Accessed 22 Sep. 2017).
Gago Jover, F. (2015). La biblioteca digital de textos del español antiguo (BiDTEA). *Scriptum Digital* 4: pp. 5–36.
García González, R. (2005). *Sonetos del siglo XVIII*. Biblioteca Virtual Miguel de Cervantes, http://www.cervantesvirtual.com/obra-visor/sonetos-del-siglo-xviii--0/html/. (Accessed 26 Nov. 2017).
García González, Ramón (2006a). *Sonetos del siglo XV al XVII*. Biblioteca Virtual Miguel de Cervantes, http://www.cervantesvirtual.com/obra-visor/sonetos-del-siglo-xv-al-xvii--0/html/ (Accessed 26 Nov. 2017).
García González, R. (2006b). *Sonetos del siglo XIX*. Biblioteca Virtual Miguel de Cervantes, http://www.cervantesvirtual.com/obra-visor/sonetos-del-siglo-xix--0/html/ (Accessed 26 Nov. 2017).
González-Blanco, E. and Rodríguez, J. L. (2014). ReMetCa: A Proposal for Integrating RDBMS and TEI-Verse. *Journal of the Text Encoding Initiative*, 8 https://jtei.revues.org/1274 (Accessed 22 Sep. 2017), doi:10.4000/jtei.1274.
Henríquez Ureña, P. (1919). El endecasílabo castellano. *Revista de Filología Española*, 6: pp. 132–157.
Herman, Ivan, Asida, B., McCarron, S., Birbeck, M. (2015). RDFa Core 1.1 - Third Edition, https://www.w3.org/TR/rdfa-core (Accessed 22 Nov. 2017).
Jewell, M. O. (2010). Semantic screenplays: Preparing TEI for Linked Data. In *Digital Humanities 2010*. http://dh2010.cch.kcl.ac.uk/academic-programme/abstracts/papers/html/ab-878.html (Accessed 22 Nov. 2017).
Marcos Marín, F. and Faulhaber, C. B. (coord.) (1992). *ADMYTE. Archivo Digital de Manuscritos y Textos Españoles*, http://www.admyte.com/admyteonline/contenido.htm (Accessed 22 Sep. 2017).
McCrae, J., Spohr, D. and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. *Proceedings of the 8th Extended Semantic Web Conference on The Semantic Web: Research and Applications*, V (Part I): pp. 245–259, Berlin: Springer-Verlag.
Navarro-Colorado, B. (2015). A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects. *ACL Workshop on Computational Linguistics for Literature*.
Navarro-Colorado, B. (2017). *ADSO project – Análisis distante del soneto castellano de los Siglos de Oro [Distant analysis of the Spanish Golden Age sonnet]*, http://adso.gplsi.es/index.php/es/proyecto-adso (Accessed 22 Sep. 2017).
Navarro-Colorado, B., Ribes Lafoz, M. and Sánchez, N. (2015). *Corpus of Spanish Golden-Age Sonnets*. Alicante: University of Alicante, https://github.com/bncolorado/CorpusSonetosSigloDeOro (Accessed 22 Sep. 2017).
Navarro-Colorado, B., Ribes Lafoz, M. and Sánchez, N. (2016). Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation. *Proceedings of the Language Resources and Evaluation Conference* http://www.lrec-conf.org/proceedings/lrec2016/pdf/453_Paper.pdf (Accessed 22 Sep. 2017)
Navarro-Colorado, B. (2017). A metrical scansion system for fixed-metre Spanish poetry. *Digital Scholarship in the Humanities.* https://doi.org/10.1093/llc/fqx009 (Accessed 22 Sep. 2017)
Quilis, A. (1964). *Estructura del encabalgamiento en la métrica española.* Consejo Superior de Investigaciones Científicas, Patronato Menendez y Pelayo, Instituto Miguel de Cervantes.
Ruiz Fabo, P., Martínez Cantón, C., Poibeau, T. and González-Blanco, E. (2017). Enjambment detection in a large diachronic corpus of Spanish sonnets. *LaTeCH-CLFL 2017, Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature.* Vancouver, Canada.
Santa María Fernández, M. T., Jiménez Fernández, C. M. (2017). *Biblioteca Electrónica Textual Del Teatro Español, 1868-1936*. Universidad Internacional de la Rioja, Spain.
Schöch, C. Henny, U., Calvo Tello, J. Popp, S. (2015). *The CLiGS Textbox*, https://github.com/cligs/textbox (Accessed 22 Sep. 2017)
Wikisource: *Categoría: Sonetos.,* https://es.wikisource.org/w/index.php?title=Categor%C3%ADa:Sonetos (Accessed 26 Nov. 2017)

Polysystem Theory and Macroanalysis. A Case Study of Sienkiewicz in Italian

Jan Rybicki
jkrybicki@gmail.com
Jagiellonian University, Krakow, Poland, Poland

Katarzyna Biernacka-Licznar
katarzyna.biernacka-licznar@uwr.edu.pl
Uniwersytet Wrocławski, Wrocław, Poland

Monika Woźniak
moniwozniak@gmail.com
Università degli Studi di Roma „La Sapienza", Italy

Introduction

Even-Zohar's polysystem theory is a well-established approach to understanding how entire translated literatures interact (or not) with the body of the receiving native literary culture. Even-Zohar identifies a number of possible interactions depending on the relative "strength" and "age" of the two (or more) literatures, and translated literatures may assume "peripheral" or "central" positions within the target literary polysystem. According to this scholar, translations are usually peripheral to native literature; but he also cites examples where a given literary polysystem places some imported subsystems in a central position, while other "foreign imports" remain in the periphery (Even-Zohar 1990).

Even-Zohar thus deals with literary creation en masse rather than, as is often the case in academic approaches to literary translation, on single books original and translated. The obvious parallel to Distant Reading has already been drawn (Helgesson and Vermeulen 2015, 25-26); but it might also be tempting to do the same for a related approach, macroanalysis, if we are to follow the distinction made by the exponent of the latter term (Jockers 2013, 48). Both bring together investigations into masses of literary material unattainable by traditional close reading; yet macroanalysis looks inside many books at once using quantitative methods applied to their lexical layers that have been called "stylometry" well before both Moretti and Jockers.

Material

From our personal mixed Polish-Italian perspective, few cases could serve as a better pretext to try to negotiate this marriage between polysystem theory and computational stylistics than that of *Quo vadis* (1896), the historical romance by Henryk Sienkiewicz, Poland's first literary Nobel Prize winner of 1905. Its international success – long gone with the wind but unparalleled by any other Polish novel to this day – resulted in a veritable explosion in terms of numbers of translations into various languages. In many countries, several different translations simultaneously vied for the public's attention. Yet "several" does not even begin to describe the situation in Italy, where at least three hundred different editions can be still found today (Woźniak 2016). In the first two years of the existence of *Quo vadis* on the Italian market (1899-1900), as many as eight different translations were already available to the readers (Berti and Gagetti 2016).

No wonder: not only was the novel set in the Italian capital and not only did it deal with a subject already very present in Italian culture old and new; the book's (and its author's) brand of conservative Catholicism must have appealed to some of the most influential circles of the country. Yet the novel was also praised by some of Italy's progressive critics, who saw, in Sienkiewicz's persecuted Christians, the struggle of their contemporary revolutionary movements, and who liked to read his depiction of Imperial Rome's decadence as a diatribe against the existing power structure (Marinelli 1984).

This profusion of Italian renderings is also the reason why building their representative selection was no easy task. Only a single translation was available online; a search in Polish and Italian libraries provided almost seventy candidate texts: signed or unsigned by a translator, published by a variety of publishers, often in several somewhat different editions. In the end, twenty-four translations produced until mid-20th c. have been identified as more or less independent of each other, although some of these still share over 50% of material, as evidenced by comparison of texts for identical 5- or longer word clusters with *WCopyFind* (Bloomfield 2011-2016). When applied to genuinely different translations, the similarity ratio is of the order of 5-7%.

The natively Italian literary polysystem was represented by close to 1300 different literary texts, mostly selected and adapted from *Progetto Manuzio*, one of the most comprehensive Internet collections of electronic texts in Italian. To include as many texts as possible, this set of Italian writing included dramas, epic poems and opera libretti as well as novels and novellas from the 15th to the 21st century. Several translations of other novels by Sienkiewicz were also added to the collection, and another big body of translations of a single author, Shakespeare, was included as well.

Methods

The stylometric method applied has been described by Eder (2017) and applied to other literary corpora by Rybicki (2014, 2016). Basing on Burrows's Delta procedure (2002), a list of most-frequent words (MFWs) is produced for the entire corpus. These words are then counted in the individual texts, and their frequencies are compared in text pairs to produce a matrix of distance measures; in this study, the distances were established by means of the modified Cosine Delta (Smith and Aldridge 2011), which is now seen as the most reliable version (Evert et al. 2017). The distance matrix then undergoes Cluster Analysis (Ward's hierarchical clustering), resulting in grouping the texts into "clus-

ters" of greatest similarity; this is repeated for reiterations from 100 to 2000 MFWs at 100-word increments, and a consensus between the individual iterations is produced to show each text's most consistent nearest neighbors, next-to-nearest neighbors and next-to-next-to nearest neighbors. The procedure is performed by means of *stylo* (Eder et al. 2016), a stylometric package for *R* (R Core Team 2016). The results are visualized by means of network analysis, applying the "Force Atlas 2" gravitational algorithm (Jacomy et al. 2008) in *Gephi* (Bastian et al. 2009) to the above-mentioned scores. Instead of applying a "human-made" classification of the resulting network of nodes and edges (i.e. identifying authors, genres and literary periods based on external and traditional literary history), the task of dividing the network into groups of greatest internal similarity was entrusted to *Gephi*'s modularity function, which finds communities within a weighted network (Blondel et al. 2008). The main experiment was conducted by successively increasing the number of communities shown until the expected separate cluster of translations of *Quo vadis* became a separate entity in the network, and the degree of its discreteness could thus be assessed.

Results

Dividing the network into just two modularity groups failed to isolate Sienkiewicz from the main Italian community. Instead, the main division was that between 19th/20th-century novels, translated or originally Italian, and everything else – the one notable exception to this rule was the prose of Pirandello, classified with the earlier texts. At three modularity groups, Italian drama detached itself from early prose. At four, the first writer became a separate community, but this was the native Deledda rather than the alien Sienkiewicz. At five, 19th- and 20th/21st-century novels became two distinct groups; at six, another native Italian, Salgari, received his own class; at seven, pre-19th-century works detached themselves from later prose. It is only at ten communities that a translated rather than an Italian author became a separate subsystem (to use Even-Zohar's term) – in fact, not one but two: Sienkiewicz (not just his *Quo vadis*) and Shakespeare (Fig. 1).

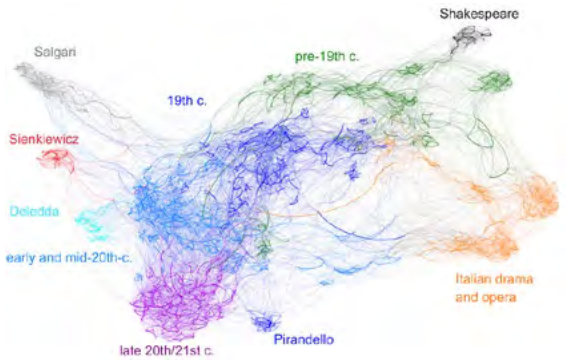


Figure 1. Network analysis of distances between most-frequent-word usage. Thick and short lines (edges) denote small distance (or high similarity). For simplicity, only the final 10-community modularity is shown.

Discussion

It seems too much of a coincidence that two major subsystems (translations of Shakespeare and translations of Sienkiewicz) become separated from the main body of literature in Italian at the same time, and that this happens only after two native authors receive their own subsystems. If such a mechanism were to be observed in even more extensive collection of texts (when they finally become available), Even-Zohar's hypothesis of the usually peripheral position of translated literature could find its stylometric illustration. At the same time, this experiment confirms not only that original novels are more similar to translated ones than the former to original drama; but also that certain original authors are more different from other original authors than those translated from another language.

Obviously, this hypothesis must be tested in the future in other literary polysystems to claim that the affinity between polysystem theory and macroanalysis is anything more than metaphorical. Even-Zohar speaks of reception of literary works within a broader national culture; macroanalysis counts context-free words. Still, in its attempts to bring distant and close reading together, stylometry has been clutching at even weaker straws. Stylometrists continue to make similar leaps (of faith?) between their graphs and trees and networks on the one hand, and traditional literary history on the other. They usually believe that frequencies of very frequent words provide insights into more abstract characteristics of texts than their mere lexical or even grammatical difference: and these abstracts so far include authorship, genre, chronology, or gender. This study might just have added a new one. At the very least, it is an invitation to apply Even-Zohar's concepts in various "distant" approaches to literature.

Acknowledgements

This research was made as part of the project: "Miejsce *Quo vadis*? w kulturze włoskiej. Przekłady, adaptacje, kultura popularna" (0136/ NPRH4/H2b/83/2016), funded by Poland's National Program for Advances in the Humanities (NPRH).

References

Bastian, M., Heymann, S., and Jacomy, M. (2009). "Gephi: an open source software for exploring and manipulating networks." *Proceedings of the International AAAI Conference on Weblogs and Social Media*, San Jose, Ca.

Berti, G. de, and Gagetti, E. (2016). "La fortuna di 'Quo vadis' in Italia nel primo quarto del Novecento." In Woźniak, M., Biernacka-Licznar K., eds, *Quo Vadis. Da caso letterario a fenomeno di massa. Ispirazioni - adattamenti - contesti.* Roma: Ponte Sisto, 50-59.
Blondel, V.D., Guillaume, J.-L., Lambiotte, R., Lefebvre, E. (2008). "Fast Unfolding of Communities in Large Networks," *Journal of Statistical Mechanics: Theory and Experiment* 10: 1000.
Bloomfield, L. (2011-2016). *WCopyFind. The Plagiarism Resource Site*, http://plagiarism.bloomfieldmedia.com. Accessed 24. Nov. 2017.
Burrows, J.F. (2002). "Delta: A Measure of Stylistic Difference and a Guide to Likely Authorship." *Literary and Linguistic Computing* 17: 267-287.
Eder, M. (2017). "Visualization in stylometry: Cluster analysis using networks." *Digital Scholarship in the Humanities* 32(1): 50-64.
Eder, M., Kestemont, M., and Rybicki, J. (2016). "Stylometry with R: A package for computational text analysis." *The R Journal* 8(1): 107–121.
Even-Zohar, I. (1990). "The Position of Translated Literature within the Literary Polysystem." In *Polysystem Studies [= Poetics Today]* 11(1): 45-51.
Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C., and Vitt, T. (2017). "Understanding and explaining Delta measures for authorship attribution." *Digital Scholarship in the Humanities* 32 (sup. 2): 4-16.
Helgesson, S. and Vermeulen, P. (2015). „Introduction. World Literature in the Making." In Helgesson, S. and Vermeulen, P. eds, *Institutions of World Literature, Writing, Translation, Markets*. London: Routledge, 1-22.
Jacomy, M., Venturini, T., Heymann, S., and Bastian, M. (2008). "ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software." *PLoS ONE* 9(6): e98679. DOI=10.1371/journal.pone.0098679.
Jockers, M. (2013). *Macroanalysis. Digital Methods and Literary History*, Champaign: University of Illinois Press 2013.
Marinelli, L. (1984). "'Quo vadis.' Traducibilità e tradimento," *Europa Orientalis* 3: 131-146.
R Core Team. (2016). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.
Rybicki, J. (2014). "Pierwszy rzut oka na stylometryczną mapę literatury polskiej," *Teksty drugie* 2: 106-128.
Rybicki, J. (2016). "Vive la différence: Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies," *Digital Scholarship in the Humanities* 31(4): 746-761.
Smith, P. and Aldridge, W. (2011). "Improving authorship attribution: Optimizing Burrows' Delta method." *Journal of Quantitative Linguistics*, 18(1): 63–88.
Woźniak, M. (2016). "Quo vadis: da caso letterario a fenomeno di massa. Dove ci ha portato Sienkiewicz?" In Woźniak, M., Biernacka-Licznar K., eds, *Quo Vadis. Da caso letterario a fenomeno di massa. Ispirazioni - adattamenti - contesti*. Ponte Sisto, Roma, 6-15

Interrogating the Roots of American Settler Colonialism: Experiments in Network Analysis and Text Mining

Ashley Sanders Garcia
asanders@cmc.edu
The Claremont Colleges, United States of America

Even as the United States fought for independence in the American Revolution, it was already in the process of becoming a settler colonial power in its own right. This short paper interrogates the origins of American settler colonialism through text mining three corpora of personal and official documents. In order to understand and address present structural inequity in the United States, scholars, policy-makers, educators, and the public need to examine the country's long history as a settler colonial society.

Through topic modeling and text mining methods, my research highlights the underlying goals and desires that prompted land acquisition, settlement, and cycles of violence between Euro-American settlers and Native Americans in the trans-Appalachian west between 1776 and 1820. This project explores three collections, or corpora, of documents, separated by the positions of the historical authors and document type: settler correspondence and records; official government documents; and writings of political elites in the eastern United States. The first corpus for this study consists of correspondence, journals, and memorials from settlers, colonial officials and military leaders in the territories (colonies) between 1776 and 1820. This is the smallest corpus of the three, at two million words. Few documents from representative settlers have been transcribed and published, so the corpus over-represents leaders in the settler communities, however the petitions from the settlers to Congress give voice to the most pressing challenges, needs, and hopes of the settlers themselves. The documents included in each corpus were transcribed and published in bound volumes during the nineteenth century and are now in the public domain. A second corpus, of approximately four million words, consists of official government records, including treaties with Native American communities, military records, documents related to public lands and governance of the territories, as well as pension and other petitions submitted to Congress in the late eighteenth and early nineteenth centuries. The third corpus is, by far, the largest of the three, at approximately 39 million words, and consists of the papers of the foremost political leaders in the eastern United States. The letters of the members of the Continental Congress are included, as are the writings of George Washington, James Madison, Thomas Jeffer-

son, Benjamin Franklin, and John Adams. Not surprisingly, these statesmen wrote far more than settlers, who were primarily concerned with agricultural cultivation, hunting, and defending their families on the frontier.

The aforementioned sources form the corpora for text mining and analysis experiments. My study extracts and compares American settler, administrator, and political leaders' perspectives on significant topics in the study of settler colonialism, such as land value; property acquisition and sales; as well as the presence, actions, and views of Native Americans. Early experiments using the LDA algorithm in MALLET to topic model the corpora and Lexos to visualize the topic clouds have already revealed significant patterns (Blei, 2012).

While recognizing that topic models are more effective with large corpora, my research began with a small experiment. Using MALLET, I created a topic model of ten topics of the twenty-five published petitions from settlers to Congress (1787-1798) from the *Territorial Papers of the United States.* This model suggests that one of the primary motivations for Euro-American emigrants to move to the western territories was to achieve what they described as a competency, or the means to rear their children "in a comfortable manner" and "raise a subsistence by their [own] industry" (Petition from the Inhabitants of Vincennes to Congress, 1787). The topic related to land reveals the dominant concerns that settlers expressed. They implored Congress to recognize their existing land claims, ensure reasonable land prices, provide military protection from Native American raids, and ensure justice through the provision of judges. These measures, they believed, would foster access to land, enable trade, establish legitimacy, and provide settlers with the means to achieve their modest goals.

Even though their objectives differed from those of the settlers, government officials both in the east and on the ground, in the western territories, were equally motivated to acquire land beyond the Appalachian Mountains. In the aftermath of the American Revolution, the government was in dire financial straits. Political leaders urged agents to obtain western lands from Native communities so that the territory could be sold to pay off the burdensome war debts. Consequently, backcountry government officials decried settler violence against neighboring Indigenous communities, even as they took advantage of the unruly settlers' actions to compel land cessions that the United States government desperately needed.

Figure 1: Topics related to land in the Continental Congress members' correspondence

There was a high price to be paid for white American independence though, as is demonstrated in the topics generated from the Continental Congress members' correspondence records (Figure 1). The words "transmitted, negotiations, ceding, extinguishment, extinguishing," and, ominously, "funeral" stand out among the more benign "northwest, lands, and western." Most of these words are more or less neutral when considered out of context, but, given their use in relation to the settler colonial endeavor, they evidence the brutal effects of American land acquisition and expropriation from Native communities.

These topics and the related documents both direct attention to specific sources for close reading, but also yield new terms of interest to explore at a distance and in a broad comparative framework. In addition to the results of topic modeling the aforementioned corpora, this presentation will also share experiments using part-of-speech tagging and collocations to explore concepts, such as land, family, independence, competency, and war to understand the ways in which settlers, and political and military leaders conceived each of these topics.

This talk offers an initial glimpse into the early stages of a much larger project that seeks to create an interactive interface for documents from the first four decades of the United States' formation as a nation and nascent empire based on topic models and text mining approaches, such named entity recognition, and collocates. The interface will eventually allow users to drill down into documents that contain specific sought-after features, such as individuals' names, gender identity, topics of interest, etc.

This interface, it is hoped, will enable historians, students, genealogists, and interested members of the public to explore some of the most important documents related to the complicated, conflicting, and, occasionally, complementary objectives of American settlers and other political actors. The policies these agents developed between 1776 and 1820 not only shaped American settler colonialism in the eighteenth and nineteenth centuries, but they continue to reverberate more than two centuries later.

References

Blei, D. M. (2012) "Probabilistic Topic Models." *Communications of the ACM* 55.4 (April 2012): 77-84.

Blei, D. M. (2012) "Topic Modeling and Digital Humanities." *Journal of Digital Humanities* 2.1 (Winter 2012). Web. http://journalofdigitalhumanities.org/2-1.

The Inhabitants of Vincennes to Congress, July 26, 1787, in *Territorial Papers of the United States,* Volume 2 (Washington, D.C.: United States Government Publications Office, 1934): 58-60.

¿Existe correlación entre importancia y centralidad? Evaluación de personajes con redes sociales en obras teatrales de la Edad de Plata?

Teresa Santa María
teresa.santamaria@unir.net
Universidad Internacional de la Rioja, Spain

Elena Martínez Carro
elena.martinez@unir.net
Universidad Internacional de la Rioja, Spain

Concepción Jiménez
concepcionm.jimenez@unir.net
Universidad Internacional de la Rioja, Spain

José Calvo Tello
jose.calvo@uni-wuerzburg.de
University of Würzburg, Germany

Resumen

¿Los nodos centrales de una red social de personajes son los protagonistas de una obra de teatro? Para responder a esta pregunta utilizamos diferentes medidas de centralidad junto con otros valores cuantitativos textuales en un corpus anotado de obras dramáticas de teatro español correspondientes a la Edad de Plata (1868-1936). Los resultados señalan que la centralidad está en correlación moderada con la importancia, siendo mayor la correlación con valores cuantitativos textuales.

Introducción

La representación de personajes literarios mediante grafos y redes sociales (Marcus 1973, Moretti 2011) aporta nuevas herramientas al estudio literario. La interpretación del concepto de centralidad en grafos (Jannidis et al., 2017) ha sido investigada en su aplicación a las obras literarias (Moretti 2011; Rochat 2014; Trilcke et al. 2015 y 2016; Jannidis et al., 2016, Rodríguez 2016; Algee-Hewitt 2017). En la tradición hispánica, se han utilizado enfoques cuantitativos para analizar la densidad versal en obras del Siglo de Oro (Hermenegildo 1994 y Espejo 2002), estudiar tanto contenido simbólico y sociopolítico de los personajes de Galdós (Menéndez 1983), así como el origen social o caracterización de los personajes de Lope de Vega (Oleza 1984 y Oleza 2013).

En este trabajo queremos evaluar cuatro preguntas:

1. ¿Qué tipo de correlación hay entre las medidas de centralidad y la importancia del personaje?
2. ¿Aparecen los personajes más importantes al comienzo del *dramatis personae*?
3. ¿Hay correlación entre importancia y valores textuales (cantidades de unidades textuales del personaje)?
4. ¿Qué valores podríamos utilizar para distinguir a los protagonistas del resto?

Textos y metadatos

A diferencia de otras lenguas europeas, el español no cuenta con un gran corpus teatral anotado en XML-TEI. El proyecto *Biblioteca Electrónica Textual del Teatro en Español de la Edad de Plata (1868-1936)* (BETTE) ha publicado veinticinco obras en XML-TEI de Lorca, Valle, Galdós, Clarín o Muñoz Seca, como repositorio GitHub (María Jiménez et al., 2017). En la versión 2.0 cada personaje ha sido anotado con diferentes metadatos:

- Sexo
- Papel en la obra (protagonista, amante, antagonista u otro)
- Naturaleza (persona, animal, no humano...)
- Importancia (personaje primordial, secundario o terciario)
- Persona individual frente a grupo

Además, se añadieron una serie de valores textuales cuantitativos de manera automática:

- Posición en el dramatis personae (castList)
- Cantidad de texto que pronuncia
- Cantidad de intervenciones
- Cantidad de referencias a su nombre
- Cantidad de escenas en las que aparece

Aquí un ejemplo de esa información en XML-TEI:

```
<person n="1" role="protagonist" sex="M" xml:id="max">
   <persName>Max Estrella</persName>
   <affiliation type="nature">person</affiliation>
   <affiliation type="importance">primary</affiliation>
   <ab>
      <measure unit="characters">15813</measure>
      <measure unit="sps">278</measure>
      <measure unit="rss">96</measure>
      <measure unit="scenes">11</measure>
   </ab>
</person>
```

Fig. 1. Metadatos de personaje en XML-TEI

El valor de importancia fue asignado según los siguientes criterios:

- Minor: si el personaje no aparece en el resumen (contenido también en el archivo TEI)
- Secondary: si aparece en el resumen
- Primary: si pertenece al grupo de entre dos y cuatro personajes esenciales

De esta manera por cada personaje (con un total de 516) tenemos:

1. Un valor de su importancia dentro de la obra (que puede ser utilizado como *ground truth*)
2. Diferentes valores cuantitativos textuales
3. Posición en dramatis personae
4. Diferentes valores según medidas de centralidad

Metodología

La implementación para extraer, analizar, evaluar y visualizar los datos se realizó en Python mediante librerías como *lxml* y *networkx*. Para la creación de las redes sociales se definió la arista no direccional como la coaparición en escenas (la definición más frecuente en trabajos de este tipo):

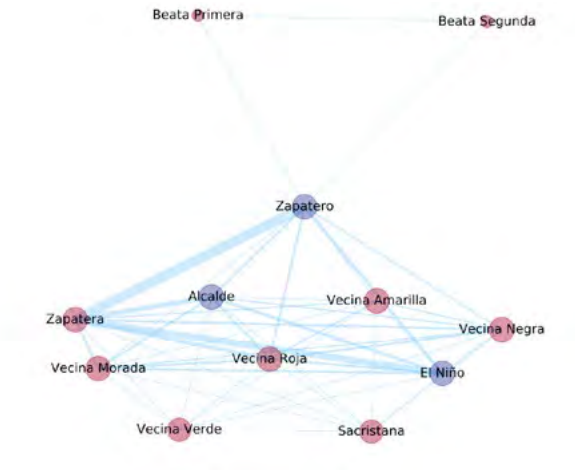


Fig. 2. Red social de personajes en La zapatera prodigiosa de Lorca

A partir de estas redes sociales, calculamos diferentes medidas de centralidad e información sobre los nodos:

- Degree
- Betweenness centrality
- Eccentricity
- Closeness centrality
- Load centrality
- Current flow betweenness centrality
- Eigenvector centrality
- Approximate current flow betweenness centrality
- Communicability centrality exp

Resultados

Analizamos la dependencia entre la importancia y el resto de valores, calculado su correlación (Spearman)

Ninguna de las medidas de centralidad tiene una correlación fuerte (> 0.6 o < -0.6 según Evans 1996). El valor máximo (0.51 en correlación negativa) es de *current flow betweenness centrality*, también conocida como *information centrality* (Brandes and Fleischer 2005; Stephenson and Zelen 1989), medida que no está entre el repertorio usual de las HD.

En cuanto a la posición en el *dramatis personae*, la correlación es solo de 0.42, con una fuerte dispersión, aunque los primeros y terceros cuartiles de personajes primarios y terciarios se posicionan en rangos totalmente diferentes. Es decir, la posición en el dramatis personae sí parece aportar cierta información sobre la importancia, aunque no podemos utilizarlo de manera exclusiva (p.ej. Muñoz Seca los ordena por sexo).

En tercer lugar, las medidas de cuantitativas textuales tienen todas correlaciones notablemente más altas, llegando hasta 0.67 en la cantidad de intervenciones.

Ante estos resultados, nos hemos preguntado si las medidas cuantitativas textuales tienen el mismo tipo de correlación con las medidas de centralidad, en concreto si la information centrality tiene una correlación más fuerte que el resto (calculando Spearman o Pearson, dependiendo si las variables son continuas u ordinales):

Como se observa *current flow betweenness (o information) centrality*, de nuevo, es la medida de centralidad con la correlación más fuerte con la cantidad de intervenciones.

Finalmente hemos observado si la distribución de centralidad o valores textuales son diferentes para los personajes protagonistas de los del resto:

La mayor diferenciación de ambos *boxplots* entre las medidas de centralidad se consigue mediante *current flow betweenness* (o *information*) *centrality*. El solapamiento menor se consigue mediante la cantidad de texto pronunciado (pers_mes_characteres). La posición relativa en el *dramatis personae* en este caso consigue diferenciar de manera bastante clara los protagonistas del resto de personajes.

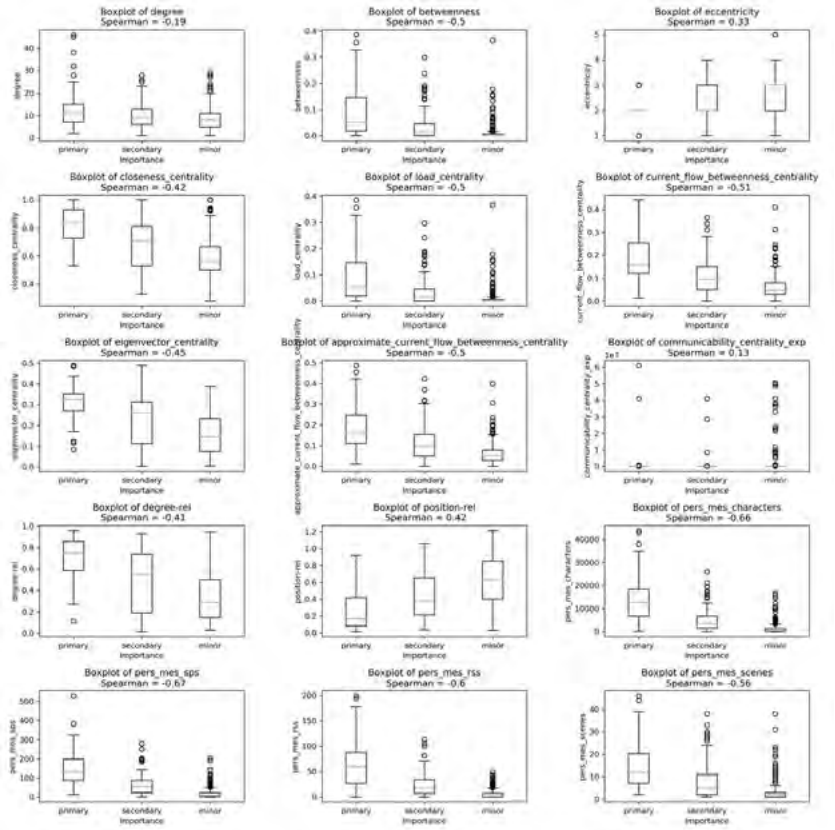


Fig. 3. Boxplots y correlaciones con importancia de todas las obras de BETTE

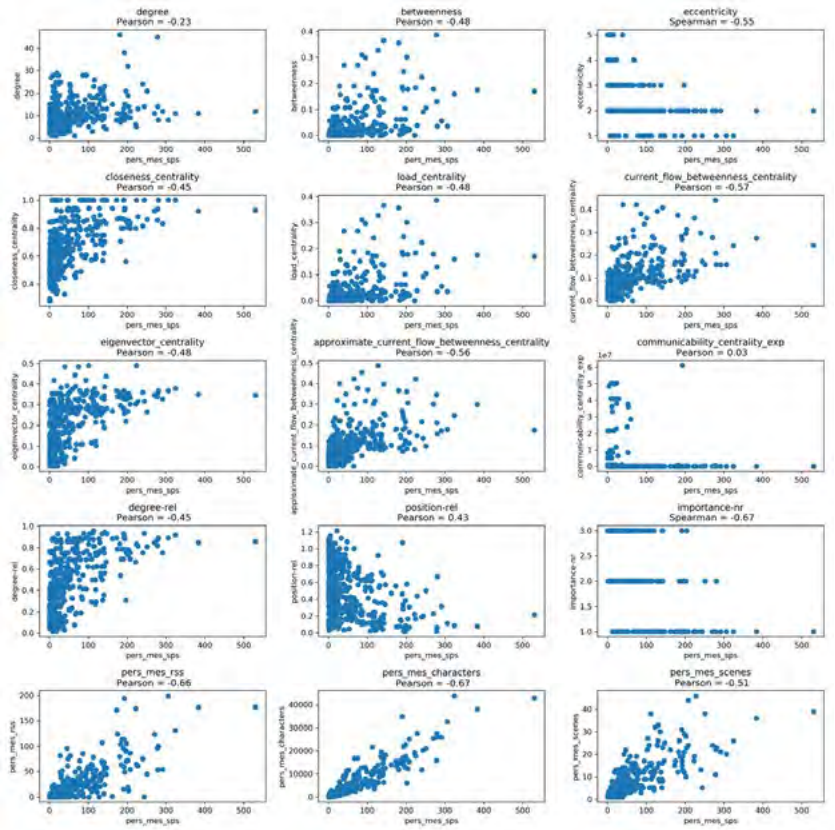


Fig. 4. Scatterplots mostrando correlación entre las veces que un personaje habla (<sp>s) y otros valores

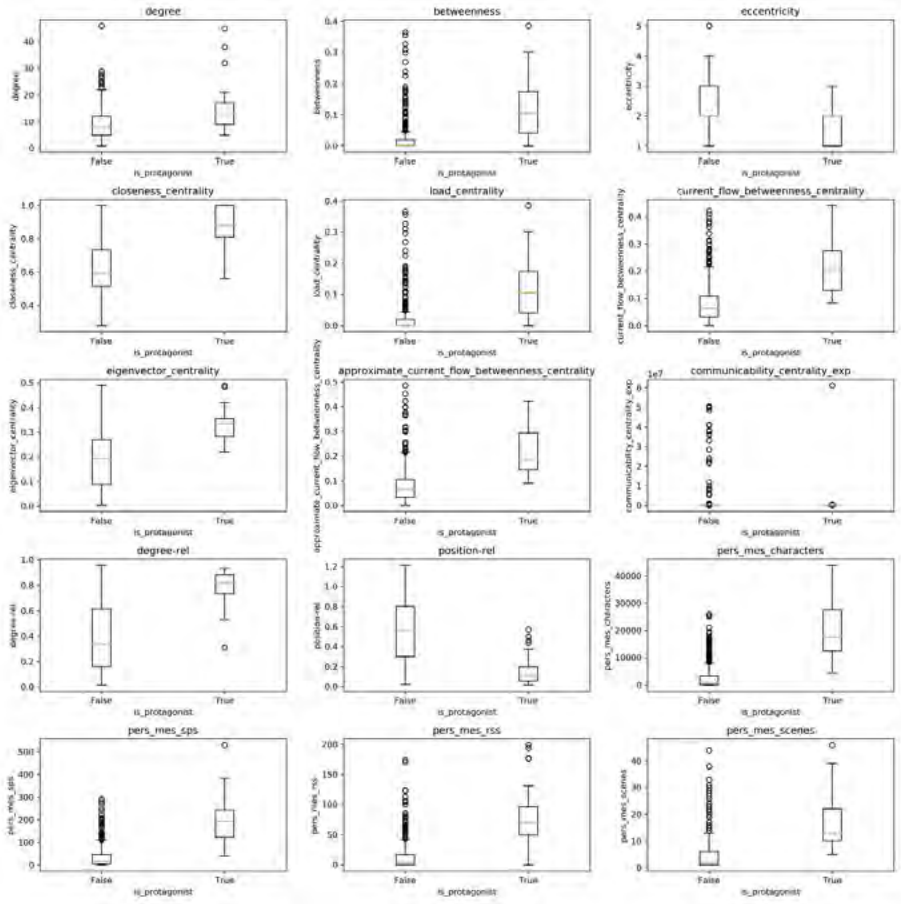


Fig. 5. Boxplots de protagonistas frente al resto de personajes

Conclusiones y futuros pasos

La anotación en detalle de información sobre los protagonistas nos permite evaluar métodos digitales. En concreto seguimos la propuesta de Moretti (2013) de abandonar la división binaria de personajes, incluyendo en nuestro caso los valores de personajes secundarios.

Nuestros resultados muestran que, para el caso del corpus BETTE y con las formalizaciones arriba explicadas:

1. La importancia tiene una correlación solamente entre débil y moderada con cualquier formalización de centralidad, teniendo la correlación más fuerte la *information centrality*
2. La posición en el *dramatis personae* puede ser un indicador sobre el protagonismo de personajes o la diferenciación entre primarios y terciarios, pero no para diferenciar a estos de los secundarios
3. Los valores cuantitativos textuales tienen correlaciones más fuertes. Este tipo de unidades son también las que mejor clasificarían personajes entre protagonistas y no protagonistas
4. Es sorprendente que unidades textuales más sencillas que la centralidad en redes aporten más información tanto sobre la importancia de los personajes, así como su papel de protagonistas.

Como otros trabajos en redes sociales (cf. Moretti 2011 y 2013; Rochat 2014) hemos trabajado con una cantidad reducida de textos. Nos gustaría comprobar estas hipótesis en mayores corpus literarios. También nos gustaría analizar los efectos que subgéneros literarios, períodos y autores ejercen sobre estos valores.

References

Algee-Hewitt, Ma. (2017). *Distributed Character: Quantitative Models of the English Stage, 1500-1920*. Montréal: McGill University & Université de Montréal, pp. 119–21.

Brandes, U. and Fleischer, D. (2005). Centrality Measures Based on Current Flow. *Theoretical Aspects of Computer Science (STACS '05)*. Springer-Verlag, pp. 533–44 http://www.inf.uni-konstanz.de/algo/publications/bf-cmbcf-05.pdf.

Espejo, J. (2002). Algunos aspectos sobre la construcción del personaje en el teatro conservado de Hernán López de Yanguas (1487-¿?). *Scriptura,* 17, pp. 113-132.

Evans, J. D. (1996). *Straightforward Statistics for the Behavioral Sciences*. Pacific Grove: Brooks/Cole Pub. Co.

Gómez, S., Calvo Tello, J., González, J. M. and Vilches, R. (2015). Hacia una biblioteca electrónica textual del teatro en español de 1868-1936 (BETTE). Texto Digital, 11(2), pp. 171–84.
Hermenegildo, A. (1995). Personaje y teatralidad: la experiencia de Juan del Encina en la Égloga de Cristino y Feba. In Pedraza, F.B. y González, R. (ed.*). Los albores de teatro español: actas de las XVII Jornadas de teatro clásico Almagro, julio de 1994*. Almagro: Universidad de Castilla-La Mancha, pp. 90-113.
Jannidis, F., Reger, I., Krug, M., Weimer, L., Macharowsky, L. and Puppe, F. (2016). Comparison of Methods for the Identification of Main Characters in German Novels. *DH2016*. Krakow: ADHO, pp. 578–82 http://webcache.googleusercontent.com/search?q=cache:LjYz88cQhboJ:dh2016.adho.org/abstracts/297+&cd=1&hl=es&ct=clnk&gl=de&client=ubuntu.
Jannidis, F., Kohle, H. and Rehbein, M. (eds). (2017). *Digital Humanities: eine Einführung*. Stuttgart: J.B. Metzler Verlag.
Jiménez, C., Martínez Carro, E., Santa María, M. T., Calvo Tello, J., Simón Parra, M., Martínez Nieto, R. B. and García Sánchez, M. (2017). BETTE: Biblioteca Electrónica Textual del Teatro en Español de la Edad de Plata. *Sociedad, Políticas, Saberes*. Málaga: HDH, pp. 88–91 http://hdh2017.es/wp-content/uploads/2017/10/Actas-HDH2017.pdf.
Marcus, S. (1973). *Mathematische Poetik*. (Trans.) Mândroiu, E. Bucureşti; Frankfurt/Main: Editura Academiei ; Athenäum Verlag.
Menéndez, C. (1983). *Introducción al teatro de Benito Pérez Galdós*. Madrid: CSIC.
Moretti, F. (2011). Network Theory, Plot Analysis. *The New Left Review (*68), pp. 80–102.
Moretti, F. (2013). "Operationalizing": or, the function of measurement in modern literary theory. *The New Left Review (*84), pp. 103-119.
Oleza Simó, J. (2013). *Biblioteca Digital Arte Lope*. Valencia: Universitat de València. artelope.uv.es/biblioteca.
Rochat, Y. (2014). *Character Networks and Centrality*. N.p. Web.
Rodríguez, D.I. (2016) *Análisis de grafos en paralelo mediante Graphx*. Trabajo de titulación. Universidad Católica de Loja. Ecuador.
Stephenson, K. and Zelen, M. (1989). Rethinking centrality: Methods and examples. *Social Networks*, 11(1): 1–37 doi:10.1016/0378-8733(89)90016-6.
Trilcke, P., Fischer, F., Göbel, M. and Kampkaspar, D. (2016). Dramen als small worlds? Netzwerkdaten zur Geschichte und Typologie deutschsprachiger Dramen 1730-1930. In Burr, E. (ed), *DHd 2016 Modellierung, Vernetzung, Visualisierung*. Leipzig: DHd/nisaba, pp. 254–57.
Trilcke, P., Fischer, F. and Kampkaspar, D. (2015). Digitale Netzwerkanalyse dramatischer Texte. *DHd-Tagung. Graz* http://gams.uni-graz.at/o:dhd2015.v.040.

Cultural Awareness & Mapping Pedagogical Tool: A Digital Representation of Gloria Anzaldúa's Frontier Theory

Rosita Scerbo
scerbo@asu.edu
Arizona State University, United States of America

This project looks at the work of American-Chicana poet and fiction writer Gloria E. Anzaldúa, author of *This Bridge we Call Home* (2002). My research proposes a digital representation of Gloria Anzaldúa's Frontier theory as part of my scholarly investigation. This study will include the creation of a mapping tool that will reflect the rhizomatic spaces analyzed by the author, raising awareness about the multiple cultural identities found in the United States. Through personal narratives, theoretical essays, poetry, letters, works of art and fiction, *This Bridge we Call Home* examines issues such as classism, homophobia, racism, political identity, native sovereignty, lesbian pregnancy and motherhood, transgender issues, Arab-American stereotypes, Jewish identities and spiritual activism. These stories are written by women and men, both of color and white, and motivated by a desire for social justice. *This Bridge We Call Home* invites feminists of all colors and genres to develop new forms of transcultural dialogues, practices, and alliances. The anthology, object of study is the last work produced by the author before she passed away and undertakes a more inclusive essence compared with her earlier writings. The book includes women and men of different classes, nationalities, races, ages and sexual orientations, reflecting the desire of inclusivity and dialogue promoted by the author and editor. This project also attempts to bring together multiethnic voices and promotes a interdisciplinary resource that interest not only the literature and culture discipline, but also other humanities fields, such as history, anthropology, sociology and gender studies.

The result of this project will be a powerful new online education and research tool for undergraduate and graduate students as well as the world community at all levels of expertise. To create this public resource I will use the mapping tool "Google Lit Trips", a site affiliated with Google. Normally this tool is used to recreate and mark the journeys of fictional characters from famous literature works. In my case I will use the various sections of Gloria Anzaldúa's anthology that reflect real life experiences of the writers. I will then provide geospatial representations of the true stories narrated by the authors that live some kind of political, racial, sexual or class struggle in the United States. In the book 87 writers are given a space to celebrate their diversity.

In the mapping tool, at each location along the journey there will be placemarks with pop-up windows con-

taining a variety of resources including relevant media, thought provoking discussion starters, and links to supplementary information about 'real world' references made in that particular portion of the text.The author voice herself emerges beyond the limits of either American or Mexican culture and provides a voice to the people of the borderlands. Her work is based on multiple experiences to create a universal history that transcends the social barriers that connect us collectively with each other. While the politics of identities requires subjecting ourselves to specific categories of identity, spiritual activism requires that we get rid of all these barriers.

This project has the objective to put the reader inside the stories, provoking reflections and awareness about contemporary social, political, sexual and racial issue that affect our modern society. The reader will travel alongside with the protagonists of the autobiographical stories through the recreation of 3D geographic tours of the narratives that have been described. At the same time the mapping tool creates an engaging and relevant literary experiences for students. At each location I will be able to include web links, videos, audios, images, annotations and critical activities related with the different sections of the anthology. The experience of the pop up windows provide a range of supplementary information, such as links that give additional information about the 87 authors or cultural traditions that have been mentioned by the characters. The students find themselves seeing the settings almost how they were there. The pop up windows provide engaging content, such as audios, videos or activities related to the story line. These activities are designed to help readers discover connections between their culture and the different cultures that have been described in the story.

One of the primary goals of this project is to emphasize the relevance of cultural diversity in the University environment in the context of the Hispanic world. My objective is to initiate contemporary debates over themes such as immigration, globalization, discrimination, acceptance and inclusion. The mapping tool will explore ways of bringing its unique materials to a wider audience inside and outside the United States. The contribution of this project is not only to continue expressing a dialogue within and between women, women of color, and among people that live in the borderlands, but also to expand visions and theoretical spaces in general. The different stories told in the anthology explore the different shades of the mixed-race identity of women and men that are often perceived as outsiders within their own country.

The digital representation of the anthology and its multiple resources proposes a new attitude towards the learning process of college students and the public sensitivity outside the academia. One of the primary intentions is to dismantle traditional forms of identity, and destroy social boundaries, by embracing difference and otherness as a unique component of every single individual part of our society. The focus on themes such as the effects of migration and globalization are evident in the transnational, transcultural and transgender identities represented though the voices of the 87 writers. The external links provided as resources bring the readers beyond the stories. The students become travelers discovering the similarities and qualities of the characters from cultures beyond their own. This could be an effective way to make students feel part of the stories and hopefully inspire them to fight against the different levels of discrimination that the writers are describing. The final goal will be to include this online platform as an integrative portion of a culture and literature class at the university level.

Corpus Linguistics for Multidisciplinary Research: Coptic Scriptorium as Case Study

Caroline T. Schroeder
carrie@carrieschroeder.com
University of the Pacific, United States of America

The Coptic language is the last phase of the Egyptian language family, descending ultimately from the ancient hieroglyphs. Coptic Scriptorium has developed a multidisciplinary research platform using core Corpus Linguistics tools and methods in collaboration with other disciplinary methods. This paper will argue that this collaborative, interdisciplinary approach allows for the creation of research resources that enrich even *disciplinary* work.

Coptic Scriptorium has created the first open source natural language processing tools for any phase of the Egyptian language family, including a tokenizer, normalizer, part of speech tagger, language of origin tagger (for loan words from Greek, Latin, and other languages), and lemmatizer. We have also contributed annotated data to the universal dependency Treebank project. A fully searchable corpus annotated with these tools is available online at copticscriptorium.org, and all tools and corpora can be downloaded from our GitHub repositories.

This paper will argue that multidisciplinary collaboration improves even disciplinary research. Three examples are provided here; these and others will be demonstrated live in the short paper.

Collaboration with Egyptologists creating a TEI Coptic lexicon file enabled the creation of an online Coptic Dictionary, in which words in our searchable database are hyperlinked to the dictionary entries. The dictionary entries likewise show frequency statistics for the terms in our database. This collaboration benefits Egyptology, by providing an open source corpus for teaching and research linked to a dictionary, and it benefits corpus linguistics, by providing clear frequency data and lexical resources for linguists.

Collaboration with Religious Studies scholars has enabled including in our corpora transcriptions of Coptic manuscripts that have never before been published in print. Scholars in Religious Studies have provided transcriptions of texts to the project, enabling scholars in other disciplines, such as Linguistics, to conduct computational corpus research on important, previously inaccessible texts. Likewise Religious Studies scholars can use the database to conduct philological and historical research on religious texts.

Coptic Scriptorium also annotates manuscript information of interest to archivists, philologists, and codicologists within a multilayer annotation model. This enables codicologists, philologists, and archivists to use the query syntax of our corpus linguistics database (ANNIS) to investigate research questions about scribal practices, spelling and morphology, and other manuscript-related issues over multiple manuscripts, including utilizing metadata such as repository information, dates and locations of the original manuscripts, etc.

We presented the very beginnings of the Coptic Scriptorium project at DH 2014 in Switzerland. This short paper will demonstrate the extensive progress made as a result of collaboration and interdisciplinary partnerships.

Extracting and Aligning Artist Names in Digitized Art Historical Archives

Benoit Seguin
benoit.seguin@epfl.ch
EPFL, Switzerland

Lia Costiner
lia.costiner@epfl.ch
EPFL, Switzerland

Isabella di Lenardo
isabella.dilenardo@epfl.ch
EPFL, Switzerland

Frédéric Kaplan
frederic.kaplan@epfl.ch
EPFL, Switzerland

The largest collections of art historical images are not found online but are safeguarded by museums and other cultural institutions in photographic libraries. These collections can encompass millions of reproductions of paintings, drawings, engravings and sculptures. The 14 largest institutions hold together an estimated 31 million images (Pharos). Manual digitization and extraction of image metadata undertaken over the years has succeeded in placing less than 100,000 of these items for search online. Given the sheer size of the corpus, it is pressing to devise new ways for the automatic digitization of these art historical archives and the extraction of their descriptive information (metadata which can contain artist names, image titles, and holding collection). This paper focuses on the crucial pre-processing steps that permit the extraction of information directly from scans of a digitized photo collection. Taking the photographic library of the Giorgio Cini Foundation in Venice as a case study, this paper presents a technical pipeline which can be employed in the automatic digitization and information extraction of large collections of art historical images. In particular, it details the automatic extraction and alignment of artist names to known databases, which opens a window into a collection whose contents are unknown. Numbering nearing one million images, the art history library of the Cini Foundation was established in the mid-twentieth century to collect and record the history of Venetian art. The current study examines the corpus of the 330'000+ digitized images.

Image Processing Pipeline

Photo/Cardboard Extraction

The records in the Cini Foundation consist of a photographic reproduction mounted on a cardboard card onto which metadata information is recorded. The initial scan of these records is a 300 dpi picture produced on a scanning table, and includes the digitized cardboard and color balance markers. The first task consists in separating the cardboard backing and the photographic reproduction from the raw scanned image.

Despite the apparent simplicity of such a task, it proved challenging on account of the multiple layouts of the metadata information on the cardboard cards, and the variations in the sizes and positions of the attached images. In the end, what proved most effective in the extraction of the image was a Convolutionnal Neural Network (CNN) architecture designed for semantic segmentation (Ronneberger, O. et al 2015). For this, an accurate model was trained on scans which had been annotated in the course of 2 hours. The details oft he approach are part of another study (Ares Oliveira, S. and Seguin, B. 2018).

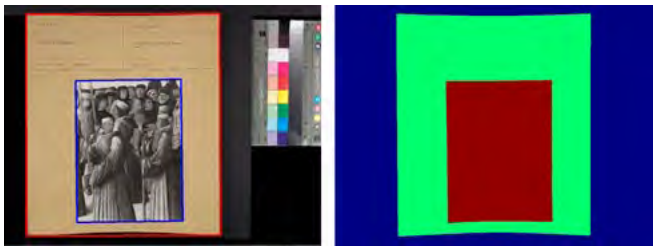


Figure 1 Left: original scan with the extracted areas highlighted with red and blue rectangles.
Right: the prediction mask generated by the neural network.

Text Extraction

The second part of the pipeline consists of extracting and reading the metadata. For this task, the open-source Tesseract toolkit and the commercial Google Vision API were tested, with the latter having better performance.

The OCR system provided a list of words and their positions, which were then clustered into blocks of text representing the different metadata fields (authorship, title of painting, location etc.). A layout model was used to represent the expected positions of these different fields. This allowed the assignment of each block of text to its corresponding metadata field.

A precise analysis of the performance of this step is presented in another publication (Seguin, B. 2018).

Figure 2 Illustration of the OCR process. The extracted words (top-left) are clustered into blocks of metadata (top-right) and then assigned to their corresponding label (bottom).

Automatic Alignment of Artist Names

In order to leverage the extracted metadata to get insights into a collection, it is important to link them to a knowledge database. This can allow, for example, city names to be placed geographically on a map. Here, we focus on aligning artist names with a knowledge database: the Union List of Artist Names (ULAN), managed by the Getty. This opens up a wealth of new information for the contextual understanding of the artwork's creation.

The alignment process is depicted on Figure3, it is a complex two-pass process that integrates automatic matching with collection specific knowledge in an efficient manner. The first pass tries to perform an exact match with a large name dictionary. For the second pass, a list of candidates are generated from the correctly matched elements of the first pass, and approximate matching is used to correct small OCR errors.

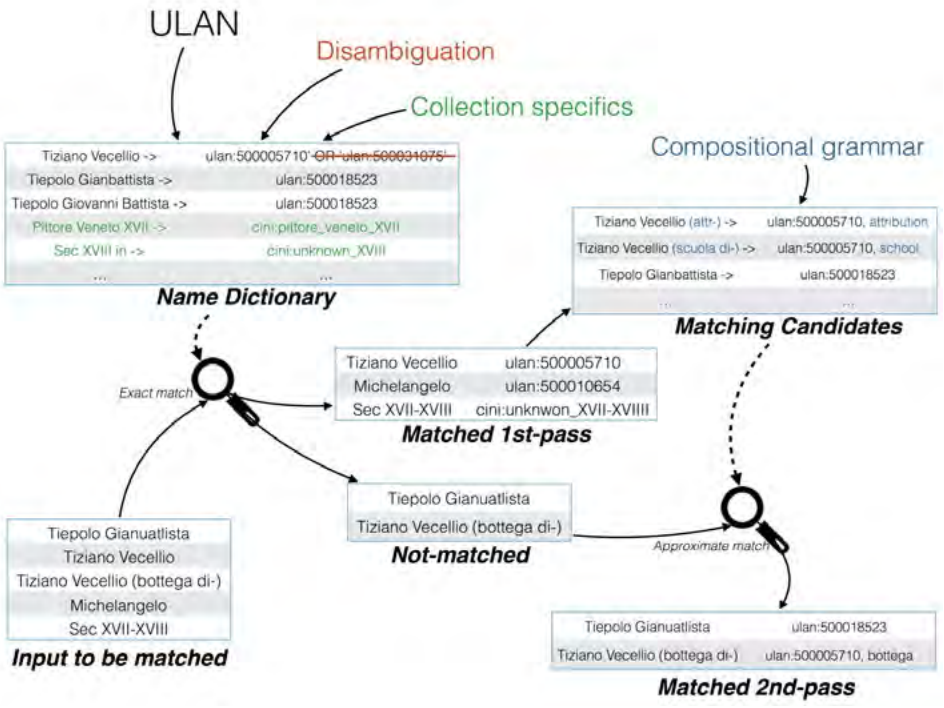


Figure 3 Alignment process. The parts in color correspond to collection-specific knowledge.

There are three challenges that needed to be tackled during this alignment process :

- *Names variation* : one major issue that arises is that a given artist may be called by different names, depending on regional variations and pseudonyms. Many variations are recorded in ULAN (i.e "*Tiepolo Gianbattista*" and "*Tiepolo Giovanni Battista*" both corresponding to the same artist), although some have to be added to the name dictionary. Furthermore, the naming conventions for elements whose dating or provenance is known but not authorship, which may be specific to a collection, can be added to the dictionary.
- *Implicit knowledge* : one related challenge is linked with the pragmatics of the annotation process. Understanding that if one archivist writes "*Leonardo*" on a file, he or she is referring to *Leonardo da Vinci* implies modeling a series of implicit assumptions which are changing depending on the evolution of local cataloging practices and that of the art historical field itself. In our case, we tackle this by disambiguating unclear names. For instance "*Tiziano Vecellio*" could technically refer to the well-known "*Tiziano*", or his relative "*Tizianello*", but the first is much more prominent than the second.
- *Compositional structure* : the last challenge is linked with the practice of archivists to describe particular unknown authors using specific syntactic process like ("*Tiziano (bottega di-)*", "*Tintoretto (Maestro di)*" or "*Michelangelo (copia da-)*"), referring to workshop productions or copies. Understanding and modeling this "grammar" permits to generate, in a compositional manner, potential matching strings to be considered when looking for possible alignments. Such strings do not only give a link to an artist but also qualify relationships (how strongly an artist was involved in the creation process of a painting, whether the piece is an original or a copy, etc.).

Results

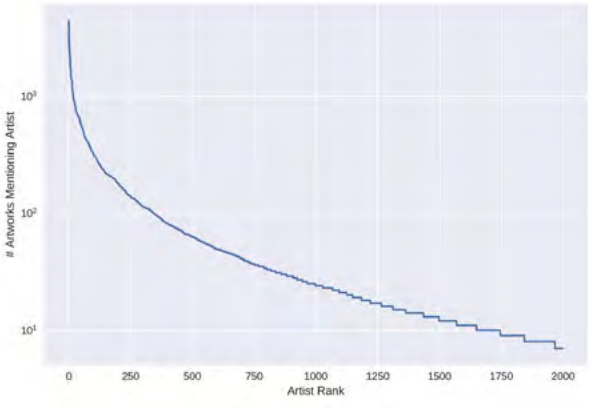
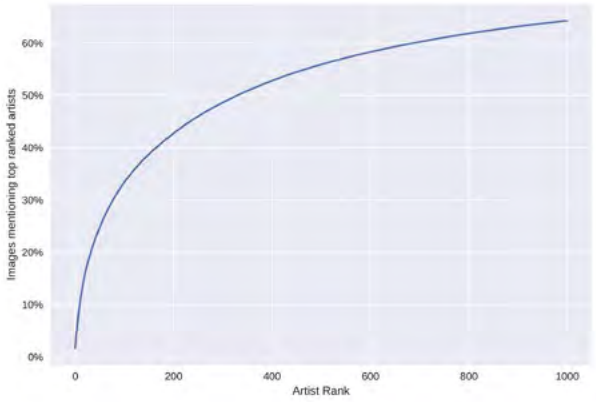


Figure 4 : Distribution of number of artworks assigned for each artist.



Proportion of images assigned with respect to the most common artists. The 200 most represented artists represent 43% of the collection.

Of the 330,078 scans composing the corpus of study, 14.6% had an empty author field, mostly because the photographs represented architecture or aerial city views. Out of the remaining 85.4% with an authorship field, 73.8% were automatically matched to an author (61.6% after the first pass), with an additional 1.4% representing ambiguous situations which could be resolved. This accounts for 208'510 elements automatically matched. At the end of pre-processing, the potential author names can be divided into three categories :

- (A) Author names which have been matched with a reference record of another database
- (B) Author names which may have been matched if the algorithm were to be improved (e.g. in terms of author name variation or possible compositional structure)
- (C) Authors undocumented in standard databases of artists.

Figure 5 shows the global matching results for category A. The geographical composition of aligned authors is dominated by Venetian artists (Tiepolo, Tintoretto, Palladio, Tiziano, Veronese, etc.) showing the rationale behind the creation of the collection. In terms of chronology, the collection is focused on the sixteenth century, as shown by the distribution of year of death of the aligned artists. This is in line with the period referred to as the "Venetian Golden Age". Figure 4 shows the very uneven representation of artists, with only 346 having more than 100 images, representing more than 50% of the whole collection.

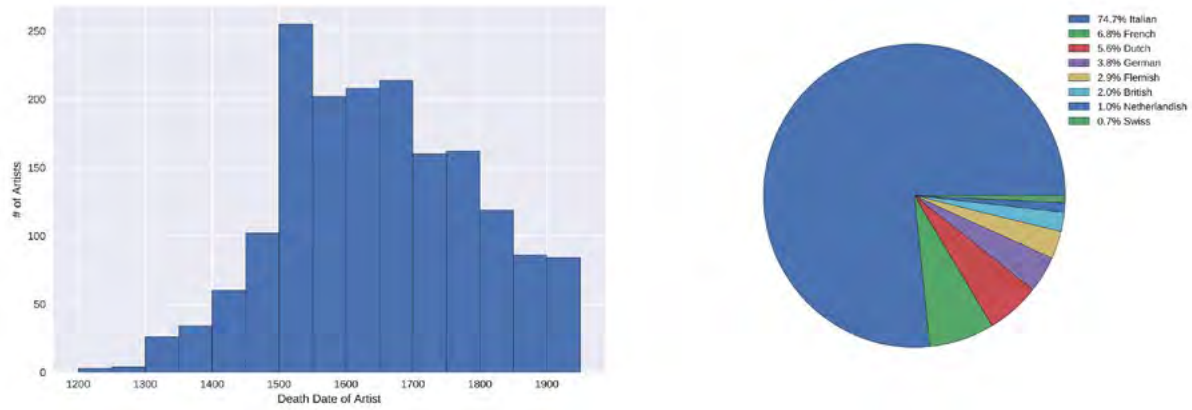


Figure 5 Spatial (right) and temporal (left) distribution of the 1'746 artists with at least 10 images assigned.

Category B is predominant in the elements that were not matched. Apart from OCR errors, the most typical unmatched string corresponds to collective works in which several authors are named. For instance, the string "*Bassano Jacopo e Francesco*" (his son) corresponds to 134 records. Adding additional parsing capabilities to the system could enable the resolution of such cases in the future.

Names in category C, which were not matched with ULAN, are in fact not a product of misalignment but represent new discoveries in the collection. In the present study, a number of artists who do not feature in ULAN were uncovered in the Cini archive. These include, Augusto Caratti, a minor artist from nineteenth-century Padua, who is represented by 65 images in the Cini collection, and Natale Melchiori an early eighteenth-century painter from Castelfranco, Veneto, represented by 39 images. Another artist who does not feature in the ULAN database but nevertheless has a significant presence in the Cini archive with 106 drawing, is Antonio Contestabile, an eighteenth-century draftsman from Piacenza.

Conclusion

These early results show the potential of the systematic processing of a large number of art historical records, leading to the mapping of unknown collections, and to new discoveries. It also highlights for the first time the challenges inherent in the process. Such challenges, it is important to note, are not purely technical but rather linked with the complexity of modeling local archiving traditions and the historical practices of art history.

References

Pharos. *PHAROS: The International Consortium of Photo Archives.* http://pharosartresearch.org/

Ronneberger, O. and Fischer, P. and Brox, T. (2015) *U-Net: Convolutional Networks for Biomedical Image Segmentation.*

D. A. Brown, D. A. and Ferino-Pagden, S. and Anderson, J. and Berrie, B. H (2006) *Bellini, Giorgione, Titian, and the Renaissance of Venetian painting*

Ares Oliveira, S.* and Seguin, B.* and Kaplan, F. (2018) *dhSegment: A generic deep-learning approach for document segmentation.*

Seguin, B. (2018) *New Techniques for the Digitization of Art Historical Photographic Archives—the Case of the Cini Foundation in Venice,* Proceedings of Archiving 2018.

A Design Process Model for Inquiry-driven, Collaboration-first Scholarly Communications

Sara B. Sikes
sara.sikes@uconn.edu
University of Connecticut, United States of America

Even as the scholarly communications field pursues the opportunities presented by digital technology, its routine operations remain anchored in print-centric regimens. For those working to evolve scholarly communications in the Internet age, particularly as it bears upon long-form scholarship, there is compelling need to productively disrupt and reconfigure the workflows and work cultures that have naturalized around the production of printed products. It is precisely this complex, systemic issue that Greenhouse Studios | Scholarly Communications Design at the University of Connecticut (UConn) addresses with its design-based, collaboration-first model of scholarly production.

With funding from the Andrew W. Mellon Foundation, the Digital Media & Design Department at UConn, the University Library and UConn Humanities Institute launched Greenhouse Studios in 2017. As a transdisciplinary collective, Greenhouse Studios employs design-thinking methodology to long-form digital scholarship. With its first two cohorts of collaborative projects, the Studios implemented an inquiry-driven approach that addresses the

divided workflows and counter-productive labor arrangements that have complicated scholarly communications in the digital age.

While the introduction of digital tools across the "information chain" model of scholarly communications has altered activities from research and writing through to preservation and reading, it has not reconfigured the larger workflow in which the various actors remain interlinked but largely independent save for key transactional, or "handoff," moments (CNI, 2016). Simply put, the "information chain" of scholarship begins with a knowledge creator, passes through to a publisher and culminates with accessibility secured by libraries and use by readers (Owen, 2002: 275-88). This transactional model has contributed to the persistence of an increasingly detrimental division of activities into those of the knowledge creation, or "domain," side and those of the production, or "build," side (Sosin, 2016).

By disrupting and reconfiguring divided workflows that have naturalized around the production of printed products,Greenhouse Studios brings together project teams on the "domain" side versus the "build" side. Each year, a new theme or problematic frames the work of the project teams, and diverse groups of collaborators are brought together, including designers, developers, editors, faculty and librarians. Starting with a problematic or issue rather than a faculty interest flattens counterproductive hierarchies and bringing in partners early in the process lends itself to the collaboration-first approach of the creation and expression of knowledge. Digital formats for the projects are not presupposed, as the format—digital or analog—that best represents the long-form scholarly work is taken under consideration. The first cohort of Greenhouse Studios teams developed projects in diverse formats including a documentary film, a virtual reality environment and an electronic decision-making novel.

Guiding the work of the teams, the Greenhouse Studios design process model provides a workflow for each project through five major sprints or phases. The design process modelwas developed through a series of exercises to elicit individual mental models of the scholarly design process from the perspective of a project manager, scholar, designer, repository manager, digital scholarship librarian, developmental editor and MFA student/research assistant. Comparisons of the mental models highlighted similar project phases for each participant, although the points of intersection were often differently identified. In looking at these points of overlap, neutral descriptors for shared activities were adopted, both for mutual intelligibility and to eliminate the kinds of value judgments that domain-specific terms may inscribe.

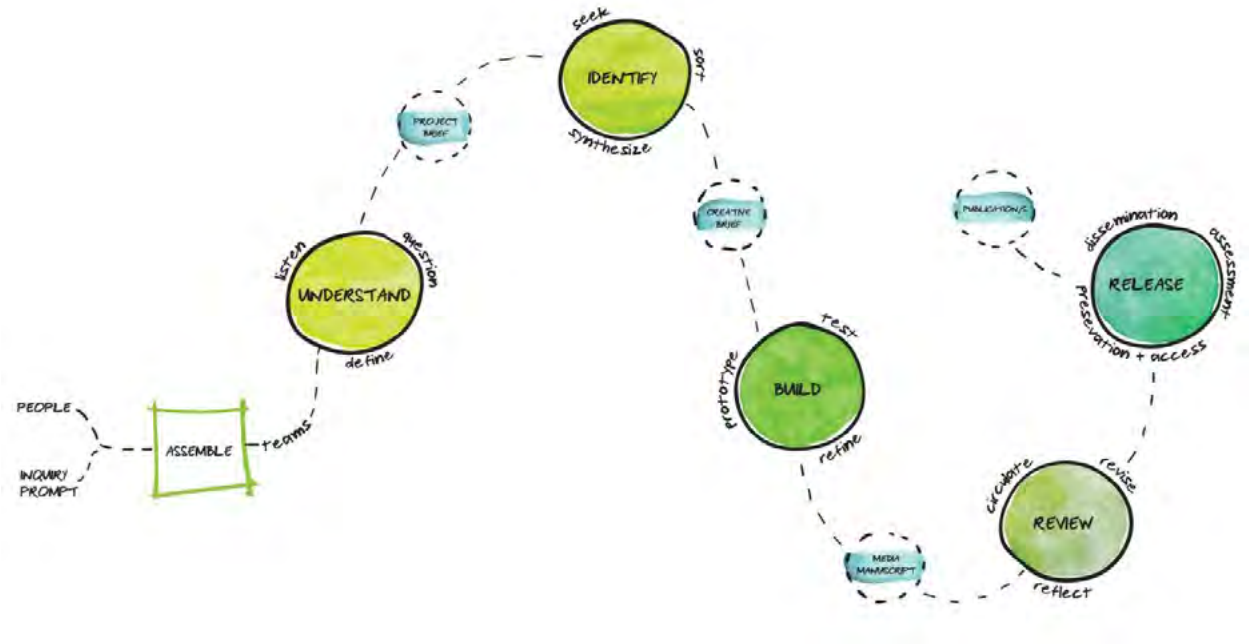


Figure 1. Greenhouse Studios Design Process Model

Being mindful of the Greenhouse Studios goals for workflow and work culture, the design process model adopts elements from the long tradition of design thinking as applicable across diverse fields. Design thinking as taught, practiced and disseminated by its most well-known and long-standing academic and corporate proponents, Stanford University's Hasso Plattner Institute of Design (aka the d.school) and the design firm IDEO, traces its roots to the 1960s' merger of a Stanford program that joined arts and mechanical engineering (Miller, 2015). Today, it has extended to endeavors as far afield as finance, films, museum exhibition, journalistic communications, education, and critical making in the digital humanities. Across various incarnations, design thinking processes

typically involve a series of iterative discovery and development cycles, each characterized by a subset of activities designed to facilitate that cycle's goal.

Work through the Greenhouse Studios design process model begins with an inquiry or prompt and brings together team members in response to a central problematic. During the catalyst phase of **Assemble,** team members gather, meet fellow participants and review the guidelines for project teams. The relevant human talents and other resources are defined during the first full sprint, or the **Understand** phase, which produces a *project brief* framing the project's aims and audiences. During the subsequent **Identify** phase, relevant sources of knowledge and inspiration are researched and synthesized. The resulting *creative brief* outlines the media formats of the project, as well as the formal peer review and assessment plans for the work. Iterative prototyping and refining of a project takes place during the following **Build** phase, producing a *media manuscript*, which could be a website, book manuscript, documentary film, exhibition or other format. During the **Review** phase, the project is revised, edited and submitted for peer review. The final phase is the **Release** or launch of the project, as well as the longer-term work of dissemination, assessment and preservation. Adjacent to this phase, there may also be *other publications* produced by individual project team members.

This design process model guides each of the Greenhouse Studios inquiry-driven, collaboration-first projects. The implementation of the process began before the launch of the first cohort of projects,and the model has undergone subsequent iterations across several development cycles. The team participants and an inquiry prompt act as catalysts for the workflow, which places collaboration at the center of the process, rather than an individual scholar's research goals. The emphasis on the "collaboration-first" nature of the process allows participants to collectively imagine scholarly projects from the outset and serves as a corrective to divided workflows, even digital-centric ones, where collaborators are only brought on board for the final implementation of projects.

References

Coalition for Networked Information. (2016). Supporting the Digital Humanities: Report of a CNI Executive Roundtable, 3. https://www.cni.org/wp-content/uploads/2016/05/CNI-SupportDH-exec-rndtbl.report.F14.pdf.

Miller, P. N. (2015). Is "Design Thinking" the New Liberal Arts? *The Chronicle of Higher Education*, *The Chronicle Review*, 61 (29). http://chronicle.com/article/Is-Design-Thinking-the-New/228779/.

Owen, J. M. (2002). The New Dissemination of Knowledge: Digital Libraries and Institutional Roles in Scholarly Publishing. *Journal of Economic Methodology*, 9 (3): 275-88.

Sosin, J. (June 29, 2016). Associate Professor, Department of Classical Studies and Director of the Duke Collaboratory for Classics Computing, Duke University. Interview by Tom Scheinfeldt, Clarissa Ceglio, and Sara Sikes.

Métodos digitales para el estudio de la fotografía compartida. Una aproximación distante a tres ciudades iberoamericanas en Instagram

Gabriela Elisa Sued
gabriela.sued@gmail.com
Tecnológico de Monterrey Ciudad de México, Mexico

Introducción

Debido a la generalización del uso de cámaras fotográficas en teléfonos móviles y a la posibilidad de su inmediata publicación en redes sociales, la fotografía compartida ha devenido una parte fundamental de la comunicación on-line. Sin embargo, ha sido menos estudiada que los objetos textuales publicados en algunas plataformas sociales, por ejemplo en Twitter. (Highfield y Leaver, 2016). Recientemente la investigación académica comienza a analizar el contenido visual generado por los usuarios. Estas temáticas y nuevos modos de producción de información son crecientemente abordados en análisis científicos y críticos con metodologías innovadoras, como la analítica cultural (Manovich, 2009) los métodos digitales (Rogers, 2009) y la visualización de información (Niederer y Taudin Chabot, 2015).

Nos proponemos investigar empíricamente en el modo en que las ciudades iberoamericanas son representadas en Instagram. A tal fin hemos recolectado un conjunto de fotografías etiquetadas como #buenosaires, #cdmx (México) y #madrid, publicadas en la mencionada plataforma durante la primera semana de octubre de 2016. El estudio profundiza en las formas en que las tres ciudades son representadas desde el punto de vista de los usuarios de la plataforma, describe las especificidades de cada una, e identifica el uso social de las etiquetas o hashtags de gran porte, donde se publican miles de fotos diariamente.

Empleamos una aproximación metodológica distante (Moretti, 2007, 2015) que considera tanto la dimensión digital de esas interacciones como una interpretación crítica que pueda identificar el papel que los objetos digitales juegan en la producción de la cultura contemporánea. Emplea una metodología de investigación mixta que combina el análisis cuantitativo, el empleo de software de procesamiento de datos textuales y numéricos y la interpretación de resultados desde una perspectiva sociocultural.

La aproximación distante como método digital

Interrogamos el corpus a través de un conjunto de técnicas que denominamos genéricamente aproximación distante. En el campo de los estudios literarios cuantitativos Moretti (2007, 2015) distingue dos tipos de lecturas: la distante y la cercana. La primera es de tipo exploratoria, opera con la masa, la generalidad y los hechos comunes. El crítico identifica en esa masa patrones de regularidad y frecuencia, grandes agrupamientos o clústers, ciclos temporales, y estructuras reticulares (Manovich, 2009). Por otro lado, la aproximación cercana usa técnicas procedentes de diferentes corrientes interpretativas a fines de atribuir un sentido a la producción analizada, y establecer relaciones con la cultura en la que estas manifestaciones tienen lugar. Las teorías interpretativas otorgan a las producciones simbólicas y de registro un lugar fundamental para la comprensión de las culturas.

La aproximación distante propone interrogar los datos y metadatos desde diferentes técnicas basadas en software. Aplicamos el análisis de contenido (Rose, 2016) y la analítica visual (Thomas y Cook, 2005 Manovich, 2011b) al corpus fotográfico con el fin de identificar temáticas recurrentes y patrones estéticos. Empleamos la analítica textual (Moreno y Redondo, 2016) para identificar las palabras frecuentes en las descripciones o Captions. El análisis de redes (Venturini, Jacomy y Carvalho, 2015) nos fue útil para establecer conexiones y clústers o agrupamientos entre etiquetas co-ocurrentes. Finalmente estudiamos las reacciones en relación al consumo activo de las fotografías una vez publicadas en la plataforma, también denominado *engagement* (Turner, 2014 y Rogers, 2016).

Hallazgos empíricos y discusión metodológica

En el corpus estudiado se evidencia la recurrencia de elementos temáticos, estéticos y textuales. Las palabras frecuentes evidencian una práctica homogénea, cuyo significado se fija en pocas redes semánticas asociadas a la fotografía, el viaje, la arquitectura, el consumo. Los patrones textuales demuestran diferentes maneras de representar y experimentar las ciudades. En *#madrid* las fotografías de personas en el ámbito urbano cobran mayor importancia que en otras etiquetas, en consecuencia merecen ser estudiadas en profundidad. El uso publicitario y la promoción del consumo también son recurrentes. En #buenosaires lo son la experiencia de práctica fotográfica y los estilos de vida, así como en #cdmx el patrón dominante es el de la estilización del entorno urbano. Las recurrencias pueden interpretarse en varias direcciones: en relación a las características propias de los objetos representados, debido a la homogeneidad de imaginarios sociales, o también como emergencia de nuevos géneros narrativos asociados a la fotografía compartida. La co-ocurrencia de etiquetas esboza redes de intereses, temáticas y comunidades acordes a la definición de la fotografía compartida como elemento de exhibición y mensaje comunicativo de intercambio efímero El análisis de de reacciones evidencia las diferencias culturales y comunicativas entre el acto de fotografiar y el de mirar una fotografía. En #cdmx se destaca la alta cantidad de fotografías de amaneceres y atardeceres, pero es la temática urbana y arquitectónica la que recibe mayores reacciones.

Las tres ciudades se distinguen por el uso publicitario de la imagen generado por los propios usuarios. Los que reciben mayores reacciones por otro lado también hacen un uso económico del hashtag pues su fin es el de la autopromoción. Además desarrollan estrategias para lograr la visibilidad de sus fotos publicando en múltiples hashtags y deslocalizando los territorios representados a partir del etiquetado. Al menos una parte de ellos concibe su práctica como parte de una comunidad, evidenciada por la aparición recurrentes de hashtags asociados a la publicación en Instagram: "instagrammers", "mextagram" y otras similares. Podemos suponer entonces que Instagram instaura una suerte de "economía de la visibilidad", donde las reacciones son la moneda con la que se paga la creatividad vernácula.

La alta homogeneidad y recurrencia sugieren la emergencia de codificaciones semióticas propias de la plataforma y demuestra la existencia de una gramática de acción (Agre, 1994), donde las acciones aisladas adquieren sentido cuando se las analiza colectivamente. Estos elementos pueden indicar el surgimiento de la fotografía compartida no sólo como práctica cultural sino como género discursivo con sus propias temáticas, estéticas y prácticas. Las producciones recurrentes de los usuarios pueden considerarse codificaciones que se siguen para ser parte de diversas comunidades de práctica materializadas en el uso de hashtags.

A partir del estudio realizado podemos observar que la aproximación distante resulta efectiva para abrir la caja negra de los medios sociales y mapear los principales temas y patrones estéticos de la fotografía compartida, aunque este abordaje debe en un futuro someterse a mayor investigación empírica y profundización epistemológica sobre varios de sus componentes. Entre los puntos que requieren mayor investigación se encuentran las técnicas de investigación de datos, la comprensión del modo en que funciona el software que se emplea en el procesamiento de los datos y metadatos, y la determinación de la importancia del volumen de datos que se producen en las redes para la investigación social.

El estudio de las mediaciones en Latinoamérica siempre ha relacionado la práctica cultural con las estructuras sociales, identificando en los consumos culturales o bien prácticas de subordinación, o bien de resistencia (Martín Barbero, 2001). En este trabajo la fotografía compartida sobre ciudades emerge como una práctica donde la búsqueda de visibilidad y el uso de autopromoción resultan evidentes. Será entonces necesario plantear su función social en el contexto de una economía global de intercambios simbólicos, línea que deberá ser profundizada tanto teórica como empíricamente.

References

Highfield, T. y Leaver, T. (2016). Instagrammatics and digital methods: Studying visual social media, from selfies and GIFs to memes and emoji. *Communication Research and Practice*. 2: 47-62.

Manovich, L. (2009). Cultural Analytics: Visualizing Patterns in the era of more media. Recuperado el 14 de mayo de 2017, a partir de http://www.manovich.net

Manovich, L. (2011b). What is visualization? *Visual Studies*, 26(1): 36–49.

Martín Barbero, Jesús. (2001). *De los medios a las mediaciones: comunicación, cultura y hegemonía*. Naucalpan, Mexico: G. Gili.

Moreno, A., Redondo, T. (2016). Text Analytics: the convergence of Big Data and Artificial Intelligence. *International Journal of Interactive Multimedia and Artificial Intelligence*, 3 (Special Issue on Big Data and AI): 57–64.

Moretti, F. (2007). *La literatura vista desde lejos*. Barcelona: Marbot.

Moretti, F. (2015). *Distant reading*. London: Verso.

Niederer, S., y Taudin Chabot, R. (2015). Deconstructing the cloud: Responses to Big Data phenomena from social sciences, humanities and the arts. *Big Data and Society*, 2(2).

Rogers, R. (2009). The End of Virtual. Digital Methods. Recuperado a partir de http://www.govcom.org/rogers_oratie.pdf

Rose, G. (2016). *Visual methodologies: an introduction to researching with visual materials*. London: Sage

Thomas K., y Cook, K., eds. (2005). Illuminating the Path: Research and Development Agenda for Visual Analytics. IEEE-Press, recuperado de http://vis.pnnl.gov/pdf/RD_Agenda_VisualAnalytics.pdf

Turner, P. (2014). The figure and ground of engagement. *AI and Society*, 29(1): 33–43.

Venturini, T., Jacomy, M., y Carvalho P. D. (2015). Visual Network Analysis. Recuperado a partir de http://www.tommasoventurini.it/wp/wp-content/uploads/2014/08/Venturini-Jacomy_Visual-Network-Analysis_WorkingPaper.pdf

Revitalizing Wikipedia/DBpedia Open Data by Gamification -SPARQL and API Experiment for Edutainment in Digital Humanities

Go Sugimoto
go.sugimoto@oeaw.ac.at
Austrian Academy of Sciences, Austria

Introduction

The Linked Open Data (LOD) community is growing In Digital Humanities (DH). Important datasets are being published in RDF. SPARQL endpoints have been progressively created in many cultural heritage organizations (Edelstein et al., 2013). However, the use of those datasets in real research is still not prevalent. Although there are several DH projects (Boer, V. de et al., 2016), SPARQL query exploitation is often limited within small technology-savvy communities (Lincoln, 2017). The situation is better for less-complicated Application Programming Interfaces (APIs) (XML and JSON). However, Sugimoto (2017b) suggests the needs of API standardization and ease of data reuse for ordinary users. In a broader context, the underuse of data, tools, and infrastructures seems to be a common phenomenon in DH. For example, the use of the Virtual Language Observatory in CLARIN is rather low (Sugimoto, 2017a). In case of the limited use of SPARQL endpoints, there could be different reasons for this:

- Lack of awareness of existence
- Lack of skills to use SPARQL
- Opened data is too narrow in scope
- Lack of computing performance to be usable
- Interdisciplinary research is not widely exercised

It is a pity that the benefit of Open Data is only partially spread, although data is available. To this end, the author has experimented with Wikipedia/DBpedia to explore the potential use of and/or the revitalization of Open Data in and outside research community.

Revitalization of Wikipedia/DBpedia by gamification

The choice of Wikipedia/DBpedia is rationalized by taking into account the above-mentioned issues. The broad scope of their datasets would solve the problem of datasets in DH being too specific to be used by third party researchers (or the researchers do not know how to use data and/or what to do with them (Edmond and Garnett, 2014; Orgel et al., 2015). In addition, interdisciplinary research could be more easily adopted, using a more comprehensive yet relatively detailed level of knowledge.

The keyword of the approach of this project is **gamification**. In order to showcase a social benefit of Open Data and DH, gamification would be a catalyst to connect the scholars and the increasingly greedy public consumers. Kelly and Bowan (2014) stated that limited attention has been paid to digital games until recently, although this is changing rapidly (see Hacker, 2015). Although there are a few projects such as Cross Cult which uses elaborate semantic technologies (Daif et al., 2017), this article contributes to this discourse from a web innovation perspective in a simplified DIY project environment.

The game developed for the project is quite simple. It is a quiz that requires users to guess the age of a randomly selected person by looking at a portrait of the person (born between 1700 and 2002) (Figure 1). Apparently, the age of a person in a particular image is provided neither

by Wikipedia, nor by DBpedia. It is, in fact, calculated programmatically by comparing the birthdate and the date of image. The random selection of data is sometimes costly for data processing, but it is the key to developing a game application. The application is intended for fun, thus, includes all types of contemporary persons such as politicians, sport athletes, musicians, actors, and businesspersons. In addition, the inclusion of historical figures is very important in DH in that the user would learn history.

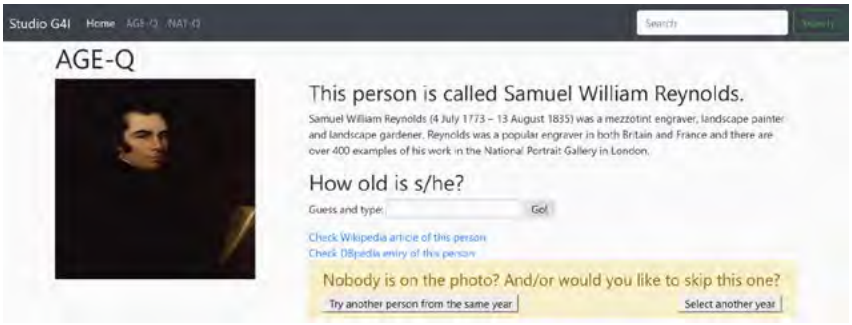


Figure 1 Quiz to guess the age of a person found in a Wikipedia article

When the user cannot guess the age, there is a help function. A hint section is equipped with a face detection API of IBM Watson, suggesting the estimate age and gender of the person in the image by machine learning. Finally, this game is extended into another quiz to guess the nationality of a person. Indeed, any interesting data of Wikipedia/DBpedia can be used for gamification, and the method is easily adoptable.

Potential for Citizen Science

As a reflection of critics of Linked Data quality, Daif et al. (2017) reckon that human supervision is needed to manage the data. In our case, the application is sometimes not able to calculate the age of a person, due to several reasons of metadata quality. For instance, data may be not numeric ("16th century") (Figure 2), malformed (not ISO compliant: "05/11/88"), confusing (the creation date of digital image is used instead of that of analogue image), inaccurate, wrong, or missing, resulting in an error message. This is normally regarded as an optimization problem of the code. However, it is possible to take advantage of this error. When it occurs, it is a sign of data quality problem. Therefore, users are persuaded to follow the provided links to Wikipedia/DBpedia and able to double-check the original data (Figure 3). This scenario creates a dual possibility. In other words, the application can be used as:

- A curation tool of Wikipedia/DBpedia for existing active editors of Wikipedia.
- A tool to transform normal users into new curators of Wikipedia

Although this scenario has not happened due to the project setting, if the users are able to correct data, the impact for data curation could be considerable. Not only is it to the benefit of correcting and/or adding data in Wikipedia, but DBpedia will also be improved, leading to the higher quality of datasets of this LOD magnet, affecting hundreds of applications worldwide. In this way, this application opens up the potential to **crowdsource the curation** of Wikipedia/DBpedia. The success of the crowd data curation has been proven in DH (see Brinkerink, (2010) and NYPL Labs).

Figure 2 Wikimedia metadata displaying "16th century"

Figure 3 The game persuades users to improve Wikipedia

Conclusion

In conclusion, this article demonstrates an experimental case study of mixing gamification (entertainment) with data-driven research (education) and the possibility for data curation (crowdsourcing), showcasing cutting-edge technologies such as SPARQL and Deep Learning API, with the help of Open Data in the framework of DH. It also displays a potential for a new digital research ecosystem among humanities research and digital technologies, connecting various stakeholders including humanities researchers and the public.

References

Boer, V. de, Penuela, A. M. and Ockeloen, C. J. (2016). Linked Data for Digital History: Lessons Learned from Three Case Studies. *Anejos de La Revista de Historiografía*(4): pp139–62.

Brinkerink, M. (2010). Waisda? Video Labeling Game: Evaluation Report. *Images for the Future – Research Blog* http://research.imagesforthefuture.org/index.php/waisda-video-labeling-game-evaluation-report/index.html (accessed 12 April 2018).

Daif, A., Dahroug, A., López-Nores, M., Gil-Solla, A., Ramos-Cabrer, M., Pazos-Arias, J. J. and Blanco-Fernández, Y. (2017). Developing Quiz Games Linked to Networks of Semantic Connections Among Cultural Venues. *Metadata and Semantic Research*. (Communications in Computer and Information Science). Springer, Cham, pp. 239–46 doi:10.1007/978-3-319-70863-8_23.

Edelstein, J., Galla, L., Li-Madeo, C., Marden, J., Rhonemus, A. and Whysel, N. (2013). Linked Open Data for Cultural Heritage: Evolution of an Information Technology. http://www.whysel.com/papers/LIS670-Linked-Open-Data-for-Cultural-Heritage.pdf (accessed 24 April 2018).

Edmond, J. and Garnett, V. (2014). Building an API is not enough! Investigating Reuse of Cultural Heritage Data. *LSE Impact Blog* http://blogs.lse.ac.uk/impactofsocialsciences/2014/09/08/investigating-reuse-of-cultural-heritage-data-europeana/ (accessed 27 February 2018).

Hacker, P. (2015). The Games Art Historians Play: Online Game-based Learning in Art History and Museum Contexts. *The Chronicle of Higher Education Blogs: ProfHacker* https://www.chronicle.com/blogs/profhacker/the-games-art-historians-play-online-game-based-learning-in-art-history-and-museum-contexts/61263 (accessed 12 April 2018).

Kelly, L. and Bowan, A. (2014). Gamifying the museum: Educational games for learning | MWA2014: Museums and the Web Asia 2014 https://mwa2014.museumsandtheweb.com/paper/gamifying-the-museum-educational-games-for-learning/ (accessed 12 April 2018).

Lincoln, M. (2017). Using SPARQL to access Linked Open Data. *Programming Historian* https://programminghistorian.org/lessons/graph-databases-and-SPARQL (accessed 12 April 2018).

NYPL Labs Whats on the menu? http://menus.nypl.org/about (accessed 12 April 2018).

Orgel, T., Höffernig, M., Bailer, W. and Russegger, S. (2015). A metadata model and mapping approach for facilitating access to heterogeneous cultural heritage assets. *International Journal on Digital Libraries*, 15(2–4): pp189–207 doi:10.1007/s00799-015-0138-2.

Sugimoto, G. (2017a). Number game -Experience of a European research infrastructure (CLARIN) for the analysis of web traffic. *CLARIN Annual Conference 2016*. Aix-en-Provence, France: CLARIN ERIC and Laboratoire Parole et Langage and Laboratoire des Sciences de l'Information et des Systèmes (LSIS) and Aix-Marseille Université and Centre National de la Recherche Scientifique (CNRS) https://hal.archives-ouvertes.fr/hal-01539048 (accessed 17 November 2017).

Sugimoto, G. (2017b). Battle Without FAIR and Easy Data in Digital Humanities. *Metadata and Seman-*

tic Research. (Communications in Computer and Information Science). Springer, Cham, pp. 315–26 doi:10.1007/978-3-319-70863-8_30.

The Purpose of Education: A Large-Scale Text Analysis of University Mission Statements

Danica Savonick
danicasavonick@gmail.com
The Graduate Center of the City University of New York, United States of America

Lisa Tagliaferri
ltagliaferri@gradcenter.cuny.edu
Fordham University, DigitalOcean

What is the purpose of higher education? In the United States, this question dates back to at least the nineteenth century with the passage of the Morrill Acts of 1862 and 1890, and has taken on new urgency in an era of manufactured austerity and neoliberal crisis. In particular, scholars of critical university studies such as Christopher Newfield, Fred Moten and Stefano Harney, Sara Ahmed, Craig Steven Wilder and Roderick Ferguson critique the ways higher education often reproduces the very conditions of inequality it claims to challenge. Often, these compelling analyses are based on the investigation of conditions at a few representative universities, but through leveraging digital methodologies we can gain a wider perspective that enables a more comprehensive analysis of what universities put forward as their purpose.

Our research advances these conversations through large-scale textual analyses of two data sets: university mission statements included in the U.S. Department of Education's Database of Accredited Postsecondary Institutions and Programs and recent demand statements put forth by activist students. Mission statements offer a public-facing proclamation that bridge universities to larger communities and educational contexts. Often, they present idealized claims that reflect the university's marketed brand. We use "university" broadly; our data set includes community colleges, public universities, private universities, research institutions, teaching-focused institutions, for-profit and nonprofit schools, and our analysis highlights the variation in their commitments to education. The second data set is a collection of student demands compiled by WeTheProtestors and the Black Liberation Collective, two social justice groups that are working to address institutional inequality across U.S. universities. In many cases, these demands are written to address the institutions of the official university mission statements we are working with, and range from private institutions like Yale University and Ithaca College, to public universities such as Iowa State and UCLA. These demands challenge existing institutional language and require analysis in their own right.

With this research, we seek to answer two questions: 1) What do contemporary U.S. universities claim as their mission and vision? 2) How do these stated aims of education intersect or diverge with the demands of activist students calling for pedagogical, institutional, and social change? In analyzing this data, we draw from the insights of critical race, gender, and sexuality studies, which have long been sites of institutional critique. Coupling digital tools with a theoretical lens informed by activist pedagogy enables us to better apprehend the power structures and social dynamics at play in public-facing institutional documents and how those interface with the communities they are tasked with serving. By better understanding the professed commitments of academic institutions, we aim to contribute to the project of making education more just, equitable, and inclusive.

This work is carried out through the web scraping of data, topic modeling, and statistical analysis in Python. Once analyzed, the raw data and findings are also rendered as interactive web-based data visualizations in JavaScript to make the research more accessible to the public and available for refactoring. Initial statistical textual analysis and data visualization that we have conducted has revealed interesting trends among public universities in contrast to demands put forth by students. Mission statements from state universities emphasize a commitment to the objectives of research, knowledge, and professionalism, and the endeavors of providing and serving, and learning and teaching. However, student demand statements have a more expansive understanding of education that stresses inclusivity and community, while also voicing concerns about race, gender, workers, and resources. By comparing and contrasting across data sets, we examine what each type of institution and group is seeking to achieve, and work to determine whether universities are serving the needs of student populations. When universities are more concerned with vocational skills training rather than challenging power hierarchies, structural inequalities, and the distribution of resources along embodied axes of race and gender, there is a clear disconnect between what institutions are offering students and what students in turn demand. If universities aren't serving their students and communities, who are they serving?

Our data set and programming files will be made publicly available in a code repository so that others who investigate higher education can perform their own research. While our focus is grounded in the specific histories of higher education in the United States, we hope that sharing this research at an international conference will encourage others to perform similar analysis of institutional and popular discourse in their countries, thus allowing for a more vibrant understanding of how higher education functions in different contexts. By inviting

others to add data to our public repository from international institutions, we can begin to consider how globalization impacts learning institutions.

In an effort to advance intercultural scholarly exchange, a Spanish translation of this research will be available online.

Digital Humanities Integration and Management Challenges in Advanced Imaging Across Institutions and Technologies
Nondestructive Imaging of Egyptian Mummy Papyrus Cartonnage

Michael B. Toth
mbt.rbtoth@gmail.com
University College London, United Kingdom; R.B. Toth Associates, United States of America

Melissa Terras
m.terras@ed.ac.uk
University College London, United Kingdom; University of Edinborough, United Kingdom

Adam Gibson
adam.gibson@ucl.ac.uk
University College London, United Kingdom

Cerys Jones
cerys.jones.15@ucl.ac.uk
University College London, United Kingdom

This rapid development and testing project brought together international partners, scholars and collections in an exploratory, pilot effort from November 2015 to March 2017. The international, multidisciplinary team demonstrated that some nondestructive digital imaging techniques and technologies (Fig. 1) have potential to make texts visible in Egyptian Ptolemaic papyrus mummy mask cartonnages. A major challenge in working across the different technologies, disciplines and institutions was integrating data from diverse technical imaging systems and work processes, requiring new and proven digital humanities data management capabilities.

Before this project, other scholars destroyed the masks to access the papyri, denying future researcher access to the primary historical artefacts (Mazza, 2014). This project capitalized on digital humanities skills and data management techniques in assessing the integration of non-destructive digital imaging technologies to make texts visible in layers of papyrus in mummy cartonnages for open research and analysis. Intermediate goals, such as detecting the presence of text, also proved valuable in highlighting the destructive techniques used to study mummy masks and offering scientifically valid approaches for documenting the initial state of objects and their production for future research.

Figure 1. Multispectral Imaging of Mummy Mask at UCL Centre for Digital Humanities, one of the advanced imaging techniques researched during this project.

A global team pulled together expertise from science and the humanities, including: digital humanities, Egyptology and papyrology, medicine, dentistry, particle physics, imaging science, data and project management, and systems engineering. Team members rapidly implemented a phased and agile approach at multiple institutions to develop and apply increasingly complex imaging, processing and data integration techniques to penetrate the paint and papyrus layers in mummy cartonnage and host all data online (Fig. 2).

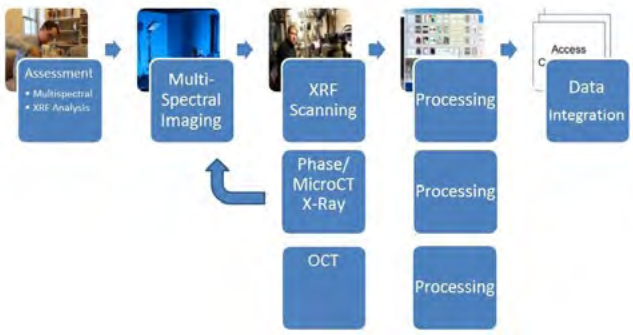


Figure 2. Mummy cartonnage advanced imaging process flow

Data Integration

Project data integration was dependent on common data and metadata standards for ease of image correlation and integration, as well as effective data and project management across disciplines, technologies and institutions. All

the different imaging modalities (Multispectral imaging, X-ray fluorescence, Optical Coherence Tomography, X-ray microCT, Terahertz and others) yielded very different data sets from each technology and institution. Integration of images from multiple imaging sources offered potential to apply the strengths of multiple imaging techniques for ease of visualization by scholars and curators. Integrating data from a variety of equipment required significant planning and collaboration across institutions and disciplines (Fig. 3). This required streamlined standardization processes and/or more time and resources to devote to this part of a program.

Imaging Technology	Imaging Institutions	Contributing Objects	Principal Investigators
Multispectral Imaging	UCL, Manchester, Duke, UC Berkeley, RB Toth Associates	UCL Petrie, UC Berkeley, Duke, UCL*	Melissa Terras, Adam Gibson, Bill Christens-Barry, Michael Toth
Spectral Domain Optical Coherence Tomography	Duke University	Duke, UCL*	Sina Farsiu, Adam Wax, Cynthia Toth
X-ray Fluorescence Scanning	SLAC SSRL	Berkeley†, UCL*	Uwe Bergmann
X-ray Micro-Computer Tomography	*University of California at Berkeley*	Berkeley†, UCL*	Dula Parkinson
X-ray Micro-Computer Tomography	*Queen Mary University of London*	Berkeley†, UCL*	David Mills, Graham Davis
Terahertz Imaging	*University of Western Australia*	UCL*	Vincent Wallace, Shuting Fan, Anthony Fitzgerald
XRF Analyzer	Bruker Scientific	Petrie, Berkeley, Duke, UCL*	Lee Drake, Adam Gibson
Fiber optic Reflectance Spectroscopy	Equipoise Imaging	Berkeley†, UCL*	Bill Christens-Barry

*UCL Phantom surrogate papyrus samples †UC Berkeley Tebtunis Center s.n. cartonnage fragment

Table 1. Participating institutions and imaging techniques used during cartonnage imaging.

The integration of data and work processes from a variety of scientific tools, disciplines and institutions required storage, dissemination, and searchable access to data from instruments that provide output in different formats, some of which were unique to the research methods and disciplines (Emery et. al., 2004). While common standards and processes across institutions were encouraged, this was difficult with data and standards from technologies as diverse as nuclear synchrotrons and optical cameras. In addition, many contributors to this project volunteered their time and equipment for imaging and basic processing, but had limited time to spare from their day-to-day responsibilities – ranging from medical personnel preventing blindness to particle physicists studying elemental changes in bone formation.

Data Storage and Management

The approximately 300 Gb of data products– including images, individual reports, captured and processed data sets, analytical data and metadata– are now freely available online at https://www.ucl.ac.uk/dh/projects/deepimaging/data. This data set comprises a core content set of digital images, analytical data and technical reports on the imaging and analysis of mummy mask cartonnage and modern surrogates from the multiple imaging institutions. UCLDH established this project website to host the project information and data at the same for scholars, scientists and the public (UCL, 2017).

Collecting, organizing and hosting data with appropriate metadata from multiple institutions and systems around the globe proved to be a complex problem. This included providing access to and sharing of timely, complete, and relevant data during the project. This was due to both different data collection standards and the wide range of output from proprietary equipment. A key strength of this program was all institutions agreed to make all data freely available under Creative Commons license. This allowed the free exchange of all data for digital processing, analysis and research.

The data structures of the Archimedes and Galen Palimpsests and the University of Pennsylvania's OPenn served as models, but had to be adapted to include the various types of data sets for each image and data collection modality. To support scientific data integration, the team also used the Library of Congress CLASS-D data

model. Some adjustments were needed to previous flat file access protocols to make the data product more accessible to users and future researchers. As an example, the large captured multispectral data sets were put in separate folders from the processed images, with the former available for follow-on digital processing and research, and the latter available for immediate visualization of our findings produced with current processing tools.

The need for quality assurance to verify and validate the data proved important. Once the data was integrated, some type of feedback mechanism was needed to validate and check the data against other data in collaboration with the collector as part of collaborative research. This highlighted the value of the data in conjunction with other data, with feedback on the efficiency and quality of the data and its reproducibility as initially structured and standardized. This significantly improved data sharing and preservation across the research team.

Conclusions

Effective data management, integration and technical support are critical enablers in any broad digital research program to ensure data availability for follow-on research, even those (like this one) with a limited budget. The ability of imaging equipment to produce a standard data output with relative ease of use by the operator and researcher is important to the visualization, storage of and access to the data. Standardized procedures and data output better allow independent imaging of the same object with multiple technologies, with subsequent integration of data to leverage the strengths of each technology and technique.

References

Emery D., France, F.G., Toth, M.B. (2009) "Management of Spectral Imaging Archives for Scientific Preservation Studies", Archiving 2009, IS&T, May 4-7, 137-141

Mazza R. (2014) "Another Indiana Jones? Josh McDowell, mummy cartonnage and biblical papyri." Faces and Voices. 2014. https://facesandvoices.wordpress.com/2014/05/05/another-indiana-jones-josh-mcdowell-mummy-cartonnage-and-biblical-papyri/. (Accessed 22 Mar 2018).

UCL (2017) "Deep Imaging Mummy Cases, The Data" https://www.ucl.ac.uk/dh/projects/deepimaging/data, (Accessed 27 Feb 2018).

Towards A Digital Dissolution: The Challenges Of Mapping Revolutionary Change In Pre-modern Europe

Charlotte Tupman
c.tupman2@exeter.ac.uk
University of Exeter, United Kingdom

James Clark
j.g.clark@exeter.ac.uk
University of Exeter, United Kingdom

Richard Holding
r.j.holding@exeter.ac.uk
University of Exeter, United Kingdom

This work-in-progress paper offers for critical review the current challenges of an ambitious project to create a digital framework for interpreting the dissolution of monasteries in Europe. The most dramatic episode of the European Reformation (*c*.1517- *c*.1648), the state suppression of monasteries, the dispersal of their populations, the re-distribution of their property and the re-deployment of their infrastructure, represented the largest and furthest-reaching re-ordering of society, economy and culture before the Industrial Revolution (Chadwick, 2001; Youings, 1971). The scale, scope, pace and reach of the process make it perhaps the most formidable of all pre-modern territories for the data-driven researcher, and have ensured that narrative histories founded on conventional methods of data analysis have consistently failed to provide perspectives of adequate breadth, depth, and accuracy.

In respect of research data, the medieval monastery presents both the best and worst of all prospects. A world in microcosm, possessed of its own demographic, economic, social, cultural and environmental imprint, in principle there are multiple layers to its source-base. It also runs deep through time, passing any polity, dynasty, and even place of settlement to reach back to the remote beginnings of Christian-occupied Europe. Yet for these same reasons, the sources of the medieval monastery are also uniquely unstable. The self-containment of the monastery was such that while the form and function of its documentary record might be comparable one to another, it is never quite the same. A durable - but not always enduring - presence in a world that was chronically disturbed, the record underwent repeated and extended interruptions. The monastery invited the manipulation of those in power, and its records are susceptible to conscious distortion. Even well-preserved monastic records can confound the researcher.

The closure and re-constitution of these centuries-old institutions brings these data complexities into collision with the records of the state, city, commune and of private individuals at a moment when these constituencies were in transition to a post-medieval world. The bare historical record may give the impression of the dissolution as an event bounded by the dates of specific acts of state, but in fact its course and consequences were a collective experience which unfolded over several generations. This means that for effective interpretation, datasets should be defined not by the intrinsic criteria of a particular monastery but rather by those that can be related to the contexts in which it was situated, relating to a

range of organisational and social networks and to physical place and space. This requires drawing from the monastery's records data that they were not originally created to document. For example, a contextualised approach to data on the monastery's population profile demands not only a raw numeral but also a measure of its geographical origin, social status and generational mix, each in relation to other neighbourhood constituencies. Because the dissolution was experienced over *la longue durée*, a wide chronological frame is needed: only by capturing data from 1450 to 1650 can the process of dissolution be traced in real time. Given the inherent characteristics of the records, this can be no conventional time-frame bringing a strict linear order to each dataset. With unequal interruptions in every category of record, instead the timeline must be drawn between irregular census points derived from individual documents.

Presently, we are applying these principles to a single case-study, the English Benedictine abbey of Battle, in the county of Sussex, dissolved in 1538. A substantial foundation, holding territory across seven counties of England and Wales, overseeing diverse agricultural, commercial and industrial interests and governing a network of satellite churches and communities, Battle presents sufficient scale and complexity to guide, and test, our emerging methodology (Evans, 1941-2; Searle, 1974). We are creating datasets which aim to measure (1) every aspect of the monastery's presence in and imprint upon its neighbourhood in the period before its dissolution and (2) the pattern and pace of change in that presence and imprint as the monastery was suppressed. We have defined data categories to evaluate its dynamic role in its neighbourhood, providing a series of key performance indicators at those census points which can be established. Although these do not always directly reflect the categories of the monastic records, generally it has been possible for data to be anchored by a specific documentary reference. However, it is sometimes necessary to make use of proxies. For example, because the family origins of monks are rarely documented, surnames are taken as an index of origin and social position, and because the precise site and proportions of monastic buildings are not consistently documented we have adopted a 'best-guess' principle, utilising historic mapping, field-, excavation and environment surveys, realising the benefits of the Archaeology Data Service (www.archaeologydataservice.ac.uk) and Heritage Gateway (www.heritagegateway.org.uk).

Our paper explores how we are addressing these complexities and challenges in our sources by combining a webapp built on an open source XML database (xQuery-based eXist-db, http://exist-db.org/) with highly customisable mapping using jQuery and GoogleMaps API to create a digital framework for analysing the process of dissolution across Europe. The framework allows researchers to interpret its events and sources at levels from regional to site-specific, utilising a comparative approach to reveal and visualise patterns that have been largely obscured within this often chaotic set of sources. We are building the webapp to be redeployed by others: its source code and documentation will be released freely on GitHub, enabling others to reuse it for their own research aims. The complexity of the process of dissolution might suggest that certainty in interpretation is an impossible goal, but our work to date suggests that far from this being a deterrent to digital approaches, it instead raises important questions about how we describe our datasets and how we can represent with honesty and clarity the uncertainty, inconsistency and gaps in our sources. These are questions with which every digital humanities scholar must grapple, and having set out the solutions we have identified so far, we are keen to invite discussion on how we might resolve or improve our approach for the benefit of current and future projects encountering similar issues.

References

Bradshaw, B., (1974). *The dissolution of the religious orders in Ireland under Henry VIII*. Cambridge University Press: Cambridge

Chadwick, O. (2001). *The early reformation on the continent*. Oxford University Press: Oxford, pp. 151-180.

Evans, A. (1941). Battle Abbey at the Dissolution. *Huntington Library Quarterly*, 4:4, pp. 393-442; 6:1 (1942), pp. 53-101

Knowles, D. & Hadcock, R.N., (1971). *Medieval Religious Houses: England and Wales*. 2nd edition, Routledge & Kegan Paul: London

Searle, E. (1974). *Lordship and community: Battle Abbey and its banlieu, 1066-1538*. Pontifical Institute of Medieval Studies: Toronto

Youings, J. (1971). *The dissolution of the monasteries.* Routledge & Kegan Paul

An Archaeology of Americana: Recovering the Hemispheric Origins of Sabin's Bibliotheca Americana to Contest the Database's (National) Limits

Mary Lindsay Van Tine
mva@upenn.edu
University of Pennsylvania, United States of America

This long paper will offer an archeology of the Gale database *Sabin Americana, 1500-1926*, tracing its origins through an earlier microfilming project to Joseph Sabin's *Bibliotheca Americana*, a monumental 29-volume "Dictionary of works related to America" begun in 1868 and completed in 1937. While Bonnie Mak, Ian Gadd, and others have explo-

red the bibliographic roots of much-used digital resources like the ESTC and EBBO, the category of Americana has a distinct bibliographic tradition whose digital implications have not been examined. While many contemporary databases derive from earlier bibliographic projects organized by language or nation, "Americana" was for Sabin and his contemporaries a transnational and multilingual category that understood "America" as the entire Western Hemisphere. Sabin and other nineteenth-century bibliographers of "Americana" ultimately produced works with an implied teleological view of a New World history that began with "discovery" and culminated in the emergence of the United States; nevertheless, they conceived of the early history of the hemisphere as a shared one, and their work emerged from an extended scholarly network that encompassed not only the Anglophone but also the Hispanophone world.

While Gale's database borrows Sabin's name and title, it is otherwise strikingly vague on the exact nature of its relationship to the original print bibliography. A close examination reveals that, although the structuring logic of the database is not dissimilar to Sabin's alphabetic schema and indexing, its selection principles and framing radically redefine America as the United States. Unlike the original bibliography, the vast majority of the works included are in English, with few in Spanish and even fewer in indigenous languages. The search interface offers "subject" options that uncritically sort the entire span of New World history into U.S.-based periodizations: colonial era, early republic, antebellum, postbellum, and so on. These silent omissions both assume and reinforce the conflation of "America" and "United States." When a database that claims to be "drawn from Joseph Sabin's famed bibliography" and, like it, to "cove[r] four centuries of life in North, Central, and South America, and the West Indies," returns overwhelmingly English-language sources from the "colonial era," or fails to produce a single hit for one of the most prominent Mexican historians of the nineteenth century while returning dozens for his U.S. counterpart, the effect is not just inaccurate but deeply pernicious. I will argue that this dramatic shift is not so much a function of digital remediation as of a changed scholarly infrastructure that cannot accommodate the capaciousness of "Americana" in its earlier bibliographic sense. The logic of nineteenth-century Bibliotheca Americanas, I suggest, invites us to think otherwise, offering an alternate bibliographic framework that might inform the development of non-proprietary digital systems for bibliographic control.

I will conclude by considering my own work towards this end in the context of the Digital Bibliotheca Americana project. It assembles a freely-available dataset that re-centers indigenous and Spanish-language texts, offers insight into the contours of Americana at scale, and enables computational analysis of the material and conceptual relocation of "Americana" to the United States over the course of the nineteenth century.

Tweets of a Native Son: James Baldwin, #BlackLivesMatter, and Networks of Textual Recirculation

Melanie Walsh
melanie.walsh@wustl.edu
Washington University in St. Louis, United States of America

In the wake of Michael Brown's murder in Ferguson, Missouri, on August 9, 2014, and the non-indictment of police officer Darren Wilson on November 25, 2014, backlashing protests and riots took to the streets of Ferguson and to other major American cities across the country. They also took to the Twittersphere. A national conversation about police brutality and the American criminal justice system exploded on Twitter during this time period, eventually elevating the hashtag #Ferguson, tweeted over 27 million times, to the most frequent in Twitter's ten-year history, and the hashtag #BlackLivesMatter, tweeted over 12 million times, to third place (Sichynsky, 2016). First coined by Alicia Garza, Patrisse Cullors, and Opal Tometi in July 2013, the hashtag #BlackLivesMatter became a banner for a national protest movement and an index for conversations about the systematic devaluing and elimination of black life. Over the last five years, literary scholars and historians have noted that, within this massive social media movement, the novelist, essayist, and civil rights literary icon James Baldwin seemed to be often and increasingly invoked (Maxwell, 2016). The perceived frequency of Baldwin-related tweets has been pointed to by many as evidence of the Harlem-born author's 21st-century resurrection and recent political resonance (Glaude Jr., 2016; Robinson, 2017). Because tweets can be digitally archived and made computationally tractable, they can be collected, measured, and analyzed at scale, and they can offer a picture of Baldwin's social media reception that goes beyond perception and anecdotal evidence. This talk will share work-in-progress from my project *Tweets of a Native Son* (http://www.tweetsofanativeson.com/), which brings large-scale social media data and computational methods to bear on Baldwin's 21st-century remediation, recirculation, and reimagination. This talk will discuss the methods and progress made in the project thus far, argue that social media analysis might usefully contribute to a growing body of computationally-assisted scholarship focused on readership, reception, and textual circulation, and finally gesture to how such an approach might change our understanding of how texts are shared between communities of people, namely through its emphasis on networks.

Methods, Analysis, Initial Findings

First I "hydrated," that is, retrieved the full JSON information for, an archive of over 32 million tweets that were sent between June 1, 2014 and May 31, 2015 and that

mentioned Ferguson, Black Lives Matter, and 20 other black individuals who were killed by the police during this time period, which was first purchased from Twitter and shared by Deen Freelon, Charlton McIlwain, and Meredith D. Clark (Freelon, McIlwain, Clark, 2016). I next searched for all the tweets that mentioned "James Baldwin" by his first and last names using the Python and command-line tool "twarc" and the command-line JSON processor "jq," which returned 7,326 tweets and retweets. By using twarc utilities, a k-means clustering algorithm, and manual tagging, I then identified the most retweeted tweets in the archive and the text that appeared most often across all tweets in the archive, which revealed that the most frequent appeal to Baldwin during this time period was through quotation and overwhelmingly through the quotation of Baldwin's 1960s-era essays, radio interviews, and television appearances.

By studying the text of the most retweeted and most frequently cited tweets, and by tracing tweeted Baldwin quotations back to their literary and historical origins, my project argues that Baldwin's appeal as a #BlackLivesMatter muse comes, at least in part, from the remediation of much of his non-fictional work into YouTube videos and free online essays; from his aphorisms with deep roots in African American written and oral traditions; and from his sympathetic proximity to but never full embrace of black radicalism. Another goal of *Tweets of a Native Son*, however, is to let others explore, hypothesize, and learn about Baldwin's #BlackLivesMatter-related social media reception through a series of interactive data visualizations on the project's website. These interactive visualizations are meant to provide a perspective on Baldwin's living legacy, a refracted vision of Baldwin's life and career through those who actively called upon him in a moment of political and emotional urgency, a means by which others can come to their own conclusions about Baldwin's resurrection.

DH Reception Studies and Networked Reading

Tweets of a Native Son most broadly hopes to join and affirm recent digital humanities work that is trained on readership, reception, and textual circulation, such as Lincoln Mullen's *American's Public Bible* and Ryan Cordell and David Smith's *Viral* Texts, and to amplify Katherine Bode's call that the digital humanities better attend to and account for the ways in which literary texts "circulated and generated meaning together at particular times and places" (Mullen, 2016; Cordell and Smith, 2017; Bode, 2017). Like the 19th-century newspaper archives used by Mullen, Cordell, and Smith, social media archives offer a window into how texts travel, how texts are used and changed by individuals, and what these texts mean in context. Social media archives additionally offer massive amounts of (relatively) clean, recent data. Though of course with these advantages, they also present more ethical challenges, since this data is often tied to corporations and produced by still-living human beings whose consent, possible harm, and creative attribution must always be considered.

Finally, however, I believe that social media data might help us better theorize and make visible the networked structures of readership, reception, and textual circulation, because social media data, such as Twitter data, is often inherently networked in structure, recording retweets, replies, follower communities, hashtag communities, and more. This networked structure emphasizes the way that texts are not only engaged with by individuals but are shared between individuals, taking on social and communal meanings. For the particular case of Baldwin and #BlackLivesMatter in 2014-2015, the quotations of Baldwin's words were often recirculated as coalition- and community-building material, helping to forge connections between individuals across space, time, and American history. During the future stages of this project, I hope to employ network science and network visualization to better understand Baldwin's significance within #BlackLivesMatter.

References

Bode, K. (2017). The Equivalence of "Close" and "Distant" Reading; or, Toward a New Object for Data-Rich Literary History. *Modern Language Quarterly*, 78(1): 94.

Cordell, R. and Smith, D. (2017).Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines, http://viraltexts.org.

Freelon, D., McIlwain, C. D., and Clark, M. D. (2016). Beyond the hashtags: #Ferguson, #Blacklivesmatter, and the online struggle for offline justice. Center for Media and Social Impact.

Glaude Jr., E. S. (2016). James Baldwin and the Trap of Our History. *Time*.

Maxwell, W. J. (2016). Born-Again, Seen-Again James Baldwin: Post-Postracial Criticism and the Literary History of Black Lives Matter. *American Literary History*, 28(4).

Mullen, L. (2016). America's Public Bible: Biblical Quotations in U.S. Newspapers, http://americaspublicbible.org.

Robinson, Z. (2017). Ventriloquizing Black Feeling, Re-Voicing Black Life: Speaking Baldwin on the Internet. *Communities in Conversation: Digital Baldwin*, Rhodes College.

Sichynsky, T. (2016). These 10 Twitter Hashtags Changed the Way We Talk about Social Issues. *The Washington Post*.

Abundance and Access: Early Modern Political Letters in Contemporary and Digital Archives

Elizabeth Williamson
e.r.williamson@exeter.ac.uk
University of Exeter, United Kingdom

Letters stand as one of the most extensive sources of information on daily life in the early modern period and the study of epistolary culture(s) is a vital and growing area in Renaissance studies (see Daybell and Gordon, 2016; Daybell, 2012; Del Lungo Camiciotti and Pallotti, 2014). Access to such archives and collections is rapidly expanding – and changing – in the wake of mass digitization, online editions, OCR and federated search. In this paper I explore the extension of the narrative of archival history and epistolary provenance into the digital realm. Specifically, I compare the contextual afterlife of early modern letters in nascent state archives to their representation in the digital world, with particular emphasis on classification and metadata, surrogacy and access. Going beyond paralleled modern and early modern anxieties of information overload (the standard comparison of the print and digital revolution), this allows me to explore issues of access, search, and retrieval; control, preservation, and loss, then and now. This is an under-studied area ripe for discussion, and this paper aims to test these ideas in preparation for a wider study that connects the gathering, transmission, and preservation of political information in the early modern period to the digital life of these material primary sources and to our lives as digital researchers. There is a ready parallel to be found between the burgeoning administrative and institutional drive to preservation found in the early modern period – essentially the evolution of state archiving – and the informational anxieties of the internet age, where that largest of archives can offer everything and nothing, excess and restriction, results or dead ends. I explore tensions around archives facilitating both preservation and forgetting, which finds its apotheosis in the endless loss and abandonment of digital data, and in digital methods of retrieval as strict gatekeepers (a roulette of keyword search, privations of metadata, and dreams of text analysis).

I will use the concept of copia, fundamental to early modern humanism and classical pedagogy, to explore these twin pressures of abundance and lack, of meaningful quantity and meaningless repetition. Copia in the early modern period referred to the abundance of language, where mastery over the myriad ways of expressing a single idea gave students the rhetorical strategies to navigate the vast expanse of language. The incessant imitation of classical models, particularly concerning letter-writing, was encouraged not least by Erasmus in the wildly popular *De Copia*, and became a ubiquitous part of humanist education. This concept, of expertise and thus authority being created by sheer mass, by repetition, is particularly apposite when considering rhetoric and knowledge creation today. In fact, this abundance was framed as both knowledge and folly, particularly from the late sixteenth century, when the drive to systematizing information and rise of scientific method pushed against classical humanism (Francis Bacon offers a good example of a writer in this transitional time who both criticized copia and performed it in his criticism). Christine Hoffman, in her recent *Stupid Humanism: Folly as Competence in Early Modern and Twenty-First-Century Culture*, has also identified a productive parallel here, and links early modern rhetorical strategy and modern excess in online 'news' and social media (Hoffman, 2017). I will focus rather on early modern and digital archive creation in order to explore how access to these constituent building blocks of knowledge shapes our historiography and thus the world we construct. I will use copia's two semantic faces, abundance and copying, to firstly think through wider preoccupations with sheer tides of (often repeated) information acting as knowledge and secondly to consider our creation and reception of digital surrogates of primary sources. The early modern relationship to copia as copying is complex: on one hand it is intimately related to the massive growth of bureaucracy and paperwork, where the copy is the transmitter of authority, on the other hand the advent of print and the selfsame abundance of paperwork led to associations of degradation and inferior quality. If print could be alternately the thing itself or the degraded copy of the original, the digital surrogate today is held intermittently as the preserved, unquestioned work and as the flat representation that has lost both the materiality and authenticity of the original.

I will combine discursive reflection with concrete examples that draw parallels between early modern and modern concerns, to consider how the preoccupations and experiences of a particularly early modern growth of the archive and associate concern with amassing, preserving and accessing information inflects our understanding of the internet age, and vice versa. I draw parallels between search and preservation concerns today and amongst early keepers of the state paper office, demonstrating that long-held anxieties around access and information overload continue through into the digital archive. I place the digital archive in a history of indexing and cataloguing, which capitalizes on recent interest in early modern construction of metadata witnessed by Oxford University's 2017 conference 'The Book Index', for example. In both this history of information management and in our relationship to archiving today, what is kept, who has access, and how meaningful and stable that access is, are all essential questions. No less do we need to engage critically with the term access in an increasingly business and market-driven university sector. I point to the open-access philosophy as increasing the availability of digital primary sources, particularly around libraries and archives releasing high-quality digital images and the IIIF initiative making sharing images increasingly possible on a practical level, and to the untold opportunity for connecting resources in the abundant meta-archive promised by the LOD philosophy. From endorsed letters held in the labelled drawers of the Elizabethan Secretary of State's office, to authority files and shared standards, metadata has often been our key to texts. It both enables and restricts

our access, it channels our vision and helps construct our understanding of our sources. Understanding the stories of what is kept and the nuance of how it is described is vital to reveal the limits of this vision. We cannot forget that archives are variously permissive, proscriptive, and problematic, constructed through a history of loss, antiquarianism, colonialism, class and power structures: those with power leave records, those with power control them. With the risk of reproducing existing and long-standing power structures in our digital representations of the early modern world, we need to ask what we want from the modern digital archive, and who will be invited in?

How people create, access, and preserve (particularly political) information can tell us much about the power structures, value systems and personal concerns of both the writers and keepers of texts and the society at large. I argue that understanding the conditions of production and preservation of early modern letters is necessary to fully comprehend their use and meaning: the natural extension of this is that in order to fully engage with these texts we must attend to the methods and conditions of our own access in an increasingly digital scholarly environment. The digital has a natural and increasingly significant place in this textual history: we need to recognize and interrogate the continuous history or provenance that connects the first moment of textual creation to the most recent instance of representation and remediation in a digital space. Reflecting on this reveals that we can read in what is preserved and how it is described the power structures inherent in the society that creates the archive: this is absolutely true for the digital.

References

Daybell, J. (2012). *The Material Letter in Early Modern England: Manuscript Letters and the Culture and Practices of Letter-Writing, 1512-1635*. Basingstoke; New York: Palgrave Macmillan.

Daybell, J. and Gordon, A. (eds). (2016). *Cultures of Correspondence in Early Modern Britain*. Philadelphia: University of Pennsylvania Press.

Daybell, J. and Gordon, A. (eds). (2016). *Women and Epistolary Agency in Early Modern Culture, 1450-1690*. (Women and Gender in the Early Modern World). Burlington, VT: Ashgate.

Del Lungo Camiciotti, G. and Pallotti, D. (2014). Letter Writing in Early Modern Culture, 1500-1750. *Journal of Early Modern Studies*, 3 doi:10.13128/JEMS-2279-7149-3. http://www.fupress.net/index.php/bsfm-jems/issue/view/1023 (accessed 27 October 2017).

Erasmus, D. and Thompson, C. R. (1978). *Collected Works of Erasmus*. Toronto; London: University of Toronto Press.

Hoffmann, C. (2017). *Stupid Humanism: Folly as Competence in Early Modern and Twenty-First-Century Culture*. S.l.: Springer International Pu.

Balanceándonos entre la aserción de la identidad y el mantenimiento del anonimato: Usos sociales de la criptografía en la red

Gunnar Eyal Wolf Iszaevich
gwolf@gwolf.org
Instituto de Investigaciones Económicas, UNAM, Mexico

La criptografía por fin está detrás de prácticamente cualquier acción que emprendamos en Internet — Hemos llegado por fin a una adopción masiva del cifrado para la mayor parte de las comunicaciones en línea. Esto puede leerse desde ángulos muy distintos — Por un lado, nuestras comunicaciones están seguras del espionaje o modificación por parte de terceros. Por otro lado, está *firmada* –por nosotros y por nuestra contraparte– de forma que permite un *no repudio*.

La criptografía puede abordarse y estudiarse desde muy distintos ángulos. El ángulo matemático propone, desarrolla, valida (o refuta) los esquemas que se van presentando; abordar el tema desde la ingeniería en seguridad informática presenta diferentes aplicaciones, analiza modelos de amenaza, estandariza algoritmos en protocolos y formatos, etcétera. Puede hacerse también un análisis legal — Es bien sabido que hasta octubre del 2000, los Estados Unidos consideraban a la *criptografía fuerte* como municiones, y prohibían su *exportación*; afortunadamente sus legisladores reconocieron que los tiempos que corrían eran distintos y rectificaron dicha ley (Export Administration Regulations 15 CFR Part 730 et seq.), pero el tema respecto a quién y cómo debe tener derecho a mantener su privacidad incluso ante entidades de gobierno se mantiene vigente, con ejemplos como el del teléfono del asesino del ataque de San Bernardino (2016).

En esta intervención, el objetivo es hacer un análisis de la criptografía desde un punto de vista social: ¿Qué se entiende por *hacker*? ¿Por qué los *hackers* y la criptografía van tan de la mano? ¿Puede verse el trabajo de éstos ya sea en la comprensión social de la criptografía o en su avance técnico? E incluso yendo un paso más allá, ¿cómo han ido definiendo la criptografía los grupos *hackers* a los diferentes movimientos sociales en que encuentran cabida, tanto en países desarrollados como en la región latinoamericana?

Abordaremos casos como el *Chaos Computer Club* alemán, con más de 35 años de existencia y un congreso que atrae a más de 10,000 personas anualmente y la *Electronic Frontier Foundation*, fundado en 1990 y enfocado a defender legalmente los derechos de libertad de expresión en ambientes en línea, pero también como el *Rancho Electrónico* en México, *Vía Libre* en Argentina, *Patio Maravillas* en España.

Pero además, todos estos grupos, con sus características únicas y sus distintos grados de formalidad/in-

formalidad/aformalidad, se han ido engranando y retroalimentando. La *cultura hacker* es, ante todo, cultura. Los espacios (presenciales o electrónicos) de reunión de hackers son, necesariamente, espacios de creación cultural. Y mostraremos cómo todos estos grupos se han convertido en referentes de la creación cultural, de la generación de movimientos sociales — Estando, en todo momento, vinculados con sus principios creadores: Con la belleza técnica y elegancia algorítmica que posibilitan la existencia de la criptografía.

A White-Box Model for Detecting Author Nationality by Linguistic Differences in Spanish Novels

Albin Zehe
zehe@informatik.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Daniel Schlör
schloer@informatik.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Ulrike Henny-Krahmer
ulrike.henny@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Martin Becker
becker@informatik.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Andreas Hotho
hotho@informatik.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Germany

Introduction

Automatic nationality detection of authors writing in the same language (such as Spanish) can be used for many tasks, like author attribution, building large corpora to analyse nationality specific writing styles, or detecting outliers like exiled or bilingual authors. While machine learning provides many methods in this area, the corresponding results are usually not directly interpretable. However, in the Digital Humanities, explainable models are of special interest, as the analysis of selected features can help to confirm assumptions about differing writing styles among countries, or reveal novel insights into country-specific formulations.

In this work, we aim to bridge this gap: Our assumption is that nationality or country of origin of an author is strongly connected to their writing style. Thus, we first present a machine learning approach to automatically classifying literary texts regarding their author's nationality. We then provide an analysis of the most relevant features for this classification and show that they are well interpretable from a literary and linguistic standpoint.

Related Work

The problem of detecting regional linguistic differences is at the core of Digital Humanities, as it touches research questions in both traditional linguistics and modern computer science. In Spanish philology and linguistics, the analysis of different regional varieties has a long tradition (see for example Alvar 1969, Eberenz 1995, Noll 2001). There are well-known differences between the Spanish spoken and written in Spain itself and the variations used in the former colonies, for example in forms of address ("vosotros/ustedes" vs. just "ustedes", voseo) and articles (le/la vs. lo).[1] More recently, these differences have been investigated with quantitative methods, for example by applying Zeta to find distinctive words for novels from Spain and from Latin America, respectively (Schöch et al. 2018).

Model

Baseline SVM-Model for classifying author nationality

We assume that writers from different countries are distinguishable by a) their vocabulary and b) phrases that are more or less popular in different regions (cf. Section "Related Work"). Thus, we choose to use an n-gram model to represent our corpus in a computer readable way: First, we determine all word n-grams of length 1 to 4 in the corpus. Then, we select the 1000 most frequent n-grams of each length. We also tried selecting the 100 or 10000 most frequent n-grams, which led to slightly worse results. We represent a piece of text as tf*idf vectors of these n-grams (see Manning 2008).

We then train a linear SVM (see Steinwart 2008) to predict the nationality of an author given a piece of text. The linear SVM is known for good results in text classification (Joachims 1998) and - essential for interpretability - allows to inspect the importance of specific features.

Enhancing Feature Interpretability

When examining our classification model, we observed an over-representation of geographical entities (e.g., frequent locations like Buenos Aires) as well as names. To instead enforce linguistic properties, we replaced all uppercase tokens by distinct UNKNOWN-tokens (except at the beginning of a sentence). For example "¡Oh, María, María! ¡Cómo deseaba triunfar, conquistar Buenos Aires [...]", becomes "¡Oh, UNK_1, UNK_2! ¡Cómo deseaba triunfar, conquistar UNK_3 UNK_4 [...]". This ensures that n-grams with proper nouns will never be frequent enough to be used as a feature in our classification task.

1 http://lema.rae.es/dpd/?key=voseo, http://lema.rae.es/dpd/?key=loísmo&lema=loísmo

Augmenting Training Examples

The success of machine learning algorithms depends largely on the amount of training data. Thus, to increase the number of training samples, we split each novel into multiple segments of equal length, assigning each segment the same label as the entire novel.The cross validation split was performed before segmentation, ensuring that no novel was present in both training and test set. The classifier is then trained and evaluated on individual segments, resulting in a set of "votes" for the nationality of each novel in the test set. The nationality is then established by majority vote.

Corpus

We use a corpus composed of 100 novels from four Spanish-speaking countries, specifically Spain, Argentina, Cuba and Mexico, written in the 19th and early 20th century (Calvo Tello 2017, Henny-Krahmer 2017). Figure 1 shows the distribution over countries and the distributions over subgenres in the countries. All countries are represented by a roughly equal number of texts. We note that our corpus may have a bias towards a specific subgenre in some countries, which will later be addressed in the analysis of the features.

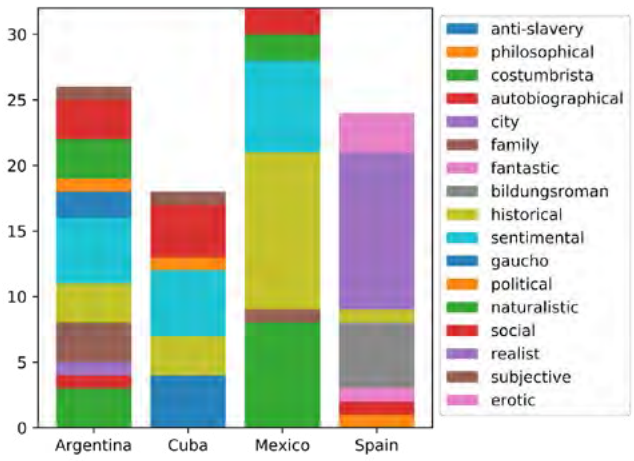


Figure 1: Distribution of countries and subgenres in our corpus

Experiments

We performed extensive experiments on the dataset to determine the accuracy of our approach. The main hyper-parameters of our model are the segment size s, determining how many words a segment contains, and the parameter C of the SVM. We performed parameter optimisation by grid search, choosing from s ∈ {100, 200, 500, 1000, 5000, 10000, 100000, ∞} and C ∈ {10^{-5}, 10^{-4}, …, 10^{5} }. The setting s = ∞ does not perform segmentation. We also varied the maximum length of n-grams: unigrams (n = 1) vs. n-grams of length 1 to 4. All scores reported below are weighted average F1-scores over 10-fold cross validation.

Generally, our model performed best when using only unigrams, removing uppercase tokens and splitting the novels into segments of length s = 1000 (see Table 1 for details).

	precision	recall	f1-score	support
0	0.800	1.000	0.889	24
1	0.923	0.667	0.774	18
2	0.824	0.875	0.848	32
3	1.000	0.885	0.939	26
avg / total	0.882	0.870	0.868	100

Table 1: Classification report for the best configuration, using only unigrams, segments of length s = 1000 and C = 10000

This can be explained by the small dataset: Unigrams are likely to occur in multiple samples even in a small corpus, while higher-order n-grams possibly only occur once and can therefore not be used for classification.

Figure 2 shows the results for varying s and C. Segments of a length around 1000 perform best, yielding F1-scores of up to 86.8 %. Very small segments fail to deliver satisfying results, while larger segments still provide reasonable classification accuracy. The value for C must be set high enough, but the specific value does not matter for C > 10.

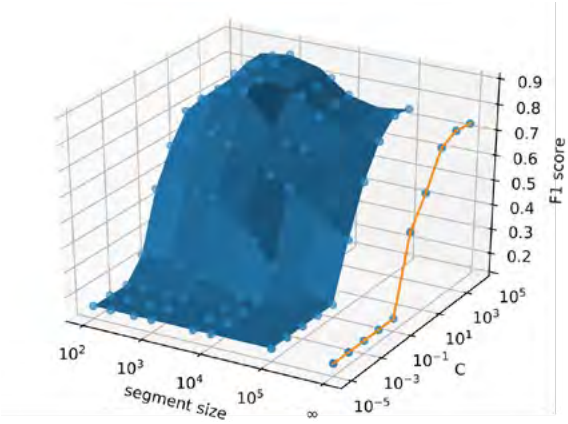


Figure 2: Weighted average F1-score depending on the segment size s and the cost parameter C of the SVM. The separated line denotes no segmentation (s = ∞). Only unigrams were used as features.

Using all n-grams of length 1 to 4 also delivered good accuracy (highest F1-score of 80.4% for C = 10000, s = 1000). Removing uppercase tokens had a positive effect when using unigrams, while it hardly influenced the accuracy using all n-grams.

A detailed view of all results can be found on GitHub.[2]

Feature Analysis

Using a linear SVM enables us to analyse the 10 n-grams that provide the strongest evidence for and against a country (according to internal weights). In the following, we focus on features that are weighted strongly in all or at least multiple folds of the cross validation.

Generally, we identify three feature groups: topical features, features related to the geographical setting and linguistic features. The presence of topical features can be explained by the bias in subgenres that is present in our corpus and is not necessarily representative. The geographical features seem to point to a tendency of the authors to base their stories in their respective home countries rather than other countries.

With regard to the different model variants, the model based on \emph{unigrams without removing uppercase tokens} tends to select names as its top-features such as the country itself or characteristic cities, for example "Madrid" for Spain. While these features are surely helpful for classification (yielding an F1-score of 81.7%), they are not particularly interesting for linguistic analysis. The features selected after removing uppercase tokens, on the other hand, seem more relevant from a linguistic viewpoint, while at the same time providing the best accuracy. Table 2 shows features that are among the highest weighted for more than 5 folds for each country in this setting.

Country	Unigrams	Comments
Spain	+ ello	linguistic (personal pronoun)
	+ seorito	linguistic (diminutive)
	+ duros	currency
	+ seores	linguistic/topical (noun)
	- pesos	currency
Cuba	+ esclavo/esclava	topical
	+ mulato	topical (ethnic group in Cuba)
	+ aadi	linguistic/narrative (verb, probably used to mark direct speech)
	- quiz	linguistic (adverb)
	- huerta	topical/linguistic (noun)
Mexico	+ hacienda	topical (haciendas are typical of Spanish colonies
	+ mexicano	
Argentina	+ entretanto	linguistic (temporal adverb)
	+ gaucho	topical
	+ misia	linguistic (form of address typical in South America)
	+ mate	topical (drink typical to Argentina)

Country	n-grams	Comments
Spain	+ se me figura que	linguistic (locution)
	+ de la huerta	topical (huerta is common in Spain)
Cuba	- de cuando en cuando	linguistic (temporal phrase)
Mexico	+ de/en la casa de	topical (probably due to a subgenre bias)
	+ al cabo de	linguistic (temporal phrase)
	+ de la hacienda	topical (typical of Spanish colonies)
	+ as es que	linguistic (locution)
	- al mismo tiempo	linguistic (temporal phrase)
	- la	linguistic (lesmo)
Argentina	+ en ese momento	linguistic (temporal phrase)
	+ se puso de pie	linguistic (verb)
	+ de vez en cuando	linguistic (temporal phrase)
	+ el hecho es que	linguistic (locution)
	+ al fin al cabo	linguistic (temporal phrase)
	- al cabo de un	linguistic (temporal phrase)

Table 3: N-grams with large weights assigned by the SVM. Features marked with + and - are signals for and against a country, respectively.

Discussion

Technical Aspects

We found that segmenting novels to augment the training data does improve results, but only if the segments are not too short and thus do not contain enough information to detect the author's nationality.

Removing uppercase tokens improves the classification accuracy and makes the selected features more interesting from a linguistic standpoint. We assume that otherwise proper nouns are picked up by the classifier as important clues on the training set, which fail to generalise to the test set.

Feature Interpretation

The words and phrases that our algorithms selects for differentiating between nationalities strongly resemble features that humans would consider given the same task. These include well-known linguistic differences (leísmo) as well as country-specific words (hacienda/huerta). However, it also finds phrases, such as temporal expressions, that are not very well known to be specific for some countries, but should be further investigated in future work. We also observe that authors in our corpus appear to have a strong tendency towards writing about their respective home countries, as evidenced by the selection of city or country names.

Conclusion and Future Work

We have presented a classifier that is able to distinguish between novels from different countries based on word n-grams. Our experiments show that this classifier is able to select features that are interpretable and reveal interesting insights into the language used in novels from different Spanish-speaking countries.

We note that our findings are only based on a limited dataset. However, the tools we have built enable us to replicate the experiments and confirm our findings as soon as larger collections of text become available.

Thus, our work is an important step towards combining machine learning with in-depth analysis and discovery of novel concepts in corpus-based linguistic studies through interpretable models.

In future work, we believe that replacing the majority vote over segments by more sophisticated methods can further improve our results. We also believe that incorporating linguistic information like parse-trees into our features can help to reveal more interesting insights into subtle linguistic differences between countries.

2 https://github.com/cligs/projects2018/tree/master/country-dh

References

Alvar, Manuel (1969). *Variedad y unidad del español: estudios lingüísticos desde la historia*. Editorial Prensa Española.

Calvo Tello, José (ed.) (2017). *Corpus of Spanish Novel from 1880-1940*. Würzburg: CLiGS. https://github.com/cligs/textbox/blob/master/es/novela-espanola.

Eberenz, Rolf (1995). "Norm und regionale Standards des Spanischen in Europa und Amerika". In: Oskar Müller, Dieter Nerius, Jürgen Schmidt-Radefeldt (eds.). *Sprachnormen und Sprachnormenwandel in gegenwärtigen europäischen Sprachen*. Rostock: Universität Rostock, 47-58.

Henny-Krahmer, Ulrike (ed.) (2017). *Collection of 19th Century Spanish-American Novels (1880-1916)*. Würzburg: CLiGS, 2017. https://github.com/cligs/textbox/master/spanish/novela-hispanoamericana/.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. *Machine learning: ECML*-98, 137--142.

Manning, C. D., Raghavan, P., Schütze, H. (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press. ISBN: 0521865719, 9780521865715

Noll, Volker (2001). *Das amerikanische Spanisch: ein regionaler und historischer Überblick*. Tübingen: Niemeyer.

Schöch, C., Calvo, J., Zehe, A., Hotho, A. (2018). Burrows Zeta: Varianten und Evaluation. *DHd 2018*.

Siskind, Mariano (2010): "The Globalization of the Novel and the Novelization of the Global. A Critique of World Literature." In: *Comparative Literature* 62 (4), 336-360. https://doi.org/10.1215/00104124-2010-021

Steinwart, I., Christmann, A. (2008). *Support Vector Machines*. Springer Publishing Company, Incorporated. ISBN: 0387772413

Media Preservation between the Analog and Digital: Recovering and Recreating the Rio VideoWall

Gregory Zinman
gzinman3@gatech.edu
Georgia Institute of Technology, United States of America

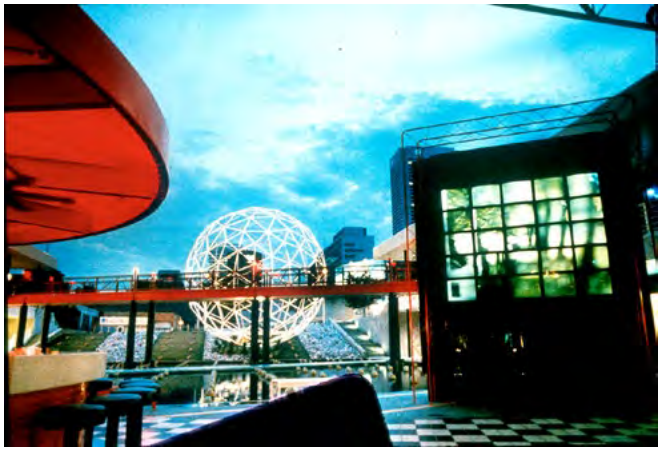


Figure 1: *The Rio VideoWall* (Dara Birnbaum, 1989), installed at the Rio Shopping Complex, Atlanta. Image courtesy of the Smithsonian American Art Museum, NEA Birnbaum collection.

"Media Preservation between the Analog and Digital" is a project that centers on the development of a digital recreation of pioneering video artist Dara Birnbaum's now-lost *Rio VideoWall* (1989), the first multi-screen artwork to be installed in a public setting in the United States. In its original instantiation, in downtown Atlanta, the work employed twenty-five identical 27" CRT monitors, stacked in a five-by-five grid, powered by eight LaserDisc players and proprietary computer code written specifically for the piece. Today, however, only a portion of the code remains; neither the CRT monitors, nor LaserDisc players, nor original computers, are in production, and to recreate the artwork—even in a lab setting—would involve a significant reimagining of the original piece. But there are additional considerations that extend beyond hardware and software: the *VideoWall* was installed in the Rio Shopping Complex, a mall in Atlanta's Old Fourth Ward, a historically African American neighborhood. The artist was attuned to this, and designed the artwork to combine scenes of the natural landscape that had been displaced by the mall with an unedited live-stream of CNN, an Atlanta-based company, all filtered through the moving silhouettes of mall patrons in real time. Neither the nature footage nor the CNN live-stream—let alone the mall patrons—presently exist (the mall was torn down in 2000), so a recreation of the artwork would need to identify footage that captures the spirit, if not the reality, of the piece. To do so

would therefore involve an analysis of the themes engaged by the original artwork: the legacy of segregation, the 24-hour media cycle, surveillance culture, the relationship between art and commerce, and the Anthropocene. This paper will provide a brief overview of the project, with an emphasis on the conceptual challenges it engages, describe the recovery work underway, and describe the current work and next steps toward the *VideoWall*'s ultimate recreation.

Recovering the Rio VideoWall: Conceptual Challenges

Recent seminars and symposia have explored the range of issues associated with doing digital art history (e.g. Harvard metaLab's "Beautiful Data," 2015, the Getty Foundation's "Art History in Digital Dimensions," 2016). At the same time, the field has devoted increasing attention to issues associated with digital preservation (e.g. the BitCurator initiative, or any number of conversations at the Digital Library Federation). And yet, with its hybrid analog/digital design, and its site-specific setting, the *Rio VideoWall* presents a unique case study for thinking through the additional conceptual challenges related to the preservation of public media art. For instance, Matthew Kirschenbaum has argued that every media artifact "leaves a "trace," by which he means a past that can be unearthed and understood. But what happens when the original artifact no longer exists, as is the case with the *VideoWall*, which was dismantled and discarded in 1999? And what is the "trace" that is left by public art that makes use of technology as a material support, as in the *VideoWall*'s CRTs and LaserDisc players, which are rarely understood in terms of their material properties? When considering the unique case of the *VideoWall*, additional questions arise: Can the artwork's original public setting can be reimagined in a creative way, perhaps online? Or should the *VideoWall* be rebuilt physically, and installed in the city of Atlanta, and if so, where? (The original site has long been built over). Or should a recreated *VideoWall* be incorporated into a museum's collection, so that it can be benefit from institutional resources? But what would be lost by limiting community access?

Recovering the Rio VideoWall: The Digital Archive

The first phase of recovering the Rio VideoWall, as well as planning for its recreation, involves the creation of a public-facing digital archive that documents the remaining materials associated with the *VideoWall's* design, construction, and eventual demise. The archive includes Birnbaum's original proposals and plans for the piece, her own 35mm slide documentation of the *VideoWall* and her recently-recovered video footage of the artwork's public opening, as well as correspondence, press clippings, and archival photographs of the site. The archive also features exhibits with demographic data and visualizations that illustrate the racial and economic makeup of the areas around the Rio Shopping Complex at the time of the mall's opening in 1987, at the time of its destruction in 2000, and today. Eventually, the site will also feature oral histories of Atlanta citizens who experienced the work. These histories will provide access to lived and felt experiences often absent from the histories provided by critics and scholars, that are nevertheless an essential component of accounting for the impact of public art within a community.

Recovering the Rio VideoWall: The Digital Recreation

Following the completion of the digital archive, the project team will begin the second phase of the project—reimagining the *VideoWall* in digital form. We are currently considering, and mocking up designs for, web-based, physical, and VR-based approaches to recreating the artwork. A web-based version might include visual overlays of the original site of the artwork on the current site. A physical recreation might involve a single, very large flat-screen display that could be separated into adjacent or tiling windows, drone footage of the chosen site, and touch-screen capabilities. A VR version of the work could be situated in a number of virtual locations, and would allow for diverse populations to interact with the artwork from around the globe. A workshop, currently being planned, will bring together scholars, conservators, designers, curators, and community organizers to discuss the computational and creative possibilities of each medium. In addition to describing the conceptual challenges associated with the piece, and highlighting certain key features of the digital archive, the paper presentation will present these mockups and solicit audience feedback in preparation for the final design and construction of the *VideoWall*.

The (Digital) Space Between: Notes on Art History and Machine Vision Learning

Benjamin Zweig
b-zweig@nga.gov
Center for Advanced Study in the Visual Arts, National Gallery of Art, United States of America

Machine vision learning and art historical practice are often poised as operations that are antithetical to one another (Spratt and Elgammal, 2014a, 2014b). A frequent criticism leveled by art historians against machine learning algorithms is that they do little that a trained art historian cannot do already (Bishop, 2017). A second criticism is that the results gleaned from machine vision

learning are heuristic exaggerations. And a third criticism is that computer scientists simply do not understand how to approach visual art, and in the process wrongly (albeit unintentionally) define the field of art history for a much larger audience than the one the humanities tend to generate. Much of the value placed on machine vision learning as it pertains to understanding artworks has been on its ability to sort, classify, and match images with similar ones through style and genre (Saleh and Elgammal, 2015, 2016), taxonomies that fail to reflect the current state of the history of art.

The above criticisms present a fair critique of the approach of machine vision learning to the history of art – but only to a certain degree. Such criticisms fail to recognize that art historians are precisely the ones who have the greatest stake in and the greatest potential for contributing to the questions raised by machine learning image analysis. Art historians simply ask different questions about artworks – questions of history, scale, tactility, surface, and representation – than the ones of which computer scientists are aware. One reason for this disjuncture is that art historians have often kept to themselves instead of engaging with other disciplines that are intensely interested in visual imagery.

Rather than simply critiquing and lamenting how computer and data scientists approach visual imagery, this short paper addresses a few "between points", as I call them, rather than intersections, where art historians can bring much critical insight into machine vision learning. For example, the issue of texture is a complex question in painting, for it can signify the texture of the paint, or the texture of the canvas weave, or how textured paint application is used to represent different physical textures, such as silk or fur. How could these distinctions be brought into machine vision learning? Another issue would be to see if a machine could identify when a painting was re-touched or repaired. Or one might compare how the descriptive terms generated by machine vision learning output correlate to the terms art historians would use when describing an object. The purpose of this paper is ultimately to pose some questions about how art historians and computer scientists might create a better dialogue in their respective practices.

References

Bishop, C. (2017). Against Digital Art History. https://humanitiesfutures.org/papers/digital-art-history/.

Spratt, E. and Elgammal, A. (2014a). Computational Beauty: Aesthetic Judgment at the Intersection of Art and Science. In Agapito, L. et al (eds), *Computer Vision – ECCV 2014 Workshops. ECCV 2014. Lecture Notes in Computer Science*8925. Cham: Springer International Publishing, pp. 35–53.

Spratt, E. and Elgammal, A. (2014b).The Digital Humanities Unveiled: Perceptions Held by Art Historians and Computer Scientists about Computer Vision Technology. *Cornell University Library arXiv Physics Archive*. http://arxiv-web3.library.cornell.edu/abs/1411.6714v1.

Saleh, B. and Elgammal, A. (2016). Large-scale Classification of Fine-Art Paintings: Learning the Right Metric on the Right Feature. *International Journal of Digital Art History*, 2: 71-94.

Posters

World of the Khwe Bushmen: Accessing Khwe Cultural Heritage Data by Means of a Digital Ontology Based on Owlnotator

Giuseppe Abrami
abrami@em.uni-frankfurt.de
Goethe-University Frankfurt, Germany

Gertrude Boden
boden@em.uni-frankfurt.de
Goethe-University Frankfurt, Germany

Lisa Gleiß
zoi-m.gleiss@gmx.de
Goethe-University Frankfurt, Germany

Poster Abstract

The Khwe are a group of former hunter-gatherers living in Bwabwata National Park in northeast Namibia. They are one of the indigenous groups in Southern Africa known as "San" or "Bushmen". The documentation of their language and cultural heritage was a mission of Oswin Köhler (1911-1996), a German scholar in African Studies.

Between 1959 and 1992 he built up an integral collection of written vernacular texts, audio files, photographs, video files, ethnographic objects, dried plants and drawings from the Khwe, currently housed in the Oswin Köhler Archive at the Goethe University Frankfurt. As his main oeuvre on the Khwe, Köhler had planned an encyclopedia on ideally every aspect of Khwe culture in vernacular texts with German translations, titled "The World of the Khwe Bushmen [Die Welt der Kxoé-Buschleute]".

Four of twelve planned parts have been published in print so far (Kölher 1989, Köhler 1991, Köhler 1997). Köhler has supplemented, revised, split, merged and moved the texts for this encyclopedia from one part or from one section within a part to another over a time period of more than thirty years.

In order to identify and visualize these processes, a team of computer scientists and anthropologists/linguists has developed an OWL ontology for the semantic use of the Köhler encyclopedia. It maps the histories of individual texts and of their position in the overall structure of the encyclopedia but also the relations between Khwe terms, subject areas, text versions, versions of table of contents, footnotes to texts describing manipulations, codes for recurrent types of manipulations to the texts, object types, specific objects, video- and audio-files, photographs, drawings, people and places. It thus allows for a more holistic or integral understanding of Khwe concepts and cultural practices by presenting them in the multiple contexts where they occurred. At the same time it allows for retracing the formation of this cultural heritage documentation by revealing the impact of individual actors in changing and manipulating the documentation, with regards to content as well as numerically, e.g. the replacement of loan words for Khwe terms or attempts to standardize syntax. All this is done with the help of the so-called *OWLnotator* (Abrami et al. 2012).

OWLnotator, as part of the eHumanities Desktop (Jussen, Mehler, Ernst 2007), is a flexible annotation system for annotating inter- and intramedial relations in multimedia corpora and can be used as an annotation platform for any project. By using OWL ontologies as an annotation scheme, arbitrary annotation tasks can be defined.

For this purpose, OWLnotator provides a generic graphical web interface that displays the available classes and properties of the underlying ontology and allows linking to arbitrary resources. These resources are provided by the integration of OWLnotator into the so-called *ResourceManager*. ResourceManager is also part of the eHumanities Desktop and provides access to various types of resources as text documents, images, audio- and video-files, as well as their individual segments and more. In addition, the OWLnotator can also import data from CSV files, provided by the ResourceManager, and assign them to the corresponding ontologies. Furthermore, in this version of the OWLnotator *Blazegraph* (https://www.blazegraph.com/) is used as the new database backend. The poster presents the challenges of understanding and designing the multiple relations between individual items within the ontology, of formally describing and transforming existing data and databases to render them readable or automatically importable, and of visualising the items and relations in the ontology. As a result, the poster will display selected Khwe concepts with all their relations. The analytical potential of the ontology will be exemplified by presenting the results of a number of queries visualization with help of OWLnotator.

References

Abrami Giuseppe and Mehler Alexander and Pravida Dietmar (2015). *Fusing text and image data with the help of the OWLnotato*r. In Sakae Yamamoto, editor, Human Interface and the Management of Information. Information and Knowledge Design, volume 9172 of Lecture Notes in Computer Science, pages 261–272. Springer International Publishing

Jussen, Bernhard and Mehler, Alexander and Ernst, Alexandra (2007). *A corpus management system for historical semantics*. Sprache und Datenverarbeitung. International Journal for Language Data Processing, 31(1-2):81–89.

Köhler, Oswin (1989). *Die Welt der Kxoé-Buschleute im südlichen Afrika: eine Selbstdarstellung in ihrer eigenen Sprache*. 1. Die Kxoé-Buschleute und ihre ethnische Umgebung. D. Reimer, Berlin, Germany.

Köhler, Oswin (1991). *Die Welt der Kxoé-Buschleute im süudlichen Afrika: eine Selbstdarstellung in ihrer eigenen Sprache*. 2. Grundlagen des Lebens: Wasser,

Sammeln und Jagd, Bodenbau und Tierhaltung. D. Reimer, Berlin, Germany.
Köhler, Oswin (1997). *Die Welt der Kxoé-Buschleute im Südlichen Afrika: Die Welt der Kxoe-Buschleute im südlichen Afrika*. 3. Materialle Ausrüustung; Wohnplatz und Buschlager. D. Reimer, Berlin, Germany.

Design on View: Imagining Culture as a Digital Outcome

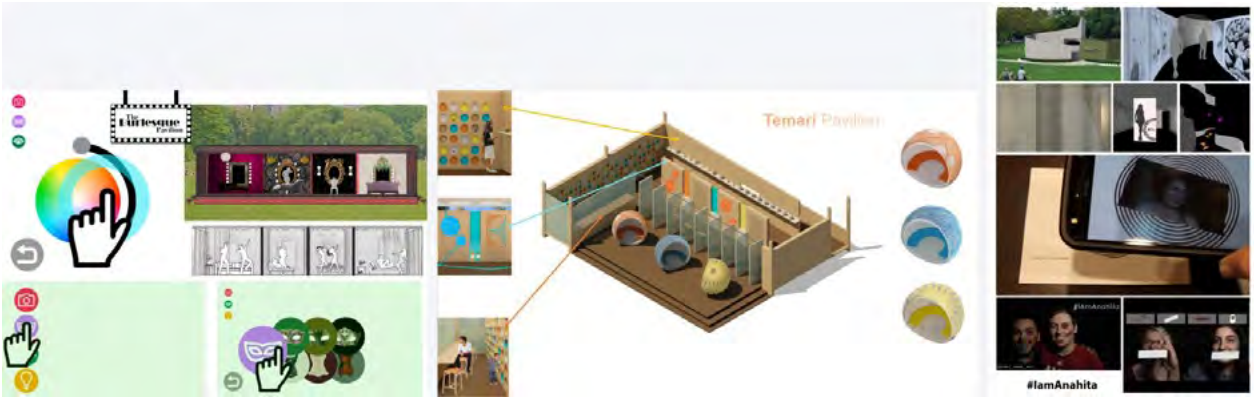
Ersin Altin
ersin.altin@njit.edu
New Jersey Institute of Technology, United States of America

Can **design** represent a culture/nation? Can the tools of digital design be used in collaboration with industrial and interior design to establish an interactive communication with culture? While design and **designwork** were seen as essential symbols of nation-based identity construction in most of the 20th century, today, the notion of design deliberately shies away from exposing its cultural/national implications because of global aspirations. Today's world, dominated by multinational corporations, with its imposition on self-centered identities seemingly curtains the close connection/flirtation of design to its cultural roots. The project that is developed as a collaborative design task at School of Art + Design at New Jersey Institute of Technology (NJIT) aims to question and build on the assumption that suggests a connection between design and culture/nation, with the emphasis on the fact that **nation** is also a social construction (Anderson, 1983).

This poster visualizes the results of the collaborative design project that I taught at NJIT in Fall 2016 and again in 2017. Throughout the semester students from different design fields were expected to work as a group on the design of a pavilion for the culture/nation of their selection that together with other teams formed an imaginary exposition center. Instead of superficial identifications, systematic research process and critical design concepts based on intellectual analysis of the findings determined a basis for the design project. By both researching and producing, teams aimed to create a digital tool that would be developed to investigate whether **designwork** can represent a culture/nation, subculture or simply a cultural issue. Three teams consist of three students from three different design fields worked on their pavilions that are imagined as interactive tools. These tools incorporating data processing software, motion capture, virtual and augmented reality establish vivid, interactive communication with the user. In doing so, instead of creating informative two-dimensional representations, projects aimed to involve users to explore their contribution to the dynamics of a culture. In other words, instead of imposing a **meaning**, pavilions ask users to build new meanings via their interactions both with the pavilion and with other users.

The poster documents three different design processes each of which produced its own interactive digital tools to communicate culture. One team envisioned a mobile pavilion for Burlesque culture that offered users to design their own shows. Augmented reality helped users/performers select and **put-on** a stage costume digitally. With a digital control panel performers were given a chance to adjust atmospheric effects such as light level and color, while physicality of the setting was conceived through a meticulous analysis of the Burlesque culture, such as heavily ornate historic furniture, wallpapers, textile, and decoration.

Second team created a digital crafting tool to educate visitors about Japanese Temari balls, which are toy balls made from embroidery may be used in handball games. Team tackled weaving as a craft with the question how and why weaving can be utilized as data analysis with an emphasis on its fabrication processes by using Japanese Temari balls as a case study. The pavilion encouraged visitors not only to learn about Temari tradition, but also share their experience with other users, who do not necessarily speak the same language or come from similar cultural backgrounds by transforming Temari making into a cultural activity that is virtually organized around a **ball game** / spectacle.



Burlesque Pavilion by Hideyoshi Azama, Emily Gutierrez, Tulio Squarcio (left); Temari Pavilion by Danielle Archibold, Wuraola Ogunnowo, Florencia Pozo (middle); Pavilion Anahita by Negaar Amirihormozaki, Albeirys Francisco-Parra, Nazifa Hamidullah (right)

The third team designed a pavilion that aimed to create a community by gathering people both physically in the space of the pavilion and virtually through social platforms such as Facebook, Twitter, Snapchat, and Instagram. The team problematized Iran's mandatory hijab law by connecting the issue to sexism in different parts of the world that creating a network on women's rights issues. Hijab's ban in some countries and its enforced use in others were carefully examined to generate a digital forum for different opinions on this specific issue.

This research was conducted to investigate culture's changing perceptions. Rather than attempting to redefine a preconceived notion of culture by simply incorporating modern technologies, digital tools, and social media, it aimed to reveal new interactive networks that culture forms with other notions and omit others when conventional relations needed replacement; for example, a new interconnectedness instead of nationality. Finally, this project highlighted areas that were defined by the conventional cultural tools and perceptions that are still relevant.

References

Anderson, B. (1983). *Imagined Communities: Reflections on the Origin and Spread of Nationalism*. London and New York: Verso.

Introducing Polo: Exploring Topic Models as Database and Hypertext

Rafael Alvarado
ontoligent@gmail.com
University of Virginia, United States of America

Since the invention of Latent Dirichlet Allocation (Blei, et al. 2003) and early demonstrations of its utility for identifying lexical clusters in collections of historical and literary texts (Block and Newman 2006, Blevins 2010), topic models have become a mainstay of the digital humanities. However, the use of topic models within the field remains narrowly conceived, restricted largely, with some exceptions, to the discovery of topics and topic trends within corpora, even though the method has been extended significantly since first introduced. One reason for this conservativism may be that, like many methods drawn from data science, both the process and the output of topic model algorithms remain interpetively opaque to the humanists (and, arguably, to the computer scientist as well). Aside from the complexity of the math involved, a contributing factor to this opacity has been the limited way in which the results of topic models are presented to the user. One the one hand, the data provided by standard topic modeling tools (whether in Java, Python, or R) are often trapped in data files or shielded by objects that cannot be queried directly or visualized freely without the use of ad hoc programming or spreadsheet software. On the other hand, the outputs typically provided by these tools, such as top words per topic (often visualized as word clouds), show a highly restricted, decontextualized, and potentially distorted picture of the model (Schmidt 2013). Recently, various tools have emerged to fill this gap, such TOME (Klein et al. 2015), which is designed to allow scholars to explore topic models more fully. In this talk I will present Polo, a topic model browser developed at the Data Science Institute at the University of Virginia designed to present topic models to users in a direct, transparent, and complete manner, so that the representational quality of models may be explored, questions, and adjusted interactively. Built on top of MALLET, Gensim, and NLTK, Polo is a Python package that provides tools to both create topic models and to inspect them by combining the source corpus with all of the data produced by the core software into a single, normalized relational database (in SQLite). This database in turn forms the foundation of an interactive web application that effectively converts the output model with associated data and the source corpus into a single hypertext relating words, topics, and documents. A key design feature of Polo is that it employs the statistical properties of the model -- such as topic entropy in documents or mutual information among topics -- not simply as readouts on a dashboard but as navigational devices that allow the user to move from a reduced dimension, high-level perspective of a corpus to its source documents, and to move laterally through the network of topics and documents that compose the model. Using examples from both newspaper and journal collections, I will demonstrate how Polo enables scholars both to investigate implied cultural newtworks in these corpora and to explore the various ways in which topics may be said to convey meaning.

References

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2002. "Latent Dirichlet Allocation." In *Advances in Neural Information Processing Systems 14*, edited by T. G. Dietterich, S. Becker, and Z. Ghahramani, 601–608. MIT Press.

Blevins, Cameron. 2010. "Topic Modeling Martha Ballard's Diary." *Cameron Blevins* (blog). April 1, 2010, http://www.cameronblevins.org/posts/topic-modeling-martha-ballards-diary/.

Klein, Lauren F., Jacob Eisenstein, and Iris Sun. 2015. "Exploratory Thematic Analysis for Digitized Archival Collections." *Digital Scholarship in the Humanities* 30 (suppl_1):i130–41.

Newman, David J., and Sharon Block. 2006. "Probabilistic Topic Decomposition of an Eighteenth-Century American Newspaper." *Journal of the American Society for Information Science and Technology* 57 (6):753–767.

Schmidt, Benjamin M. 2013. "Words Alone: Dismantling Topic Models in the Humanities." *Journal of Digital Humanities*. April 5, 2013.

El primer aliento. La expedición de los lingüistas Swadesh y Rendón en las ciencias computacionales (1956-1970)

Adriana Álvarez Sánchez
adralvsan@gmail.com
Universidad Nacional Autónoma de México, Mexico

El póster tiene como objetivos aportar conocimientos sobre los orígenes de las HD en América Latina y mostrar los primeros procesos computacionales realizados por los investigadores de la UNAM, Maurice Swadehs y Juan José Rendón, dedicados a la lingüística comparada de las lenguas mayas utilizando una computadora IBM 650 de alquiler.

Se abordan tres aspectos que serán ilustrados con documentos originales conservados en los archivos de la propia universidad: la gestión de recursos, el proceso técnico y los resultados de la investigación de estos lingüistas. Las cartas enviadas por Swadehs al director del Centro de Cálculo Electrónico de la UNAM para hacer uso de la computadora evidencian las gestiones realizadas por el lingüista, así como la descripción del proyecto. El proceso computacional ha quedado descrito en cartas pero también en los informes que este investigador hacía anualmente, así como en el archivo personal de Rendón, donde además he encontrado diagramas de flujo, matrices y otros documentos institucionales sobre el proceso realizado. Finalmente, parte de los resultados de la expedición de estos lingüistas en la aplicación de las ciencias computacionales al análisis de las lenguas quedó registrada en la revista *Estudios de Cultura Maya*. La lectura de esta publicación permite conocer también parte de los debates de aquella época acerca del uso de técnicas y tecnologías computacionales para el estudio de las lenguas. Ejemplo de ello fue el trabajo de los epigrafistas de la Sección Siberiana de la Academia de las Ciencias de la URSS sobre escritura maya antigua.

Los documentos consultados y reproducidos en el póster proceden de los Fondos Documentales Alfonso Caso del Instituto de Investigaciones Antropológicas de la UNAM, donde se encuentra el Fondo Juan José Rendón que consta de más de 100 cajas. En las Colecciones Especiales de la Biblioteca Juan Comas, Sección de Antropología del mismo Instituto se encuentran copias de los expedientes institucionales de ambos académicos.

La presente investigación se encuentra en un nivel avanzado de desarrollo y busca reconstruir la manera en la que los humanistas hicieron uso de los avances tecnológicos en sus disciplinas. Me interesa postular hipótesis acerca de si fueron las condiciones institucionales, el desinterés por metodologías de esta naturaleza o las propias corrientes de la lingüística, la razón por la cual no hubo continuidad en el desarrollo de estas investigaciones. Un proceso semejante se dio en los estudios históricos a nivel mundial, a raíz de la publicación de obras que dieron mayor peso a la estadística, dejando de lado la interpretación histórica.

Conocer la manera en la que las disciplinas humanísticas enfrentaron y/o aprovecharon los cambios tecnológicos de mediados del siglo XX contribuye a comprender el estado en el que hoy interactúan las distintas áreas del conocimiento, incluidas las ciencias computacionales.

Las aportaciones de este trabajo se concretan en ofrecer conocimientos sobre las primeras experiencias de aplicación de las ciencias computacionales que se llevaron a cabo en distintas latitudes y que forman parte de los orígenes de las HD. También se rescatan investigaciones olvidadas, incluso para la historia de la lingüística, que exploraron la relación entre la tecnología y el estudio de las sociedades. Se reconstruyen procesos computacionales del tercer cuarto del siglo XX aplicados al estudio de las lenguas indígenas y, finalmente, se ponen de manifiesto aspectos que hoy continúan conformando condicionantes para el desarrollo y sostenimiento de proyectos digitales en las universidades.

References

Barrera Vásquez, A. (1962). Investigación de la escritura de los antiguos mayas con máquinas calculadoras electrónicas: síntesis y glosa, *Estudios de cultura maya*, II: 319-342.

Beltrán, S. (1959). Carta del Ingeniero Sergio Beltrán al Dr. Maurice Swadesh, 5 de diciembre de 1959". *Colecciones Especiales de la Biblioteca Juan Comas*, Sección de Antropología, Caja 05,Exp. 06, Carácter C, núm. de registro 30, ff. 147-148.

Knorozov, Y. (1963). Ap.licación de las matemáticas al estudio lingüístico, *Estudios de Cultura Maya*, III: 169-185.

Rendón, J. J. (1967-68). Plan de trabajo de Juan José Rendón, 1967-68, *Colecciones Especiales de la Biblioteca Juan Comas*, Sección de Antropología, Caja 05,Exp. 05, Carácter C, núm. de registro 29, ff. 21-24.

Rendón, J .J. (1973). Epistolario, *Fondo Juan José Rendón*, caja 1.

Redón, J. J. (s.a.). Listas diagnósticas. Léxico-Estadística, *Fondo Juan José Rendón*, caja 12.

Redón, J. J. (1971) Reseña Breve introducción a la computación lingüística de Paul L. Garwin, *Anales de Antropología*. 8: 313-314.

Swadesh, M. (1958-60). Plan de Trabajo y Desarrollo, *Colecciones Especiales de la Biblioteca Juan Comas*, Sección de Antropología, Caja 05,Exp. 06, Carácter C, núm. de registro 30, ff. 200-201.

Swadesh, M. (1960). Interrelaciones de las lenguas mayenses, *Anales del Instituto Nacional de Antropología e Historia*, XIII: 231-267.

Swadesh, M. (1961). Carta del Dr. Maurice Swadehs al Ing. Pablo Martínez del Río", 16 de febrero de 1961, *Colecciones Especiales de la Biblioteca Juan Comas*, Sección de Antropología, Caja 05,Exp. 06,

Carácter C, núm. de registro 30, f. 113.
Swadehs, M. (1966). Carta del Dr. Maurice Swadehs al Dr. Miguel León Portilla", 29 de agosto de 1966, *Colecciones Especiales de la Biblioteca Juan Comas*, Sección de Antropología, Caja 05,Exp. 06, Carácter C, núm. de registro 30, ff. 41-42.
Swadesh, M. Curriculum Vitae del Dr. Mauricio Swadesh Talnoper, *Colecciones Especiales de la Biblioteca Juan Comas*, Sección de Antropología, Caja 05,Exp. 06, Carácter C, núm. de registro 30, ff. 99-100.
Swinggers, P. (2016). Tras las huellas de Mauricio Swadehs: en búsqueda de una lingüística total, *Revista de investigación lingüística*, 19: 107-130.

The Spatial Humanities Kit

Matt Applegate
mapplega@gmail.com
Molloy College, United States of America

Jamie Cohen
jamesncohen@gmail.com
Molloy College, United States of America

This poster session showcases "the spatial humanities kit": a combination of gear, open source code, and teaching materials for narrative GIS projects (http://spatialhumanitieskit.org/). The kit was derived and assembled from two international mapping projects executed by students and guided by faculty at Molloy College and Hofstra University. The kit includes the following gear and code, all of which will be available for faculty to interact with at DH 2018: an introduction to GeoJson with code also applicable for Mapbox, Open Street Map, and ArcGIS, an Insta 360 Camera, Snapchat Spectacles, two Garmin ETreX 20x GPS devices with preinstalled maps, a Samsung 360 Camera, a chicken foot tripod, Samsung Gear Oculus HMD, user cell phone cameras, a Skyroam Global Hotspot, a GoPuck Qualcomm Charge 3.0, and a GoPro Session.

Project Description & Framework:

The spatial humanities kit is a durable toolset designed to fit in a backpack. The gear and code that it features are meant to combine and enhance two approaches to GIS related work in the humanities. First, the combination of gear included in the kit is designed for their user to narrativize the spaces that they map. Following Jason Farman's approach to locative media, the gear's use is predicated on two concepts in particular, "site specificity" and "urban markup." Site specificity, as Farman defines it, pertains to "the unique qualities of a unique location that cannot be transferred onto another place," whereas urban markup refers to "the various ways that narrative gets attached to a specific place in a city." The spatial humanities kit is designed to capture both.

In the summer of 2016 students at Molloy College traveled to Northeastern Ireland and documented their trip under faculty guidance via the spatial humanities kit and an Omeka archive (http://molloymediaarchaeology.org). Students documented and narrativized their experience of urban and rural space, historical sites, and religious sites, combining the unique qualities of each location (GPS coordinates, landmarks, etc.) with a linear telling of their site specific experiences. In the summer of 2017, the project was refined and expanded to Hofstra University. Students used the kit under faculty guidance in Italy to research and report on social inequality, government corruption, recovery and revitalization, and media change in earthquake damaged L'Aquila, the Naples region of Scampìa, and the Roman town of Frascati (http://lhscmediaarchaeology.org).

Ultimately, both projects, especially in their map's function as an artifact, play with the spatial humanities use and function. Where our use of the kit has emphasized autoethnography, social good, and bringing accountability to historical narratives, the spatial humanities kit augments the discipline's preoccupation with space-time. Consider Ian Gregory's engagement with Doreen Massey's work in "Exploiting Time and Space: A Challenge for GIS in the Digital Humanities": "Time is needed to tell the story of how an individual place developed to become what it is now, however without space there is only one story and thus the risk that it is seen as the only possible story and the inevitable story." Thus far, the spatial humanities kit has expanded the narrative possibilities of humanities GIS projects by multiplying narratives about spaces that are mapped.

Interactive Experience:

Our proposed poster session will offer faculty the opportunity to learn what the spatial humanities kit is, how they can adopt it, and how students can operationalize it. In addition to the kit itself, we will offer faculty syllabi, access to the Molloy and Hofstra University projects, as well as the source code for our maps. Our goal is to maximize the use and function of the kit by making our work, and our student's work, more broadly available. Further, we aim to approach the interaction between tools for digital storytelling and approaches to spatial humanities differently by combining both toolsets with their attendant pedagogical applications.

References

Farman, Jason. *The Mobile Story: Narrative Practices with Locative Technologies.* Routledge, 2013.
Gregory, Ian. "Exploiting Time and Space: A Challenge for GIS in the Digital Humanities." *The Spatial Humanities*. Eds. David J. Bodenhamer, John Corrigan, and Trevor M. Harris. Indiana University Press, 2010

The Magnifying Glass and the Kaleidoscope. Analysing Scale in Digital History and Historiography

Florentina Armaselu
florentinaa@zoomimagine.com
University of Luxembourg, Luxembourg

Introduction

What is the meaning of scale in historical writings and migration narratives? Can digital tools and methods assist the detection of scale-related patterns in these categories of documents? May this enquiry be formalised into a system for scale analysis in texts? To address these questions, the paper combines theoretical background from historical, historiographical, linguistic and literary studies with digital tools and methods for text analysis and visualisation. The project is in an early phase; theoretical hypotheses and preliminary experiments are presented.

Methodology

Two types of corpora were considered: (1) historiographical - history writings mingling micro and global perspectives; (2) historical - migration narratives (autobiography). The first, in which variations of scale are clearly present, will serve to develop a prototype. The second, where representations of scale are more difficult to assess, will be used to test the approach.

Corpora

Although recent research in "global microhistory" (Trivellato, 2011) draws attention to the variable scale representation in history, the question of how this phenomenon is expressed through language in historians' discourse is less studied. Research enquiries may be related to: topics distribution pertaining to scale (local to global, micro to macro); "story" versus "study" distinctions (Kracauer, 2014: 122); epistemological explorations (Boudon, 1991). Corpus (1) samples: Brook (2009), Rothschild (2013), Wills (2001).

Corpus (2) is intended to East-West migration narratives, e.g. Kaminer (2011), Kassabova (2009), Verboczy (2017). Potential queries: representation of space and its scale-related particularities, e.g. the intimate, symbolic meaning, inspired by Bachelard (1957), of the old and new "home" (interior objects, house, street, city, country, continent) and its connections to geo-historical or cultural spaces, and a certain sense of belonging. Other elements could be considered: relations, names, events, time references.

Approach

The aim is to bridge "distant" and "close" reading, using zooming metaphor as an interpretative tool (Armaselu and Heuvel, 2017). Thus, a corpus/text can be explored via the hypothetical schema:

Level1: topic_X (obj_1, obj_2, …, obj_n)
Level2: topic_X.1 (obj_1.1, obj_1.2, …), topic X.2 (obj_2.1, obj_2.2, …), …
Level3: topic_X.1.1 (obj_1.1.1, obj_1.1.2, …), topic_X1.2, …, topic_X2.1, etc.

Where, 'obj_topic[.subtopic]' represents a whole/section/fragment of a document associated to a topic and a scale-related logic. The system will allow zooming-in/out the different topics, traversing the conceptual space, e.g. from general to specific, and accessing the corresponding objects. One of the challenges is that the levels hierarchy and the degree of granularity may not be unique but depend on different "perspectives". Corpus (1) can imply different viewpoints and objects grouped by topics on levels 1, 2: (a) world history – 17th, 18th century; (b) world history – trade routes, slavery; (c) world history – Europe, Asia, America. Some fragments generalise on world history, others discuss world trade routes between Europe, America and Asia, others narrow down to family history or paintings description. Like in a kaleidoscope, by rotating the device (changing the "magnifying-glass"), new patterns can emerge.

Proof of Concept (PoC)

The PoC phase (in progress) will test these hypotheses on corpus (1). Two experiments on Brook (2009) are presented below.

Figure 1 illustrates Paper Machines topics for each chapter. It is assumed that by combining these groupings with an analysis of the contexts where the corresponding words appear, e.g. co-occurrences, lexical chains, paths in a lexical-semantic hierarchy, a scale-related model of the text can be derived. Its levels may reflect how knowledge is organised, from synthesising, manipulating abstractions, through intermediate descriptions, to in-detail accounts referring to particular facts, persons, objects or quotations of sources.

Figure 2 shows a visualisation via Z-editor (Armaselu, 2010). The scalable layout in chapter 2 (created manually) is explored by zooming through the European hatters history in the fifteenth and sixteenth century, the opening of the beaver pelts Canadian supply and Champlain's fight with the Mohawks, the customs of wearing a hat and the rules of courtship in seventeenth century Netherlands, and, Vermeer's painting, *Officer and Laughing Girl,* illustrating these practices.

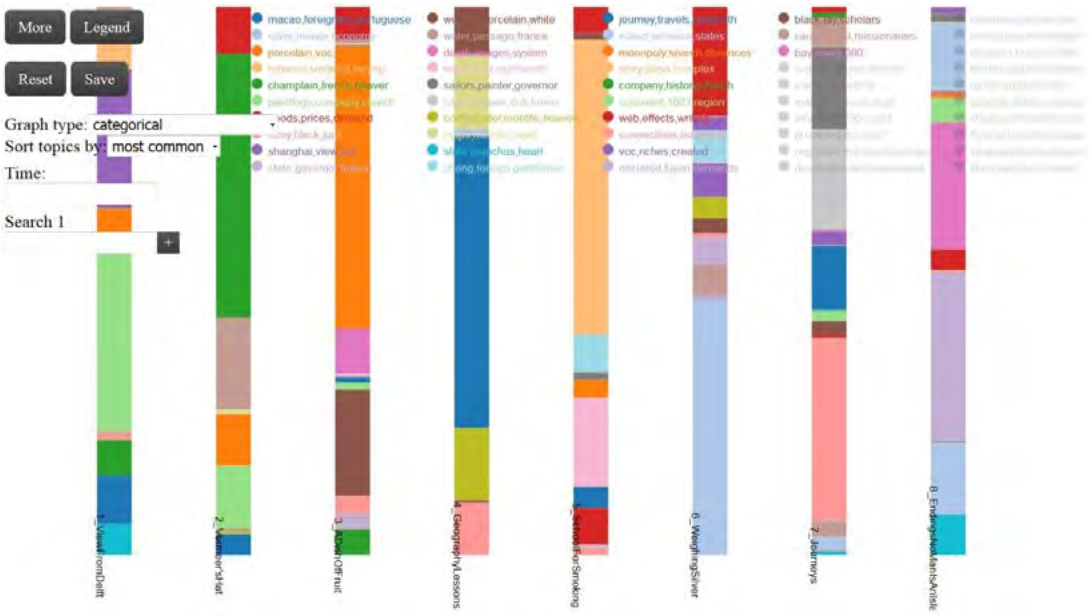


Fig. 1. *Vermeer's Hat*. Zotero - Paper Machines (topic modelling by subcollection/chapters)

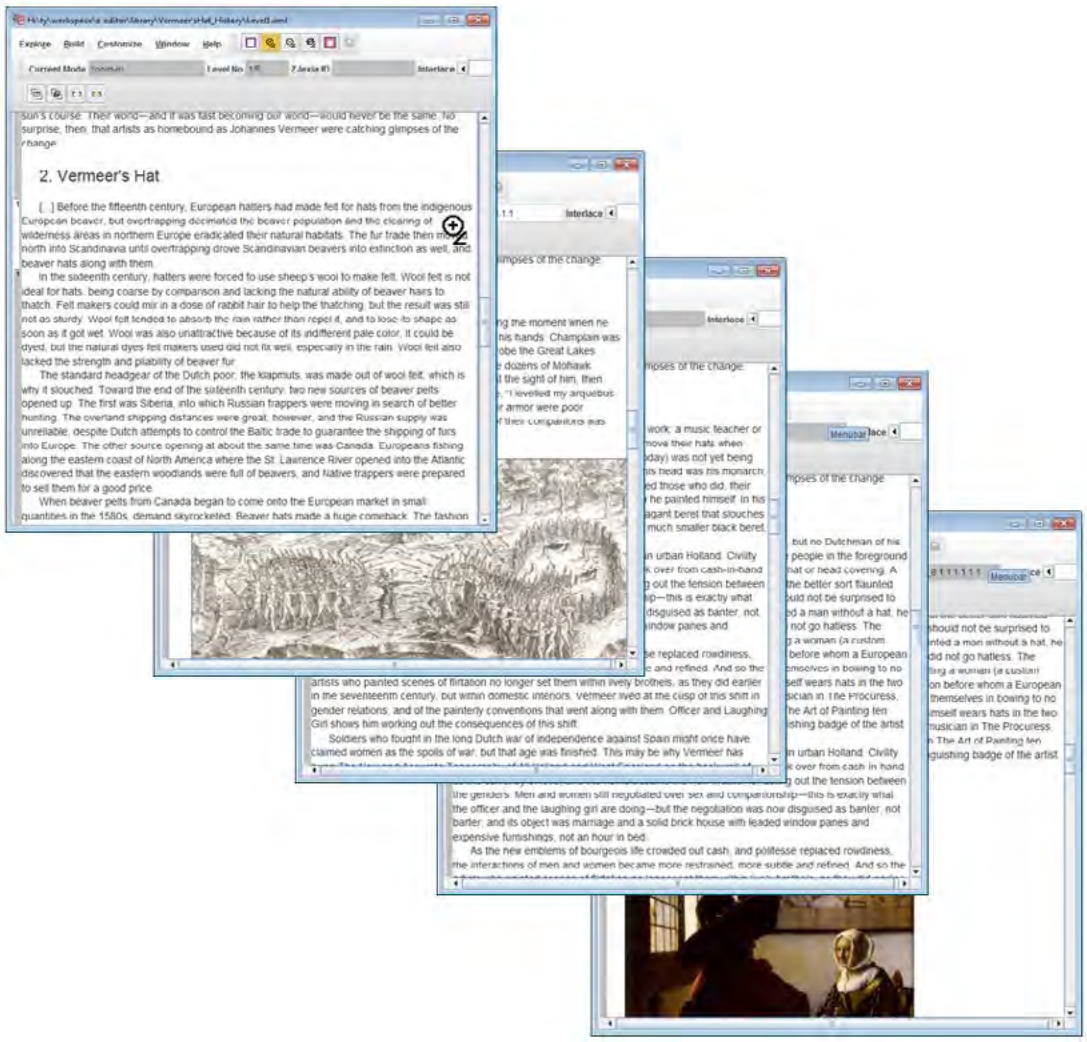


Fig. 2. *Vermeer's Hat*. Z-editor (zoomable text)

Tools/methods currently under testing: topic modelling (MALLET), textometry (TXM), lexical-semantic resources (WordNet), Named Entity Recognition (GATE), lexical chains and text structure (Morris and Hirst, 1991), visualisation (graphs, textual zooming). The PoC outcome will consist of insight into the advantages/limitations of these tools/methods in building a prototype for scale analysis.

Conclusion

The paper presents theoretical points and experiments for a system dedicated to scale analysis in historical/historiographical texts. By a combined approach, evoking the metaphors of the magnifying glass and the kaleidoscope, the system may allow both scale-related patterns detection and perspective change.

References

Armaselu (Vasilescu) F. (2010).Ph.D. Thesis, *Le livre sous la loupe : Nouvelles formes d'écriture électronique*, Papyrus, University of MontrealInstitutionalRepository.

Armaselu, F. and Heuvel, C. van den. (2017). "Metaphors in Digital Hermeneutics: Zooming through Literary, Didactic and Historical Representations of Imaginary and Existing Cities", In *Digital Humanities Quarterly (DHQ)*, Volume 11, Number 3.

Bachelard, G. (1957). *La poétique de l'espace*, PUF.

Boudon, P. (1991). *De l'architecture à l'épistémologie. La question de l'échelle*, PUF.

Brook, T. (2009).*Vermeer's Hat.The Seventh Century and the Dawn of the Global World*, Profile Books.

Kaminer, W. (2011). *Russian Disco,* Translated by Michael Hulse, Ebury Press.

Kassabova, K. (2009). *Street Without a Name: Childhood and Other Misadventures in Bulgaria*. New York: Skyhorse Publishing.

Kracauer, S. (2014). *History. The Last Things Before The Last,* Markus Wiener Publisher.

Morris, J. and Hirst, G. (1991). "Lexical Cohesion Computed by Thesaural Relations as an Indicator of the Structure of Text", In *Computational Linguistics,* Volume 17, Number 1, Association for Computational Linguistics.

Rothschild, E. (2013). *The Inner Life of Empires. An Eighteenth Century History*, Princeton University Press, 2011, paperback 2013.

Trivellato, F. (2011). "Is There a Future for Italian Microhistory in the Age of Global History?", *California Italian Studies*, 2(1).

Verboczy, A. (2017). *Rhapsody in Quebec. On the Path of an Immigrant Child.* Translated by Casey Roberts. Montréal: Baraka Books.

Wills Jr., J. E. (2001). 1688. *A Global History*, New York, London: W.W. Norton & Company.

Tools

GATE - General Architecture for Text Engineering, https://gate.ac.uk/.

MALLET - MAchine Learning for LanguagE Toolkit, http://mallet.cs.umass.edu/topics.php.

Paper Machines - http://papermachines.org/.

TXM – Textométrie project, http://textometrie.ens-lyon.fr/?lang=en.

Z-editor - http://www.zoomimagine.com.

WordNet - https://wordnet.princeton.edu/.

Encoding the Oldest Western Music

Allyn Waller
awalle18@g.holycross.edu
College of the Holy Cross, United States of America

Toni Armstrong
toarmstrong@clarku.edu
Clark University, United States of America

Nicholas Guarracino
nmguar18@g.holycross.edu
College of the Holy Cross, United States of America

Julia Spiegel
jrspie19@g.holycross.edu
College of the Holy Cross, United States of America

Hannah Nguyen
hnguye19@g.holycross.edu
College of the Holy Cross, United States of America

Marika Fox
marfox@clarku.edu
Clark University, United States of America

Problems of Encoding

This project describes a system for digitally encoding neumes and corresponding text in parallel aligned documents in order to create a digital, diplomatic edition of chant texts with neumes. Neumes are graphic marks denoting relative changes in pitch; they predate staff notation and are written above text. Each neume can be marked with performance variations, called episema or liquescence. They also include musical directions abbreviated in Latin, with 15 significative letters such as 't' for 'tenere' to indicate holding a note longer. There are at least a dozen styles of neumes, each of which has its own set of graphical symbols, like different fonts, to represent the same neumes.

A diplomatic edition of a neumed chant text must record the neumes as characters, not as absolute pitches.

It also must align neumes with text, as they are visually aligned by syllable in chant manuscripts.[1]

The 'Virgapes' system is based on a four-part encoding scheme for neumes that is flexible, extensible, and universal.[2] We have also developed a parallel document structure to align separate documents of text and neumes.

The 'Virgapes' System

In the Virgapes encoding, each neume is represented with a four-part code point. Each part is an integer standing for an aspect of the neume. The first integer denotes the number of pitches in the neume. The second integer is an arbitrarily assigned identifier within that group.[3] The system is flexible; it can expand to accommodate new or lesser known neumes. The third integer indicates the presence of episema, 1 for presence, 0 for absence. Likewise, the fourth notes liquescence in the same binary pattern.

For example: virga is a one-pitch neume, encoded as 1.1.0.0 in absence of episema or liquescence; pes is a neume of two ascending pitches, encoded (if liquescent) as 2.2.0.1.

The inclusion of episema and liquesence allows editors to note graphic marks indicating performance changes without imposing meaning. Our system also allows for specified searching: for all instances of virga or only instances of virga with episema, depending on the needs of analysis.

Parallel Aligned Documents

In addition to encoding neumes, we align transcriptions of neumes with transcriptions of texts. In a manuscript, this is done graphically: the neumes appear above the text.

In our digital editions, we create two parallel documents aligned by canonical citation using a Canonical Text Services (CTS) URN system to uniquely identify each passage.[4] With this, the two documents share a work hierarchy and a passage hierarchy. Consider the URN: 'urn:cts:chant:antiphonary.einsiedeln121.text:11.introit'. The CTS namespace is 'chant' for the domain of chant texts. The group is 'antiphonary' for the type of chant book. The specific work is Einsiedeln 121. The last section notes the version, text or neume.

The second portion of the URN system is a passage hierarchy, which subdivides the work hierarchy. A parallel would be the act, scene, and line in plays. It first identifies the feast day using numbers delineated in the *Antiphonale Missarum Sextuplex*.[5] Then, the subsection: introit, verse, etc, with further identifying numbers for graphically separated passages.

The URN system provides a citation scheme to align the texts; within the documents they are aligned by syllables, as each syllable must have at least one neume. This also provides a check for our encoding—there must be equal syllables in the text and neume document.

Digital encoding of neumes allows for advanced searching and analysis. With our two-part encoding solution, it is possible to search for repeated musical sequences, to determine if Zipf's law applies to neumes, or to analyze musical texture based on the neume:text ratio.

1 Among chant scholars, the most important digital resources are the manuscript databases of the Cantus Index network. (See http://cantus.uwaterloo.ca/) These datasets include information about the manuscripts themselves in addition to the encoding of text and music. Unlike the Cantus system, however, we encode staffless neumes without imposing interpreted equivalences to later musical notation on staffs.
Of the XML systems, the most significant is the Music Encoding Initiative (MEI) (http://music-encoding.org/documentation/3.0.0/neumes). It is largely inspired by work at Tübingen. Our system also allows encoding of basic neumes with extended properties (liquescence, episema) in a specified syntax, enabling us to take account or ignore these properties in computational manipulation. Neither the current XML schemes nor the Cantus Index allow these properties to be optional.

2 Called 'Virgapes' for the first one and two-note neumes, *virga* and *pes*.

3 These are available from our Github repository: https://github.com/HCMID/chant.

4 This system was developed as part of the CITE Architecture for the Homer Multitext Project 2010-18, and applied to this project: http://cite-architecture.github.io/ctsurn/.

Creating a Digital Edition of Ancient Mongolian Historical Documents

Biligsaikhan Batjargal
biligsaikhan@gmail.com
Research Organization of Science and Technology
Ritsumeikan University, Japan

Garmaabazar Khaltarkhuu
garmaabazar@gmail.com
Mongolia-Japan Center for Human Resources Development,
National University of Mongolia, Mongolia

Akira Maeda
amaeda@is.ritsumei.ac.jp
College of Information Science and Engineering
Ritsumeikan University, Japan

Introduction

In this poster, we introduce a digital edition of the Altan Tobchi, a Mongolian historical manuscript written in traditional Mongolian script. The Text Encoding Initiative (TEI) guidelines were adopted to encode the named entities. A web prototype was developed for digital humanities scholarship for utilizing digital representations of ancient Mongolian historical manuscripts as scholarly tools. The proposed prototype has the capability to display and search TEI encoded traditional Mongolian text

5 A standard chant reference work compiled by Dom Hesbert in the early 20th century. It contains transcriptions of six important sources of Gregorian chant.

and its transliteration in Latin letters along with the highlighted named entities and the scanned images of the source manuscript. This poster discusses how to develop a digital edition of Mongolian historical documents.

Mongolian manuscripts

Mongolian historical documents have been written in numerous scripts, i.e., the traditional Mongolian script, Square or Phags-pa script, Todo or Clear script, Soyombo script and Horizontal square script (Shagdarsuren, 2011). Among them, the traditional Mongolian script is the most popular and longest-surviving script for over 800 years. This research focuses on the traditional Mongolian script.

In 1946, Mongolia has made language reforms to eliminate a difference between written and spoken Mongolian language, and the Cyrillic script was adapted to Mongolian. The spelling of modern Mongolian was based on the pronunciations in the Khalkha dialect, the largest Mongol ethnic group (Sečenbagatur et al., 2005; Svatesson et al., 2005). Such a radical change separated the Mongolian people from their historical archives written in traditional Mongolian script. Reading traditional Mongolian documents by using literacy in modern Mongolian is not a simple task. Thus, a digital text representation that explains a given manuscript in a modern Mongolian is helpful for users who want to read, search and browse ancient Mongolian manuscripts.

Mongolian manuscripts in the digital age

To the best of our knowledge, there are a small number of digital texts of ancient Mongolian manuscripts. A few ancient Mongolian manuscripts including (1) 'Qad-un ündüsün-ü quri-yangyui altan tobči neretü sudur' (The Golden Summary: Short history of the Origins of the Khans) a.k.a. "Little" Altan Tobchi and (2) the 'Asarayči neretü-yin teüke' (The Story of Asragch) have been converted to digital texts and made publicly available (Batjargal et al., 2012).

Information processing of Mongolian manuscripts

Batjargal et al. have developed the traditional Mongolian script digital library (TMSDL) (Batjargal et al., 2012), which can be used to access and retrieve historical manuscripts written in traditional Mongolian script using a query in modern Mongolian (Cyrillic). Moreover, Batjargal et al. also proposed a named entity extraction method (Batjargal et al., 2016), which extracts proper nouns from digitized text of ancient Mongolian documents using Support Vector Machine with 0.6993, 0.5679 and 0.6268 of precision, recall and F-measure respectively. These researches have motivated us to create a digital edition that reflects ancient Mongolian historical manuscripts.

Digital edition of Mongolian manuscripts

We utilized Edition Visualization Technology (EVT) for creating a digital edition of Mongolian manuscripts, which is encoded according to the TEI XML schemas and guidelines (Del Turco et al., 2014). As shown in Figure 1 and Figure 2, all the personal names and place names (Figure 3) in the Altan Tobchi are highlighted by using the results of a named entity extraction method (Batjargal et al., 2016) and the named entities' indices obtained from the "Qad-un ündüsün quriyangyui altan tobči –Textological Study" (Choimaa, 2002). We made the following customizations in EVT to make it suitable for Mongolian manuscripts in traditional Mongolian script.

Figure 1. Image-to-text link and personal names' highlights

Parallel-text editions with transliteration

The proposed prototype can present scanned image-based editions with two edition levels: (1) diplomatic interpretative and (2) transliteration. Transliteration is helpful for those who are not familiar with a script of a certain language but understands that language. Transliteration in Latin letters of Mongolian historical documents is popular among scholars.

There is a limited recommendation to encode transliterations in TEI. Soualah and Hassoun (Soualah & Hassoun, 2012) proposed to implement transliteration by using a specific model, which uses the *<ref>* element with the *@xml:lang*, *@target*, and *@type* attributes. However, we consider transliteration as a separate edition and use it as parallel-text editions as shown in Figure 2.

Supporting the traditional Mongolian script

A unique feature of traditional Mongolian script is displaying vertically, from top to bottom, in columns advancing from left to right. Due to poor support for traditional Mongolian script at the EVT, we customized it to display the scanned images at the top and the corresponding text in traditional Mongolian script below with the direction top to bottom and left to right (Figure 1). We also set to dis-

play text in traditional Mongolian script on the left, and the corresponding transliteration in Latin letters on the right that can be used to compare them.

Additionally, we added a simple virtual keyboard composed of 22 traditional Mongolian letters and their corresponding Latin letters to help users to input a Mongolian keyword to benefit free-text search and keyword highlighting (Figure 4).

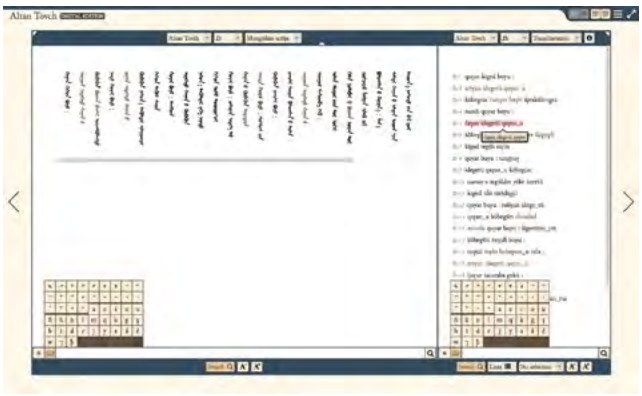


Figure 2. Parallel-text editions with transliteration

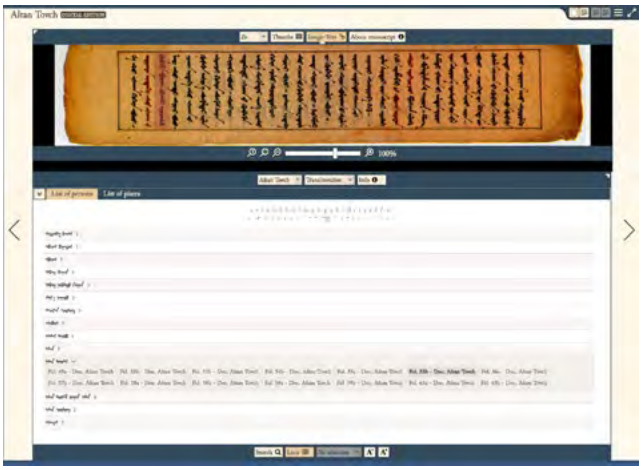


Figure 3. List of personal names in traditional Mongolian script

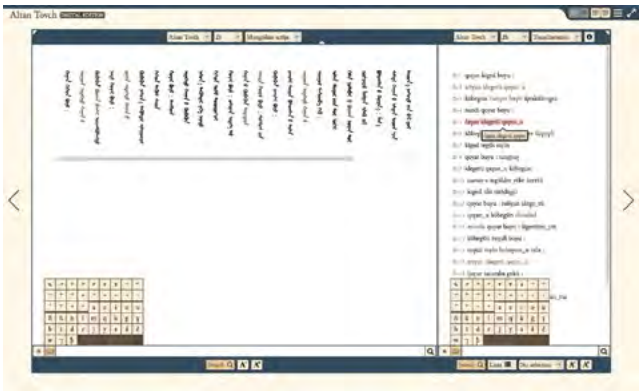


Figure 4. A simple virtual keyboard in parallel-text editions with transliteration

Summary and future directions

In this poster, we discussed our development of creating a digital edition (http://www.dl.is.ritsumei.ac.jp/AltanTovch/) of Mongolian historical manuscripts of the 13-16th century. The proposed method could be applied to other documents in Todo, Manchu, and Sibe, which are the derivative scripts of traditional Mongolian. We will further improve the proposed prototype by adding features to support critical editions.

We believe the proposed digital edition will enable users to search and browse ancient Mongolian manuscript with the highlights of historical figures and ancient place names.

References

Batjargal, B., Khaltarkhuu, G. and Maeda, A. (2016). *Named Entity Extraction from digitized texts of Mongolian Historical Documents in Traditional Mongolian Script, Conference Abstracts of Digital Humanities 2016*, pp. 734-735.

Batjargal, B., Khaltarkhuu, G., Kimura, F. and Maeda, A. (2012). Developing a Digital Library of Historical Records in Traditional Mongolian Script, *International Journal of Digital Library Systems*, 3(1): 33–53.

Choimaa, Sh. (2002). *Qad-un* ündüsün *quriyangyui altan tobči (Textological Study)*. vol. 1. Ulaanbaatar: Centre for Mongol Studies, National University of Mongolia, Urlakh Erdem, 2002. (in Mongolian).

Del Turco R. R., Buomprisco G., Pietro C. D., Kenny J., Masotti R., and Pugliese J. (2014) Edition Visualization Technology: A Simple Tool to Visualize TEI-based Digital Editions. *Journal of the Text Encoding Initiative*, Issue 8, DOI: 10.4000/jtei.1077.

Sečenbagatur Q., Tuyag-a B., and Ying U. (2005). *Monggul kelen-ü nutug-un ayalgun-u sinǰilel-ün uduridqal,* Hohhot: Öbür Monggul-un arad-un keblel-ün qoriy-a.

Shagdarsuren, T. (2011). *Study of Mongolian Scripts (Graphic Study of Grammatology)*, National University of Mongolia, Ulaanbaatar: Urlakh Erdem Khevleliin Gazar.

Soualah M. O., and Hassoun M. (2012). A TEI P5 Manuscript Description Adaptation for Cataloguing Digitized Arabic Manuscripts. *Journal of the Text Encoding Initiative*, Issue 2, DOI: 10.4000/jtei.398.

Svantesson J., Tsendina A., Karlsson A., and Franzén V. (2005). The Phonology of Mongolian, New York: Oxford University Press.

Shedding Light on Indigenous Knowledge Concepts and World Perception through Visual Analysis

Alejandro Benito
abenito@usal.es
University of Salamanca, Spain

Amelie Dorn
amelie.dorn@oeaw.ac.at
Austrian Centre for Digital Humanities Austrian Academy of Sciences, Austria

Roberto Therón
theron@usal.es
University of Salamanca, Spain

Eveline Wandl-Vogt
eveline.wandl-vogt@oeaw.ac.at
Austrian Centre for Digital Humanities Austrian Academy of Sciences, Austria

Antonio Losada
alosada@usal.es
University of Salamanca, Spain

The way we conceptualise our world is dependent on various aspects, differing with culture, time and language, and may even be subject to change over the years [5,6]. In this paper, we introduce a visual analysis tool that supports the exploration of indigenous knowledge concepts of a historic language collection, the Database of Bavarian dialects in Austria (DBÖ, dboe@ema), originally and partially collected by means of systematic questionnaires in the area of the former Austro-Hungarian empire. The collection we focus on in this work consists of 109 (original-conceptual) and 9 (supplementary) questionnaires, designed between 1913 and 1920, with answers (about 5 million paper slips). Around 11.100 persons of regional importance with various professional backgrounds and different roles in the compilation process were involved for almost a century [further info c.f. 1,8].

Our tool results from a series of iterations [3] of a custom-made, agile and collaborative workflow inspired by work from other authors [4] that was especially designed for the Digital Humanities (DH). The workflow places data visualisation as the main dialogue facilitator between the different stakeholders participating in the project. By applying user-centered design [2] techniques such as design probes [7], we can direct the development of several micro-prototypes towards the answering of fine-grained research questions. This prototype comprises the results of a full iteration of this iterative and incremental software development cycle.

Attending to the technical aspect of our approach, we employ different distant reading techniques to provide the user with a realistic view of the contents of the questionnaire and with visual mechanisms to help her form a mental image of the cultural connections of the terms at the time the questionnaires were made.

Our visualization plays with lights, colours and shadows to display related concepts, a relationship that is obtained by analysing coincident terms in the questions: the more times two or more terms appear together, the more important they all look in the visualization. The main visual component of our pilot tool is an adjacency matrix tweaked to meet the needs of the multivariate analysis task at hand. This matrix represents one single questionnaire of the collection and its rows and columns the questions conforming it. Each cell is colored to show the number of different concepts two questions have in common (richer coincidences are coloured in darker colours), forming different visual patterns that inform the user about the general distribution and importance of the concepts across the questionnaire.

The main matrix view is escorted by two other views placed on its right and at the bottom respectively: The first one offers an overview of the individual concepts in the questionnaire attending to the number of times they appear, each one represented by a coloured circle. Less frequent (and therefore, less important in our approach) concepts are moved to the top of the visualization, whereas the more important ones are placed at the bottom. Whenever the user hovers over one element, the cells in which that concept appears are in turn highlighted in an effect that imitates refraction of light, allowing for a rapid identification of particularities in the exploration process. At the bottom, the specific concept associations can be found in a similar way. More populated associations appear bigger in the visualization, whereas the more common are placed to the left. We provide an example below related to the use of colour terms:

Although thematically restricted to a single questionnaire (Q53), colours occur in questions throughout the entire collection offering valuable insights on their connection to cultural concepts. Within a single questionnaire, concept patterns/groupings across questions are revealed (see Figure 1). Interestingly, in the case of Q53 the most frequently occurring colour term *bleich* (pale) groups across questions towards the end of the questionnaire.

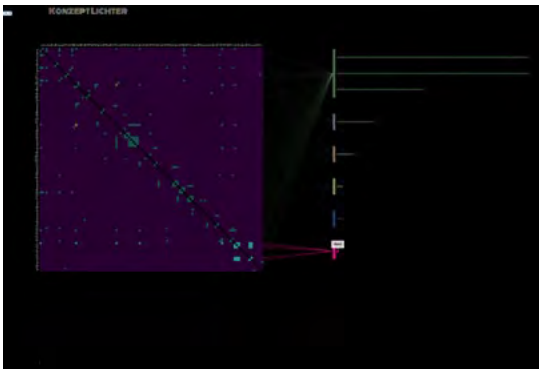


Figure 1: Visual distribution of 'bleich' (pale) grouped across questions in questionnaire 53.

Additionally, yellow (*gelb*) is the term/concept occurring most frequently across questions in questionnaire 85, thus playing an important role in the description of "The flora of our meadows / *Die Pflanzenwelt unserer Fluren* " (Q85) (see Fig. 2). Further, frequent collocations of colour terms in questions are revealed, which also shed light on the structuring of language and part of the conceptualisation of certain topics (see Fig. 3).

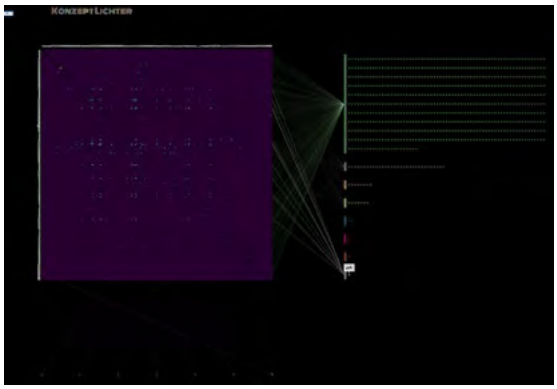


Figure 2: Distribution of 'gelb' (yellow) across questions in questionnaire 85.

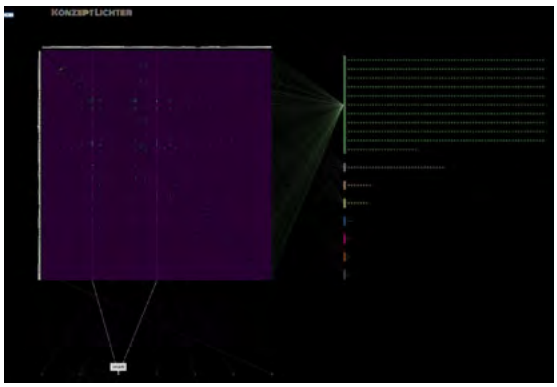


Figure 3: Visualisation of co-occurrence of terms 'rot-gelb' (red-yellow) across questions in questionnaire 85.

Note: Note: Preview of the prototype: https://concept-lights.herokuapp.com/ (Google Chrome only). Please share your remarks with us at explore@oeaw.ac.at. Thanks.

Data:

Datenbank der bairischen Mundarten in Österreich (DBÖ) | Database of Bavarian Dialects in Austria (DBÖ). Austrian Academy of Sciences: 11.2017.

Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema) | Database of Bavarian Dialects in Austria electronically mapped (dbo@ema). Ed. by Eveline Wandl-Vogt: Austrian Academy of Sciences: 2012 / 11.2017.

References

Abgaz, Yalemisew, et al.: "A Semantic Model for Traditional Data Collection Questionnaires Enabling Cultural Analysis." P*roceedings of W23 - 6th Workshop on Linked Data in Linguistics: Towards Linguistic Data Science, LREC 2018*, 21-29. http://lrec-conf.org/workshops/lrec2018/W23/index.html [last accessed: 26.04.2018]

Abras, C., Maloney-Krichmar, D. and Preece, J., 2004. User-centered design. Bainbridge, W. Encyclopedia of Human-Computer Interaction. *Thousand Oaks: Sage Publications*, 37(4), pp.445-456.

Benito, A., Therón, R., Losada, A., Wandl-Vogt, E. and Dorn, A., Exploring Lemma Interconnections in Historical Dictionaries. *2nd Workshop on Visualization for the Digital Humanities*. October 2017 - Phoenix, Arizona, USA.

Bernard, J., Daberkow, D., Fellner, D., Fischer, K., Koepler, O., Kohlhammer, J., Runnwerth, M., Ruppert, T., Schreck, T. and Sens, I., 2015. VisInfo: a digital library system for time series research data based on exploratory search—a user-centered design approach. *International Journal on Digital Libraries*, 16(1), pp.37-59.

'Concepts of the World': Publishing in Mexico's Indigenous Languages. https://publishingperspectives.com/2017/08/mexico-indigenous-language-publishers/ [last accessed: 26.04.2018]

De Beule, J. and De Vylder, B., 2005, January. Does language shape the way we conceptualize the world?. In *Proceedings of the Annual Meeting of the Cognitive Science Society* (Vol. 27, No. 27).

Gaver, B., Dunne, T. and Pacenti, E., 1999. Design: cultural probes. *interactions*, 6(1), pp.21-29.

Wandl-Vogt, Eveline. "…wie man ein Jahrhundertprojekt zeitgemäß hält: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I DINAMLEX) (mit 10 Abbildungen)" P. Ernst (Ed.), Bausteine zur Wissenschaftsgeschichte von Dialektologie / Germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen, Wien, 20. – 23. September 2006. Wien: 2008. Praesens, (pp. 93–112).

The CLiGS Textbox

José Calvo Tello
jose.calvo@uni-wuerzburg.de
Universität Würzburg, Germany

Ulrike Henny-Krahmer
ulrike.henny@uni-wuerzburg.de
Universität Würzburg, Germany

Christof Schöch
schoech@uni-trier.de
Universität Trier, Germany

Katrin Betz
katrin.betz@uni-wuerzburg.de
Universität Würzburg, Germany

Introduction

This poster presents the textbox published by the CLiGS group both from the perspective of creating the textbox and of using it for research. The poster will highlight key aspects of the manner in which the collections of literary texts included in the textbox have been compiled, annotated and published. Furthermore it suggests several ways in which the text collections can be used for research in literary studies. This poster aims to showcase the unique XML-TEI-based collections we make available and to encourage their re-use by others.

What is the textbox?

The CLiGS textbox is dedicated to making collections of literary texts in Romance languages freely available. It currently contains novels, novellas and short stories published between 1830 and 1940 in France , Spain, Italy, Portugal, and Spanish-America as well as plays published between 1640 and 1680 in France with a total of 357 texts or about 14 million words. The texts are published in XML-TEI as well as in plain text versions and include detailed document-level metadata. All texts are in the public domain and the XML-TEI markup including the metadata is published with a Creative Commons Attribution license (CC-BY) or in case of the Italian novels with a NC-SA-BY. The text collections are curated and published using a public GitHub repository. In addition, main releases are automatically archived on Zenodo.org, a long-term data and publications archiving service for researchers across Europe managed by OpenAire and supported by CERN (see Nielsen, 2013). Each release receives a DOI (Digital Object Identifier), providing the unambiguous identification and long-term availability of the resource.

Text selection

The individual text collections were created with various usage scenarios in mind, and each collection has been compiled in a slightly different manner. For example, the two collections of Spanish novels, the *Corpus of Spanish Novels (1880-1940)* and the *Collection of 19th century Spanish-American Novels (1880-1916)*, have been prepared to be used for authorship attribution. Accordingly, the two collections have been balanced with regard to the number of texts from different authors. The poster will give an overview of the sub-collections of the textbox and also about the principles guiding their compilations.

File Formats

Independently of their original source format (e.g. html or EPUB), the texts are prepared (with Python scripts or XSLT) according to a common TEI schema established by the CLiGS group . In addition to that reference format, each collection is made available in a simple plain text format automatically derived from the XML-TEI version, containing only the text included in the body of the novels and plays (in particular, excluding prefaces, other paratext, or notes) and with external metadata provided in tabular format.

Moreover, the collections of French, Spanish, Spanish-American, and Portuguese novels as well as the Italian short stories are made available in a version combining basic structural markup (chapter and sentence divisions) with token-level linguistic annotation, including lemma, part-of-speech, morphology, and basic semantic annotation using Freeling (cf. Padró and Stanislovsky, 2012) and WordNet (see Figure 1). Finally, the collection of French plays is not only available in XML-TEI, but also in the "Zwischenformat" developed by the DLINA group (Kampkaspar et al., 2015).

```
<s>
    <w form="Temía" lemma="temer" tag="VMII3S0" ctag="VMI" pos="verb" type="main"
     mood="indicative" tense="imperfect" person="3" num="singular" wnsyn="01780202-v"
     wnlex="verb.emotion">Temía</w>
    <w form="un" lemma="uno" tag="DI0MS0" ctag="DI" pos="determiner" type="indefinite"
     gen="masculine" num="singular" wnsyn="xxx" wnlex="xxx">un</w>
    <w form="despertar" lemma="despertar" tag="NCMS000" ctag="NC" pos="noun" type="common"
     gen="masculine" num="singular" wnsyn="05678745-n" wnlex="noun.cognition">despertar</w>
    <w form="lúgubre" lemma="lúgubre" tag="AQ0CS00" ctag="AQ" pos="adjective"
     type="qualificative" gen="common" num="singular" wnsyn="xxx" wnlex="xxx">lúgubre</w>
    <w form="." lemma="." tag="Fp" ctag="Fp" pos="punctuation" type="period" wnsyn="xxx"
     wnlex="xxx">.</w>
</s>
```

Linguistic annotations in an XML format that is a minimal departure from the TEI standard to allow multiple token-level annotations

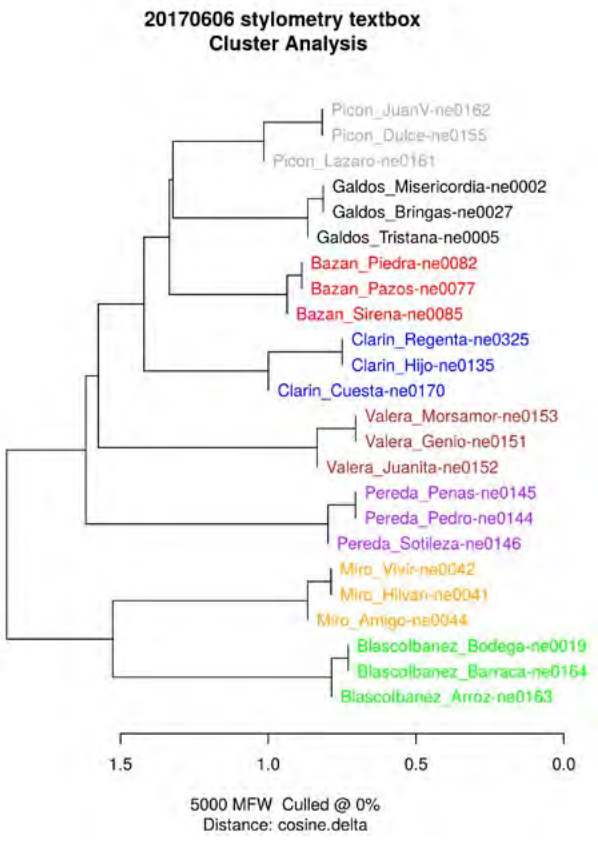
Metadata

Besides the administrative metadata like license, responsibility etc. the collections focus on descriptive metadata. There are four main areas about which information is documented: metadata concerning the authorship (VIAF, name, country, gender), metadata concerning the literary

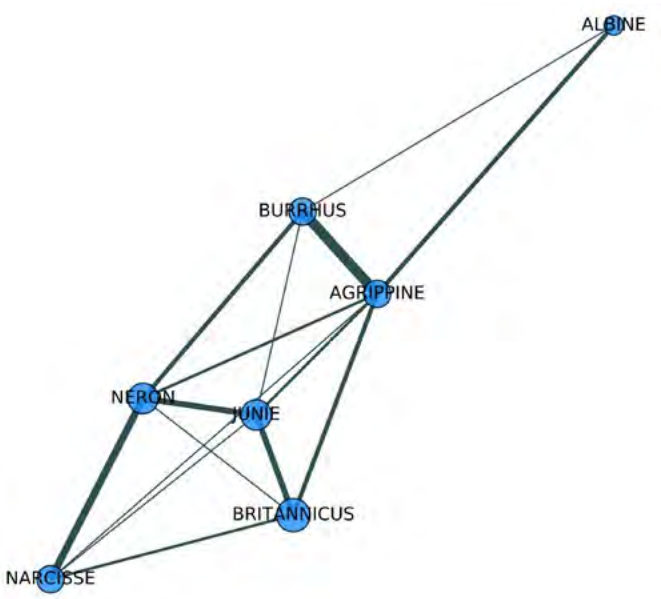
work and editions (VIAF or other identifier, extent of the texts, print and the digital source), and finally metadata concerning the genre: Since the main focus of the project is literary genre, a considerable part of the metadata is directly connected to it. Any reference to genre in the title of the work is collected as a genre label. Besides that, a hierarchical system is used, comprising supergenre (e.g. "narrative" or "drama"), genre (that is, novels or novellas), subgenre (the subtype of the novel, for example "adventure novel" or "political novel") and subsubgenre (optional, used for further differentiations like "war novel").

Usage Scenarios

There are many possible use cases for the textbox collections. The poster will demonstrate some results of these methods from the areas of authorship attribution (using the stylo package for R; Eder et al., 2016), network analysis (using NetworkX in Python), and topic modeling (using MALLET with "tmw" for Python). These scenarios are intended not only as examples of analyses conducted within the CLiGS group, but also as suggestions for potential users of the CliGS textbox, Figure 2 and 3 demonstrate some results for authorship attribution and network analysis.



Authorship attribution, results of cosine delta on the Corpus of Spanish Novels (cf. Smith and Alridge, 2011; Evert et al., 2017)



Character network based on number of words spoken in mutual presence (represented by the thickness of the lines), for Jean Racine's tragedy Britannicus (1669)

References

Eder, M., Kestemont, M. and Rybicki, J. (2016). Stylometry in R: A package for computational text analysis. In *The R Journal*, 16 (1): 1-15.

Evert, S., Proisl, T., Jannidis, F., Reger, I., Pielström, S., Schöch, C. and Vitt, T. (2017). Understanding and explaining Delta measures for authorship attribution. In *Digital Scholarship in the Humanities*, 32 (suppl_2): ii4-ii16. doi: 10.1093/llc/fqx023 https://academic.oup.com/dsh/article/doi/10.1093/llc/fqx023/3865676/Understanding-and-explaining-Delta-measures-for (accessed April 26 2018).

Kampkaspar, D., Fischer, F. and Trilcke, P. (2015). Introducing Our 'Zwischenformat'. In *Network Analysis of Dramatic Texts*. https://dlina.github.io/Introducing-Our-Zwischenformat/ (accessed April 26 2018).

Nielsen, L. H. (2013). ZENODO – An innovative service for sharing all research outputs. In *Zenodo*. doi: 10.5281/zenodo.6815 http://doi.org/10.5281/zenodo.6815 (accessed April 26 2018).

Padró, L. and Stanislovsky, E. (2012). FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference (LREC 2012)*. Istanbul, Turkey: ELRA. http://nlp.lsi.upc.edu/publications/papers/padro12.pdf (accessed April 26 2018): 2473-2479.

Smith, P. W. H. and Alridge, W. (2011). Improving Authorship Attribution: Optimizing Burrows' Delta Method. In *Journal of Quantitative Linguistics*, 18(1): 63-88. doi: 10.1080/09296174.2011.533591.

CITE Exchange Format (CEX): Simple, plain-text interchange of heterogenous datasets

Christopher William Blackwell
christopher.blackwell@furman.edu
Furman University, United States of America

Thomas Köntges
thomas.koentges@gmail.com
University of Leipzig

Neel Smith
nsmith@holycross.edu
The College of the Holy Cross, United States of America

Introduction: Sharing text libraries and data collections for teaching and research in the humanities

Source text collections and other complex datasets can be very difficult to share and reuse, and especially difficult to aggregate and disaggregate. CEX, CITE Exchange is a plain-text, self-documenting, technology-agnostic format for capturing citable texts, data collections, and arbitrary relationships among citable data at any level of granularity. CEX is based on the CITE/CTS architecture[1] and it positions itself as an alternative and complement to TEI XML and relational database schemas. TEI XML is a great archival format for storing textual information and metadata of individual editions. Managing and sharing text collections, however, can be cumbersome, especially if you only want to share a collection of excerpts based on hundreds of individual TEI XML files. When teaching text-heavy humanities disciplines, such as history, literature or classics, scholars are constantly faced with the problem of creating a source-text collections (that is, a corpus of excerpts of a bigger corpus that is deemed a representative sample able to answer a scholarly investigation), and the challenge of easily sharing this newly generated collection with students and colleagues. Based on current forms of data exchange, scholars and their students facing this task needed to have intimate knowledge of either database solutions like eXistDB[2] or of API-calls[3]. CEX circumvents this problem by simplifying the format of exchanging texts and related objects following the OCHO2 principles laid out in the CITE/CTS architecture[4].

Likewise, data collections (coins, geo-spatial data, manuscript folios, etc.) are efficiently served intact by relational databases. Extracting subsets, sharing datasets in whole or in part, and aggregating disparate collections with schemas can be very difficult. CEX, as an exchange format, simplifies this.

This paper is directed to two types of scholars: technology-savvy colleagues who want to discuss simple interchange formats for data-sets and colleagues who want to build, analyze, and exchange soure text collections with fellow researchers and students. The paper will introduce CEX, its design, utilities, and code libraries for creating, validating, and manipulating it, and examples of two types of end-user applications: applications that help to build CEX collections and applications that enable students and scholars to perform natural language processing on exchanged CEX collections. In the first part of the paper we will describe the format and structure of CEX, while the second part showcases sample applications.

The CEX format

CEX is based on clearly defined data models for texts and data collections. These data models define semantics of scholarly primitives. CITE and CTS URN citations capture the semantics of the objects they identify. CEX defines catalogs documenting repositories of texts and collections, and blocks of data capturing the data of the texts and collections themselves.

A CEX file is a plain text, UTF-8 file, containing blocks for distinct types of data. The CEX specification provides blocks for:

- Text Catalogs
- Textual Data
- Collection Catalogs
  - Collection Property Definitions
  - Collection Data
- Extensions to Collections, e.g. "Image Collections"
- Relations among citable objects
- Data models formally specifying further aggregation of primitives

Each text-block consists of a header line, followed by data records. Each line is a record, and fields within the lines are separated by a delimiting character ("#" is the default, but this is configurable).

Blocks are optional. A CEX file may contain only textual data, only collection data, or a combination of these. We will demonstrate using CEX files that contain millions of words of textual data and hundreds of thousands of data-records for collections.

In this paper, we will present these blocks, and the clearly defined abstract generic data models that they implement. ## Sample applications

We will demonstrate a sampling of utilities, services, and applications for creating, valdating, browsing, and analyzing scholarly data from CEX-formatted text files. All of these are openly licensed, with source code freely available on GitHub.

1 http://cite-architecture.github.io.
2 http://exist-db.org/exist/apps/homepage/index.html.
3 http://capitains.org.
4 http://cite-architecture.github.io.

Corpus building software: Brucheion and TEItoCEX

TEItoCEX[5] is a CTS TEI XML parser. It is written in Go and can be compiled to any operating system, including OS for Raspberry Pi [6]. It can be copied to the parent folder of multiple TEI CTS XML files. Once executed it will produce a valid CEX collection from all the XML files. **screenshot1**

Brucheion is a browser-based transcription desk written in Go and based on ICT2. **screenshot2** While the backend is based on the fast BoltDB, users can export and import CEX files to share transcriptions or feed them into analytical applications, such as (Meletē)ToPān v.0.3. **screenshot3 and screenshot4**

Microservices based on CEX: Citemicroservices written in Go and Scala

There are currently two implementations of CEX-based microservices. The first implementation we created is a prototype that helps the end-user retrieve textual nodes from CEX files via an HTTP API by simply using the barebone file system of the server as a backend.[7] It is written in Go and can be compiled to any operating system. The gitHub repository also includes executables for all major operating systems and documentation for usage.[8] A useful feature is that one can also retrieve textual nodes from remote CEX repositories by modifying config.json. Simplicity of implementation and usage was the driving force in developing this prototype. The work on this prototype spawned another implementation in Scala that could make use of all the features of the already existing Scala CTS/CITE library.[9]

Another implementation written in Go is currently under development. This implementation includes features of the Scala library and some additional natural language processing capabilities and that make use of BoltDB in its backend.[10]

Exploring CEX collections through Topic Modelling: (Meletē)ToPān v.0.3 and Metallō

(Meletē)ToPān is a GUI for topic modelling introduced at DH2016 in Krakow. It was written in R and is published on Zenodo[11] and Github[12]. Initially ToPān supported the predecessor of CEX, 82XF, which was developed in discussion at DH2016.[13] Since v.0.3, it also supports CEX. Using the CEX files for topic modelling gives users easy but precise control over which documents should be included in the analysis. It has become possible for any scholar to build a source text corpus with TEItoCEX and then feed it into ToPān. The researchers can then use Metallō[14], a commandline based topic model exploration tool, to have a closer look at the created result.

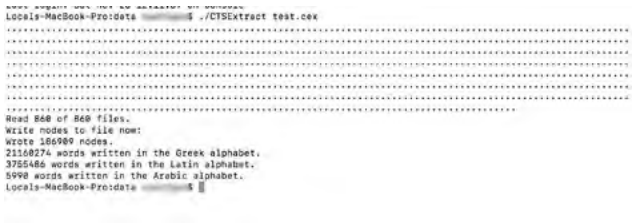
The paper will include a live demonstration of TEItoCEX, ToPān, and Metallō. Scholars visiting this talk will be able to use and build CEX collection for analysis, teaching, and research.

Browsing Integrated Text & Data Libraries with the CiteApp

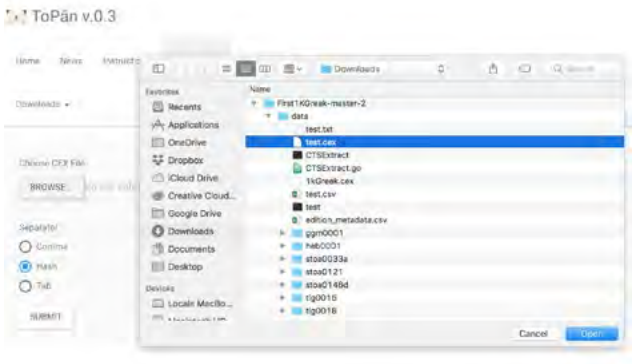
CiteApp is a "zero-config" web application written in ScalaJS that allows browsing and searching of libraries of text, collection data, and images. CiteApp itself consists of a single HTML file and can run locally or on a server. CiteApp reads data in the form of CEX files (with, optionally, an accompanying directory of image files). CiteApp allows browsing texts in a library, browsing parallel texts and translations, querying texts for strings, tokens, and NGrams, browsing data-collections and querying their records, and browsing collections of images, querying metadata, and interacting with zoomable views of high resolution images. Images may be local or hosted remotely.[15]

The audience for this talk will be able to take home a copy of CiteApp with sample libraries in CEX format of byzantine manuscript images and editions of the texts they contain.

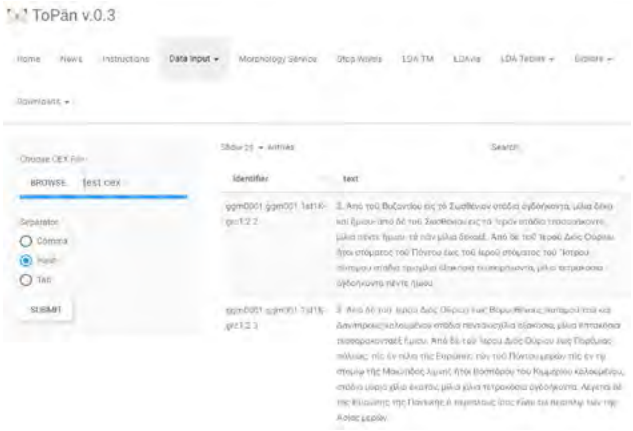
Screenshots of CEX-Capable Applications



TEItoCEX: Converting libraries of TEI documents to easily shared CEX files

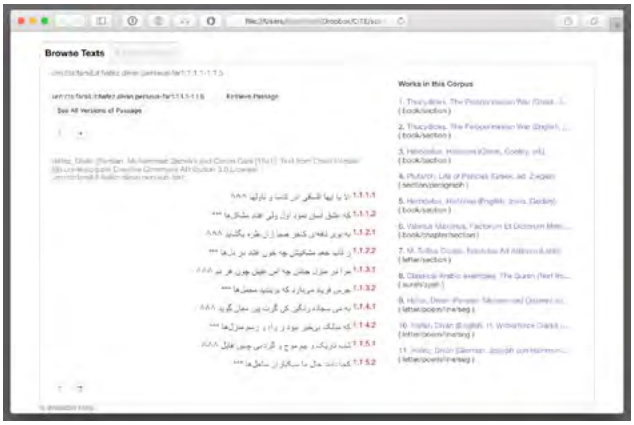


5 https://github.com/ThomasK81/TEItoCEX.
6 https://www.thepolyglotdeveloper.com/2017/04/cross-compiling-golang-applications-raspberry-pi/.
7 https://github.com/cite-architecture/citemicroservices.
8 https://github.com/cite-architecture/citemicroservices.
9 https://github.com/cite-architecture/scs-akka.
10 https://github.com/ThomasK81/Nucleus.
11 https://zenodo.org/record/556708#.WhhosLbMwll.
12 https://github.com/ThomasK81/ToPan.
13 http://neelsmith.github.io/2016/07/13/82xf_v2/.
14 https://github.com/ThomasK81/TopanTerminalBrowser.
15 https://github.com/cite-architecture/CITE-App.

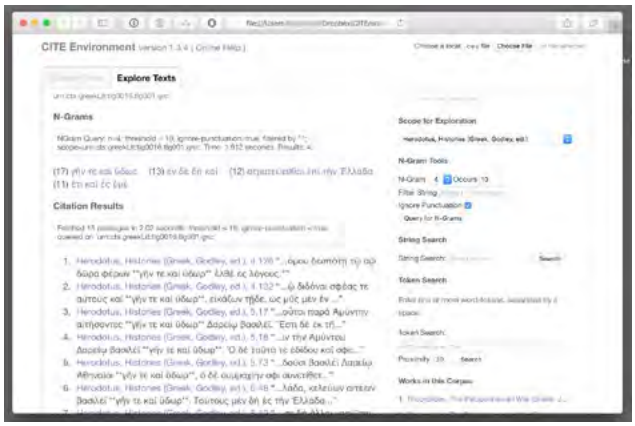


TEItoCEX
(Meletē)ToPān: Topic Modeling files in CEX format

Brucheion: Integrated Image and Textual data



Brucheion
CiteApp: Browsing a multilingual text library
CiteApp
CiteApp: Searching for NGramsTEItoCEX



CiteApp

References

Smith, D. Neel, and Gabriel Weaver. "Applying Domain Knowledge from Structured Citation Formats to Text and Data Mining: Examples Using the CITE Architecture." Text Mining Services, 2009, 129.

Blackwell, C., and D.N. Smith. "A Gentle Introduction to CTS & CITE URNs." Homer Multitext Project Documentation, November 2012. http://www.homermultitext.org/hmt-doc/guides/urn-gentle-intro.html.

DeRose, Steven J., David G. Durand, Elli Mylonas, and Allen H. Renear. "What Is Text, Really?" ACM SIGDOC Asterisk Journal of Computer Documentation 21, no. 3 (August 1997): 1–24. https://doi.org/10.1145/264842.264843.

Digitizing Whiteness: Systemic Inequality in Community Digital Archives

Monica Kristin Blair
mkb4rf@virginia.edu
University of Virginia, United States of America

Introduction

In recent years, the digital revolution has transformed the idea of the archive. Once associated with grand library buildings filled with ancient books and artifacts, today scholars are making archives out of social media, metadata, and all things born digital. Traditional archives are also revolutionizing the way that users interact with their objects by taking digitization to the next level with techniques like linked data and photogrammetry.

However, archivists and scholars are not the only ones experimenting with digital curating. Online communities are making their own virtual collections. Librarians and digital humanists, including those at the United States Library of Congress, have encouraged and assisted people

in creating these "community digital archives" (LeFurgy). But how should scholars respond when these community digital archives are linked to institutions with extremely fraught histories of white supremacy, ableism, homophobia, transphobia, xenophobia, or sexism? This study explores that broad question by analyzing community digital archives created by alumni of historically segregated K-12 private schools in Virginia, USA to investigate the form, function, and ethics of studying community digital archives attached to historically prejudiced institutions.

Research questions

How does institutional inequality manifest itself in digital community archives?

How should scholars read community digital archives that are public, but may not intended for outside audiences?

What commonalities and differences exist between these repositories, traditional archives, and digital archives curated by scholars and archivists? How do those similarities and differences affect how scholars should interact with these communities?

What are the ethics of using these archives for scholarly research?

Literature

Scholars like Bergis Jules and Piia Varis have worked to define the ethics and best practices of archiving digital sources (Jules, 2016; Varis, 2014). Moreover, several digital projects such as DocNow and Take Back the Archives have modeled how scholars can engage with digital archiving methods to advance scholarly questions and social justice. Most of this scholarship has focused on archiving the experiences and activism of marginalized groups. That work is both vital and admirable. This study contributes to this literature by examining the opposite end of the spectrum. By looking at how white communities that have supported segregated education use community digital archives, I illuminate how these groups remake and reaffirm systemic inequality in the digital landscape. In the process, I also examine the ethics of analyzing and writing about digital communities that have white supremacist roots.

Case study

This study uses the historical subject of segregation academies as its basis. Segregation academies are private schools founded in the southern United States during the 1950s and 1960s in order to provide segregated education for white students whose families refused to comply with court-ordered school desegregation following the United States Supreme Court case Brown v. Board of Education.

White supremacy was at the heart of these schools' foundation, and this study examines how whiteness is reflected in the digital community archives of these schools' alumni pages on Facebook and Classmates.com. Both Classmates and Facebook are for-profit businesses, but it is former students, teachers, and administrators who post old photographs, pamphlets, yearbooks, and personal memories of their times at these intuitions on these websites. The memorabilia they gather and publish serves as an important window into the past, and their contemporary comments reveal the ways that white southerners navigate their personal ties to this history of white supremacy in the contemporary digital landscape.

Facebook and Classmates do not follow the practices of traditional archives as outlined by archivist Kate Theimer (Theimer, 2012); nonetheless, both platforms exhibit some of the classic characteristics of archives. They are repositories for institutional and personal histories. Donors contribute to these archives by providing digital copies of their personal papers, photographs, and yearbooks. The aim of these groups is to preserve the history of an institution and, in doing so, craft historical narratives about said institutions. Ultimately, the content, organization, and narratives on these websites are fundamentally shaped by the motive of the curators of these archives. The patrons who create these archives do so out of sentimentality about their former-schools. The web hosts, Classmates and Facebook, profit from this nostalgia, and thus have no incentive to challenge the whitewashed school histories their users promote. Thus, sanitized, color-blind versions of these schools histories prevail on these digital community archives, thereby erasing decades of systemic inequality and prejudice from view.

References

Documenting the Now. http://www.docnow.io (accessed 29 November 2017).

Jules, Burgis. (2016). Some Thoughts on Ethics and DocNow. *Medium*. https://news.docnow.io/some-thoughts-on-ethics-and-docnow-d19cfec427f2 (accessed 29 November 2017).

LeFurgy, Bill. (2013). Resources for Community Digital Archives. Library of Congress. https://blogs.loc.gov/thesignal/2013/06/10-resources-for-community-digital-archives (accessed 29 November 2017).

Take Back the Archive. University of Virginia. http://takeback.scholarslab.org (accessed 29 November 2017).

Theimer, Kate. (2012). Archives in Context and as Context. *Journal of Digital Humanities,* Vol. 1, No. 2, Spring 2012. http://journalofdigitalhumanities.org/1-2/archives-in-context-and-as-context-by-kate-theimer/ (accessed 30 January 2018).

Varis, Piia. (2014). Digital Ethnography. *Tilburg Papers in Cultural Studies*. https://www.tilburguniversity.edu/upload/c428e18c-935f-4d12-8afb-652e19899a30_TPCS_104_Varis.pdf (accessed 30 January 2018).

How to create a Website and which Questions you have to answer first

Peggy Bockwinkel
peggy.bockwinkel@ilw.uni-stuttgart.de
University of Stuttgart, Germany

Michael Czechowski
mail@dailysh.it
University of Stuttgart, Germany

Initial situation and target group

Our target group are (digital) humanities scholars who want to present small, non-funded projects on their own website but have little or no experience with web design and development. Always the same questions have to be asked at the beginning of a website project to find out how complex or easy it is to implement the project. These circumstances are excellent conditions to offer a "how to"-flowchart with different applications as decision support. The poster is therefore designed as a decision flowchart and gives an overview of the possibilities and limitations of certain applications. All applications are differentiated according to whether static, dynamic or semi-dynamic websites can be built with them. Depending on which content should be presented and how, even beginners with little or no previous knowledge can create a website or get a feeling for when the support of a web developer is necessary. The questions that are answered in the flowchart concern...

1. … the complexity of the site (see 2.1. for details).
2. … the reason why the website is build, e.g. to publish a digital edition or to present research.
3. … the format of the texts to be published: Do they have a uniform format? Can they be read dynamically?
4. … the level of knowledge / technical affinity of the person creating the website, but also the capacity in terms of time and/or manpower and/or budget, i.e. for example whether it is possible to employ a student assistant or even a professional web developer.
5. … the scalability: is it possible to estimate how the project will develop, e.g. with regard to complexity? To what extent can the website be expanded and where are the limits, e.g. how far can you get with the individual technology stacks?
6. … hosting (see 2.2. for details).
7. … sustainability.

Questionnaire

Of all the relevant questions, two are presented here as examples. Nevertheless, the flowchart is made available on **github** so that as many scholars as possible can benefit from it.

What kind of data should be published?

If it is data that can be copied and pasted onto the website, a static website can be created with the website generator **jekyll**, **omeka** or the content management system **wordpress**. This can look like this, for example: http://www.germanliteratureglobal.com/

If it is data, that requires several tabs and contains recurring queries, it is a dynamic website. This type of website is also required if data is to be made available for downloading (a database is required for larger amounts of data). This can look like this, for example: http://www.berliner-intellektuelle.eu/

If your data is available in TEI format, the **TEIpublisher** or the **EVT** are good choices for publication. If you are moving in the dynamic area, the effort can quickly become very high and extensive knowledge of website development is necessary. In this case you should contact a web designer.

How can the website be published (Hosting)?

If you have chosen **jekyll**, hosting is very easy via **github** or **githubpages**. There are no additional costs. If you have access to a webserver, you can always use it. Often universities offer such server systems for its employees or even students.

Discussion

In the course of this project, we were confronted with various issues that all revolve around sustainability in the broadest sense:

What if the formats presented here are obsolete? This risk exists for any technology application. This project is meant to be a pragmatic guide. We cannot solve the problem, but we use tools that are freely available and extendable.

If a website should be sustainable, i. e. available in the long term, where should it be hosted to guarantee long-term accessibility? Again, there is no guarantee how long the services presented here are available. As far as the sustainability of humanities websites, i. e. cultural knowledge in any form is concerned, we see it as universities, libraries and archives duty to provide and maintain the corresponding infrastructure.

References

Baillot, A. (Ed.). (n.d.). *Briefe und Texte aus dem intellektuellen Berlin um 1800*. Retrieved April 27, 2018, from http://www.berliner-intellektuelle.eu/

Edition Visualization Technology. (2013). Retrieved April, 27, 2018, from http://evt.labcd.unipi.it/

GitHub. (2018). Retrieved April 27, 2018, from https://github.com/

GitHubPages. (2018). Retrieved April 27, 2018, from https://pages.github.com/
Jekyll. (2018). Retrieved April 27, 2018, from https://jekyllrb.com/
Omeka (n.d.). Retrieved April 27, 2018, from https://omeka.org/
Richter, S. (Ed.). (2017). Retrieved April 27, 2018, from http://www.germanliteratureglobal.com/index.php/Hauptseite
TEIpublisher. (n.d.). Retrieved April 27, 2018, from https://teipublisher.com/index.html
Wordpress (n.d.). Retrieved April 27, 2018, from https://wordpress.org/

La Aptitud para Encontrar Patrones y la Producción de Cine Suave (Soft Cinema)

Diego Bonilla
bonilla.diego@gmail.com
California State University, Sacramento, United States of Americas

La computadora e Internet han fomentado cambios significativos en la forma en la que las narrativas cinematográficas son construidas, teniendo un impacto no sólo en los medios digitales que serán recibidos en el propio computador, sino también en los medios análogos tradicionales (Buckland, 2009). El cine hiperliga (o *hyperlink cinema*) se refiere a películas en las que la narrativa no sigue un arco específico, presentando una historia de forma no lineal. El uso del término *hiperliga* proviene de los textos digitales en donde se pueden especificar ciertas palabras como referencias directas a otros textos. En el ámbito hipertextual de la red mundial de computadoras, los lectores desarrollan aptitudes de lectura diferentes a las de los lectores de libros impresos; por ejemplo, los lectores de hipertexto desarrollan la habilidad de encontrar patrones y conexiones en múltiples textos hiperligados (Landow, 1992). El acceso en masa a un medio de comunicación basado en procesos computacionales ayuda a redefinir las formas en la que se conceptualiza el contenido, cómo se lleva a cabo su autoría y cómo éste es recibido por las audiencias. Por lo tanto, la proliferación del cine hiperliga, en el cual los arcos narrativos tradicionales no son seguidos, se puede entender, en parte, como resultado de la adopción de la computadora como medio de comunicación (Buckland, 2009).

El término "cine suave," *o soft cinema,* es un compuesto de las palabras *software* y *cinema,* y se refiere al cine que está basado en principios computacionales (Manovich, 2005). El cine suave, de igual forma que ocurre con la lectura de hipertexto, se refiere a la capacidad de alterar la secuenciación de una narrativa cinematográfica a través de procesos computacionales. A diferencia del cine tradicional o "rígido," el cine suave presenta una narrativa en un gran número de secuencias diferentes determinadas por algoritmos.

Una característica importante del cine suave es que el vidente no tiene que interactuar de forma constante con el computador a lo largo de la presentación de la narrativa audiovisual. Como ocurre con el cine tradicional, el público recibe la obra de forma "pasiva." Esta característica es fundamental para separar a las narrativas del cine suave de las narrativas presentes en los videojuegos. Algunos ejemplos de cine suave son *A Space of Time* (Bonilla, 2003), *Soft Cinema* (Manovich, 2005) y *Accidental Occurrence* (Bonilla, 2017). En el caso de *A Space of Time*, los módulos narrativos se presentan a través de un algoritmo que determina la secuenciación y la longitud de cada versión de la película antes de que ésta sea presentada. *A Space of Time* es similar a *Soft Cinema* ya que en ésta última la narrativa se cuenta con un audio lineal y la secuencia de los elementos visuales es determinada por algoritmos. *Accidental Occurrence* también sigue una serie de algoritmos que establecen la secuenciación de cada versión antes de que la obra pueda ser vista.

Una diferencia sustancial entre *Accidental Occurrence* y las obras de cine suave citadas anteriormente es que ésta ofrece al vidente la capacidad de alterar la forma en la que los algoritmos re-editan la película. En otras palabras, la obra ofrece cierto nivel de interactividad al vidente/usuario al inicio de la experiencia cinematográfica: Se puede alterar la longitud de la película desde un mínimo de 6 minutos hasta un máximo de 70 minutos y se puede dar más énfasis a un personaje que a otro. Los dos tipos de variaciones llegan a ofrecer 9.11E+124 versiones diferentes de la obra; la experiencia de esta variabilidad evoca un sueño recurrente en el que se tiene la misma "vivencia" de una narrativa aunque ésta siempre se lleva a cabo de forma diferente.

La construcción de una narrativa que será presentada como cine suave conlleva un nivel alto de complejidad y requiere de una audiencia acostumbrada a arcos narrativos no tradicionales. Es decir, no sólo la obra debe de ser creada de tal forma que pueda variar, sino también se debe de contar con un público dispuesto a "solucionar" la película tal y como lo hace con el cine hiperliga. En otras palabras, los videntes que están acostumbrados a la no linealidad, ya sea por el uso habitual de la red mundial de computadoras o por frecuentar el cine hiperliga, son más capaces de apreciar las narrativas ofrecidas por el cine suave.

References

Bonilla, D. (2017). *Accidental Occurrence* [Program]. Retrieved November 23, 2017, from https://www.modular.film/ Programa para generar narrativas audiovisuales.

Bonilla, D. (2003). *A Space of Time (CDROM)* Hypergraphia, LLC. URL: http://www.diego.today/a-space-of-time
Buckland, W. (2009). *Puzzle films: complex storytelling in contemporary cinema*. Malden, MA: Wiley-Blackwell.
Landow, G. P. (1992). *Hypertext. The convergence of contemporary critical theory and technology* (p. 134). Baltimore: The John Hopkins University Press.
Manovich, L. (n.d.). Soft Cinema (project description). Retrieved November 23, 2017, from http://manovich.net/index.php/projects/soft-cinema
Manovich, L., & Kratky, A. (2005). *Soft cinema (DVD)*. Cambridge, MA: MIT Press.

Women's Faces and Women's Rights: A Contextual Analysis of Faces Appearing in Time Magazine

Kathleen Patricia Janet Brennan
kpjbrennan@gmail.com
SUNY Polytechnic Institute, United States of America

Vincent Berardi
berardi@chapman.edu
Chapman University, United States of America

Aisha Cornejo
corne129@mail.chapman.edu
Chapman University, United States of America

Carl Bennett
bennetca@sunyit.edu
SUNY Polytechnic Institute, United States of America

John Harlan
harlanj@sunyit.edu
SUNY Polytechnic Institute, United States of America

Ana Jofre
jofrea@sunyit.edu
SUNY Polytechnic Institute, United States of America

We are developing a methodology for exploring and finding meaning in large corpuses that contain images, such as archives of periodic publications. We focus this work on *Time* magazine, and in particular on images of faces in *Time*. We use computer vision analysis, combined with contextual research and methods from the humanities, to elucidate trends and patterns in the visual culture reflected by the publication. In particular, we are examining how representations of the human face have changed over time, and seeking relationships between the visual features we discover and their corresponding socio-political contexts. Specifically, we are interested in gaining insight about how the form and context of representations of women and ethnic minorities have changed over time. Our preliminary research focuses on the correlation between changes in facial representations in *Time* magazine and the Women's Liberation movement in the United States in the 1970s and 1980s. The main outcome of this project will be a meaningful and accessible web-based platform through which both researchers and the general public can explore the archives of *Time* magazine to discover insights into our cultural history. We expect that we will be able to apply our methodology to any periodical publication, but we chose *Time* because it stands as a record of the many pulses of U.S. and world politics and their intersections with American culture. We believe that because it is such a culturally important and ubiquitous publication much can be learned from these archives about how Americans perceived politics and culture throughout the twentieth and early twenty-first centuries.

Our methodology combines computational processes, such as computer vision analysis, with contextual research, such as the history of the magazine's production process, as well as the cultural and political climate in which each issue appears. A brief summary of our methodology is as follows. Using the entire *Time* magazine corpus (about 4800 issues spanning over 93 years), we are identifying and extracting every facial image within the corpus, and running computational analyses on the images to quantify their visual features (such as RGB pixel values). We are building a database of the images that includes their associated metadata (year, issue, page number), as well as the extracted visual feature data. Within this database, we are also including more detailed metadata for each image: the face's gender, race, the context in which the face appears (ad or feature story), whether or not the face is smiling, and whether it is an individual portrait or belongs to an image that contains more than one face. In parallel to building this database, we are developing timelines of significant contextual information, which includes a timeline of the evolution of printing technologies used by *Time* magazine, a timeline of culturally impactful geo-political events, a timeline of civil rights movements, and a timeline of women's movements. Our image database will be connected to our contextual timelines with visual analytics. The visualizations we create will be interactive, web-based, and open to the general public.

We present here compiled preliminary results using our methodology and samples from our private collections of *Time* magazine, along with a contextual timeline of the women's rights movement in the US. In the work presented here, we used human labor to extract face images from sample issues and to tag each face image with the metadata described above. We are using the data harvested through human labor to improve our facial recognition algorithms and to train new algorithms to identify gender, race, smiling, and context. There has been a great deal of interest in sentiment analysis and facial recognition across academia and the general public, and we feel

this project will allow us to examine how these complex categories interact with each other. For example, how do our understandings of race and gender impact how humans classify sentiment? How do these understandings impact algorithmic classifiers? This complexity is one of the primary motivations for developing a methodology that consciously moves back and forth between human and computer analysis.

The metadata extracted by human labor has been particularly insightful, especially when put into the context of our historical timelines. Specifically, we noticed an increase in the number of female faces in the 1970s, coincident with the many milestones in the Women's Rights movement. Interestingly, our preliminary data also suggests that as the number of women represented in the magazine increases, the proportion of women in advertisements decreases. Our poster will focus on a close examination of the data and sociopolitical context of 1965-1990 in order to fully explore this potential correlation. We will also discuss our methodology and include a few examples of our visualizations.

This project aims, not only to gain insights from an analysis of *Time* magazine and to make these insights publicly accessible, but also to establish novel methodologies for the visual analytics of large data sets, particularly of image-based corpuses, which we hope to use for years to come and to share with other researchers.

The ultimate goal of this project is to create a website with contextualized interactive visualizations based on the entire archive. Our initial approach was inspired by Manovich's Selfie-city and Photo-trails work, and by his team's use of direct visualization (Crockett, 2016), which we see as a way to engage broad audiences into complex corpuses. We also draw inspiration from *Robots Reading Vogue* (King and Leonard) and *Neural Neighbors* (Leonard), which are projects based in the Yale University library system. By exploring specific, humanities-based research questions in this early phase of our project we will be able to make meaning and better contextualize the interactive visualizations in the end.

References

Crockett, D. (2016). Direct visualization techniques for the analysis of image data: the slice histogram and the growing entourage plot. *International Journal for Digital Art History*, 0(2) doi:10.11588/dah.2016.2.33529. http://journals.ub.uni-heidelberg.de/index.php/dah/article/view/33529 (accessed 8 November 2016).

King, L., and Leonard, P. (2018). Robots Reading Vogue : Colormetric Space http://dh.library.yale.edu/projects/vogue/colormetricspace/ (accessed 5 January 2017a).

King, L., and Leonard, P. (2018). Robots Reading Vogue http://dh.library.yale.edu/projects/vogue/ (accessed 8 November 2016b).

Lauridsen, H. (2014). What's in Vogue? Tracing the evolution of fashion and culture in the media *Yale News* http://news.yale.edu/2014/09/05/what-s-vogue-tracing-evolution-fashion-and-culture-media (accessed 8 November 2016).

Manovich, L. (2011). Mondrian vs Rothko: footprints and evolution in style space http://lab.softwarestudies.com/2011/06/mondrian-vs-rothko-footprints-and.html (accessed 30 December 2016a).

Manovich, L. (2010). One million manga pages http://lab.softwarestudies.com/2010/11/one-million-manga-pages.html (accessed 30 December 2016b).

Manovich, L., Hochman, N., and Chow, J. (2013). Phototrails: Visualizing 2.3 M Instagram photos from 13 global cities http://lab.culturalanalytics.info/2016/04/phototrails-visualizing-23-m-instagram.html (accessed 30 December 2016).

Rushmeier, H., Pintus, R., Yang, Y., Wong, C. and Li, D. (2015). *Examples of challenges and opportunities in visual analysis in the digital humanities*. vol. 9394. pp. 939414-939414–19 doi:10.1117/12.2083342. http://dx.doi.org/10.1117/12.2083342 (accessed 5 January 2017).

softwarestudies.com (2009). *Timeline: 4535 Time Magazine Covers, 1923-2009*. Photo https://www.flickr.com/photos/culturevis/3951496507/ (accessed 30 December 2016).

Decolonialism and Formal Ontology: Self-critical Conceptual Modelling Practice

George Bruseker
bruseker@ics.forth.gr
Centre for Cultural Informatics,
Institute of Computer Science-FORTH, Greece

Anais Guillem
aguillem@ucmerced.edu
School of Social Sciences, Humanities and Arts, University of California Merced, United States of America

Digital humanists taking up the challenge of the decolonialist approach face, with regards to information management, the question of how to structure their data in a way which escapes the confines of the repressive episteme that they seek to challenge. And yet, the database and the data form have enormous potential to replicate and even intensify, in a new medium, the colonial intersection of knowledge and power. A data model operates, at least potentially, on its 'subject' as an authoritative power, disenfranchising the epistemological constellations of those it chooses to represent and submitting them to a colonial order of knowledge. Such subjugation can be argued to be represented in classic arrangements of knowledge like the 'tombstone' data explaining objects in museums

and archaeological collections. In such data models, the analyses that go along with the object and which tie these objects into a web of knowledge privilege the interpretation of the scholars who speak of and for the object. It is often the agency of the 'discovering' or 'gifting' agent that is most associated to the object over/above the cultures, groups and individuals for whom the object was a living part of life and practice. (saywhatnathan, 2017) The digital humanist would work on a corpus of well-formatted data in order to build up a new knowledge, contesting colonial representations, but the epistemic, ethical and pragmatic challenge comes together here: what can be the form of this representation and how to conceptualize and maintain it, without re-introducing imperialistic paradigms?

In this question, the theoretical and practical interests of decolonialism and the discipline of knowledge engineering / formal ontology overlap and have the chance for a fruitful methodological dialogue. By decolonialist thought we intend the theoretical tendency building up from the post-colonialism of Said (1979) on to the work of Mignolo (2011), Borgstede (2010), and Smith (2012) amongst others. This movement looks to challenge the identification drawn between scientific practices originally developed in the West, meaning the traditions of European scholarship, and a universalist objectivity. The critique is undertaken in order to identify and recognize limits of the Western, scientific project, towards the end of opening a space for the self-articulation of suppressed modes of discourse, so that they can reach expression and be understood as autonomous spaces of potential truth disclosing, as elaborated under non-dominant conceptual paradigms. By knowledge engineering and formal ontology, we intend methods proposed since the 1990s (Gruber, 1995; Guarino, 2997; Smith, 2003) as a means to make better data structures within information systems by engaging in an interdisciplinary practice to build these latter through a disciplined dialogue between computer science, philosophy and the domain practitioners concerned. Established and well known applications of this method are known in the areas of linguistics with DOLCE and cultural heritage CIDOC CRM as described in Gangemi et al. (2002) and Doerr (2003).

The general aim of adopting a formal ontological approach in a research discipline or community is to serve as a means to robustly model data and create more accurate digital representations in a way that creates community consensus around the generic representational form. Within the digital humanities, formal ontology is an important tool to solve the long term data integration and data provenance problems that are correlate to the creation of ever greater datasets by scholars. A formal ontology offers much that the decolonialist digital humanist would need in their toolset. Can it, however, meet their epistemic and ethical requirements?

Here we would argue that decolonialist thinking and well founded formal ontological thinking share fundamental theoretical commitments which are mutually beneficial. The potential for enrichment is two-way, offering a path forward for an information structure suitable to decolonialist studies but also providing to formal ontology research an important control point. In particular, what binds together these two approaches is a shared commitment to a radical and critical approach to known epistemic structures. Both are committed to a self-reflexive critique which does not accept the given epistemic prejudice of the 'form' of scientificity but rather aims to critique it in order to understand a wider form. This is expressed in a radical empiricism in the sense of a deliberate openness to understanding from practice rather than from the formalisms of science. What is to be modelled is not what is said but what is done. This commitment on the part of formal ontology leaves the final model of information representation always open to modification. The work of the decolonialist scholar brings material that can continually challenge the prejudice in a model and cause its redesign. On the other hand, the open ended design of the formal ontological model which does not follow the logic of the data form but of an open ended graph of knowledge, allows for the representation of multiple perspectives and the multi-participation of objects in different epistemological formations.

An illustration of this self-reflexive and openly critical practice in action can be taken from the modelling of 'discovery' activities in the CIDOC CRM extension, CRMsci. (Doerr et al., 2017) Classic data representation and inbuilt cultural prejudice would offer the 'intuitive' category of 'discovery' to describe scientific observation activities such as ethnography, archaeology, botany and so on. Such categorizations, however, are one-sided and privilege the 'discoverer' while decentering and subjecting the 'discovered'. Extensive, long-term dialogue and conversation over this issue, led to the elaboration of a general class of the ontology called 'Encounter'. 'Encounter' avoids one-sidedness of representation and the implication that something comes to be known through the encounter event. It shifts the representation to a third party point-of-view, and allows modelling the fact that some group met some thing. This encounter finds an object and may produce new knowledge, for the group that has initiated an encounter activity, but not as such.

The intersection of decolonialist thought and knowledge engineering in the practice of digital humanism offers the opportunity to lift the tombstone off cultural knowledge and open it to expression and contention with the dominant episteme by means of the construction of open graphs of knowledge that empower the representation, reconstruction and expression of suppressed knowledge by the actors from whom it originates.

References

Borgstede, G., Cipolla, C. N., Gullapalli, P., Lilley, I., Jiménez, J. R. P., Patterson, T. C., Preucel, R. W., et al. (2010). *Archaeology and the Postcolonial Critique*. (Ed.) Liebmann, M. & Rizvi, U. Z. Reprint edition. AltaMira Press.

Doerr, M. (2003). The CIDOC conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI Magazine*, 24(3): 75.

Doerr, M., Kritsotaki, A., Rousakis, Y., Hiebel, G. and Theodoridou, M. (2017). *Definition of the CRMsci: An Extension of CIDOC-CRM to Support Scientific Observation*. Technical Report Crete: ICS-FORTH.

Gangemi, A., Guarino, N., Masolo, C., Oltramari, A. and Schneider, L. (2002). Sweetening Ontologies with DOLCE. *Knowledge Engineering and Knowledge Management: Ontologies and the Semantic Web*. (Lecture Notes in Computer Science). Springer, Berlin, Heidelberg, pp. 166–81 doi:10.1007/3-540-45810-7_18. https://link.springer.com/chapter/10.1007/3-540-45810-7_18 (accessed 25 April 2018).

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing?. *International Journal of Human-Computer Studies*, 43(5): 907–928.

Guarino, N. (1997). Understanding, building and using ontologies. *International Journal of Human-Computer Studies*, 46(2): 293–310.

Mignolo, W. D. (2011). *The Darker Side of Western Modernity: Global Futures, Decolonial Options*. edition. Duke University Press Books.

Said, E. W. (1979). *Orientalism*. 1st Vintage Books ed edition. New York: Vintage.

saywhatnathan (2017). Maker unknown and the decentring First Nations People *Archival Decolonist [-O-]* https://archivaldecolonist.com/2017/07/21/maker-unknown-and-the-decentring-first-nations-people/ (accessed 25 April 2018).

Smith, B. (2003). Ontology. In Floridi, L. (ed), *Blackwell Guide to the Philosophy of Computing and Information*. Oxford: Blackwell, pp. 155–166.

Smith, L. T. (2012). *Decolonizing Methodologies: Research and Indigenous Peoples*. 2 edition. London: Zed Books.

Rules against the Machine: Building Bridges from Text to Metadata

José Calvo Tello
jose.calvo@uni-wuerzburg.de
University of Würzburg, Germany

Introduction

Digital literary studies advance in their research, requiring more specific metadata about literary phenomena: narrator (Hoover 2004), characters (Kastorp et al. 2015), place and period, etcetera. This metadata can be used to explain results in tasks like authorship attribution or genre detection, or to evaluate digital methods (Calvo Tello 2017). What could be the most efficient way to start annotating this information in corpora of thousand of texts in languages, genres and historical periods for which many NLP tools are not trained for? In this proposal, the aim is to identify specific literary metadata about entire texts with methods that are either language-independent or easily adaptable for humanists.

Two Ways from Text to Metadata

The two approaches to classify unlabeled samples applied here are rule-based classification and supervised machine learning. In rule-based classification (Witten et al. 2011), domain experts define formalised rules that correctly classify the samples. For example a rule based on a single token can be defined for each class to predict whether a text is written in third person (83% of the corpus) or first person using tokens for the two values are the Spanish words *dije* ('I said') and *dijo* ('he said'), and the rule:

1. if *dijo* appears 90% more than *dije*, the novel is written in third person
2. if *dijo* appears less, in first person

The results of applying this rule can be presented as a confusion matrix:

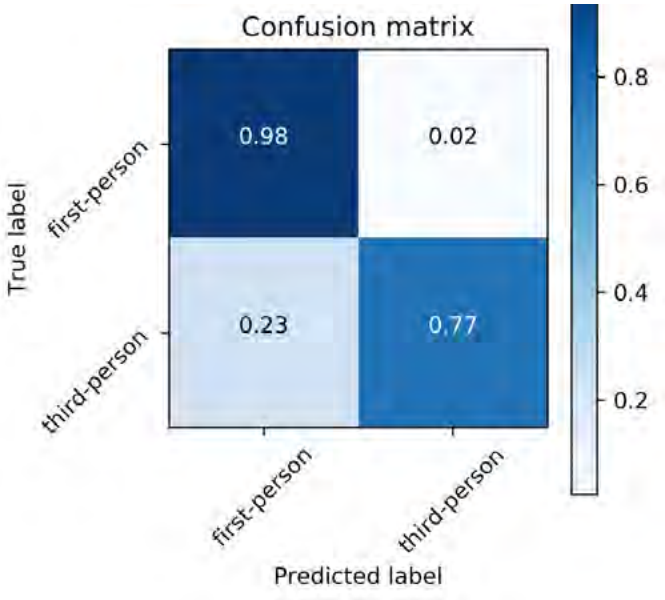


Fig 1. Confusion Matrix of rule-based results about narrator

For supervised methods (Müller and Guido 2016; VanderPlas 2016), we need labeled samples to train and

evaluate the method. In the following table, the different classifiers and document-representations achieve different accuracy scores:

	raw	relative	tfidf	zscores
SVC	0.90	0.83	0.83	0.88
KNN	0.83	0.88	0.81	0.81
RF	0.88	0.88	0.86	0.90
DT	0.84	0.83	0.84	0.82
LR	0.88	0.83	0.83	0.17
BN	0.72	0.72	0.72	0.82
GN	0.72	0.80	0.80	0.81

Fig 2. Accuracy (F1-score) for narrator

Corpus and Metadata

The data is part of the *Corpus of Spanish Novels of the Silver Age (1880-1939)* (used in Calvo Tello et al. 2017), with 350 novels in XML-TEI by 58 authors. Each text has been annotated manually with metadata and its degree of certainty has been assigned. 262 texts with either high or medium certainty have been used to create a gold-standard with the following classes:

1. protagonist.gender
2. protagonist.age
3. protagonist.socLevel
4. setting.type
5. setting.continent
6. setting.country
7. setting.name
8. narrator
9. representation
10. time.period
11. end

Modelisation and Methods

The scripts have been written in Python (available on GitHub) (https://github.com/cligs/projects2018/tree/master/text2metadata-dh). The features have been represented as different document models (Kestemont et al. 2016):

- raw frequencies
- relative frequencies
- tf-idf
- z-scores

Different classify algorithms (cross validation, 10 folds) and amount of Most Frequent Words have been evaluated. For each class a single token was used to represent each class value and a ratio was assigned for the default class value (see repository in GitHub for rules). Both approaches were compared to a "most populated class" baseline, quite high in many cases.

Results

The results of both approaches are as following:

Class	F1 baseline	F1 Rule	F1 Cross Mean	F1 Cross Std	Algorithm	Model	MFW	Winner
end	0.60	0.54	0.60	0.02	LR	tfidf	100	Baseline
narrator	0.83	0.80	**0.91**	0.04	**RF**	**tfidf**	**1000**	**ML**
protagonist.age	0.55	0.25	0.55	0.01	LR	tfidf	100	Baseline
protagonist.gender	0.80	0.68	0.80	0.01	BN	tfidf	100	Baseline
protagonist.socLevel	0.63	0.49	0.64	0.07	SVC	zscores	5000	Baseline
representation	0.88	0.80	0.88	0.01	LR	tfidf	100	Baseline
setting.continent	0.95	0.94	0.96	0.01	SVC	zscores	5000	Baseline
setting.continent.binar	0.95	0.95	0.95	0.19	LR	zscores	500	Baseline
setting.country	0.93	0.38	0.94	0.01	SVC	zscores	1000	Baseline
setting.country.binar	0.87	0.47	0.88	0.03	SVC	zscores	1000	Baseline
setting.name	0.64	**0.85**	0.71	0.02	SVC	zscores	1000	**Rule**
setting.type	0.48	0.46	**0.71**	0.05	**SVC**	**zscores**	**5000**	**ML**
time.period	0.95	0.95	0.97	0.01	BN	zscores	5000	Baseline

Fig 3. Results

In many cases the baselines are higher than the results of both approaches. The rule outperformed the baseline in the case of name of the setting with very good results. In two cases (narrator and setting's type), Machine Learning is the most successful approach and its F1 is statistically higher than the baseline (one sample t-test, = 5%). The algorithms Supported Vector Machines, Logistic Regression and Random Forest are most successful, while tf-idf and speacilly z-scores got the best results, the last one a data representation "highly uncommon in other applications" different from stylometry (Kestemont et al, 2016).

Conclusions

In this proposal I have used simple rules and simple features in order to detect relatively complex literary metadata in many cases with high baselines. While Machine Learning showed a statistically significant improvement in detection for two classes (type of setting and narrator), rules worked better for the name of the setting. This is a promising point to continue researching in order to annotate the rest of the corpus.

References

Calvo Tello, J. (2017). What does Delta see inside the Author?: Evaluating Stylometric Clusters with Literary Metadata. III Congreso de La Sociedad Internacional Humanidades Digitales Hispánicas: Sociedades, Políticas, Saberes. Málaga: *HDH*, pp. 153–61 <http://hdh2017.es/wp-content/uploads/2017/10/Actas-HDH2017.pdf>.

Calvo Tello, J., Schlör, D., Henny-Krahmer, U. and Schöch, C. (2017). Neutralising the Authorial Signal in Delta by Penalization: Stylometric Clustering of Genre in Spanish Novels. Montréal: *ADHO*, pp. 181–83 <https://dh2017.adho.org/abstracts/037/037.pdf>.

Hoover, D. L. (2004). Testing Burrows's Delta. *Literary and Linguistic Computing*, 19(4): 453–75.

Kastorp, F., Kestemont, M., Schöch, C. and Bosch, A. Van den (2015). *The Love Equation: Computational Modeling of Romantic Relationships in French Classical Drama. Sixth International Workshop on Computational Models of Narrative*. Atlanta, GA, USA. <https://zenodo.org/record/18343>.

Kestemont, M., Stover, J., Koppel, M., Karsdorp, F. and Daelemans, W. (2016). Authenticating the writings of Julius Caesar. *Expert Systems with Applications*, 63: 86–96 <http://dx.doi.org/10.1016/j.eswa.2016.06.029>.

Müller, A. C. and Guido, S. (2016). *Introduction to Machine Learning with Python: A Guide for Data Scientist*. Beijing: O'Reilly.

VanderPlas, J. (2016). *Python Data Science Handbook: Essential Tools for Working with Data*. First edition. Beijing Boston Farnham: O'Reilly.

Witten, I., Frank, E. and Hall, M. (2011). *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd edition. San Francisco: Morgan Kaufmann.

Prospectiva de la arquitectura en el siglo XXI. La arquitectura en entornos digitales

Luis David Cardona Jiménez
lcardona@ucm.edu.co
Universidad Católica de Manizales, Universidad de Caldass, Colombia

Desde la irrupción y aplicación de los adelantos de los medios digitales en los procesos de diseño y creación en arquitectura, los ámbitos académicos y profesionales han encontrado alternativas, posturas y desarrollos para la adopción de la tecnología digital, la cual, ha propiciado impactos en las formas de trabajo y en los abordajes del problema arquitectónico definiendo interesantes posibilidades en la imaginación y creación arquitectónica.

Se ha establecido por varios autores como, Carpo, Piccon, Frazer, Menges Cache, entre otros, que estos procesos de cambio en la imaginación y procesos de diseño y creación del espacio arquitectónico se dieron a principios de la década de los 90's con la aparición de los primeros programas comerciales de CAD (Computer-Aided Design) los cuales ofrecieron oportunidades de transformación y manejo en la proyectación geométrica de espacios y edificios.

Es importante mencionar que la tradición del pensamiento arquitectónico desde sus orígenes se ha basado en la geometría euclidiana y en los sólidos platónicos, prisma, cilindro, cubo, pirámide, esfera, son figuras que se encuentran en las arquitecturas de todas las civilizaciones antiguas, las cuales se podían identificar claramente como arquetipos únicos y aislados (Fernández-Álvarez, 2014)

El giro post-digital

El termino post-digital es relativamente reciente y aún en construcción, sin embargo, una postura es no entenderlo como "después" de lo digital o lo "anti" digital, más bien se debe pensar como en la relación y el dominio de lo humano sobre lo tecnológico. El término "post-digital" apunta la llamar la atención sobre "una actitud que se preocupa más por ser humano que por ser digital" (Zreik & Gareus, 2012)

Reiteradamente se ha mencionado a los años 90 como el momento en el que se evidencia la aplicación de medios y tecnólogas digitales en arquitectura. De acuerdo con Buchanan (1992), el reposicionamiento de nuevas ideas y planteamientos desencadenados por las pociones e intenciones de interpretación de nuevas preguntas y prácticas en todo al diseño incorporando tecnologías alternativas con entornos de simulación buscando productos y materiales innovadoras.

La incorporación de medios digitales al pensamiento, imaginación y creación arquitectónica expresadas en la representación y visualización del espacio arquitectónico. Aquí cabe mencionar de nuevo a Buchanan (1992) el cual, sin ser arquitecto, pero con una consciencia y formación en diseño, nombra la arquitectura deconstructivista como una de las iniciativas arriesgadas y agresivas que contribuirían a recuperar el significado que trasmite la obra arquitectónica.

El futuro de la arquitectura, el Siglo XXI

La arquitectura como disciplina de tradicional arraigada a principios y postulados casi inmutables, viene des-

de hace tres décadas presentado cambios en la forma de abordar el problema del diseño del espacio habitable representado en la ciudad, edificios o viviendas. Con el constante cambio tecnológico, el fortalecimiento de las relaciones humanas a través de los digital, la arquitectura empieza a responder acertadamente a las demandas de nuevas formas de abordar la trasformación de la realidad.

En un mundo hiperconectado, con una producción diaria de datos incalculables, el Big Data se convierte en una herramienta que permite crear plataformas de trabajo colaborativo, fortaleciendo la relación entre usuarios y diseñadores.

Phil Bernstein, arquitecto y profesor de Yale University visualiza el futuro de la arquitectura a través avatares para el análisis de comportamiento de usuarios en entornos construidos virtualmente desde la visualización del Big Data.

> Big Data ya está transformando la forma en que los arquitectos diseñan edificios. Cambiando las potencias Big Data y la realidad virtual, se avanzará en la práctica arquitectónica a pasos agigantados (Phil Bernstein)

Esta visión prospectiva de la arquitectura es emergente y esta en proceso de convergencia. Hoy, aunque incipiente, en Latinoamérica se empieza a mostrar un interés por avanzar en el entendimiento y aplicación de conceptos de investigación e innovación a través de las posibilidades que las tecnologías digitales ofrecen para el desarrollo de una arquitectura que responda a las expectativas del mundo en constante proceso de cambio.

References

Alexenberg, M. (2011). *The Future of Art in a Postdigital Age: From Hellenistic to Hebraic Consciousness.* Chicago: Intellect Ltd.

Allen, S. (2009). *"Velocidades terminales" en La digitalización toma el mando.* Barceolna: Gustavo Gili.

Amado, R. (2007). La arquitectura como interfaz. En *Arte, Arquitectura y Sociedad_ Digital* (págs. 107-109). Barcelona: Universitat de Barcelona/ESARQ UIC.

Baltazar, A. (2009). *Cyberarchitecture: the virtualisation of architecture beyond, Tesis doctoral.* University College London, UCL.: The Bartlett School of Architecture.

Buchanan, N. (1992). Wicked Problems in Design Thinking paper. *Design Issues*, 5-21.

Carpo, M. (2013). *The Digital Turn in Architecture 1992 - 2012.* Chichester, UK: John Wiley & Sons Ltd.

Cross, N. (1982). Designerly Ways Of Knowing paper. *Design Studies*, 221-227.

Fernández-Álvarez, Á. J. (2014). Riding the cloud. Information and architectural representation in the post-digital age. *EGE: Revista de Expresión Gráfica en la Edificación,* , 159-166.

Ortega, L. (2009). *La digitalización toma el mando.* Barcelona: Gustavo Gili.

Períes, L. (2016). *Estereotomía y topología en arquitectura.* Buenas Aires: Editorial de la Universidad Nacional de Córdoba.

Piccon, A. (2010). *Digital Culture in Architecture.* Basilea: Birkhäuser.

Sandoval Vizcaíno, M. (2014). Herramientas de diseño y arqutiectura, la relación intrínseca entre herramientas y diseño . *Revista Legado de Arquitectura y Diseño*, 39-56.

Zreik , K., & Gareus, R. (2012). *PostDigital Art - Proceedings of the 3rd Computer Art* . Paris: Europia Productions.

Visualizando Dados Bibliográficos: o Uso do VOSviewer como Ferramenta de Análise Bibliométrica de Palavras-Chave na Produção das Humanidades Digitais

Renan Marinho de Castro
renan.castro@fgv.br
Fundação Getulio Vargas, Brazil

Ricardo Medeiros Pimenta
ricardopimenta@ibict.br
Instituto Brasileiro de Informação em Ciência e Tecnologia, Brazil

O objetivo dessa pesquisa é mapear, através da identificação de termos de palavras-chave, quais as principais atividades presentes nas humanidades digitais construindo e visualizando mapas bibliométricos oriundos de uma revisão de literatura deste tema. Dessa forma é proposta uma análise desses dados a partir da utilização do software VOSviewer para construção de redes de relacionamento dos termos provenientes das bases: Web Of Science (WoS) e Scopus. Assim, foram gerados grafos de palavras-chave baseados nos termos atribuídos à literatura registrada nessas duas bases de dados. Buscamos dessa forma combinar essas duas análises a partir da construção de dois mapas distintos e possibilitar seu cotejamento.

Partindo dessa proposta, elaboramos uma expressão de busca[1] para dar conta de recuperar a publicação sobre *digital humanities* em inglês, espanhol e português nas bases de dados eleitas para esta revisão. Adotou-se como padrão a opção de filtro que contemplasse o '*abstract*', sendo o campo 'resumo' escolhido como foco da recuperação por apresentar maiores concentrações de termos relacionados à indicação temática dos documentos. Os resultados reportados pelas buscas foram expor-

1 A expressão de busca aplica às bases selecionadas pode ser representada pela *string* ((((((„Digital Humanities")) OR (("Humanidades Digitales")) OR (("Humanidades Digitais"))))))

tados no formato compatível com o VOSviewer, no caso da Web Of Science, 'Tab Delimited (Win)' e no caso da Scopus, o formato 'CSV'. Foram recuperados na Web Of Science 1067 documentos e, na Scopus, 1575.

De posse dos arquivos extraídos, utilizamos do recurso de criação de grafos baseados em co-ocorrência de palavras-chave. Essa análise oferece as opções *'Author´s keywords'* e *'Keywords Plus'*, por isso elegemos a opção *'all keywords'* que engloba essas duas modalidades, além do método de *full counting* que atribui o mesmo peso para cada link em co-ocorrência. Na WoS foram totalizadas 2826 palavras-chave com exigência mínima de 8 ocorrências para integrar a análise, essa filtragem resultou em 38 núcleos conectados. No caso da Scopus também elegemos a opção *'all keywords'* para contemplar as palavras-chave atribuídas pelos próprios autores (*Author´s Keywords*) além da opção *'index keywords'*, cuja atribuição é proveniente da base. Foram, assim, identificadas 5195 palavras-chave e a nota de corte elevada à recorrência mínima de 15 vezes. Essa configuração produziu um grafo com 64 (após desambiguação: 61) termos com núcleos de conexão entre si. Este grafo também considerou o método *'full counting'*.

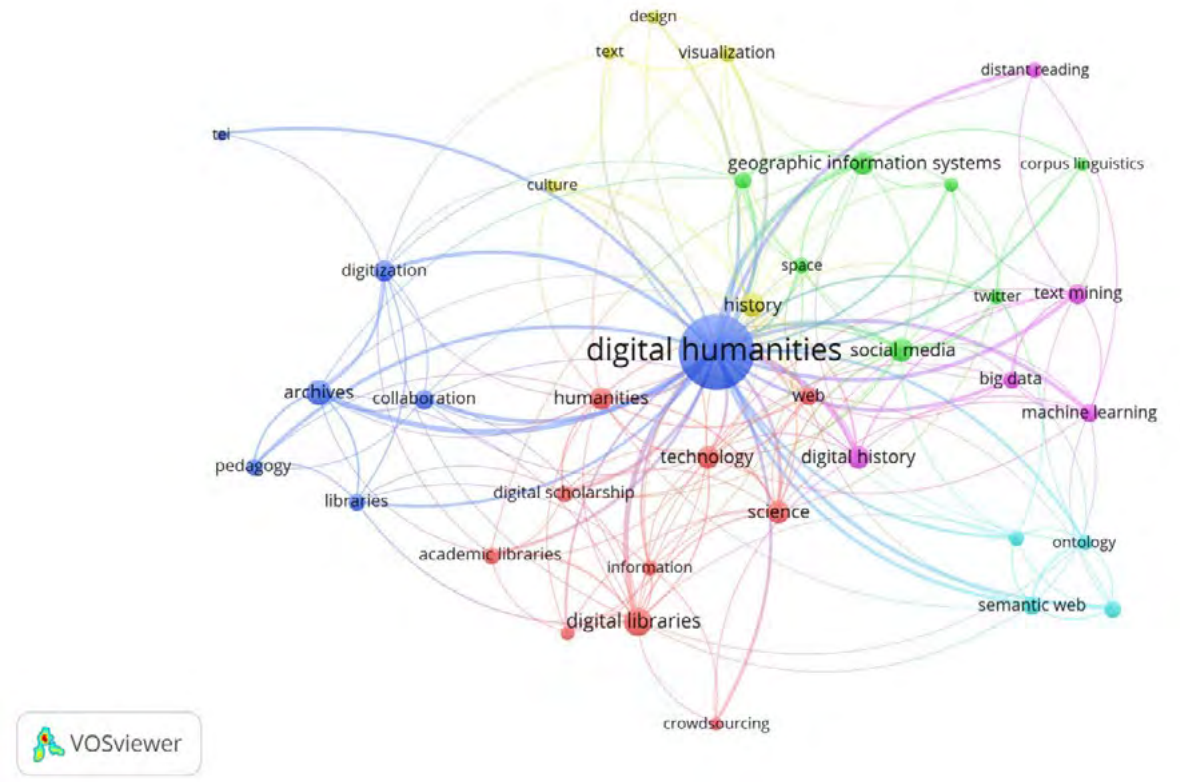


Figura 1 Grafo de palavras-chave da produção registrada na Web Of Science com ‚nós' calculados segundo seu link total de força.

Na sequência produzimos dois mapas baseando-se nas respectivas fontes de literatura e, baseados nestas, geramos, além dos dois mapas, a mineração dos termos mais recorrentes que servem de base para construção do grafo. O grafo respectivo à WoS possui 6 clusters compostos por 10 termos no de maior tamanho e 4 no menor. A análise de clusters permite identificar que dentre estes há clusters estreitamente relacionados às bibliotecas digitais e à ciência da informação (C.I.) (por exemplo termos como *information* e *technology*), como no caso do cluster 1. Também há um cluster relacionado às técnicas de visualização (cluster 2). O cluster 3 volta a apresentar termos relacionados à C.I. como *archive, digitization e libraries*. Já o cluster 5 volta-se às técnicas das humanidades digitais como *text mining* e *machine learning*.

O grafo com dados da Scopus também possui 6 clusters tendo no maior deles 16 termos e, no menor, 7. Também é possível perceber a recorrência de um cluster voltado às técnicas de visualização (cluster 4: *visualization, data visualization e gis*) bem como a reverberação da presença da C.I. com os termos *digital libraries, digital archives e digital collections* (cluster 2). Outras técnicas das humanidades digitais reincidem como *data mining* e *text mining* (cluster 3), além de outros termos relacionados à ciência da informação: *archives, libraries* e *digitization* (cluster 5). Vale destacar que o termo com maior peso foi *digital libraries* tanto na WoS como na Scopus.

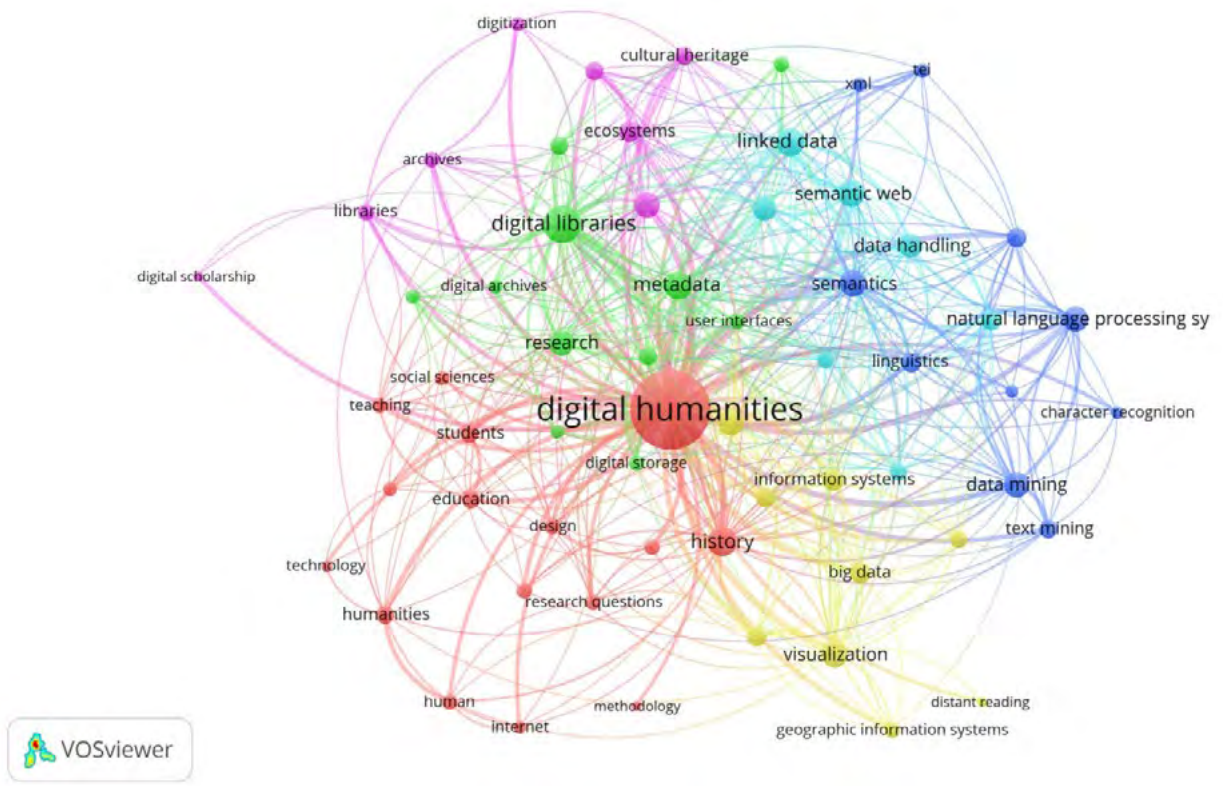


Figura 2 Grafo de palavras-chave da produção registrada na Scopus com ‚nós' calculados segundo seu link total de força.

Dessa forma os mapas permitem visualizar termos e conceitos mais presentes na literatura e, consequentemente, possibilitam a clarificação da relação entre eles. Apesar da grande rede de relacionamento que os mapas exibem é possível, mesmo interpretando apenas os clusters criados, contemplar, por exemplo, as áreas principais que interagem para formar a ideia de humanidades digitais na literatura. Além disso, sobretudo, o cotejamento dos grafos provenientes de cada repositório de literatura permite corroborar quais termos se consolidam através de sua reincidência nos mapas.

References

DACOS, Martin. (2011). Manifesto das Humanidades Digitais. *ThatCamp Paris*, [S.l.] 26 mar. 2011. Disponível em: <https://tcp.hypotheses.org/497> Acesso em 10 out. 2016.

ECK, Nees Jan Van; WALTMAN, Ludo. (2016) *VOSviewer Manual*. Disponível em http://www.vosviewer.com/documentation/Manual_VOSviewer_1.6.5.pdf Acesso em 10 de julho de 2017

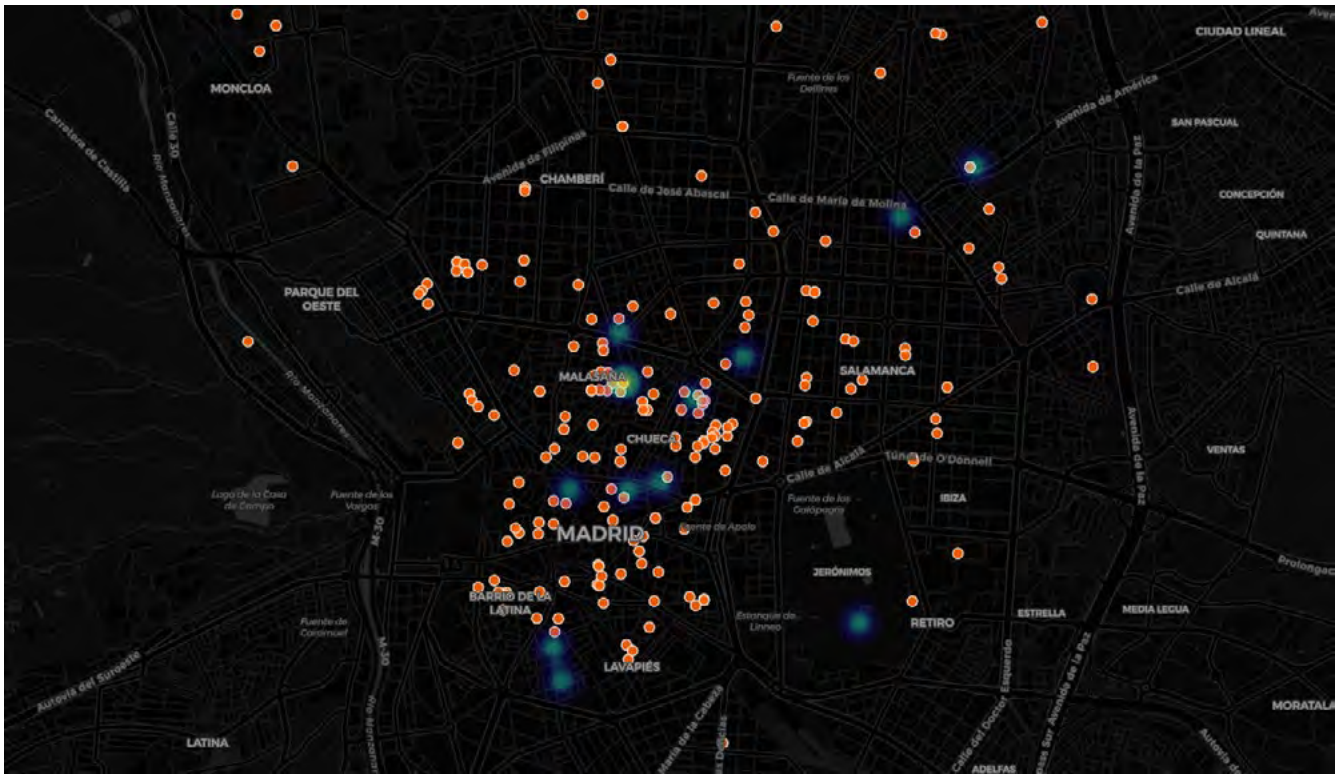
KOLTAY, Tibor. (2016) Library and information science and the digital humanities: perceived and real strengths and weaknesses. *Journal of documentation*, 72(4), pp. 781-792.

TANG, Muh-Chyun; CHENG, Yun Jen; CHEN, Kuang Hua. (2017) A longitudinal study of intellectual cohesion in digital humanities using bibliometric analyses, *Scientometrics*, v.113, n.2, pp.985-1008, nov. 2017.

Mapping the Movida: Re-Imagining Counterculture in Post-Franco Spain (1975-1992)

Vanessa Ceia
vanessa.ceia@mcgill.ca
McGill University, Canada

Mapping the Movida is an open web archive and geo-spatial project that visualizes the cultural and creative hubs and networks of the *Movida madrileña*, a sociological phenomenon and cultural renaissance that emerged in the first decade of Spanish democracy (roughly 1976-1986), most notably in central Madrid. Among the Movida's most well-known artists are filmmakers Pedro Almodóvar and Iván Zulueta, photographers Alberto García-Alix and Ouka Leele, illustrators El Hortelano, Nazario, and Ceesepe, poet Eduardo Haro Ibars, novelist Eduardo Mendicutti, fashion designers Jesús del Pozo, Manuel Piña and Agatha Ruiz de la Prada, and musicians Ana Curra and Alaska (Olvido Gara), among many others. One of the most striking characteristics of those who have been historicized as so-called "artists of the Movida" is that



they are, with few exceptions (Almodóvar, Ruiz de la Prada, Curra, Alaska), men of upper middle-class upbringing. Additionally, when examining the canonized geographies of the Movida—that is, the cultural hubs and culturally productive spaces of the period—we find that these canonical artists are primarily centered around Madrid's core neighborhoods, such as Malasaña, and, in few instances, the affluent north-central sector of the Spanish capital. In the canonized Movida, peripheral, and often working-class neighborhoods are largely excluded from the cultural map and countercultural histories of this period.

This project is a scholarly response to the limited scope of artists—mostly male and professionally active in central Madrid—historically associated with the Movida in mainstream press and scholarship. In its mission to bring to light and build "BRIDGES/PUENTES" with uncharted human geographies of the period, Mapping the Movida aims to: (1) re-create the Madrid of the Movida using a range of visual, textual and spatial media, data, and thick (Presner, Shepard, Kawano 2014) and deep mapping technologies that document the Madrid of the past; (2) visualize creative networks and cultural hubs of the Movida through various cultural and critical lenses—including mainstream Spanish media outlets (*El País*, *ABC*, *El Mundo*), scholarly articles, and subcultural publications from the period (*La Luna de Madrid*, *El Víbora*, *Ozono*, *Madrid Me Mata*)—to reveal how each lens represents the Movida in different, divergent, and/or similar ways and "provoke negotiation between insiders and outsiders, experts and contributors, over what is represented and how„ (Bodenhamer, Corrigan, Harris 2015: 4); (3) create a public archive and searchable database of Movida events and artists' documented movements in Madrid during the Movida; and, perhaps most importantly, (4) de-colonize the geographies of the Movida by revealing spaces, artists, and socio-economic groups that problematize the cultural and spatial canon of the period.

This poster, grounded in archival research from Brown University's Revistas de la Movida Collection, will exhibit the methodology and tools (Carto, Esri Story Maps) that have been used, the archival and theoretical concerns that have arisen, and the revelations that have been made during the various stages of project development. It will also demonstrate how Mapping the Movida's marriage of archival research and technology questions and queers the scope of what has been historicized and canonized as the "culture of the Movida" over the last nearly 40 years. At stake in this project is our understanding of the cultural and human geographies of Madrid during this period as well as our knowledge of artists and cultural products that have rarely, if at all, been studied and imagined within the corpus of so-called Movida artists and texts.

References

Bodenhamer, D.J. and J. Corrigan, T.M. Harris. (2015). *Deep Maps and Spatial Narratives*. Bloomington: Indiana University Press.

Presner, T. and D. Shepard, Y. Kawano. (2014). *HyperCities: Thick Mapping in the Digital Humanities*. Cambridge: Harvard University Press.

Intellectual History and Computing: Modeling and Simulating the World of the Korean Yangban

Javier Cha
javiercha@gmail.com
Seoul National University, Korea

This poster presentation demonstrates the use of computational methods to discover hidden collectives and communities from Korean historical data. The overarching question is derived from the intellectual history of early modern Korea, which was defined by the coalescence of several schools of Neo-Confucian thought and literary movements. Such developments took place at a time of increasing localization of population, material resources, state institutions, and culture. In the existing body of research, the connections between the material and ideational aspects of the yangban aristocracy have been unclear, owing in large part to the undue attention given to a small number of famous personalities, source materials, and locations. Can this skewed picture be redrawn from the bottom-up, through a more balanced and fuller use empirical data? Fortunately for social scientifically-minded historians of Korea, the government of South Korea has aggressively funded the digitization of cultural heritage. Access to this "big data" has allowed me to embark on a critique of existing reified generalities with large-scale data analysis. This kind of data also demands a new type of research concerning social, cultural, and historical entities which may not yet have been identified and therefore not yet been given a label. The data are drawn from two sources: (1) 50,000 civil service examination degree holders and their extended kin and (2) 198 million Sinitic characters of writing extracted from 1200 collected works. The pilot run has already revealed a surprising assemblage of *yangban* aristocrats interconnected via complex ties of patronage and marriage. As the method gets refined, and more data gets added and cleaned, I expect to discover other hidden entities and groupings. Finally, I will explain the theoretical and philosophical implications of historical entity discovery through computing by engaging with the works of social scientists and philosophers such as Gilles Deleuze, Manuel DeLanda, Norbert Elias, Zhuangzi, and Su Shi.

In addition to sharing this digital project's historical and philosophical contributions to East Asian Studies, I will share my experience with the uses of software tools to address key issues in early modern Korean history. Computational history entails the processing of digitized or born-digital sources using software packages and algorithms designed for use in another discipline or industry. Moreover, historians of East Asia may need to consider the support for Unicode encoding or rare Sinitic characters. I will explain the strategies I developed to collate genealogical data and scrape a large amount of text with the aid of a macro program. Thereafter, I will discuss my adaptation of Cytoscape, a network visualization platform designed for bioinformatics, to analyze the robust ties of marriage that contributed to the self-perpetuation and regional division of the early modern Korean *yangban* aristocrats. A highlight of this demonstration will be my linking of multiple data sources and the subsequent extraction of a subnetwork (~300 nodes) from a large network (~20,000 nodes). The marriage networks and subnetworks will be compared against the patterns of localization discovered through spatial data and text analysis. The presentation will consist of large-format prints as well as digital media shown on a monitor or a projection screen (which I will bring with me).

More Than "Nice to Have": TEI-to-Linked Data Conversion

Constance Crompton
constance.crompton@uottawa.ca
University of Ottawa, Canada

Michelle Schwartz
michelle.schwartz@ryerson.ca
Ryerson University, Canada

For developers of TEI-based projects, linked data is often much-desired but nonessential, an added output that would be nice to have, but that is not critical to ultimate success of the project. The recent catalyzation of interest in linked open data in the context of TEI (including the revitalization of ADHO's LOD SIG and the TEI's Ontologies SIG) is, however, a promising sign of our field's engagement with linked data, and our readiness to join international efforts to produce and publish linked data (Huber et al.; Pattuelli et al.; Lehmann et al.; Shadbolt et al.; Hellmann et al). Currently linked data only makes up 1% of the web, and much of that 1% is used for commercial rather than scholarly purposes (Simpson and Brown). The conversion of existing digital humanities data into linked data offers humanities scholars an opportunity to intervene in the semantic web as it is being built. It allows the power of the semantic web to be harnessed for more than just commercial purposes, and offers rich and readily accessible information about the research topic of the liberal arts: the human record. The underlying assumption of the semantic web is the same as the underlying assumption of humanities research—we can never assume ourselves to be in a full state of knowledge; there is always new information that may come to light. The creation and exposure of linked data from the vast number of existing authoritative TEI projects could enable scholars to embrace linked cultural data at scale. But what is the path to success? Our poster reflects on the technical and

institutional challenges to linked data creation, and proposes a workflow and toolset for the creation of linked data from TEI.

Despite calls in the digital humanities for TEI-linked data compatibility (Simpson and Brown, Ciotti and Tomasi), scholars have yet to develop best practices for creating linked data from richly encoded TEI resources. For many projects, the production of linked data is an ancillary goal, one that would be gratifying to achieve, but one that is secondary to the encoding itself, or only necessary to facilitate aggregation. We propose the development of XSLT-backed tools to convert and connect otherwise incommensurable data sets. The tools will require human checks, since mapping the unique usages of hierarchical elements by TEI-based projects onto existing ontologies—including CIDOC-CRM, FOAF, SKOS, schema, dcterms, and others—is hardly one-to-one. Furthermore, the historical primary source material that the TEI permits encoders to so diligently represent requires significant contextualization, since the conditions of its production were often underpinned by historical worldviews that today may be read as racist, sexist, ableist, or homophobic. Without machine and human-readable contextualization, historic intents, biases, and worldview may be reified by the inferencing that linked data permits. The ideal outcome would instead be an understanding, without valourization, of those worldviews. We are testing our tools and workflow against data sets that present exactly these challenges. We are working with four sample TEI-based data sets representing four hundred years of Atlantic cultural production, including manuscripts, books, periodicals, biographies, art works, legislation, places, and events, representing 45,000 entities. The data spans four hundred years, two regions (Europe and the Americas), five religions, three languages, all with particular historical-contextual specificity. The upcoming phases of our work will involve testing the tools against more diverse TEI sets. We are especially interested in the poster format, as we are keen to solicit feedback from peers on the balance between granularity and generality in the representation of people, places, time, and cultural production as linked data.

References

Ciotti, F., Tomasi, F. (2016). Formal Ontologies, Linked Data, and TEI Semantics. *Journal of the Text Encoding Initiative*. https://doi.org/10.4000/jtei.1480

Hellmann, S., et al. (2014). Knowledge Base Creation, Enrichment and Repair, in: Auer, S., Bryl, V., Tramp, S. (Eds.), *Linked Open Data — Creating Knowledge Out of Interlinked Data, Lecture Notes in Computer Science*. Springer International Publishing, pp. 45–69.

Huber, J., Sztyler, T., Noessner, J., Murdock, J., Allen, C., Niepert, M. (2014). LODE: Linking Digital Humanities Content to the Web of Data. I*EEE/ACM Joint Conference on Digital Libraries*. http://arxiv.org/abs/1406.0216.

Pattuelli, M.C., Miller, M., Lange, L., Fitzell, S., Li-Madeo, C. (2013). Crafting Linked Open Data for Cultural Heritage: Mapping and Curation Tools for the Linked Jazz Project. *The Code4Lib Journal*.

Shadbolt, N., O'Hara, K., Berners-Lee, T., Gibbins, N., Glaser, H., Hall, W., Schraefel, M.C. (2012). Linked Open Government Data: Lessons from Data.gov.uk. *IEEE Intelligent Systems* 27, 16–24. https://doi.org/10.1109/MIS.2012.23

Simpson, J., Brown, S. (2014). Inference and Linking of the Humanist's Semantic Web, in: Implementing New Knowledge Environments. Presented at the *Building Partnerships to Transform Scholarly Publishing*, Whistler, BC.

Animating Text Newcastle University

James Cummings
james.cummings@newcastle.ac.uk
Newcastle University, United Kingdom

Tiago Sousa Garcia
tiago.sousa-garcia@newcastle.ac.uk
Newcastle University, United Kingdom

Animating Text Newcastle University

This *DH 2018* poster will provide an introduction to a new kind of digital humanities research network and the pilot projects it is building. Animating Text Newcastle University (ATNU) is a three year interdepartmental research project exploring new frontiers at the cross-roads between traditional scholarly textual editing, digital editing, digital humanities and computer science. It is a collaboration between humanities researchers and computing scientists that is exploring research questions raised by pre-1860 editing projects. The poster at *DH 2018* will introduce the ATNU network, the successes and failures of the project so far, and the individual pilot projects it has undertaken.

ATNU connects original historical research from across Newcastle University from the School of English Literature, Language and Linguistics, the School of Arts and Cultures, the School of Modern Languages, and the School of History, Classics and Archaeology, with the transformational research of the Digital Institute. The intention is to share expertise and intellectual resources and to work to deliver ambitious, future-facing research that will nurture future large-scale collaborative projects. The network is hosting invited expert workshops, visiting speakers, and undertaking pilot digital projects informed by editing challenges. It is hoped that this will not only increase familiarity with DH methodologies and technologies inside the institution but foster partnerships outside it.

Why pre-1860 texts?

In these earlier periods the characteristics of manuscript and the printed book (and their relationship with one another) are fundamentally distinct from how they are in the period from the late nineteenth century to the present. Yet the ways in which pre-1860 texts are re-presented in current print and digital editions often fails to recover their vital, distinctive contexts (the relations between authors, copyists, printers, publishers and booksellers), and the way the printed page is meant to facilitate particular experiences. ATNU is contributing to a vital debate not just about the history of the text and the future of the book, but also about the place of historically-focussed editorial scholarship in the story of the humanities and its digital future.

Funding Streams and Resistance to Failure

A frustrating aspect of many research projects is the tendency to promote their successes and ignore failures. These projects may produce excellent outputs which benefit the humanities, but in discussing their projects they often count the hits and ignore the misses. It is completely understandable when highlighting the success of their projects to those who funded them. However, ATNU is fortunate in being slightly different: it is funded by Newcastle University's Research Investment Fund specifically to bolster digital humanities research at the institution. Part of the ATNU mission is the development of additional grant applications for cutting edge projects that specifically have their basis in more risky blue skies thinking. Moreover, in order to develop these funding bids ATNU is undertaking a series of pilot projects but because these are funded internally they are allowed to be more experimental. They do not have to be successes -- failure is indeed an option! Where the pilot projects succeed they will go on to be the base for external funding bids, but where these projects are less successful, their failures can be publicly documented and projects can be re-oriented towards more successful techniques.

Pilot Projects

The network's pilot projects are in three categories: "Manuscripts and Print", "Performance", and "Translation". The projects in each of these have a set of shared interests, methodologies, and an overlap of possible technological solutions.

- **Manuscript and Print:** the projects in this area investigate topics such as scholarly digital editing, the process of collaborative editing, the presentation of editions, and the handling of variation across multiple versions. The first pilot is a prototype digital edition of the Sarum Hymnal involving text, image, and music encoding.
- **Performance:** many texts have a life beyond the page, and include acoustic and visual experiences. ATNU is exploring how best to represent and enable these performative and interactive dimensions. One pilot in this looks at a visual, interactively animated, view of James Harrington's early modern proposal for reforming voting systems, another experiments with the acoustic effect of punctuation in early modern texts.
- **Translation:** investigating pre-modern texts and their translations, how these entities relate, and developing tools for researchers comparing texts in translation. A pilot under this theme is examining the concept of the social translation.

The poster will provide more details about the network and its pilot projects.

Una Investigación a Explotar : Los Cristianos de Alá, Siglos XVI y XVII

Marianne Delacourt
marianne.delacourt@univ-tlse2.fr
Université Toulouse Jean Jaurès, France

Véronique Fabre
veronique.fabre@univ-tlse2.fr
Maison des Sciences de l'Homme et de la Société de Toulouse/ CNRS, France

En los siglos XVI y XVII, el Mediterráneo fue el reto geopolitico entre la Monarquia Española y el Imperio Otoman. Entre batallas e incursiones, muchos cristianos fueron reducidos a la esclavitud por los Berberiscos. Unos,para suavizar sus condiciones de vida o por fuerza, se convertieron al Islam, y fueron llamados *Renegados* Ellosfueron *puentes* entre las dos civilisaciones y religiones.

Cuando regresaban a la vida cristiana, fueron juzgados por la Inquisición.

Bartolomé Bennassar, historiador frances, hizo, al fin de los 80, fichas de papel sobre más de 1550 renegados, basadas en las fuentes de los archivos de la Inquisición.

Nuestro proyecto es digitalizar esas fichas y crear una base de datos, albergada en la plataforma francesa de humanidades digitales del CNRS : *HUMA-NUM*

El poster que queremos presentar da cuenta del método y de las etapas de un proyecto de *Humanidades digitales* entre dos instituciones que no suelen trabajar juntas.

La numerisacion fue bastante facil... construir la base de datos que permitara interrogar a las fichas es muchas mas dificil.

El Profesor Bennassar habia preparado fichas dactilograficas con datos fijos tipo nombre, lugar de nacimeinto, condiciones de renegacion, etc...Pero, leyendo el archivo de los procesos, añadia muchas informaciones

manuscritas que vienen enriquecer el perfil de vida del renegado, pero que no son «normalizadas».

Es decir que para construir la base de datos teniamos que ser pertinente con las estructuras de interrogacion y decidir cual informaciones adicional teniamos que tomar en cuenta para aclarar unos datos biograficos del renegado y sucitar el interes del investigador .

Entonces, pedimos a historiadores de validar cada etapa de la elaboracion de la base de datos.

Esperamos que esos datos seran explotados: que los datos geograficos un dia sean explotados por un logiciel de visualisacion y analisis de redes, que la informacion «se casa con uzansas de moros», permiten investigar sobre la vida intima de los renegados, etc…

Asi que como lo dice el titulo del poster : « los critianos de Ala, una investigacion a explotar»

The Iowa Canon of Greek and Latin Authors and Works

Paul Dilley
paul-dilley@uiowa.edu
University of Iowa, United States of America

This poster will introduce the Iowa Canon of Greek and Latin Authors and Works, which aims to be the most comprehensive list of classical texts from the origins of Greek and Latin literature through the end of the Antiquity (the 6th century CE), and associated metadata, made available for researchers through an innovative online interface. The Iowa Canons are affiliated with the Big Ancient Mediterranean Project, for which I am a co-PI with Sarah Bond, with lead developer Ryan Horne, which seeks to provide an interface for the coordinated exploration of linked textual, geospatial, and network data relating to the ancient world. Both BAM's interface and the Iowa Canons are in development; a beta-version of the Iowa Canon of Latin Authors and Works is available at http://bam.lib.uiowa.edu/iclaw/. The Iowa Latin Canon currently stands at over 5,400 works; a more extensive version, paired with the Iowa Canon of Greek Authors and Works, which currently includes over 9,000 entries, will be published in May 2016. I have been assisted in data collection by students in my graduate seminars on distant reading, as well as undergraduate and graduate research assistants.

The goal of both Iowa Canons is to integrate existing canons of Greek and Latin Literature, especially the Perseus Catalog, the Thesaurus Linguae Graecae (TLG) Canon, the Packard Humanities Institute (PHI) Classical Latin Texts, the Brepols Library of Latin Texts (LLT-A), and other resources such as the Clavis Apocryphorum; to increase their granularity and the amount of associated metadata; and to make this data collection searchable in an interface that integrates Greek and Latin texts, which none of the previous Canons do. None of the existing Canons include lost works, and they group fragmentary works under a single entry (e.g. "Fragmenta"), with no functionality to search for individual titles within it, which sometimes number in the hundreds. The Iowa Canons, in contrast, will include all known lost or fragmentary works, and include additional metadata, such as time and place of composition, genre (using the same "in-house" classification system for both Greek and Latin texts), meter (if poetic), and Christian/non-Christian content. Finally, the Iowa Canons will cross-reference each work to existing canons (when possible), as well as to the Perseus Catalog, which will provide stable reference urns for Greek and Latin works, a project with which we are collaborating.

The Iowa Canon of Greek and Latin Authors and Works will make this data available to users through an interface, which will provide faceted search of available metadata, for example, by selecting all works of a particular genre, in a specified time period and/or location. The results of the search are displayed geospatially, with circles around all locations with relevant works, their diameters proportionate to the number of "hits" in that location. Clicking on the circles reveals those "hits." When combined with the extensive records of lost and fragmentary titles, this search functionality will greatly facilitate research into Greek and Latin literary history beyond the usual focus on canonical works, which will themselves be contextualized. Jockers has described this sort of research metadata as the "lowest hanging fruit of literary history" (Jockers 2013: 35); his work, as well as Franco Moretti's (Moretti 2009), have explored the possibilities of this approach for studying certain genres of 19th and 20th century literature in English, which is of course far more extensive than surviving ancient Greek and Latin literature. But the cumulative metadata that will be accessible through the Iowa Canons will offer a unique picture of an entire literary field, with over 60 genres, as it developed over centuries, and in several languages. The poster will be of interest not only to digital classicists, but to literary scholars working in other languages and eras, from whom I will solicit feedback about its functionality, as well as its potential for distant reading.

References

Jockers, Matthew, *Macroanalysis: Digital Methods and Literary History* (University of Illinois Press, 2013)

Moretti, Franco, "Style, Inc. Reflections on Seven Thousand Titles (British Novels, 1740-1850)," *Critical Inquiry* 36 (2009): 134-58.

Packard Humanities Institute Latin Author List: http://latin.packhum.org/browse

Perseus Catalog: http://catalog.perseus.org/

Thesaurus Linguae Graecae: http://stephanus.tlg.uci.edu/index.php

Digital Storytelling: Engaging Our Community and The Humanities

Ruben Duran
ruben.duran@hccs.edu
Houston Community Collegeo, United States of America

Charlotte Hamilton
charlotte.hamilton@hccs.edu
Houston Community Collegeo, United States of America

Houston Community College is one of the leading two-year colleges in the United States incorporating digital storytelling into the curriculum while reaching out to the community to achieve the history of the diverse communities with vibrant background that provides such a rich tapestry that makes Houston the city it is today.

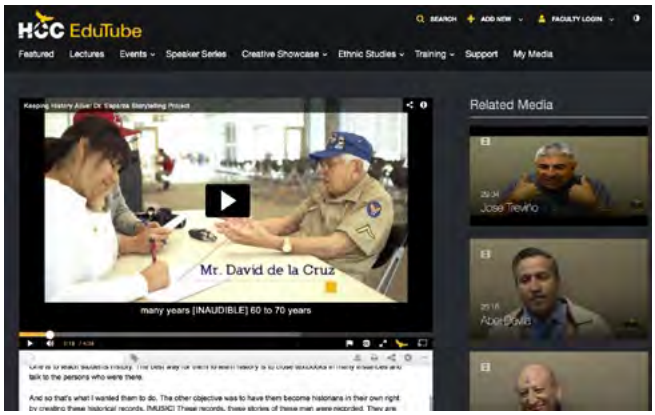
Working with the Center for Digital Storytelling we have trained our faculty and staff to incorporate these initiatives into the instructional curriculum.

Digital Storytelling supports projects that bring ideas and insights of the humanities to life for general audiences. Our past projects engage humanities scholarship to analyze significant themes in disciplines such as history, literature, and art history. Our projects support and encourage activities that involve members from the many Houston cultural communities through collaboration with humanities scholars and students. We have also invited contributions from the community in the development and delivery of humanities programming.

These presentations provide video examples of the following initiatives:

History of Latino war veterans from Korean and Vietnam Wars

Students from a Mexican American history class interviewed veterans from the Korean and Vietnam Wars. The veterans expressed pride in their contributions to the war, some of them for the first time since returning home from their deployments many years ago. Students shared their excitement while developing insight into history through stories not contained in their textbook.

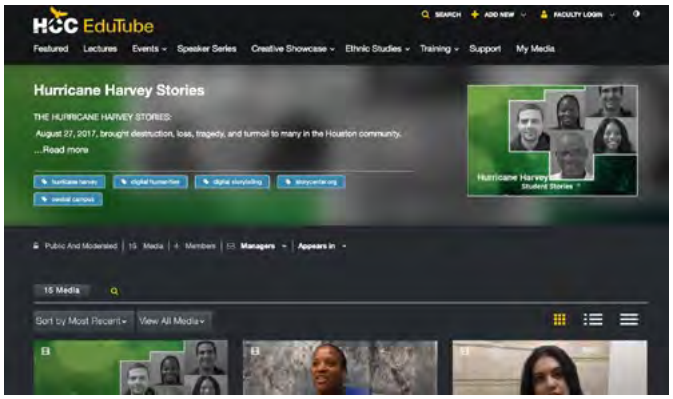


MECA grant for outreach to K-12 students

MECA is a community-based non-profit organization committed to the development of under-served youth and adults through arts and cultural programming, academic excellence, support services, and community building. Under the tutorage of library trainers, using staff and equipment resources from our institution, students produced short videos of their interviews with members of their community. This project fostered in the students a better understanding of the significant contributions of their community peers, and it helped them to develop discipline, self-esteem, and increased cultural pride.

Harvey Listening Stations provide support through student and staff stories

Following the disastrous hurricane that stuck Houston in August, 2017, we provided equipment and the opportunity for staff and counselors to capture stories reflecting the impact on individual students of the devastation of the flooding. The objective was both to allow students to tell their stories, as well as to determine whether there were specific actions we could implement to provide support for their continued success in their educational activities.



Current project is the collaboration with HHA 2018 Year of the Woman in Houston

Working through the Texas Historical Association and the Houston Historical Alliance we are providing class assignments that include developing short videos and scholarly research on women in the greater Houston community who affected or influenced the history of our region and the state. Based on established guidelines this initiative allows our students to complete the project as a class assignment in history, government or other disciplines. The women include pilots, activists, oil magnates, storytellers, scientists, ranchers, daughters, and mothers who have made significant contribution to the richness of our diverse communities. The digital stories should include notable women, as well as lesser known figures. These videos will be hosted on our Media Space as a reference tool and

will be eligible for selection by a peer jury for inclusion in the online *Handbook of Texas Women*.

The poster session showcases tools from our storytelling arsenal that includes the Listening Stations and iPads displaying referencing videos from our initiatives. All the projects were developed using WeVideo, a collaborative cloud editing application that serves as the online video editor that makes it easy to capture, create, view and share the stories.

The stories are shared in Edutube, HCC's media community tube. The Learning Station was designed as a public kiosk for people to share stories with a listener, hold a conversation, or be part of an interview. The included app automates the upload and delivery of files to participating organizations and the participants. The apps integrates the metadata collection, registration, release and transcription processes, making it a state-of-the-art tool for gathering primary source material for documentary projects.

References

Center for Digital Storytelling. Storycenter.org
WeVideo. Wevideo .com
Edutube. https://edutube.hccs.edu

Text Mining Methods to Solve Organic Chemistry Problems, or Topic Modeling Applied to Chemical Molecules

Maciej Eder
maciejeder@gmail.com
Pedagogical University in Kraków,
Institute of Polish Language, Poland

Jan Winkowski
jan.winkowski@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences), Poland

Michał Woźniak
michal.wozniak@ijp.pan.pl
Institute of Polish Language (Polish Academy of Sciences), Poland

Rafał L. Górski
rafal.gorski@ijp.pan.pl
Pedagogical University in Kraków,
Institute of Polish Language, Poland

Bartosz Grzybowski
nanogrzybowski@gmail.com
Pedagogical University in Kraków,
Institute of Polish Language, Poland; Ulsan National Institute of Science and Technology, Korea

Introduction

The Renaissance Humanism was probably the last moment in the history of ideas when the development of exact sciences was shaped according to the intellectual paradigms of the humanities (the Liberal Arts, to be precise). After the advent of the Scientific Revolution in the 17th century – with its empiricism, experimental reasoning, mathematical apparatus, and so forth – the exact sciences became the point of reference for all the other disciplines, in terms of scientific inference and its methodology. The imbalance between the humanities and the sciences has been growing ever since. Nowadays, statistical analysis is routinely applied in social sciences, cognitive linguistics tries to take advantage of the fMRI technology, text analysis studies are overwhelmed by numerous machine-learning techniques, ranging from hierarchical cluster analysis to Support Vector Machines classification and Deep Learning. The exact sciences have affected the humanities to a considerable extent, but at the same time they continue to be rather resistant to any methodological inspirations coming from the "soft" scholarship. This study is an example of such a reversed influence, since we propose to apply text mining methods to study chemical molecules. Arguably, the phrase "If an atom is a letter, then a molecule is a word", even if popular in chemistry, sounds rather naïve for anyone who has some expertise in linguistics. Nonetheless, despite a shallow similarity between language structures and organic chemistry at first glance, the methodology developed in text mining proves very promising as a way to discover internal molecule structures.

The problem

One of the biggest issues in contemporary organic chemistry is an enormous number of different molecules

and their fragments that play role in chemical reactions. To cut a long story short: any reaction involves certain changes in molecules' structures, which usually means that certain bonds are disjoined, and particular atoms change their positions within each molecule. On theoretical grounds, these changes can be predicted and/or controlled. In practice, however, predicting optimal bond cuts requires high-level expert knowledge, due to the extreme complexity of the problem, or an enormous computer power to run brute-force combinatoric algorithms. This is, however, still far beyond our capabilities, because completing a task that involves testing billions of billions of combinations would require decades if not centuries. For that reason, the big question at stake is how to optimize the entire process of identifying relevant molecule substructures (Ruddigkeit et al., 2012).

Splitting complex chemical molecules into "meaningful" substructures is the first problem to be overcome. In this context, "meaningful" means groups of atoms that are local centers of reactions. The nature of bonds between atoms is very well understood since the first half of the 20th century. However, it is still unclear why certain clusters of atoms tend to keep together while reprehend some other groups. Being one of the most crucial issues in organic chemistry, this question has been approached using different methods, which are aimed at finding repetitive fragments of molecules. It can be assumed that methods derived from text mining can be adopted to (partially) solve the task.

Chemical "words"

Let us assume that a molecule is a sentence (with some obvious caveats in mind, non-linearity of molecules being the most important one). If so, then a list of known molecules can be considered a corpus. Quite striking is the fact that a commonly used convention of describing chemical structures (referred to as SMILES) uses sequences of characters, what makes any comparisons to corpora even more natural. E.g., caffeine is coded as follows: CN-1C=NC2=C1C(=O)N(C(=O)N2C)C.

To make the language–chemistry parallel complete, one has to define "words" as well, keeping in mind that there are no explicit substructure boundaries in molecules. To this end, we adopt the idea of Cadeddu et al. (2014), who compared a few thousands of molecules pairwise, in order to extract their maximum common substructures, with the belief that they represent chemical "words"; this step was followed by a term frequency–inverse document frequency (tf/idf) heuristic.

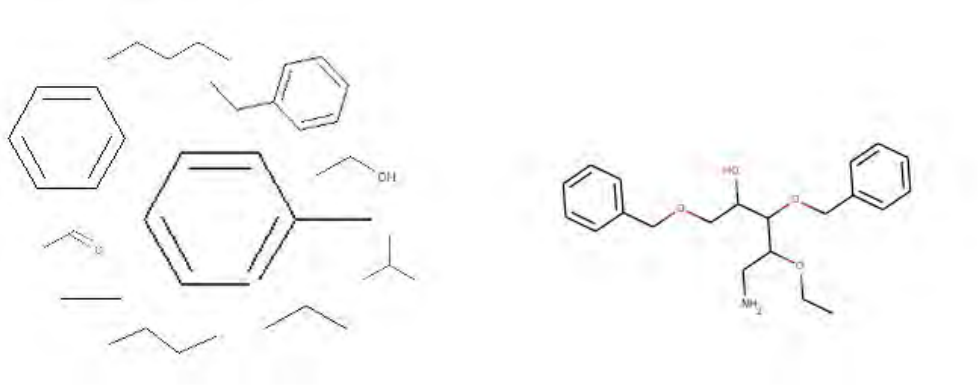


Fig. 1: Chemical "words" defined as maximum common substructures shared by chemical molecules: 10 most frequent chemical "function words" (left), and an example of an unfrequent "content word" (right).

Using the above idea of extracting "words", we picked randomly 50,000 reactions from the Reaxys database (www.reaxys.com), and computed the pairwise comparison, resulting in a corpus of >800,000 word types and 2.5 * 10^9 tokens. Interestingly enough, the chemical "words" share the characteristics of a typical natural language, e.g. they follow the Zipf's law, but they also exhibit the behavior of function and content words in their relation to frequency (see Fig. 1). Moreover, the chemical "words" can be subject to time-proven text mining methods such as keywords analysis, as has been demonstrated in our previous study (Woźniak et al., 2018).

Topic modeling

In order to identify any relations between chemical "words", we analyzed our corpus using topic modeling (Blei et al., 2003), a technique that attracted a good share of attention in Digital Humanities, but has never been popular beyond text-centric applications. Topic modeling belongs to a group of distributional semantics methods, which are based on a general assumption that the meaning of a word is defined by its lexical context (Firth, 1962). In its extended form, the distributional hypothesis says that the degree of semantic similarity between words can

be modeled as a function of the degree of overlap among their linguistic contexts (Miller and Charles, 1991; Baroni and Lenci, 2010). Topic modeling, usually computed via the LDA algorithm (Blei et al., 2003) assumes the "bag-of-words" type of context, which means that the sequence of words in a sentence is irrelevant. This feature allows for computing chemical "words", which, essentially, do not follow any linear sequence.

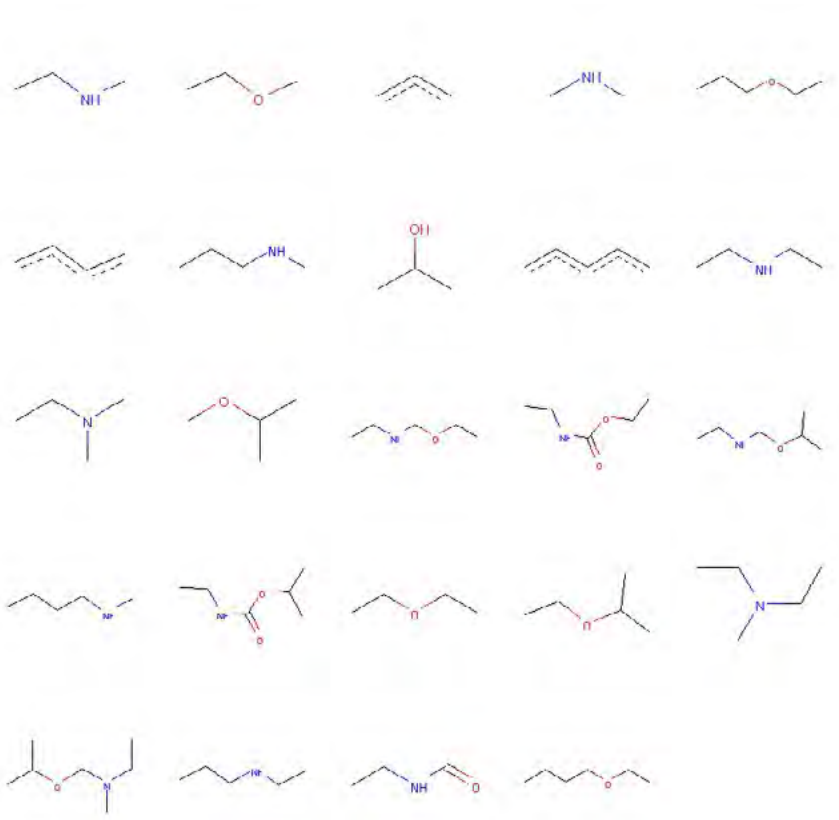


Fig. 2: Topic 47 extracted from the corpus of chemical "words".

We trained a few models ranging from 50 to 200 topics, using the LDA technique. Therefore, we were able to substantially reduce the enormous number of >800,000 "word" types into a small number of word constellations (topics) that contain meaningful information about co-occurring chemical fragments. One of the topics is shown in Fig. 2. Among the 24 most distinctive "words" one can recognize some amines, fragments of aromatic rings, fragments containg carboxyl functional groups, and so on. Inspected by trained practitioners in organic chemistry, the topics revealed several collocations that seemed meaningful, and could not have been identified in the original (raw) collection of molecules. Despite the intuitive interpretation via close-reading, however, such an outcome inevitably leads to a more serious question, namely if one can define *meaning* in organic chemistry, in the context of distributional semantics.

Classification

Interesting as they are, the chemical topics cannot solve any real-life problem *per se*, even if they seem to be meaningful from the naked eye's perspective (note that the same holds for topic modeling based on texts). Specifically, one cannot discover any general structure of, say, natural products by manual inspection of their prominent topics, nor can one predict if a given substance is likely to be toxic. There is a plethora of similar classification (or prediction) tasks where topics might prove useful, provided that the analysis goes beyond the close-reading perspective. If the topics' proportions are indeed significantly different across the corpus – i.e. if they really keep some information about semantic differentiation between the molecules – they should be applicable as a set of input features for machine-learning classification.

To test this hypothesis, we designed a controlled experiment on a (somewhat artificial) problem of classifying molecules as potential drugs. Again, we used the same Reaxys database to extract relevant training material: 1,800 known drugs and a similar number of known non-drugs. Our two-class supervised setup involved a simple neural network (implemented via Keras with Tensorflow backend), the input layer being the most probable topics for each chemical molecule. The final results varied de-

pending on a topic model used for prediction, nevertheless they turned out to be fairly optimistic. The best accuracy was: 0.7851 (the model for 200 topics), the worst: 0.7135 (the model for 50 topics). Even if preliminary, these results suggest that some semantic information can be indeed extracted from chemical corpora using text mining algorithms.

Acknowledgements

This research is part of project UMO-2014/12/W/ST5/00592, supported by Poland's National Science Centre.

References

Baroni, M. and Lenci, A. (2010). Distributional memory: A general framework for corpus-based semantics. *Computational Linguistics*, 36(4): 673–721.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993–1022.

Cadeddu, A., Wylie, E. K., Jurczak, J., Wampler-Doty, M. and Grzybowski, B. (2014). Organic chemistry as a language and the implications of chemical linguistics for structural and retrosynthetic analyses. *Angewandte Chemie*, 126(31): 8246–50.

Firth, J. R. (1962). A synopsis of linguistic theory 1930-55. In Firth, J. R., *Studies in Linguistic Analysis*. Oxford: Blackwell.

Miller, G. A. and Charles, W. G. (1991). Contextual correlates of semantic similarity. *Language and Cognitive Processes*, 6(1): 1–28.

Ruddigkeit, L., Deursen, R. van, Blum, L. C. and Reymond, J.-L. (2012). Enumeration of 166 billion organic small molecules in the chemical universe database GDB-17. *Journal of Chemical Information and Modeling*, 52(11): 2864–75.

Woźniak, M., Wołos, A., Modrzyk, U., Górski, R. L., Winkowski, J., Bajczyk, M., Szymkuć, S., Grzybowski, B. and Eder, M. (2018). Linguistic measures of chemical diversity and the 'keywords' of molecular collections. *Scientific Reports*, 8: forthcoming.

Studying Performing Arts Across Borders: Towards a European Performing Arts Dataverse (EPAD)

Thunnis van Oort
t.vanoort@uva.nl
University of Amsterdam, The Netherlands

Ivan Kisjes
i.kisjes@uva.nl
University of Amsterdam, The Netherlands

EPAD is an emerging network of scholars investigating the history of European performing arts (theatre, music, cinema) using digital methods and (shared) datasets. EPAD builds on the infrastructure and expertise collected in existing projects at the CREATE research program at the University of Amsterdam with a data-driven approach to the history of cinema, theatre and music performances, functioning as point of departure from where more extensive European cross-sectorial cooperation can develop.

Cultural performances in theatre, music and film have contributed vividly to the formation of individual and social identities in the European past. Cinemas, theatres and concert halls are places par excellence to examine how modern notions of identity like nation, class or gender were forged in a collective, 'live' appropriation of ideas, images and experiences (Balme, 2014; Furnée, 2012).

Traditionally, scholarship in music, theatre and film history has prioritized the study of the artwork over its consumption. Since the 1980s, the prevalent text-oriented perspectives have been complemented by a substream of historiography contextualizing the distribution and reception of performing arts (Allen & Gomery 1985; Booth 1991; Fischer-Lichte 1997; Gerhard 1992; Johnson 1995; Staiger, 1992; Weber 1975 Wollenberg and McVeigh 2004). This research tradition is dominated by qualitative approaches often based on distinct case studies, the results of which have proven hard to compare or generalize beyond the local scale (Biltereyst et al. 2018; Cowgill and Rushton 2006; Maltby 2006; Müller 2014). More advanced digital methods and larger datasets can push the research agenda beyond the prevailing particularism by providing wider comparative frameworks and new levels of generalization. Upscaling the scope yields largely uncharted possibilities for transnational perspectives on the relations between cultural consumption and the formation of shared identities (Balme 2015; Charle 2008; Garncarz 2015; Hall-Witt, 2007; Sedgwick 2000). Furthermore, EPAD's interdisciplinarity promises rare insights in the extent to which audiences of theatre, music and film overlapped and shared socio-cultural characteristics (Engelen et al. 2017; Furnée 2017; Röttger 2017).

Current historiography on the consumption of performing arts is predominantly conceived in local or national frameworks, often limited to the discipline-specific object music, theatre or film. In relative isolation, European musicologists, film and theatre scholars are confronting similar historical questions and methodological and technical issues. Joining forces opens up an agenda of transnational and cross-sectoral comparative research, that does justice to the capacity to travel across geographical, social and medium boundaries that is so characteristic of the performing arts. Moreover, data-driven historical audience research has the capacity for significant revisions of established cultural canons or genre hierarchies (Blom and Van Marion 2017; Garncarz 2015; Nieuwkerk 2017; Weber and Newark forthcoming).

Dozens of performing arts databases are scattered across Europe (Baptist et al. forthcoming). These multiform online data collections contain a variety of information on programming, and/or the venues, locations, people and organisations involved in theatrical presentation. Aggregated and combined with socio-economic data, these data can generate new insights in the social meanings of the cultural exchanges in European theatres and concert halls, for instance by delineating taste patterns and (other) socio-spatial audience characteristics.

To realize a data-driven history of the performing arts we need to join forces. The EPAD network strives to open up an exchange of expertise, data and technical know-how. To develop this research agenda, collaborating scholars need to find solutions in three (interlocking) domains:

1) address methodological-ontological questions. To facilitate comparative research into the socio-cultural dynamics of performing arts audiences, we need to reflect on the definitions of the objects of study. What exactly constitutes a performance, a venue? Can we agree on shared ontologies for structuring our data?
2) develop and refine a theoretical-historiographical framework for comparative, transnational and interdisciplinary research into the performing arts that addresses the relation between cultural consumption and social identity formation.
3) confront technical-infrastructural issues: outline the conditions for data interoperability. How can existing facilities and tools best be utilized for creating a virtual research infrastructure for comparative transnational research on the history of performing art cultures? We aim to build upon the CLARIAH infrastructure and tools for harmonizing and querying socio-economic datasets based on a linked data approach (CLARIAH Structured Data Hub). The work involves developing ontologies, shared data models and thesauri containing internationally shared terms for performing arts data, as well as building the actual infrastructure within the context of the European Digital Research Infrastructure for the Arts and Humanities DARIAH.

References

Allen, R. and D. Gomery. *Film History: Theory and Practice*. New York, 1985.

Balme, C. *The Theatrical Public Sphere*. Cambridge, 2014.

Balme, C. 'The Bandmann Circuit: Theatrical Networks in the First Age of Globalization,' *Theatre Research International* vol. 40 no. 1 (2015) pp. 19-36.

Baptist, V., T. van Oort and J. Noordegraaf. 'Mapping European Performing Arts Databases: An Inventory of Online Historical Data Projects,' in: N. Leonhardt ed. *The Routledge Companion to Digital Humanities in Theatre and Performance*. Abingdon, forthcoming.

Biltereyst, D., T. van Oort and P. Meers, 'Comparing Historical Cinema Cultures: Reflections on New Cinema History and Comparison with a Cross-National Case Study on Antwerp and Rotterdam,' in: R. Maltby, D. Biltereyst and P. Meers eds. *The Routledge Companion to New Cinema History*. Abingdon, 2018 (in press).

Blom, F. and O. van Marion. 'Lope de Vega and the Conquest of Spanish Theater in the Netherlands,' *Prolope. Anuario Lope de Vega. Texto, literature, cultura* no. 23 (2017) pp. 155-177.

Booth, M. *Theatre in the Victorian Age*. London, 1991.

Charle, C. *Théâtres en capitales. Naissance de la société du spectacle à Paris, Berlin, Londres et Vienne, 1860-1914*. Paris, 2008.

Cowgill, R. and J. Rushton. *Europe, Empire, and Spectacle in Nineteenth-Century British Music. Music in 19th-Century Britain*. Aldershot, 2006.

Engelen, L., R. Vande Winkel and L. Van de Vijver eds. Spektakelcultuur in de Lage Landen. Special Issue *Tijdschrift voor mediageschiedenis* vol. 20 no. 2 (2017).

Fischer-Lichte, E. *Die Entdeckung des Zuschauers. Paradigmenwechsel auf dem Theater des 20. Jahrhunderts*. Tübingen and Basel, 1997.

Furnée, J. *Plaatsen van beschaafd vertier. Standsbesef en stedelijke cultuur in Den Haag, 1850-1890*. Amsterdam, 2012.

Furnée, J. 'Cultuurliefhebbers. Sociale structuren en persoonlijke voorkeuren,' Inaugural Lecture Radboud University, Nijmegen (24 March 2017).

Garncarz, J. *Wechselnde Vorlieben:* Über *die Filmpräferenzen der Europäer, 1896-1939*. Frankfurt and Basel, 2015.

Gerhard, A. *Die Verstädterung der Oper: Paris und das Musiktheater des 19. Jahrhunderts*. Stuttgart, 1992.

Hall-Witt, J. *Fashionable Acts: Opera and Elite Culture in London, 1780-1880*. Hanover, 2007.

Johnson, J. *Listening in Paris: A Cultural History*. Berkeley, 1995.

Maltby, R. 'On the Prospect of Writing Cinema History from Below', *Tijdschrift voor mediageschiedenis* vol. 9 no. 2 (2006), pp. 85–7.

Müller, S. *Das Publikum macht die Musik. Musikleben in Berlin, London und Wien im 19. Jahrhundert*. Göttingen, 2014.

Nieuwkerk, M. van. 'The Felix Meritis Concert Program Database. Work-in-progress in Research and Data Curation,' Paper at CREATE ACHI Conference, October 2016, Amsterdam.

Röttger, K. 'Technologies of Spectacle and "The Birth of the Modern World": A Proposal for an Interconnected Historiographic Approach to Spectacular Culture,' *Tijdschrift voor mediageschiedenis* vol. 20 no. 2 (2017) pp. 4-29.

Sedgwick, J. *Popular Filmgoing in 1930s Britain: A Choice of Pleasures*. Exeter, 2000.

Staiger, J. *Interpreting Films: Studies in the Historical Reception of American Cinema*. Princeton, 1992.

Weber, W. *Music and the Middle Class. The Social Structure of Concert Life in London, Paris and Vienna*. London, 1975.
Weber, W, and C. Newark, eds. *The Oxford Handbook of the Operatic Canon*. Oxford, forthcoming.
Wollenberg, S. and S. McVeigh eds. *Concert Life in Eighteenth-Century Britain*. Aldershot, 2004.

The Archive as Collaborative Learning Space

Natalia Ermolaev
nataliae@princeton.edu
Princeton University, United States of America

Mark Saccomano
mss2221@columbia.edu
Columbia University, United States of America

In 2014, the Columbia University Rare Book & Manuscript Library (RBML) acquired a unique archival collection. The Serge Prokofiev Archive, which contains materials related to the twentieth-century Russian composer Sergei Prokofiev (1891-1953), contains more than 17,500 diverse items: music manuscripts, letters, financial documents, scores, concert programs, notebooks, monographs, articles, journals, photographs, audio and visual recordings, and ephemera in original, photocopy, and digital formats. The archive was first established in 1994 at Goldsmiths College, London (Mann, 2008). In the twenty years that it grew, a complex, intricate, and item-level descriptive apparatus evolved alongside. By the time the collection came to Columbia, the archival items were accompanied by hundreds of metadata files in formats such as spreadsheets, Word documents, text files, PDF, Endnote databases, Access database, MARC records, and various XML encodings. Our poster describes how we — an archivists and digital humanities researcher — curated, explored, and analyzed, this dense and diverse body of data.

Our first steps were to satisfy the immediate need of funders and stakeholders: making records of the Prokofiev Archive publicly available through the finding aid on the RBML website. Though the goal was clearly defined — records in XML using Columbia's EAD (Encoded Archival Description) schema and style guide — the process was complex. Records from Goldsmiths differed in both structure and content depending on the item catalogued. For example, data about books was captured in EndNote and MARC, while information about music manuscripts was kept in Excel spreadsheets, and correspondence records were in an Access database. We worked to transform all data into XML, and then ran customized XSLT transformations to generate standard EAD. However, what we gained in standardization we lost in information richness: this custom EAD schema didn't allow the encoding elements at the level of granularity we had in the original records. Significant scholarly information was lost. In addition, the conventional finding aid interface limits the user's options for exploring a large archival collections: content is presented in blocks of narrative, long lists of items, and search and browse organized by series and sub-series that does not allow for easy cross-collection discovery.

Thus, our next task was to find alternatives for the analysis and representation of the Serge Prokofiev Archive. We decided to pivot our approach, and moving away of EAD, transformed structured XML into a series of CSV files that could be manipulated with various data analysis and visualization tools. Not surprisingly, both the processes and our results deepened our understanding of the archive and of Prokofiev's work and legacy: an alluvial chart of the music manuscript series, for example, showed patterns in the way Prokofiev used multiple languages for different types of annotations and markings as he wrote his scores; a map using location data of Prokofiev's letters revealed his correspondence with Russian-American composers who had emigrated to China; a network graph using metadata about the secondary literature on Prokofiev (books, journal articles) showed surprising connections between editors and authors in Soviet and Western publications.

Our experience demonstrated the value of creative engagement with archival data; through experimentation and play, the Serge Prokofiev Archive became a site of collaborative research and learning. Our work was guided by two important conceptual shifts in the library and archives profession: one is the "Collections as Data" movement, which encourages reframing the digital object as data (Padilla, 2016), and the second is the move away from locating value exclusively in the *objects* of a collection to the impact collections have on *people* and *communities*. In Kate Theimer's notion of "archives as platform," tools and technologies help users interact with archives in creative ways that add value to their lives and experiences. Work that takes place "behind the scenes" (Theimer, 2014) by archivists and their collaborators helps define the archive as a dynamic cross-disciplinary learning space.

References

Noëlle Mann, "The Serge Prokofiev Archive in London - A Complex Story," *Fontes Artis Musicae*,Vol. 55, No. 3 (July-September 2008), pp. 543-547.
Thomas Padilla, "On a Collections as Data Imperative," conference report, Collections as Data: Stewardship and Use Models to Enhance Access, Library of Congress, Washington, DC, September 27, 2016,
Kate Theimer, "The Future of Archives is Participatory: Archives as Platform, or A New Mission for Archives," April 3, 2014. http://archivesnext.com/?p=3700&cpage=1#comment-4180873

Tensiones entre el archivo de escritor físico y el digital: hacia una aproximación teórica

Leonardo Ariel Escobar
leonardo.ariel11@gmail.com),
Universidad Autónoma del Estado de Morelos, Mexico

En mi investigación doctoral exploro el "archivo de escritor" como un artefacto que cambia la manera en que los lectores se relacionan con los textos de determinado escritor, entendiendo como dispositivo la: "disposición de una serie de prácticas y de mecanismos (conjuntamente lingüísticos y no lingüísticos, jurídicos [...]) con el objetivo de hacer frente a una urgencia y de conseguir un efecto" (Agamben). En este orden de ideas, una cita a Derrida es acertada y ahí radica lo arcóntico del archivo: "No solo aseguran la seguridad física del depósito y del soporte sino que también se les concede el derecho y la competencia hermenéuticos. Tienen el poder de *interpretar* los archivos" (Derrida 1997 10).

En la tesis se ha tomado como materia de estudio el archivo del escritor Gabriel García Márquez, repartido entre la Universidad de Texas en Austin y la Biblioteca Nacional de Colombia, aunque el segundo no sea muy numeroso y se trate solo de cierto material bibliográfico especifico (La Nación, 2014).

El estudio se emprende en medio de un panorama teórico que no resulta muy numeroso respecto a las definiciones del archivo de escritor. Se sabe que el archivo de escritor puede tener diferentes significados, no obstante, debemos aclarar que se tomará en la siguiente de sus acepciones dentro de esta propuesta:

> un conjunto organizado de documentos, de cualquier fecha, carácter, forma y soporte material, generados o reunidos de manera arbitraria por un escritor a lo largo de su existencia, en el ejercicio de sus actividades personales o profesionales, conservados por su creador o por sus sucesores para sus propias necesidades o bien remitidos a una institución archivística para su preservación permanente (Goldchluk y Pené 13).

Hoy en día una de las principales preguntas que se realizan a la hora de postularse a una beca de estancia en un archivo físico de escritor es justificar el porqué es obligatoria la consulta del archivo del escritor en físico, aunque se encuentre gran parte de dicho legado en forma digital (Harry Ransom Center, 2017). La idea del presente escrito es precisamente observar qué tensiones se encuentran presentes entre una y otra forma del artefacto, ya que aunque se pudiera decir que son equiparables y que equivalen a lo mismo, están lejos de cumplir una misma función en común, debido a que sustancialmente pienso que funcionan de maneras distintas. Encontrar una aproximación teórica en torno a estas tensiones es precisamente el fin de este escrito. Interesa explorar estas cuestiones porque el archivo no es un artefacto inocente, sino que: "[...] se constituye como el espacio físico que resguarda los documentos, pasando por su institucionalidad arcóntica que ejerce su poder de custodia y autoridad hermenéutica legitimadora [...]" (Nava 96).

Hay que decir que en muchas ocasiones las opciones digitales son tomadas como las más amenas, precisamente por su disposición pública y su libertad, no obstante, a través de este escrito pienso que esto debe verse con sumo cuidado:

Con el advenimiento de las tecnologías vinculadas a la información y la comunicación, y la generación de espacios virtuales donde se pueden almacenar y consultar volúmenes considerables de documentos, [...]. Entra en escena el concepto de domiciliación, definido por Derrida (1997) como el lugar donde los documentos residen de modo permanente, transitando el camino institucional que va de lo privado a lo público. Esta domiciliación implica algo más que una simple noción espacial, es el reconocimiento de ese espacio dentro de una dimensión jurídica que le asigna determinadas características específicas" (Goldchluk y Pené 14).

Así que la domiciliación de los documentos se convertiría en un primer escollo de esta problemática. Esta se hace patente sobre todo cuando se decide aquello que se digitaliza y se pone en público y qué se deja en privado, resguardado a la parte física del archivo. El domicilio se apropia de la materia de los documentos, y ésta sería una primera tensión.

En un segundo momento la domiciliación que se aúna a la desterritolización, porque aquello que se posee está localizado y resguardado y solo se consulta con permiso institucional. La segunda de las tensiones que se presentan entre una y otra forma del archivo pienso que va por el lado de la desterritolización de las literaturas, precisamente porque opera un dispositivo, es decir, una conjunción entre el poder y la institucionalidad (Agamben). En últimas es un ejercicio de poder el que determina qué nación se apropia de un archivo. Dicha desterritolización no se presenta únicamente con nuestras literaturas, también pasa lo mismo con otras literaturas, por ejemplo, sobre los diversos ejemplares literarios del dadaísmo francés (Iowa University), de tal manera que la labor arcóntica de los archivos estadounidenses ha estado presente desde hace algún tiempo, y va en aumento constante.

Se conoce que un archivo de escritor está básicamente poblado de documentos y es obvio que el presente debate también va en la vía de las tensiones y la actualización obvia de dicho concepto. Se puede decir en cierta media que los verdaderos documentos se encuentran en la versión física y que muchas veces los archivos digitales se limitan a ser solo una muestra. Esto es notorio en la descripción que se puede leer en la página del Ransom Center sobre el archivo de García Márquez: de más de 1000 documentos guardados solo 33 están para la consulta pública en línea. De esta manera queda difícil emprender una labor como la que propone Foucault:

> ahora bien, por una mutación que no data ciertamente de hoy, pero que no está indudablemente terminada aún, la historia ha cambiado de posición respecto del documento: se atribuye como tarea primordial, no el interpretarlo, ni tampoco determinar si es veraz y cuál sea su valor expresivo, sino trabajarlo desde el interior y elaborarlo. […] (Foucault 9-10).

Así, ¿cómo es esto posible si los archivos no se poseen? Es la pregunta que queda en el aire para nuestra propia tradición crítica.

En tercer lugar, encuentro que los países que no tienen en su poder los archivos de sus escritores tienen menos opciones de poder proceder a ediciones críticas de sus literaturas que tengan en cuenta el modelo de la genética textual, puesto que dichos manuscritos y demás son tenidos en cuenta meramente como materia para especialistas que se puedan desplazar hasta estos lugares de consulta, la mayoría de las veces más accesible para aquellos que se encuentren dentro del ámbito lingüístico al que pertenece el archivo. Un ejemplo claro de esto es el documento de las galeradas corregidas de la versión de conmemoración que hizo la RAE en el año 2007 de *Cien años de soledad*, si no fuera por estas galeradas que reposan en Austin, entonces no sabríamos los cambios (casi imperceptibles) que tuvo la novela en su edición revisada, lo cual sería una tarea titánica de comparación de ediciones (Harry Ransom Center, 2017). Se piensa así que el acceso a los archivos físicos da mayor opción a cierta actualización editorial de la obra.

Para señalar la última de las tensiones, citamos de nuevo a Foucault: "reconstituir, a partir de lo que dicen esos documentos – y a veces a medias palabras- el pasado del que emanan y que ahora ha quedado desvanecido muy atrás de ellos; el documento seguía tratándose como el lenguaje de una voz reducida ahora al silencio: su frágil rastro, pero afortundamente descrifrable (Foucault, 9).

Dicha idea de Foucault está muy conectada precisamente con la idea de iterabilidad de Derrida, es decir, poder reconstruir el enunciado del emisor aunque no se cuente con su presencia:

La posibilidad de repetir, y en consecuencia, de identificar las marcas está implícita en todo código, hace de éste una clave comunicable, transmisible, descifrable, repetible por un tercero, por tanto por todo usuario posible en general. Toda escritura debe, pues, para ser lo que es, poder funcionar en la ausencia radical de todo destinatario empíricamente determinado en general (Derrida 1998 364).

Obviamente si no se tiene un acceso físico a los archivos, la capacidad de su iterabilidad se desvanece, sobre todo en lo que tiene que ver con la genética de los textos. Al ordenar y definir qué se da al público y qué se conserva privado se está resguardando de cierta forma la capacidad de iterabilidad que podría tener tal documento, en ese orden de ideas, su capacidad de iterable se disminuye. ¿Hasta qué punto son más iterables aquellas obras que se ponen en público y en digital y aquellas a las que se les guarda con más celo?

Finalmente, lo que se quiere lograr con esta aproximación es observar qué contrastes existen entre ambas formas de presentación del archivo de escritor y las diversas tensiones que se producen entre una forma y otra de presentación de los archivos, a pesar de su supuesto carácter de equiparabilidad.

References

Agamben, Giorgio. "¿Qué es un dispositivo?". *Arte y pensamiento*. Universidad Internacional de Andalucía. Web. 20 de noviembre de 2017. http://ayp.unia.es/r08/IMG/pdf/agamben-dispositivo.pdf

"Colombia: polémica por la venta del archivo personal de Gabriel García Márquez a la Universidad de Texas". La Nación. 24 de noviembre de 2014. Web. http://www.lanacion.com.ar/1746618-colombia-polemica-por-la-venta-del-archivo-personal-de-gabriel-garcia-marquez-a-la-universidad-de-texas

Derrida, Jacques. "Firma, acontecimiento, contexto". Márgenes de la filosofía, Madrid, Cátedra, 1998, pp. 347-372. Impreso.

Derrida, Jacques. *Mal de archivo: una impresión freudiana*. Madrid: Trotta, 1997. Impreso.

Foucault, Michel. *La arqueología del saber*. Buenos Aires: Siglo XXI Editores, 2002. Impreso.

Goldchluk, Graciela y Pené, Mónica Gabriela. "Archivos de escritura, génesis literaria y teoría del archivo". Repositorio institucional. Universidad Nacional de la Plata. Web. 20 de noviembre de 2017. http://www.memoria.fahce.unlp.edu.ar/trab_eventos/ev.772/ev.772.pdf

Harry Ramson Center. "2018–2019 Research Fellowships Application Instructions". Universidad de Texas en Austin. Web. 20 de noviembre de 2017. http://www.hrc.utexas.edu/research/fellowships/application/

Harry Ramson Center. "Gabriel García Márquez: Un Inventario de sus documentos en el Harry Ransom Center". Universidad de Texas en Austin. Web. 20 de noviembre de 2017. http://norman.hrc.utexas.edu/fasearch/findingAid.cfm?eadid=01084

Iowa University. "Digital library". Web. 20 de noviembre de 2017. http://digital.lib.uiowa.edu/

Nava Murcia, Ricardo. "El mal de archivo en la escritura de la historia". *Historia y grafía*, núm. 38, enero-junio 2012, pp. 95-126. Impreso.

Using Linked Open Data To Enrich Concept Searching In Large Text Corpora

Christine Fernsebner Eslao
eslao@fas.harvard.edu
Harvard Library, United States of America

Stephen Osadetz
osadetz@fas.harvard.edu
Harvard University, United States of America

This poster presents the library metadata aspects of a web-based text mining application for sifting corpora of unstructured text in order to find particular passages that deal with a concept of interest. In addition to overcoming the limitations of vendor-supplied search platforms, which tend to be based on simple keyword searches that place the burden of interpreting, refining, and iterating on search results on the laborious grunt work of scholarly users (De Bolla, 2013), this tool demonstrates the utility of reconciling named entities with external structured data to refine its results and to enrich its output for use in research, visualizations, and secondary analytic tools by leveraging demographic (Hwang, 2015), temporal, and geographic data from the linked open data cloud. This necessitates the creation of entity resolution workflows with both automated matching tools and practices for manual reconciliation and maintenance, exploring a variety of open-source tools including OpenRefine (Van Hooland, 2014; Hwang, 2017), Python, and Mix'n'Match (Knoblock, 2017) and contributing to the development of "functional requirements for how [library] systems use and maintain these identifiers and associated data" (Folsom, 2017) by metadata librarians and researchers and "the complexities inherent in managing both locally-created and externally-assigned identifiers" in the context of library infrastructure (Tarver, 2017). Our goal is to integrate a tool catering to advanced researchers into library discovery platforms by "[exploring] partnerships with external entities to create game changing discovery" (Wones, 2017) and leveraging those users' domain expertise to "interrogate corpora of resources directly ... to discover new patterns that exist across the literature, perform their own ranking of relevance against particular parameters, and find new pathways for discovery more efficiently than could be enabled through existing information portals" (MIT Libraries, 2016). The process is as follows:

1. Combine vendor metadata for large corpora with bibliographic metadata from Harvard Library collections
2. Reconcile authors, including persons and organizations, in those metadata resources, with external URIs, including those of ISNI (International Standard Name Identifier), Wikidata, and Geonames entities, generating batches of new entities in external resources at scale as needed (Mika, 2017)
3. Integrate data from external URIs into a text mining tool for sifting large corpora to drive filters and enrich data extracted from that tool
4. Work with library technology staff and metadata librarians to facilitate retrieval of rare materials in Harvard Library collections, as well as their electronic reproductions, based on results of text mining tool and integration of URIs in library metadata
5. Export resulting data to produce visualizations and secondary analytic tools

Through this process, we hope to enable the serendipitous discovery (Bourg, 2017) of relevant but unknown works in library collections: traditional reading of the "great unread" (Cohen, 1999) facilitated by distant reading (Moretti, 2013). Our poster includes: an explanation of the linked data principles underlying the metadata aspects of the text mining tool, our entity reconciliation workflow, implications for library metadata and name authority practices in support of digital research projects, and an example of combined and enriched metadata for a work of eighteenth century literature, and an example of an iterative concept search and its output presented both as a static flowchart on the poster as well as an interactive prototype on a laptop.

References

Bourg, Chris. (2017). *Serendipity as prick* https://chrisbourg.wordpress.com/2017/02/11/serendipity-as-prick/ (accessed 18 November 2017).

Cohen, Margaret. (1999). T*he Sentimental Education of the Novel*. Princeton, N.J.: Princeton University Press.

De Bolla, Peter. (2013). *The architecture of concepts : The historical formation of human rights* (First ed.). New York: Fordham University Press.

Folsom, Steven. (2017). New Models Require New Action Plans: Implementing Linked Data within the PCC. PCC (Project for Cooperative Cataloging) *Strategic Planning Meeting Keynote*, 1 November 2017. https://docs.google.com/presentation/d/11DHY-Ry24F4aQjYbPsVmInO2ovcb_pujBPIwJBQ0RrAQ/edit?usp=sharing (accessed 27 November 2017).

Hwang, Karen. (2015). *Enriching the Linked Jazz Name List with Gender Information* https://linkedjazz.org/enriching-the-linked-jazz-name-list-with-gender-information/ (accessed 1 November 2017).

Hwang, Karen. (2017). *Using OpenRefine to Reconcile Name Entities* http://mnylc.org/fellows/2017/03/17/using-openrefine-to-reconcile-name-entities/ (accessed 10 October 2017)

Knoblock, C.A., et al. (2017). Lessons Learned in Building Linked Data for the American Art Collaborative. In: d'Amato C., et al. (eds) *The Semantic Web – ISWC 2017 : 16th International Semantic Web Conference*, Vienna, Austria, October 21-25, 2017, Proceedings, Part II (Lecture Notes in Computer Science, 10588). Cham: Springer International Publishing : Imprint: Springer.

Mika, Katie. (2016). *The Role of Librarians in Wikidata and Wikicite*. https://library.mcz.harvard.edu/blog/role-librarians-wikidata-and%C2%A0wikicite (accessed 1 November 2017).

MIT Libraries, *Ad Hoc Task Force on the Future of Libraries*. (2016). Institute-Wide Task Force on the

Future of Libraries—Preliminary Report https://future-of-libraries.mit.edu/sites/default/files/Future-Libraries-PrelimReport-Final.pdf (accessed 25 November 2017).
Moretti, Franco. (2013). *Distant Reading*. London: Verso.
Tarver, Hannah, & Phillips, Mark. (2017). *Identifier Usage and Maintenance in the UNT Libraries' Digital Collections* http://dcevents.dublincore.org/IntConf/dc-2016/paper/download/458/546 (accessed 27 November 2017).
Van Hooland, Seth, & Verborgh, Ruben. (2014). *Linked data for libraries : How to clean, link and publish your metadata*. Chicago, IL: Neal-Schuman.
Wones, Suzanne. (2017). *Harvard Library Digital Strategy, Version 1.0*. http://projects.iq.harvard.edu/files/overseers/files/vc_3_hl_digital_strategy_v2.pdf (accessed 10 November 2017).

Pontes into the Curriculum: Introducing DH pedagogy through global partnerships

Pamela Espinosa de los Monteros
espinosadelosmonteros.1@osu.edu
Ohio State University Libraries, United States of America

Joshua Sadvari
sadvari.1@osu.edu
Ohio State University Libraries, United States of America

Maria Scheid
scheid.31@osu.edu
Ohio State University Libraries, United States of America

This poster proposes a discussion on the challenges and lessons learned in the integration of digital humanities pedagogy into a traditional graduate foreign language course through a heterogeneous collaboration among global DH scholars and North American experts in geographic information systems (GIS), copyright, digital humanities, and area studies. Significant barriers of entry exist for humanities students and faculty attempting to introduce DH into their departments and classrooms. Uneven institutional infrastructure and programmatic presence of DH at universities leave faculty with the dilemma of simultaneously learning DH methods themselves and integrating DH pedagogy into the curriculum for their students. As such, DH methods can present real and perceived psychological and cultural barriers (Battershill and Shawna, 2017) that surpass the digital and technology competencies of students or faculty. Humanities departments recognize the value of DH methods, research, and the need to develop the next generation of DH scholars, but may lack in-house expertise to design the initial curriculum.

Partnership with the community of DH scholars, and the research library, may assist faculty member to overcome technological infrastructure and subject expertise lacking in their own departments. The proposed poster and case study will highlight a team based approach to introduce DH research and DH curriculum from the Lusophone world. Three DH methods were introduced including text-analysis, GIS, and text-encoding and transcription. Each method was paired with course content, DH literature, mediated exploratory assignments, and current DH research by scholars in the field. Sessions were team taught in workshop and lecture settings to provide students with both experimental learning models and theoretical background. DH curriculum was customized to meet the subject content of the Portuguese literature course and taught in both English and Portuguese.

The most significant and time-intensive DH assignment students completed during the course was the collaborative creation of an ArcGIS Story Map on the African diaspora of Lisbon. With the advent of web-based mapping platforms, user-friendly on-ramps exist for humanities scholars to integrate geovisualization and location-based storytelling into their research (Presner & Shepard, 2016), and this assignment was designed for students to recognize the utility of such a platform for their own work. After a brief introduction to some key GIS concepts and a hands-on tutorial, students collaboratively identified images and text associated with course topics to overlay points on a georeferenced historical map of Lisbon. In this way, students combined a growing knowledge of course subject matter, copyright considerations when identifying and incorporating suitable content, and newly-developed digital mapping skills to create an end product that differed from the more traditional written paper to which they might be accustomed. In collaboration with the faculty instructor, adjustments were made throughout the project to accommodate humanities students' varying levels of technical and information literacy proficiencies in the classroom.

In this poster, we will address challenges faced by the team to blend and balance traditional and DH pedagogy, multilingual limitations of existing DH tools, and design of an exploratory assignment with specific disciplinary content. A focal point of the poster will be the role of each participant and timeline for the project's implementation. By sharing our experiences in developing this introductory intervention, we hope to explore with attendees the ways in which DH methods, tools, and dispositions can be introduced into traditional foreign language humanities courses. This poster will outline lessons learned and promote discussion on unique challenges in curriculum design, collaborative instruction, and delivery of GIS DH instruction for a foreign language course.

References

Presner, T., & Shepard, D. (2016). Mapping the geospatial turn. In S. Schreibman, R. Siemens, & J. Unsworth

(Eds.), *A new companion to digital humanities*. Chichester, UK: John Wiley & Sons, Ltd., pp 201-212.
Battershill, C., & Ross, S. (2017). *Using digital humanities in the classroom: a practical introduction for teachers, lecturers, and students*. London: Blomsbury Academic.

Milpaís: una wiki semántica para recuperar, compartir y construir colaborativamente las relaciones entre plantas, seres humanos, comunidades y entornos

María Juana Espinosa Menéndez
mj.espinosam@uniandes.edu.co
Universidad de los Andes, Colombia

Camilo Martinez
gemartin@uniandes.edu.co
Universidad de los Andes, Colombia

Milpaís, proyecto de tesis para la maestría en Humanidades Digitales de la Universidad de los Andes de Colombia, nace como iniciativa del colectivo Savias y Sabias quienes en sus trabajos con comunidades expertas y no, han encontrado la necesidad de apropiarse de herramientas digitales que permitan democratizar el acceso al conocimiento experto sobre plantas, visibilizar el conocimiento tradicional y local y sobre todo defender este saber en tanto bien común (Bollier, 2016; Zuluaga Ramírez, 1994). Con especial énfasis discutimos los aspectos éticos y legales que tuvimos que sopesar al formular este trabajo en Humanidades Digitales sobre conocimientos tradicionales en el contexto global y en particular en el caso colombiano (Organización Mundial de la Propiedad Intelectual (OMPI), 2017; Gómez Madrigal, 2013). Al respecto, debimos considerar estrategias para prevenir la expropiación indebida de conocimientos mediante la definición de la catalogación y la visibilización en la wiki de los territorios, las personas y comunidades que cuidan, siembran y trabajan con las plantas. Este mapeo permite construir elementos probatorios de la pertenencia cultural de conocimientos colectivos circunscritos a territorios.

El *prototipo se desarrolla a partir de una wiki semántica del software libre Media Wiki* (*semantic-mediawiki.org*, 2018) para la gestión del conocimiento etnobotánico de comunidades que usan, defienden y comparten saberes sobre las plantas. Diseñada a partir de una reflexión ética y legal de lo que implica documentar, catalogar y difundir conocimientos tradicionales y locales, el prototipo cuida de no exponer contenidos susceptibles de expropiación indebida tales como componentes, fórmulas, técnicas y rituales. Es así que en este prototipo nos interesa conectar qué lugares, qué personas (comunidades) y de qué maneras se construyen las relaciones con las plantas, entendidas estas como uno de los bienes comunes que sostienen y equilibran entornos como el cuerpo y el medio ambiente (Lafuente, 2007).

En términos técnicos, la SMW permite estructurar una Base de Datos Relacional (BDR) mediante el uso de plantillas que integran notación semántica y vocabularios controlados. Para esta wiki utilizamos el estándar de metadatos FOAF(*The FOAF Project*, 2018) y un conjunto de metadatos propios y vocabularios controlados alimentados de diversas fuentes de catalogación etnobotánica (Royal Museum From Central Africa, 2017; *BRIT - Native American Ethnobotany Database*, 2003). Igualmente, se ha tenido y se tendrá en cuenta la información que colaboradores y posibles usuarios han reportado necesaria. Para recuperar la información y que se integre la notación semántica, el prototipo implementa los formularios de Semantic Media Wiki (*Page Forms - MediaWiki*, 2018). Estos formularios permiten a los colaboradores/creadores de la wiki ingresar la información mediante una interfaz amable sin necesidad de hacer notación semántica manual. Una vez se ingresa la información la SMW permite recuperar información relacional (qué personas son amigos de una planta, qué plantas sirven a las personas para hacer artesanía, qué comunidades resguardan una semilla en particular, quiénes y dónde hay médicos tradicionales, yerbateras, investigadores, médicos alópatas que trabajan con plantas, etc.) así como visualizar datos tales como los geográficos.

Finalmente, el prototipo tiene una fase piloto anterior a la implementación (2018-II) en la cual empezamos a trabajar la campaña de difusión "Adopta una planta y cultiva su conocimiento en la web". Dicha estrategia se enmarca en el trabajo que se adelanta con comunidades potencialmente usuarias en zonas aledañas y urbanas de Bogotá y en la cual se llevó a cabo un primer rastreo sobre la información que consideran importante documentar, compartir y defender. Así, la herramienta digital se dispondrá al servicio de procesos educativos con comunidades que quieran intercambiar saberes, investigar y defender el conocimiento tradicional y local sobre las plantas.

References

Biblioteca digital de la medicina tradicional mexicana (2009). Available at: http://medicinatradicionalmexicana.unam.mx/index.php (Accessed: 27 October, 2017).
Bollier, D (2016). *Pensar desde los comunes*. Madrid: Traficantes de Sueños.
BRIT - Native American Ethnobotany Database (2003). Available at: http://naeb.brit.org/ (Accessed: 27 March 2017).
Gómez Madrigal, L. S. (2013). *Protección de la tradición. Los derechos no tradicionales de la propiedad intelectual. Comité intergubernamental de recursos genéti-*

cos, conocimientos tradicionales y folclore de la OMPI. *Revista La Propiedad Inmaterial*. Available at: http://revistas.uexternado.edu.co/index.php/propin/article/view/3581/3798 (Accessed: 2 February 2017).
Lafuente, A. (2007). Los cuatro entornos del procomún. *Archipiélago*: *Cuadernos de crítica de la cultura*, (77), pp. 15–22. Available at: https://dialnet.unirioja.es/servlet/articulo?codigo=2500491 (Accessed: 11 September 2017).
Organización Mundial de la Propiedad Intelectual (OMPI) (2017). *Guía para la catalogación de conocimientos tradicionales*, *OMPI*. Available at: http://www.wipo.int/edocs/pubdocs/es/wipo_pub_1049.pdf (Accessed: 8 May 2017).
Page Forms - MediaWiki (2018). Available at: https://www.mediawiki.org/wiki/Extension:Page_Forms (Accessed: 27 April 2018).
Royal Museum From Central Africa (2017). *Prelude*, *Medicinal Plants Data Base*. Available at: http://www.africamuseum.be/collections/external/prelude/plant_collection (Accessed: 27 April 2018).
The FOAF Project (2018). Available at: http://www.foaf-project.org/ (Accessed: 5 February 2018).
Semantic-mediawiki.org (2018). *Semantic MediaWiki*. Available at: https://www.semantic-mediawiki.org/wiki/Semantic_MediaWiki (Accessed: 5 March 2018).
Zuluaga Ramírez, G. (1994). *El aprendizaje de las plantas*: *en la senda de un conocimiento olvidado*. Bogotá: Seguros Bolívar.

Cataloging History: Revisualizing the 1853 New York Crystal Palace

Steven Lubar
lubar@brown.edu
Brown University, United States of America

Emily Esten
emily_esten@brown.edu
Brown University, United States of America

Steffani Gomez
steffani_gomez@alumni.brown.edu
Brown University, United States of America

Brian Croxall
brian.croxall@brown.edu
Brown University, United States of America

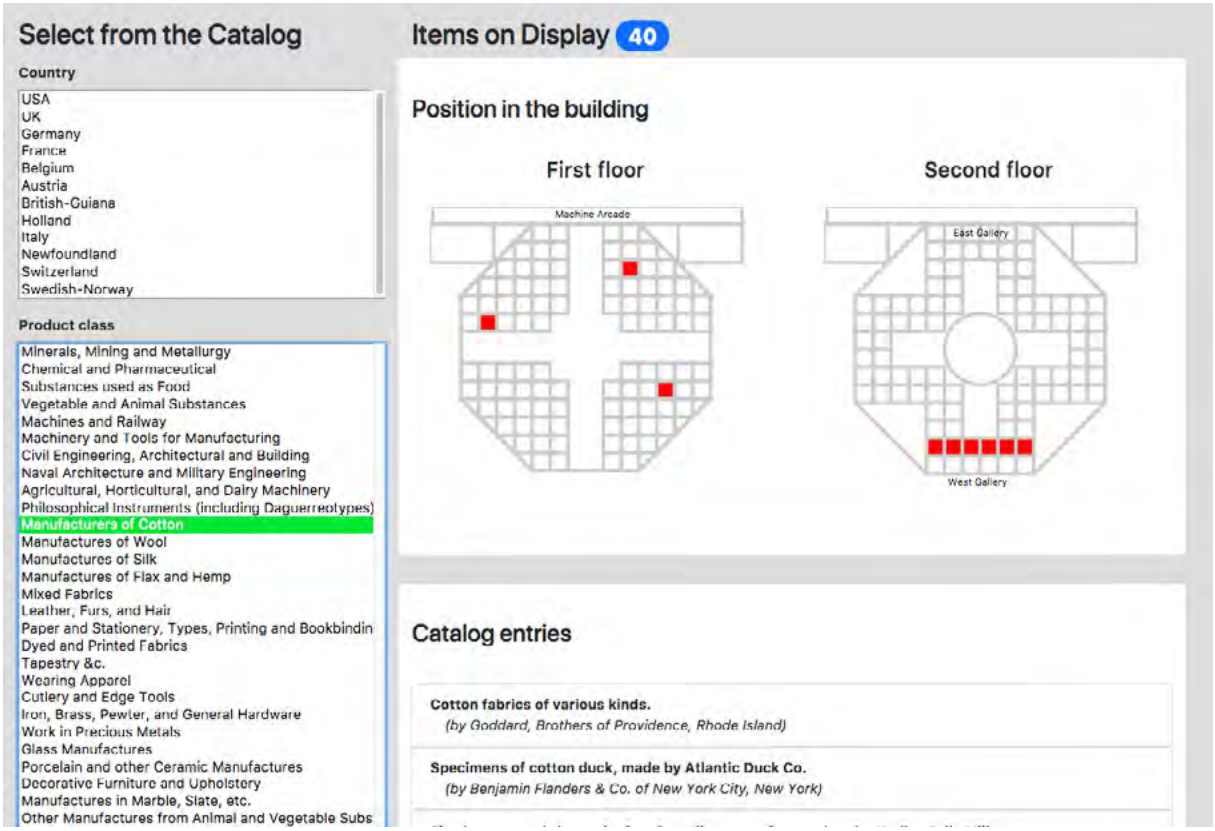
Patrick Rashleigh
patrick_rashleigh@brown.edu
Brown University, United States of America

The 1853 New York Crystal Palace, also known as the Exhibition of the Industry of All Nations, was the center of America's first World's Fair. Modeled on the Great Exhibition held at the Crystal Palace in London, the exhibition sought to "draw forth such a representation of the world's industry and resources as would enable us to measure the strength and value of our own, while it indicated new aims for our enterprise and skill." While the exhibition burnt down by the end of the decade, it survives through catalogues documenting its success and breadth.

This poster considers the catalogues from the 1853 New York Crystal Palace exhibition in physical and digital forms: as book, file, and database. Originating with the Museum Wormianum and the Louvre, catalogs were published by early museums to visualize and document their work for the world. The New York Crystal Palace exhibition generated multiple print publications to record the museum-like experience through databases and narratives. In the digital era, however, these traditional exhibition catalogs can be used in new ways. As both a physical and digital object, the cataloging forms encourage and allow relationships among user, data, and experience to come to fruition. This project created a dataset and subsequent database from a digitized copy of the New York Crystal Palace catalog to explore the artifacts documented inside. Explored through digital tools such as OpenRefine, Tableau, Palladio, D3.js, and Google Fusion, the Crystal Palace catalogs aid us in viewing catalogs, and their modern database descendants, more generally. How can looking at the Crystal Palace through digital tools let us see not only what others have seen, but also to see things better, see things differently?

Databases, by their nature, lend themselves to exploration – the 1853 Crystal Palace exhibition catalog is no exception. Creating a database from the catalog frees the information inside. The curators of the exhibition thought hard about the best way to arrange the *Official Catalog*, but once they decided, it couldn't be changed. Now, the data is alive and fluid. It can be analyzed, represented in new ways. It can be searched, sorted, faceted, mapped, and turned into networked and nodes. Revisualized and revitalized in digital scholarship, this database and subsequent visualizations offer non-narrative perspectives to the exhibition's construction and imagined possibilities. Contributing both to historical and museological scholarship, *Cataloging History* considers histories of technology through the catalog to understand exhibition construction and the ways in which we can reconstruct it in the digital age.

Ultimately, visualizations of any kind open up the catalog to a new form of interrogation. But they also ask us to reimagine the relationships connections embedded inside. What if we wanted to re-curate the Crystal Exhibition by role, showing off inventors and agents in the major divisions? What if we wanted to use this as an opportunity to tease out specific class categories? What if we wanted to reorganize the catalog by class first, instead of country? Visualizations offer us the ability to think through both



the construction of the exhibition and the catalog, while also allowing us the chance to reconstruct its data to new ends. We can use the database and visualizations as way to look into the past, evolving this nineteenth-century exhibition along with new forms oft he catalogue.

Developed in cooperative collaboration with representatives of Brown's Center for Digital Scholarship, *Cataloging History* examines the ways in which traditional museum data can be mobilized to reimagine historical spaces. Using the catalogue as a piece of technology for understanding the past, it also opened a new dialogue for thinking about catalogues of the future. Building on conversations around "collections as data," this project uses a historical example to pose to both scholars and museums about how cultural heritage may work to be more readily open to computation. How does digital humanties help us unpack historical collections? These visualizations highlight how digital tools can unearth relationships among data to better understand what was there. *Cataloging History* challenges us to think more deeply about what information is contained in a catalog, about what remains when an exhibition is gone, and about how datasets and tools like these promote the evolution of exhibitions.

References

Croxall, B., Esten, E., Gomez, S., Lubar, S., and Rashleigh, P. (2017). 1853 New York Crystal Palace accessed April 25, 2018. http://cds.library.brown.edu/projects/crystalpalace/.

Esten, E. (2017). "Visualizing the Crystal Palace." https://github.com/sheishistoric/Visualizing-the-Crystal-Palace.

Lubar, S. (2017). "A brief history of American museum catalogs to 1860." https://medium.com/@lubar/cataloging-history-eac876941db6.

Lubar, S. and Esten, E. (2017). "Catalog as Book, File, and Database." https://medium.com/@lubar/catalog-as-book-file-and-database-ac954096152e.

Lubar, S. (2017). "The New York Crystal Palace Catalogs." https://medium.com/@lubar/the-new-york-crystal-palace-catalogs-b09d1f2bd20e.

Lubar, S. and Esten, E. (2017). "Revisualizing the Crystal Palace." https://medium.com/@lubar/revisualizing-the-crystal-palace-d239e50d9e12.

New York Exhibition of the Industry of All Nations, New York, N.Y. (1853). *Official Catalogue of the New-York Exhibition of the Industry of All Nations. 1853*. New York: G.P. Putnam & Co.

Crowdsourcing Community Wellness: Coding a Mobile App For Health and Education

Katherine Mary Faull
faull@bucknell.edu
Bucknell University, United States of America

Michael Thompson
michael.thompson@bucknell.edu
Bucknell University, United States of America

Jacob Mendelowitz
jpm061@bucknell.edu
Bucknell University, United States of America

Caroline Whitman
alw001@bucknell.edu
Bucknell University, United States of America

Shaunna Barnhart
sb060@bucknell.edu
Bucknell University, United States of America

In response to the widely reported increase in obesity and related health problems in the US, a team of faculty, staff, and students at Bucknell University have authored a mobile app that incentivizes exercise through the use of crowdsourced public-facing humanities content of local interest. ReadySetFit, available on both Apple and Android phones, is a completely student-coded app that leverages a Google Maps platform and the Google My Maps application. (http://www.readysetfitapp.org) The user can select from a set of walking paths that have been created using the Google My Maps app, which contain points of interest that present cultural/historical information to the user as he or she approaches the physical location of each point. Once a user has reached all points of interest, or manually clicked a button to finish a workout, the distance covered is saved to the handheld device and can be reviewed at a later date.

Key components of the success of ReadySetFit have been the ease of use and the localized and crowdsourced nature of the information provided. Griffiths and Barbour (2016) argue that the creation of "smart cities" greatly enhances the sense of place among local citizens. Our university collaboration with a local civic group (The Improved Milton Experience) in the post-industrial central Pennsylvania town of Milton has engaged in local history through crowd-sourcing content for specific points of interest while incentivizing citizens to walk around the town. Users receive rewards and discounts at local shops when they earn "Milton Bucks" by walking on set paths in the borough.

Furthermore, partnering with the statewide system of parks (DCNR) and its "Think Outside" higher-education partnership program has promoted the app to a wide user-base who are already visiting the parks but who want to know about the history and environment through which they are walking. (http://www.dcnr.pa.gov/Education/ThinkOutside/Pages/default.aspx) Newly launched to the public, ReadySetFit has shown potential to overcome the major obstacle to maintaining an exercise routine--incentive (Harris and Roushanzamir 2014: Conroy et al) 2014. The app's incentive is multi-dimensional: engaging with new and interesting place-based content in realtime, collecting completed pathways, obtaining fitness levels for financial rewards through local business partnerships, and contributing to the creation of new pathways. Through crowdsourcing content, user participation promotes both individual wellness and community buy-in. The place-based content that is provided to the user is created by members of the community and fosters active engagement in creating a sense of place (Lepofsky and Fraser, 2003). The poster presentation will demonstrate the app itself and also show the process undergone by the students in terms of technology and content development. We will also demonstrate the path creation-guidelines that have been shared with local organizations and can be adopted for creating pathways anywhere in the world with cellular data connectivity.

References

Conroy, David E., Chih-Hsiang Yang, Jaclyn P. Maher, "Behavior Change Techniques in Top-Ranked Mobile Apps for Physical Activity", In American Journal of Preventive Medicine, Volume 46, Issue 6, 2014, Pages 649-652, ISSN 0749-3797, https://doi.org/10.1016/j.amepre.2014.01.010.

Griffiths Mary and Kim Barbour. "'Imagine If Our Cities Talked to Us': Questions about the Making of 'responsive' Places and Urban Publics." In Making Publics, Making Places, 27-48. South Australia: University of Adelaide Press, 2016, http://www.jstor.org/stable/10.20851/j.ctt1t304qd.8

Harris,Felicia and Elli Lester Roushanzamir. "#Blackgirlsrun: Promoting Health and Wellness Outcomes Using Social Media." Fire!!! 3, no. 1 (2014): 160-89. doi:10.5323/fire.3.1.0160.

Leipert, Beverly D., Belinda Leach, and Wilfreda E. Thurston, eds. Rural Women's Health. Toronto; Buffalo; London: University of Toronto Press, 2012. http://www.jstor.org/stable/10.3138/j.ctt2tv021.

Lepofsky, Jonathan, and James C. Fraser. "Building Community Citizens: Claiming the Right to Place-making in the City." Urban Studies 40, no. 1 (2003): 127-42. http://www.jstor.org/stable/43084177

Bad Brujas Only: Digital Presence, Embodied Protest, and Online Witchcraft

Amanda Kelan Figueroa
browngirlsmuseumblog@gmail.com
Harvard University, United States of America

Ravon Ruffin
ravonruffin@gmail.com
National Museum of African American History and Culture, United States of America

Over a year before Donald Trump took the presidency in 2016, a group of self-identified Cuban brujas, latina practitioners of witchcraft and/or indigenous rituals, led by Yeni Sleidi released an online video titled "Brujas Hex Trump." From this platform on YouTube, the video called for fellow witches in both the digital and physical realms to intervene in the presidential candidate's campaign through a type of activism not previously considered political — ritual, witchcraft, hexing. Since this initial call via YouTube, monthly hexes have continued among Brooklyn-based latinas, organized via social media.

The organization of social justice activism through interpersonal networks on Facebook, Twitter, Instagram, and other platforms is not an unusual phenomenon for marginalized communities, evidenced by such movements as #BlackLivesMatter and #MeToo. However, the model of integration between online and offline practices demonstrated in social media witchcraft or brujería communities is worthy of note, as a reclamation of the female-identified body and indigeneity in this current political climate. Witchcraft in its traditional forms would seem to be the antithesis of digital media due to its emphasis on materiality, embodied presence, and physically-enacted rituals. However, these networked communities of digital brujas transcend this divide, as a politicized tool for empowerment, and decolonization of history and the female body.

Groups like Brujas Hex Trump capitalize on the ability of personal practices to have overarching political impacts, while an organization called Witch Cabinet creates workshops and digital courses for femme- and queer-identified people to learn self-care through magic ritual, and social media astrologer Danielle Ayoka gives horoscopes and personalized tarot readings via Twitter and Instagram. The intersection of the embodied rituals of witchcraft and the digital space of social media appear to be irreconcilable, and for this reason, digital expressions of witchcraft and magic are widely considered to be cheap, commercialized, or inconsequential. However, we will examine these points of apparent conflict, between medium and message that occur in these examples of witchcraft, in order to demonstrate a method for seeing the social media space as a mediator, and not an obstacle, for these practices.

Using a theoretical frame based on Chela Sandoval's work on dissident coalition building, Chon Noriega's understanding of museological power structures, and the investments of black digital studies in a radical black archival practice, we build from embodied theories of ethnic studies and art ecosystems to find a method for considering race and ritual in the digital sphere. Undertaken primarily by women of color, digital witchcraft is successful at translating online presence into embodied action in ways that perhaps offer strategies for other social, institutional, and cultural communities and their activism. By studying the ways in which these digital witchcraft communities make use of social media platforms in order to bridge these divides, sometimes by using them against their designed purposes, the potential of digital activism, and its implications for studies of chicana and black feminism, indigenous studies, and other branches of ethnic studies, digital or otherwise, can be considered.

La geopólitica de las humanidades digitales: un caso de estudio de DH2017 Montreal

José Pino-Díaz
jpinod@uma.es
Universidad de Málaga, Spain

Domenico Fiormonte
domenico.fiormonte@uniroma3.it
Università Roma Tre, Italy

Resumen: Las discusiones y reflexiones sobre los desequilibrios culturales, políticos, lingüísticos y de género en las Humanidades Digitales se han concentrado sobre aspectos generales y, con pocas excepciones (Dacos, 2016; Fiormonte 2017b; Grandjean, 2014; Weingart 2014; Weingart and Eichmann-Kalwara 2017), no han ofrecido un análisis de datos y casos concretos. Nuestra propuesta intenta aportar una contribución al debate a través del análisis de las 420 colaboraciones del congreso DH2017 de Montreal. Los datos archivados nos permitieron realizar mapas de colaboración entre países y entre países y centros académicos o de investigación; así como mapas de temas de investigación (palabras clave) y de redes de autores. El resultado es una imagen real de lo que son hoy en día las Humanidades Digitales a nivel global, y donde parece confirmado el papel hegemónico del Norte global, y sobre todo de los países anglosajones, en la comunidad internacional.

El conjunto de comunicaciones presentadas en el último congreso global Digital Humanities 2017 (DH2017), celebrado, del 8 al 9 de agosto de 2017 en Montreal (Canadá), constituye, hasta la fecha, el corpus de conocimiento cooperativo más actual sobre Humanidades Digitales. Este corpus documental constituye el elemento más actual y necesario para indagar en las relaciones de asociación que se establecen entre países y centros (centros de investigación, universidades, etc.). A partir de la afiliación de los autores de los documentos, el estudio multidisciplinar de la colaboración en la investigación se ha abordado comúnmente desde diversos campos científicos: Historia de la Ciencia, Filosofía, Documentación, Bibliometría o Sociología (González Alcaide y Gómez Ferri 2014).

Este trabajo, planteado desde la óptica de la geopolítica del conocimiento (Adriansen 2016 y 2017; Canaga-

rajah 2002; de Sousa Santos 2010; Graham et al. 2011; Fiormonte 2017a; Mignolo 2011)[1], evidencia los desequilibrios culturales, institucionales o políticos existentes en el ecosistema de las Humanidades Digitales. El estudio de las relaciones de coautoría establecidas en las comunicaciones presentadas al congreso DH2017, pone de manifiesto las desigualdades entre países y centros.

La información disponible de cada registro, accesible on-line en el sitio web del congreso[2], consta de: nombre, apellidos y correo electrónico de los autores; título, resumen y palabras clave de la comunicación; y, centro y país de procedencia. A partir de la información proporcionada por la web del congreso se ha elaborado un archivo de texto en formato *Research Information System* (RIS), base de conocimiento y partida para realizar los análisis de asociación.

VOSviewer[3] (Van Eck y Waltman 2010, 2011 y 2014), herramienta desarrollada en la Universidad de Leiden, facilita el análisis de la colaboración científica mediante la visualización y la clasificación de las redes bibliométricas (Waltman, Van Eck y Noyons 2010) de autores, palabras clave, países o centros, existentes, pero no explícitas, en archivos RIS. El método de normalización de los enlaces elegido ha sido el del "valor de asociación".

Se han analizado las relaciones de asociación entre países, entre países y centros, entre palabras clave y entre autores, y se han obtenido tres tipos de redes y mapas de calor, según se haga proporcional el tamaño de los nodos a los valores de las ocurrencias, al número total de enlaces o al peso total de los enlaces. Veánse los siguientes ejemplos:

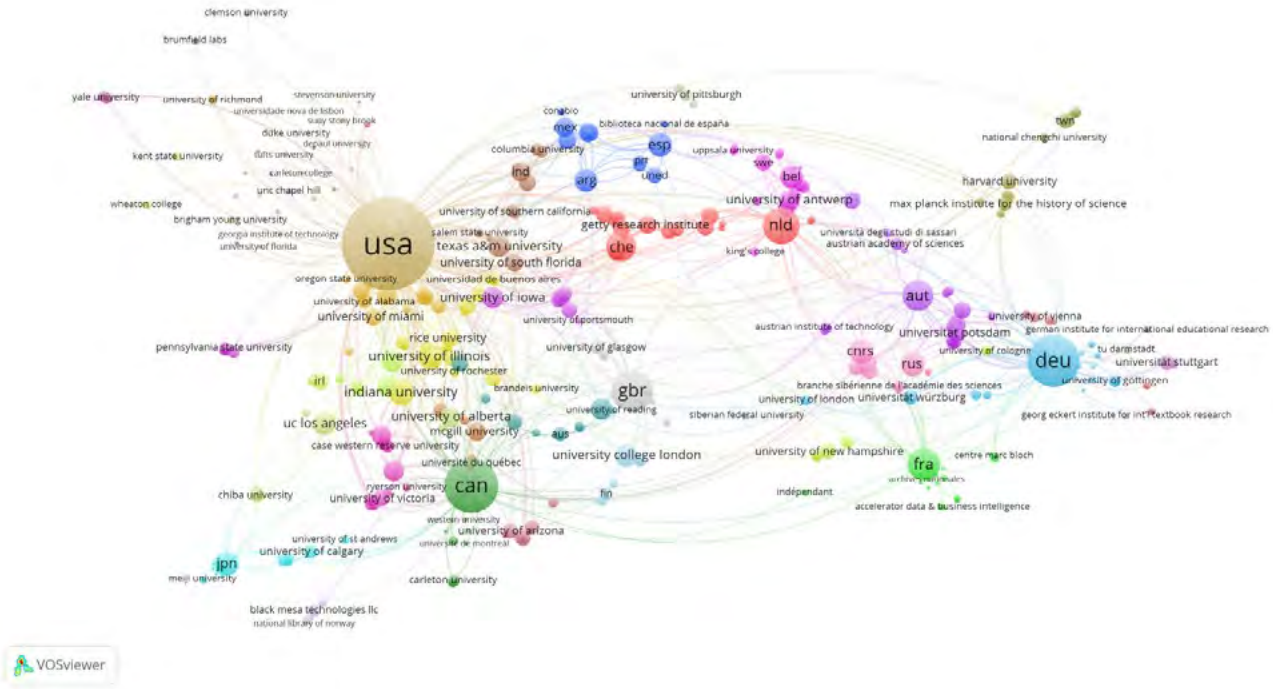


Figura 1.- Mapa de relaciones de asociación entre países y centros. Sólo aparecen los nodos conectados (modo de visualización "total link strength"; factor de variación de tamaño x0,5).

1 En el ámbito académico ver el reciente proyecto http://knowledgegap.org/

2 https://dh2017.adho.org/program/abstracts/

3 http://www.vosviewer.com/

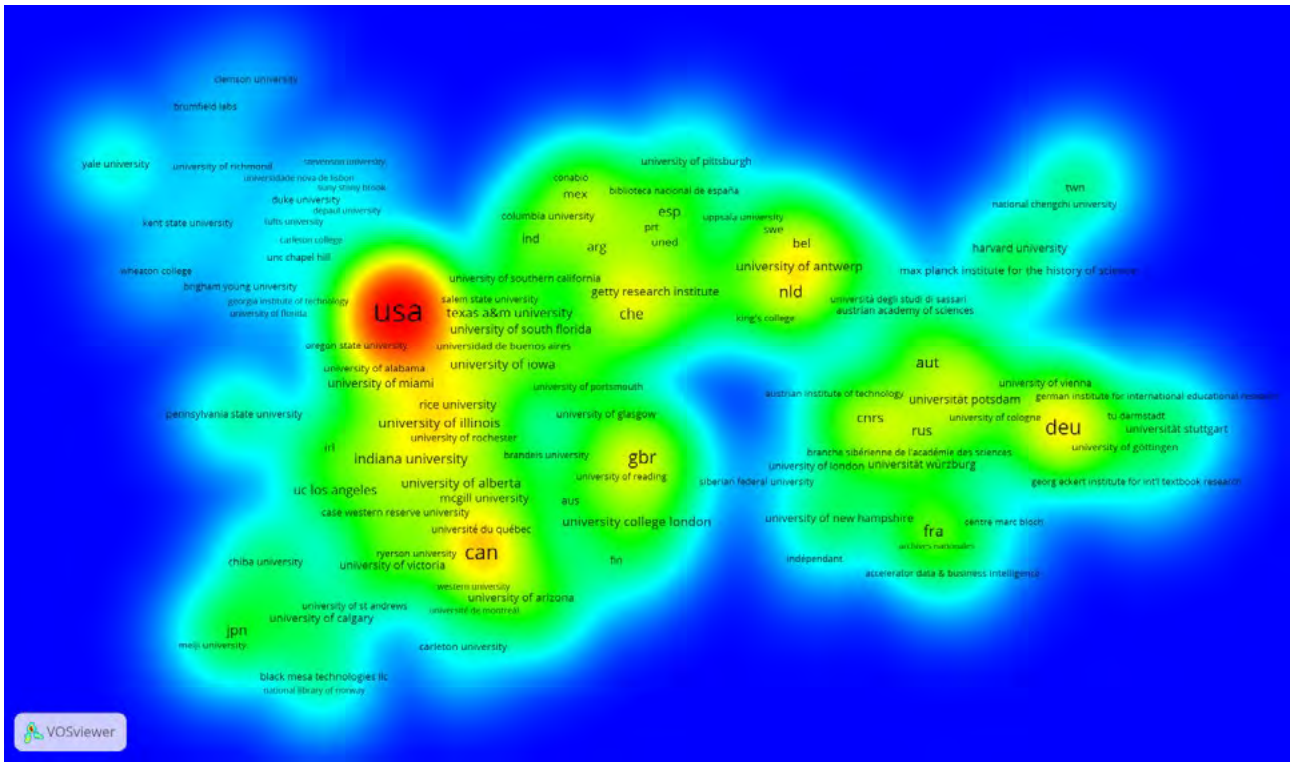


Figura 2.- Mapa de calor de las relaciones de asociación entre países y centros. Sólo aparecen los nodos conectados "total link strength"; factor de variación de tamaño x0,5).

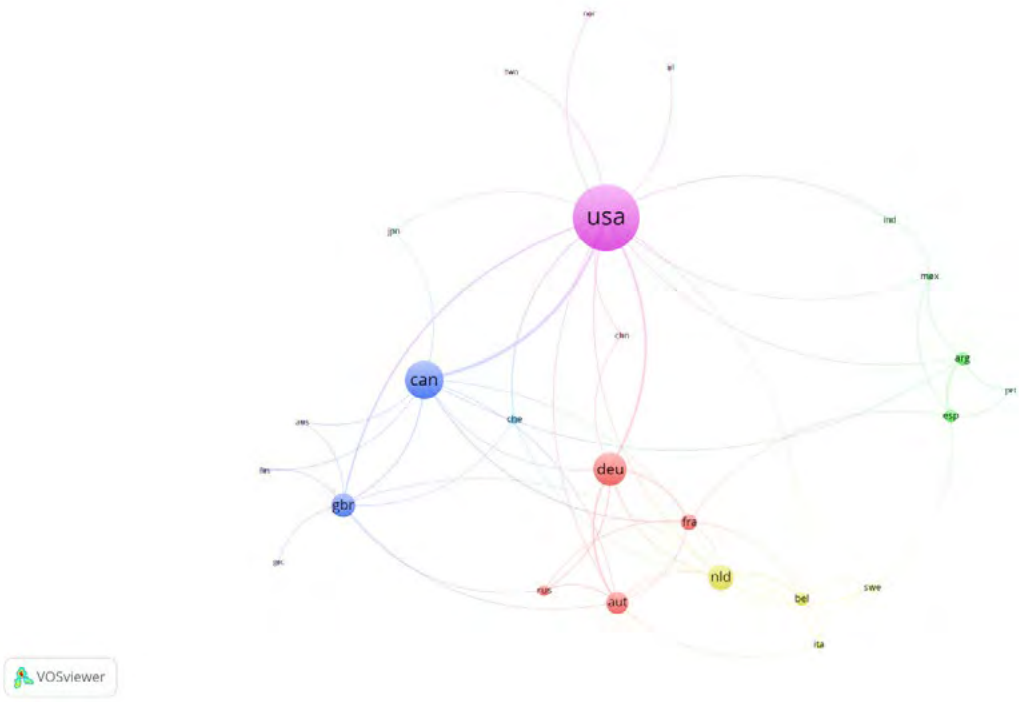


Figura 3.- Mapa de relaciones de asociación entre países. Sólo aparecen los nodos conectados (modo de visualización "total link strength"; factor de variación de tamaño x1).

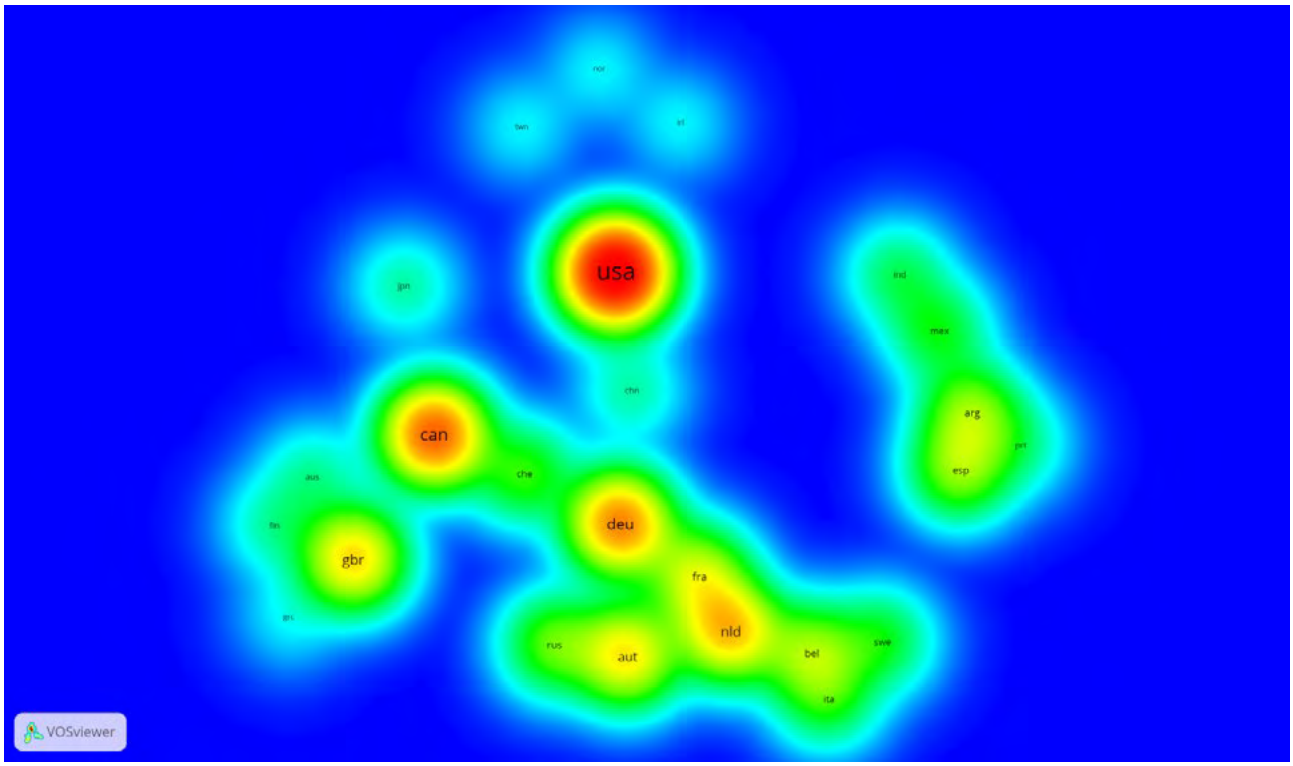


Figura 4.- Mapa de calor de las relaciones de asociación entre países. Sólo aparecen los nodos conectados (modo de visualización "total link strength"; factor de variación de tamaño x1).

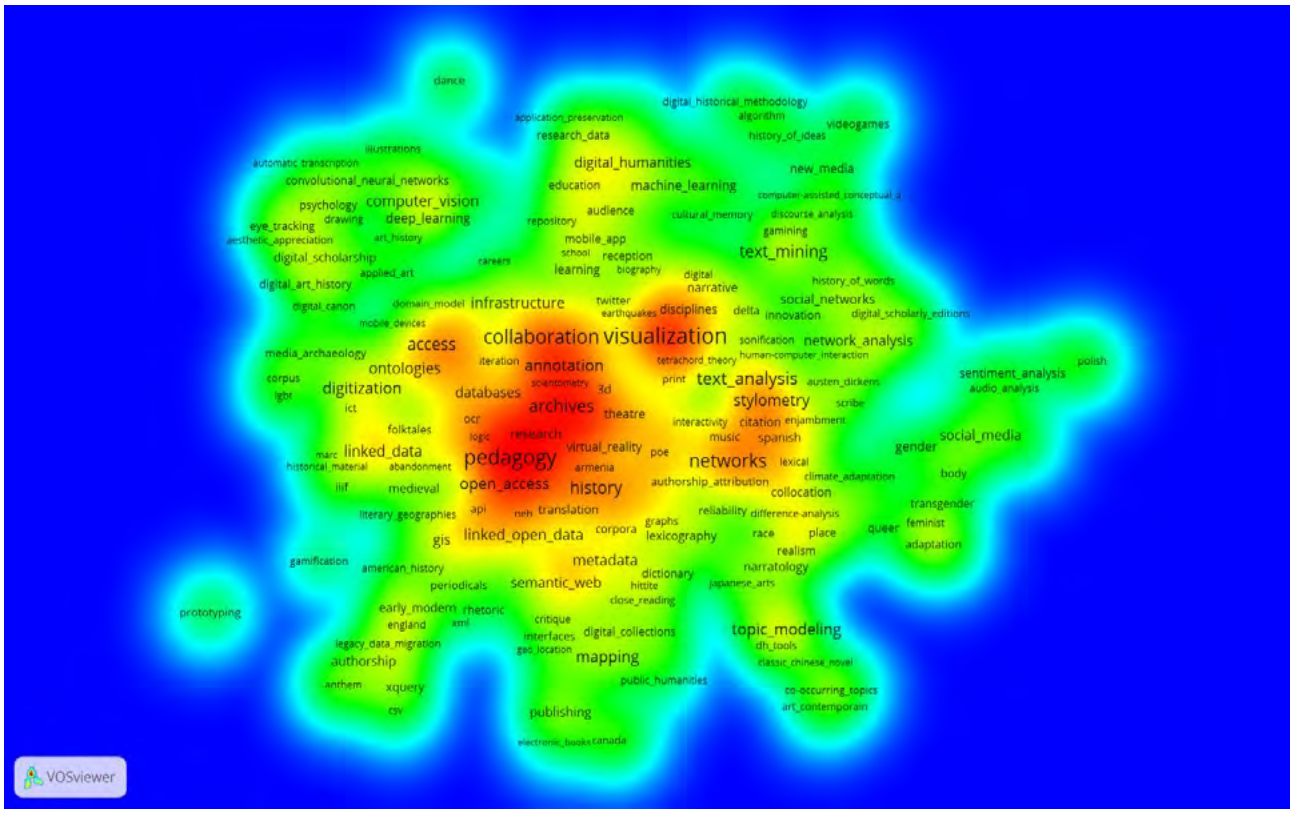


Figura 5.- Mapa de relaciones de co-ocurrencia entre palabras clave. Sólo aparecen los nodos conectados (modo de visualización "total link strength"; factor de variación de tamaño x0,5).

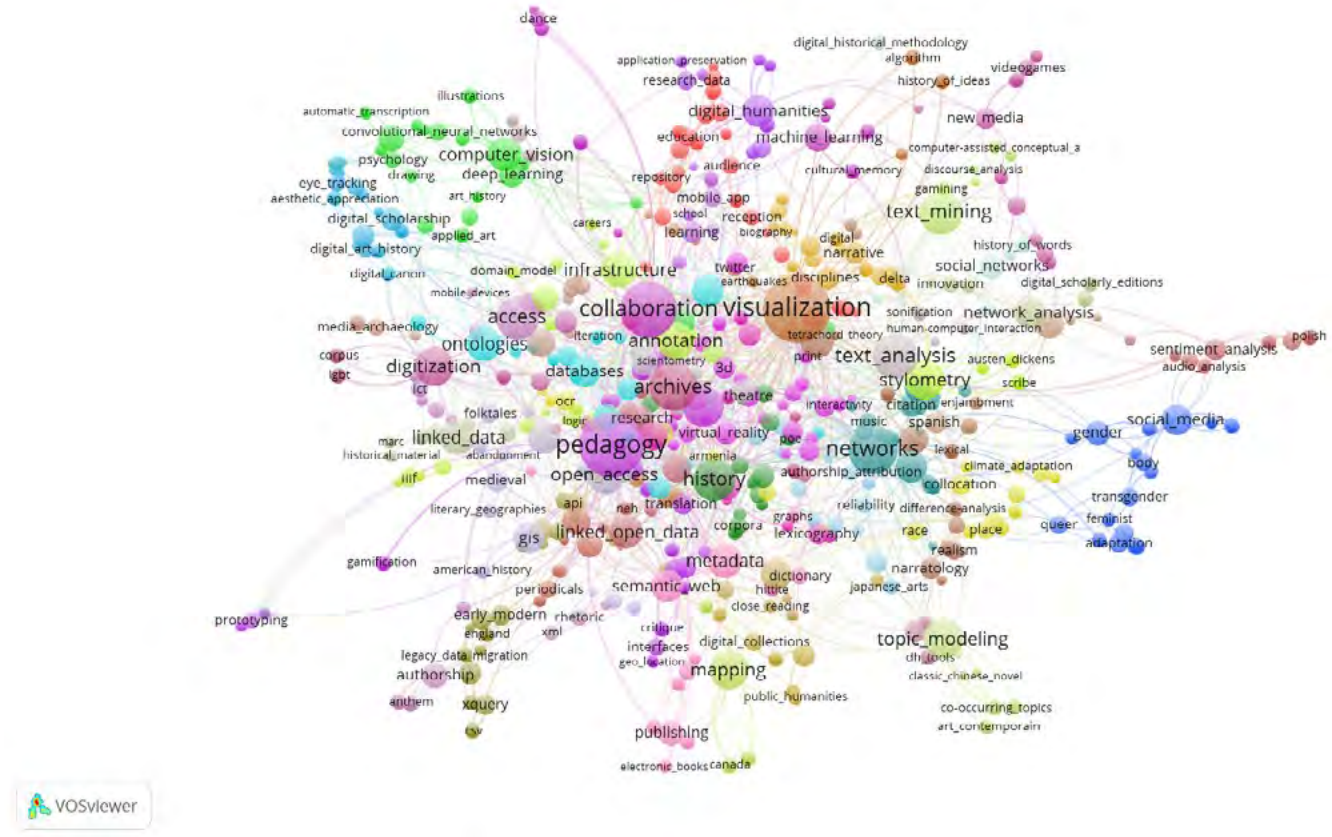


Figura 6.- Mapa de calor de las relaciones de co-ocurrencia entre palabras clave. Sólo aparecen los nodos conectados (modo de visualización "total link strength"; factor de variación de tamaño x0,5).

References

Adriansen, H. K. (2017). The power and politics of knowledge: what African universities need to do. *The Conversation*, September 28 2017. https://theconversation.com/the-power-and-politics-of-knowledge-what-african-universities-need-to-do-84233

Adriansen, H. K. (2016). Global Academic Collaboration: A New Form of Colonisation? *The Conversation*, July 8 2016. https://theconversation.com/global-academic-collaboration-a-new-form-of-colonisation-61382

Alcaide, G. G., & Ferri, J. G. (2014). La colaboración científica: principales líneas de investigación y retos de futuro. *Revista española de Documentación Científica*, *37*(4), 062.

Canagarajah, A. S. (2002). *A Geopolitics of Academic Writing*. Pittsburgh: University of Pittsburgh Press.

Dacos, M. (2016). La stratégie du sauna finlandais: Les frontières des Digital Humanities. *Digital Studies/Le champ numérique*. http://doi.org/10.16995/dscn.41

de Sousa Santos, B. (2010). *Descolonizar el saber, reinventar el poder*. Montevideo: Extensión, Universidad de la República-Ediciones Trilce.

Fiormonte, D. (2017a). Digital Humanities and the Geopolitics of Knowledge. *Digital Studies/Le champ numérique*, 7(1). http://doi.org/10.16995/dscn.274

Fiormonte, D. (2017b). Lingue, codici, rappresentanza. Margini delle Digital Humanities. *Filologia digitale: problemi e prospettive* (pp. 113-140). Accademia Nazionale dei Lincei. Contributi del Centro Linceo Interdisciplinare "Beniamino Segre", 135. Roma: Bardi.

Graham, et al, M. (2011). Visualizing the uneven geographies of knowledge production and circulation. *Global Higher Education*, 14.9. https://globalhighered.wordpress.com/2011/09/14/visualizing-the-uneven-geographies-of-knowledge-production-and-circulation/.

Grandjean, M. (2014). Le rêve du multilinguisme dans la science : l'exemple (malheureux) du colloque #DH2014. http://www.martingrandjean.ch/multilinguisme-dans-la-science-dh2014/

Mignolo, W. (2011). *The darker side of Western modernity*. Durham (N.C.): Duke University Press Books.

van Eck, N. J., & Waltman, L. (2010). Software survey: VOSviewer, a computer program for bibliometric mapping. *Scientometrics*, 84(2), 523-538. https://doi.org/10.1007/s11192-009-0146-3

van Eck, N. J., & Waltman, L. (2011). Text mining and visualization using VOSviewer. *arXiv:1109.2058 [cs]*. http://arxiv.org/abs/1109.2058

van Eck, N. J., & Waltman, L. (2014). Visualizing Bibliometric Networks. En Y. Ding, R. Rousseau, & D. Wolfram (Eds.), *Measuring Scholarly Impact: Methods and Practice* (pp. 285-320). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-319-10377-8_13

Waltman, L., van Eck, N. J., & Noyons, E. C. M. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics*, *4*(4), 629-635. https://doi.org/10.1016/j.joi.2010.07.002

Weingert, S. B. (2014). Submission to DH2015. http://scottbot.net/submissions-to-digital-humanities-2015-pt-1/

Weingart, S. B., & Eichmann-Kalwara, N. (2017). What's Under the Big Tent?: A Study of ADHO Conference Abstracts. *Digital Studies/Le champ numérique*, 7(1), 6. DOI: http://doi.org/10.16995/dscn.284

Using Topic Modelling to Explore Authors' Research Fields in a Corpus of Historical Scientific English

Stefan Fischer
stefan.fischer@uni-saarland.de
Universität des Saarlandes, Germany

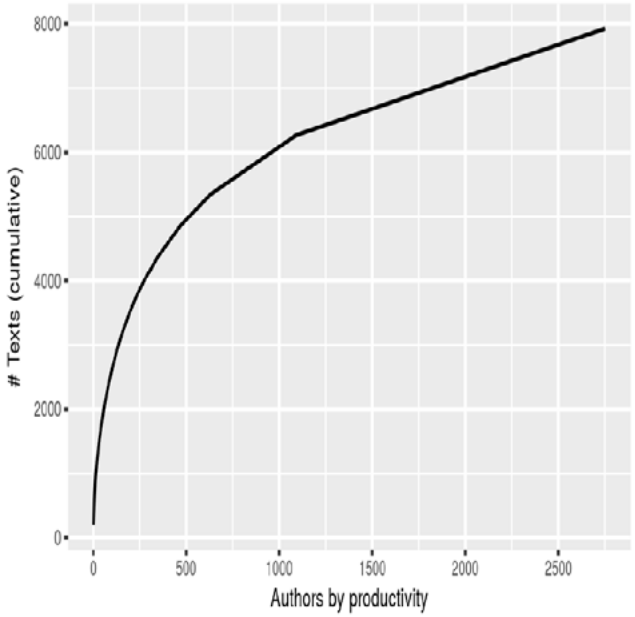
Jörg Knappen
j.knappen@mx.uni-saarland.de
Universität des Saarlandes, Germany

Elke Teich
e.teich@mx.uni-saarland.de
Universität des Saarlandes, Germany

Introduction

In the digital humanities, topic models are a widely applied text mining method (Meeks and Weingart, 2012). While their use for mining literary texts is not entirely straightforward (Schmidt, 2012), there is ample evidence for their use on factual text (e.g. Au Yeung and Jatowt, 2011; Thompson et al., 2016). We present an approach for exploring the research fields of selected authors in a corpus of late modern scientific English by topic modelling, looking at the topics assigned to an author's texts over the author's lifetime. Areas of applications we target are history of science, where we may be interested in the evolution of scientific disciplines over time (Thompson et al., 2016; Fankhauser et al., 2016), or diachronic linguistics, where we may be interested in the formation of languages for specific purposes (LSP) or specific scientific "styles" (cf. Bazerman, 1988; Degaetano-Ortlieb and Teich, 2016).

We use the *Royal Society Corpus* (RSC, Kermes et al., 2016), which is based on the first two centuries (1665–1869) of the *Philosophical Transactions* and the *Proceedings* of the *Royal Society of London*. The corpus contains 9,779 texts (32 million tokens) and is available at https://fedora.clarin-d.uni-saarland.de/rsc/. As we are interested in the development of individual authors, we focus on the single-author texts (81%) of the corpus. In total, 2,752 names are annotated in the single-author papers, but the activity of authors varies. Figure 1 shows that a small group of authors wrote a large portion of the texts. In fact, the twelve authors used for our analysis wrote 11% of the single-author articles.



Productivity of writers of single-author papers

Approach

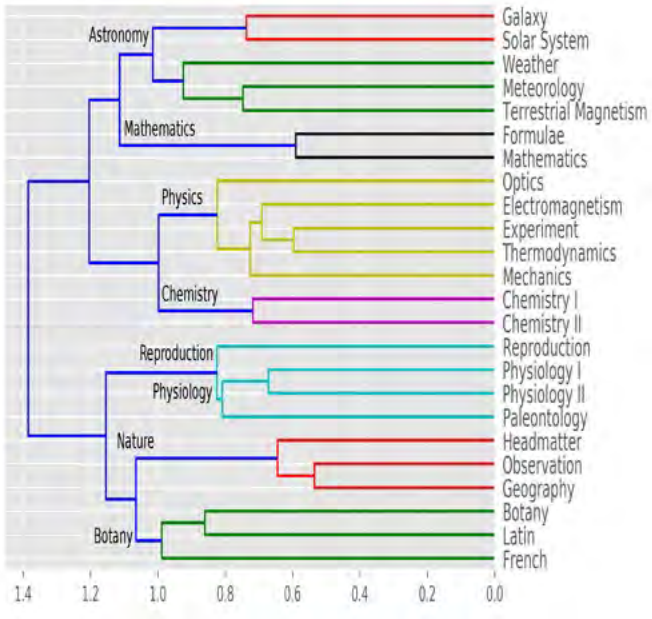
The topic modelling approach we take as a basis is Latent Dirichlet Allocation (LDA, Blei et al., 2003). LDA assumes that corpora contain a number of recurring topics and it treats texts as bags of words. Topics, which can be regarded as groups of semantically related words, are represented as probability distributions over words and each text is treated as a mixture of topics. Typically, topics are displayed as lists of the most probable words and labels are assigned manually. We also considered author-topic models (Rosen-Zvi et al., 2010) but their author-topic matrix implies that authors' topics are fixed over time.

As disciplines were not part of the original metadata of the RSC, we applied topic modelling to approximate disciplines. Using MALLET (McCallum, 2002), we built a model with 24 topics, which are shown along with their characteristic words in Figure 2.

Label	Words	%
Botany	plant leaves plants tab tree foliis folio seeds flowers bark seed species le...	2.0
Chemistry I	water acid grains quantity salt iron solution air experiments found lime col...	6.8
Chemistry II	acid water solution hydrogen oxygen obtained action salt cent alcohol gave s...	5.9
Electromagnetism	wire iron electricity experiments current experiment made end electric coppe...	4.3
Experiment	author present general subject state results nature similar case place great...	3.6
Formulae	cos sin oo tan ab sine axis ac io nt cd aa log vi cc arc al bc ef	4.3
French	la les le des en dans du par qui une qu il ou pour ce je sur au ne	4.2
Galaxy	stars distance position star obs equatorial diff small vf double magnitudes ...	1.1
Geography	sea water great miles found north part time river south side earth land east...	5.3
Headmatter	years year society time royal life age great number made letter part work pu...	0.3
Latin	quae quam sed ab sit vero hoc sunt ac qui esse etiam autem pro erit inter qu...	4.0
Mathematics	equation equations series number form terms values case equal order point cu...	0.7
Mechanics	force motion equal point surface velocity axis line plane body direction ang...	0.5
Meteorology	observations time hours tide water station hill height diurnal made stations...	4.0
Observation	made great found parts part make time small water body account long nature m...	5.5
Optics	light rays glass eye spectrum red lines colour colours surface blue white le...	1.3
Paleontology	bone part bones teeth surface upper lower side anterior length posterior jaw...	4.7
Physiology I	blood time animal day urine parts hours heart found food part days quantity ...	19.1
Physiology II	fibres nerves nerve part muscles vessels side muscular posterior anterior le...	2.8
Reproduction	cells form species surface structure cell membrane found part shell animal s...	1.6
Solar System	sun time observations made moon distance observed observation telescope limb...	7.8
Terrestrial Magnetism	needle magnetic ship observations direct force compass north made dip erebus...	6.7
Thermodynamics	air water heat temperature experiments tube experiment gas time made mercury...	0.1
Weather	rain cloudy ditto fair wind weather clear sw day fine ne cy se rn winds apri...	3.2

Topic labels and top words

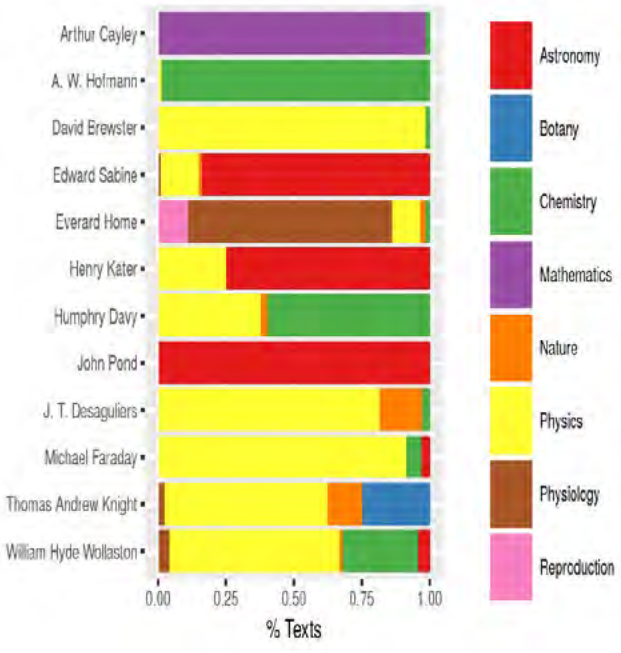
Following the approach of Fankhauser et al. (2016), we clustered the topics using Jensen–Shannon divergence. Figure 3 shows the resulting topic hierarchy. Based on this clustering, we identified broader research areas, which we marked on the branches of the dendrogram.



Hierarchical clustering of the 24 topics

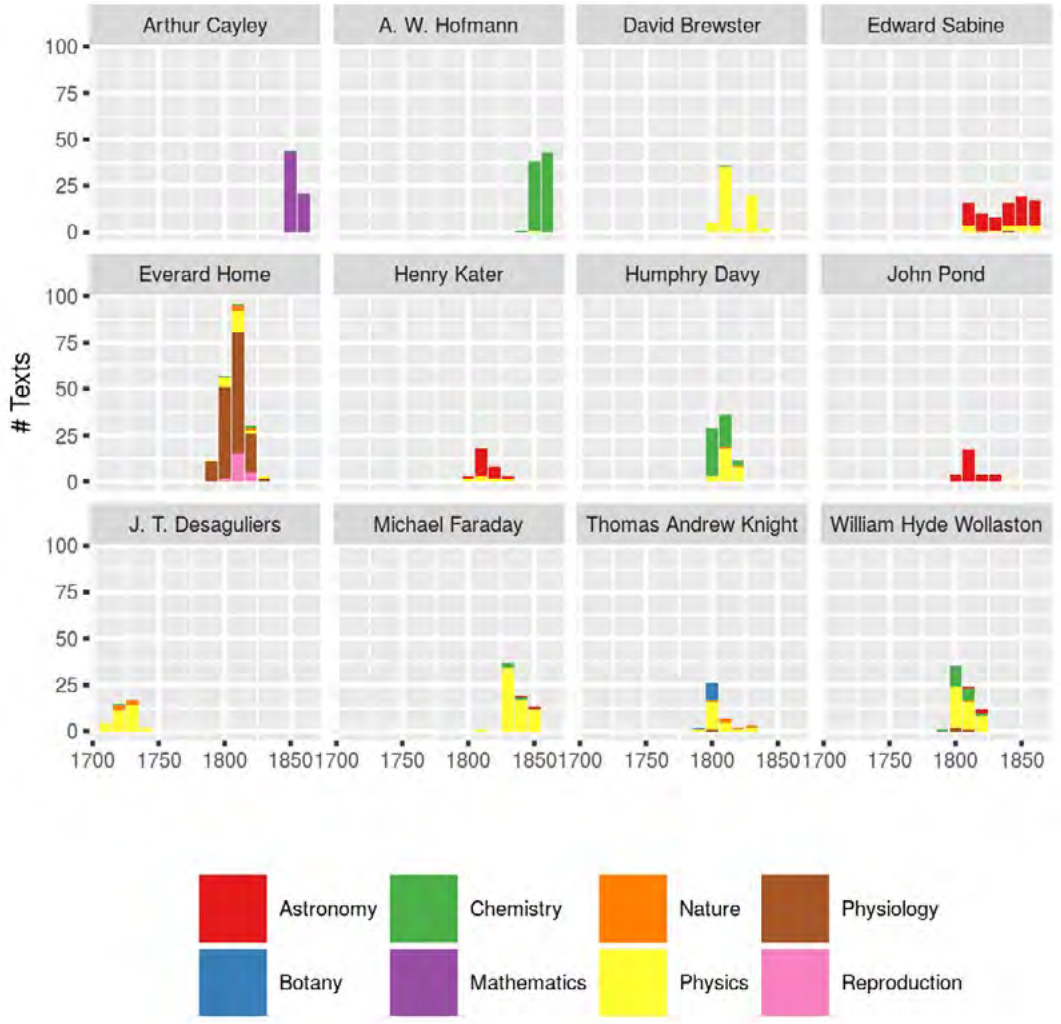
Results

Using these broader categories, we explore whether individual authors stayed in the same area or shifted their focus during their time of scientific production. For this purpose, we selected the most prolific authors (29–198 articles) in the RSC and tracked their topics over time (see Figures 4 and 5). We excluded names if we could not identify the author in the *Virtual International Authority File* or if publication years did not match the author's lifetime.



Comparison of topics of most prolific authors

Figure 4 shows the topics used by twelve authors during their career. We can see two groups of authors. Authors like *Arthur Cayley* dedicated their life to a single research area whereas *Humphry Davy* worked on two topics or in an interdisciplinary area. Figure 5 shows the development of the same authors over time. Overall, the authors' interests did not change dramatically over their professional life. However, one can identify a peak of productivity for most authors.



Development of individual authors over time

Conclusion

We proposed to use topic modelling as a method of exploring the development of the scientific orientation of individual authors over time. Taking topic as an approximation of discipline, our approach can be used to explore the contribution of a particular author to a given discipline over time or find authors with potentially interesting production profiles (e.g. authors shifting topics). In our future work, we will improve our models (e.g. avoid potential confusion of namesakes) by better metadata on the authors which we will obtain from the Royal Society.

Acknowledgement

We acknowledge the support of DFG (Deutsche Forschungsgemeinschaft) through the Cluster of Excellence *Multimodal Computing and Interaction* (MMCI).

References

Au Yeung, C. and Jatowt, A. (2011). Studying How the Past is Remembered: Towards Computational History Through Large Scale Text Mining. *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. (CIKM '11). Glasgow, Scotland, UK: ACM, pp. 1231–1240.

Bazerman, C. (1988). *Shaping Written Knowledge: The Genre and Activity of the Experimental Article in Science*. Madison: University of Wisconsin Press.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3: 993–1022.

Degaetano-Ortlieb, S. and Teich, E. (2016). Information-based Modeling of Diachronic Linguistic Change: from Typicality to Productivity. *Proceedings of the 10th SIGHUM Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*. Berlin, Germany: Association for Computational Linguistics, pp. 165–173.

Fankhauser, P., Knappen, J. and Teich, E. (2016). Topical Diversification over Time in the Royal Society Corpus. *Digital Humanities 2016: Conference Abstracts*. Kraków: Jagiellonian University & Pedagogical University, pp. 496–500.

Kermes, H., Degaetano-Ortlieb, S., Khamis, A., Knappen, J. and Teich, E. (2016). The Royal Society Corpus: From Uncharted Data to Corpus. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia: European Language Resources Association (ELRA).

McCallum, A. K. (2002). *MALLET: A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu (accessed 1 April 2018).

Meeks, E. and Weingart, S. B. (2012). The Digital Humanities Contribution to Topic Modeling. *Journal of Digital Humanities*, 2(1): 1–6.

Rosen-Zvi, M., Chemudugunta, C., Griffiths, T., Smyth, P. and Steyvers, M. (2010). Learning Author-topic Models from Text Corpora. *ACM Transactions on Information Systems (TOIS)*, 28(1): 4:1–4:38.

Schmidt, B. M. (2012). Words Alone: Dismantling Topic Models in the Humanities. *Journal of Digital Humanities*, 2(1): 49–65.

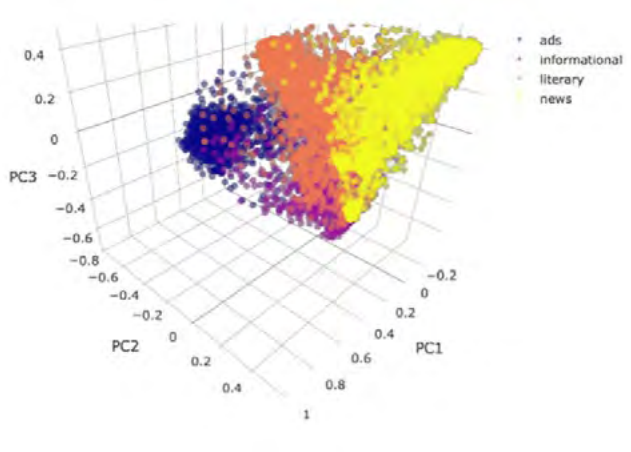
Thompson, P., Batista-Navarro, R. T., Kontonatsios, G., Carter, J., Toon, E., McNaught, J., Timmermann, C., Worboys, M. and Ananiadou, S. (2016). Text Mining the History of Medicine. *PLOS ONE*, 11(1): 1–33.

Stranger Genres: Computationally Classifying Reprinted Nineteenth Century Newspaper Texts

Jonathan D. Fitzgerald
fitzgerald.jo@husky.neu.edu
Northeastern University, United States of America

Ryan Cordell
r.cordell@northeastern.edu
Northeastern University, United States of America

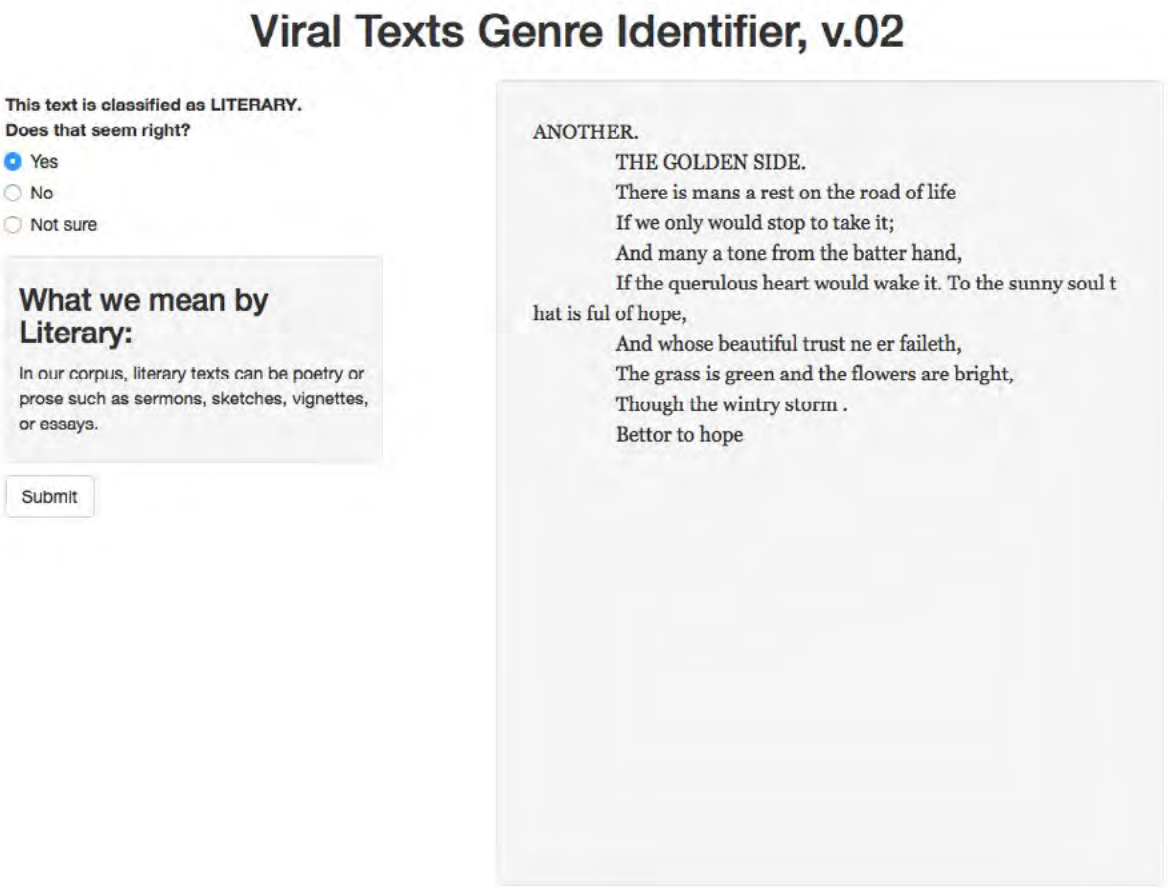
Since its inception in 2012, the *Viral Texts Project* has identified several millions of reprinted texts from corpora of nineteenth-century newspapers. The project began with the aim of isolating texts worthy of closer academic scrutiny from the "big data" of scanned newspapers, but the project's derived data is itself now so big that it cannot be effectively studied through browsing and reading alone. This poster describes our efforts to theorize and implement one solution to this challenge, through computational classification that identifies reprinted texts by genre. The poster will also share a prototype crowd-sourcing experiment that creates a bridge between computational research and various publics by encouraging scholars, students, journalists, and others to explore the strange genres of the nineteenth-century newspaper while enhancing our ground-truth data for training improved classifiers. Following other scholars who affirm the importance of human judgment in computational text analysis (Underwood, 2017; Klein, 2014; Long and So, 2015), our classification method employs unsupervised and supervised modelling: topic modeling and principal component analysis to group similar texts within a training set and generalized linear modelling to sort additional texts from the larger corpus. When the PCA data are visualized in three dimensional space, they cluster around four centers, which, upon closer inspection, can be described as four discrete but overlapping genres: news, advertisements, informational pieces, and literary pieces. Our GLM-based classifier—trained on data derived from PCA and confirmed by human readers—has been successful at finding and identifying thousands of previously unclassified texts in each of these genres.



These early experiments are helping our team more effectively isolate particular genres of texts for deeper literary-historical study, but these experiments are perhaps more valuable for the ways they are helping us reconsider our notions of genre itself in nineteenth century newspapers. Genres, as noted by other scholars who use computational methods to classify texts by genre (Schöch, 2017), are highly complex and fluid through time. In an effort to avoid presentist or anachronistic readings of genre, we

dispense with conceptions of journalistic genres drawn from twentieth- and twenty-first-century newspapers, and attend instead to the much more complex reality of the nineteenth century newspaper. For example, among the texts found in the "literary" category, we've identified many examples of what we name "vignettes"—short prose pieces that are a hybrid of fact and fiction, moral lesson and humorous anecdote. Vignettes of this kind are quite remote from contemporary journalistic genres, and yet we theorize that vignettes encapsulate the hybrid nineteenth century periodical press.

To make our classification efforts accessible to wider publics—and following other scholars who have done likewise in recent years (Beals, 2017; Mullen, 2016)—we have created a crowd-sourcing web application. This app, "The Amazing Generic Automaton," creates accessible paths into our work by allowing users to read a text alongside its most probable genre according to our classifier, asking users to determine whether our classifier has correctly identified the genre. If a user agrees with the classification, she simply clicks "Yes" to confirm; if, however, the genre does not appear to describe the text, the user may select "No" and a list of other genres, listed in the order of their probability as determined by the classifier, appear. The user can then select another genre, or instead choose "other," with a prompt to specify how she might classify the text. The results are saved as CSVs, which, when combined, constitute a new training set for *Viral Texts*. This app, in addition to confirming some of our classification efforts and providing a larger set of ground-truth data, fulfills a major goal of our work: it makes relatively complex computational work more accessible, thus adding a public face to our scholarship. For other humanities scholars less familiar with computational approaches, this app helps them see classification not as a "binary" decision, but instead as a constellation of overlapping generic probabilities.



The poster we propose will outline our process, describe what we're learning about genre in the nineteenth-century periodical press, present early results and visualizations, and offer conference attendees the opportunity to try out "The Amazing Generic Automaton." We expect our presentation will lead to meaningful conversations about innovative approaches to genre classification, the nature of literary genre situated in specific historical periods, and the benefits of creating bridges between complex computational text analysis work and the public.

References

Beals, M. H. (2017). Scissors-and-Paste-O-Meter Officially Launched for 1800-1900 http://mhbeals.com/scissors-and-paste-o-meter-officially-launched-for-1800-1900/ (accessed 28 November 2017).

Klein, L. F. (2014). Talk at Digital Humanities 2014 *Lauren F. Klein* http://lklein.com/2014/07/talk-at-digital-humanities-2014/ (accessed 28 November 2017).

Long, H. and So, R. J. (2015). Literary Pattern Recognition: Modernism between Close Reading and Machine Learning. *Critical Inquiry*, 42(2): 235–67 doi:10.1086/684353.

Mullen, L. (2016). America's Public Bible: Biblical Quotations in U.S. Newspapers http://americaspublicbible.org/ (accessed 17 April 2018).

Schöch, C. (2017). Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. *Digital Humanities Quarterly*, 11(2) http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html.

Underwood, T. (2017). We're probably due for another discussion of Stanley Fish *The Stone and the Shell* https://tedunderwood.com/2017/07/13/were-probably-due-for-another-discussion-of-stanley-fish/ (accessed 28 November 2017).

Humanities Commons: Collaboration and Collective Action for the Common Good

Kathleen Fitzpatrick
kfitz@msu.edu
Michigan State University, United States of America

Humanities Commons is an open-source, open-access not-for-profit social network and scholarly communication platform founded by the Modern Language Association and supported by a collective of scholarly organizations. Scholars and practitioners across the humanities and around the globe can create a professional profile, discuss common interests with colleagues, develop new publications, and share their work with other scholars and with the world.

Humanities Commons grew out of the MLA's experiences with its January 2013 launch of MLA Commons; the earlier platform was designed to serve the needs of MLA members by providing a range of types of open, networked communication. Early adopters, however, exhibited a strong desire to collaborate with scholars in fields other than those represented by the MLA. At the same time, the MLA was approached by several other ACLS member societies seeking similar networked communication solutions. Further, increasing concerns among scholars about the future disposition of commercial scholarly networks, given the sale of both Mendeley and SSRN to Elsevier and the problematic profit models being developed by ResearchGate and Academia.edu, revealed a strong desire for a sustainable not-for-profit alternative.

Given its successful prior work in the area, the MLA was well-positioned to explore the development of a federated platform that might be jointly supported by multiple scholarly societies, bringing together proprietary membership-oriented spaces with a range of fully public functions. With the support of the Andrew W. Mellon Foundation, the MLA met with a group of societies to discuss the possibility and then designed a pilot project to test the technical assumptions behind the federated network. Working with three partner organizations — the Association for Jewish Studies; the Association for Slavic, East European, and Eurasian Studies; and the College Art Association — the MLA launched Humanities Commons in beta in December 2016.

The network currently comprises four primary functions:

- a profile system permitting humanities practitioners to create a professional presence in a non-for-profit online space where they can easily connect with others in their fields;
- an open-access repository that allows members to archive and share the many products of their work, and to notify other members of their availability;
- a community platform, permitting members to join groups, share ideas, and discuss common interests;
- a publishing platform, permitting individuals or groups to create articles, books, teaching materials, Web sites, and blogs, to make their research public and to seek feedback on work in progress.

The network is built on the Commons In A Box (CBOX) platform, developed by the CUNY Graduate Center; CBOX is in turn based on WordPress and BuddyPress, which bring together a flexible publishing engine with rich social networking capabilities. The network's repository system is Fedora/Solr-based, with a WordPress front-end, developed in collaboration with the Columbia University Libraries and with the support of the National Endowment for the Humanities. Additionally, the network uses a federated authentication and identity management system, primarily based on COmanage and other Internet2-based systems, that communicates with the membership databases of participating scholarly organizations, thus allowing members to access all the organizations to which they belong through a single sign-on mechanism.

As of mid-April 2018, Humanities Commons has over 13,500 members who are actively developing their professional profiles. In order for the network to thrive, however, it must develop in a sustainable fashion. The planning and development for Humanities Commons were

undertaken by the MLA as a service to the profession, as well as to its sister societies, with the goal of providing an open-source, scholar-governed alternative to the available commercial services. That development was partially supported through grant funding, as noted above, but grant funding is not a business model; funders expect a project such as this to develop a sustainability plan to ensure its future. Humanities Commons is thus working toward collective action by and shared services for scholarly societies and other kinds of scholarly organizations who want to work together to provide a rich scholarly communication infrastructure for their members and for the profession at large.

This poster presentation will include an active demo of Humanities Commons as well as discussions of its platform, its community, its sustainability plan, and its development roadmap. We want to encourage members of the ADHO community to join the network and connect with one another across the conference and throughout the year. We also want to invite active participation among ADHO members and constituent organizations in the network's development process.

Making DH-Course Together

Dinara Gagarina
dinara@psu.ru
Perm State University, Russian Federation

The project involves the development of the MA-course "Concepts and Approaches of Digital Humanities", which is one of the basic courses in the MA-program "Digital Technologies in Sociocultural and Art Practices." Currently, the approbation of the experimental methodology in one of the universities is underway. The total volume of the discipline is three credits. The course is placed in the first semester, after which courses in specific areas of the DH are studied.

The goal of the course is to review the existing concepts of Digital Humanities, approaches to defining the subject and methods of Digital Humanities, theoretical and methodological foundations of using IT in various humanities, to get an idea of the relevant directions, tools, and projects.

The relevance of the project is primarily due to the growing importance of Digital Humanities as an interdisciplinary area. The developed MA program and this course as its part, allow combining the knowledge and methods accumulated in the application of IT in separate humanities disciplines, and train specialists of a new type. This association also takes place at the organizational level and facilitates the interaction of faculties, the implementation of joint projects.

Our approach to constructing the course is to actively involve students not only in research activities and projects, but also in the discussion and formation of the structure of the course, and then in filling it. The means of implementing this approach is the dynamic creation of the course site during the entire training period. We want to involve students in the practice of shaping the course, the constant retention of the focus of attention.

At the same time, we solve several problems and obtain a number of didactic opportunities and advantages.

Firstly, there is a problem of the lack of educational literature on DH in general, mainly prevailing literature and pedagogical developments in selected areas of DH. There is a language problem, for example, the only one reader by DH in Russian is released (Digital Humanities: A Reader, 2017). We use various types of materials - video, MOOC-courses, web resources, and actively read scientific literature.

Secondly, it is important for us to show a common landscape and a multitude of DH-directions on a scale. We work with masters who have different backgrounds in the bachelor's degree: historians, culture and art studies, philologists, PR, etc. There are also students in the group without humanities education at the bachelor's level.

Third, we want to combine teaching with the formation of a set of important DH skills: designing and retrieving information for web resources and databases, working with maps and timelines, corpora of texts. We use the students' conscious and active approach to learning and suggest that they become co-authors of the training course, first discussing possible course structures with them, and then jointly creating a special course site (Gagarina, 2017).

All links, texts related to the work of students during the course, are posted on the special site of the course. The entire group has access to viewing and commenting on all sections of the site and editing their own materials. Students learn to work in a team.

In the first lesson after the opening remarks, students are divided into pairs and offer their own model of the course. Since we are now conducting an experiment, we can already confidently say that students see DH with a skewing in its background or experience. During the discussion of these models, we jointly design a common framework and plan. For the convenience of combining the models, I suggest that students make a structure of 3-5 modules with a possible division into topics within the module. At this stage, one can clearly see the feature - students see as the first block the definition and history of DH, as 2-3 modules, most of them suggest considering directions within DH (computer linguistics and digital history, for example). Then many students propose to do their project. Almost no one talks about the consideration of common approaches, the general DH methodology, the DH infrastructure, and the classification of DH by technologies and tools.

The skill of the teacher is at this stage in combining your pre-designed curriculum and the vision of students.

It is quite possible to do this by shifting the emphasis somewhat.

At the conference, we plan to present the results of the experiment and the project as a whole.

The study is supported by Vladimir Potanin Foundation.

References

Gagarina D. (2018). *Studying Digital Humanities,* https://dhumanities.ru/.

Digital Humanities: A Reader (2017). Ed. M. Terras, J. Nyhan, E. Vanhoutte, I. Kizhner, Krasnoyarsk, 352 p. (in Russian).

Standing in Between. Digital Archive of Manuel Mosquera Garcés.

Maria Paula Garcia Mosquera
mpgarcia10@brown.edu
Brown University, United States of America

How have the life and achievements of preeminent Afro-Colombians been depicted in digital spaces? Which aspects of their lives have been highlighted in those efforts? What do those projects talk about the way these peoples have been remembered? Starting from these questions, *Standing in Between. The Digital Archive of Manuel Mosquera Garcés* is an initiative aiming to deepen in the history of Afro-Colombian politicians and intellectuals from the mid 20th century by creating an extended (digital) narrative of Manuel Mosquera Garcés.

Born in the Pacific coast in 1907, Mosquera Garcés was among the first Afro-Colombians to reach prominence in the Colombian government between the 1940s and 1970s. A leader of the Conservative party, Mosquera Garcés was part of a generation of politicians coming from the periphery who actively worked towards the inclusion of their home region into national dynamics. His story, however, has been blurred within the historical narratives of the country. Mosquera Garcés' legacy does not easily fit into the dominant narratives typical of a Colombia's official and centralized history (white, conservative, wealthy, eager to replicate Western and Catholic values), nor the mainstream narratives of the Afro-Colombians (black, liberal, underprivileged, eager to claim their African roots). His story in sharp contrast against those narratives, as he was a conservative politician from a marginalized region of the country who believed profoundly in Catholic principles. Additionally, he was black, a lawyer and a passionate reader of intellectuals of the Western tradition. He worked in Bogotá (capital city) while he was standing for his people in Chocó. The project is designed as a digital repository that will publicly display— for the first time — Mosquera Garcés' personal archive, along with additional documents related to his work, contextualizing the whole set as a curated collection.

Based on Kim Gallon's work on the "politics of recovery"(1) and the ways historiographical reinterpretations could be considered political enterprises to restore the "humanity" of black people as historical, political, and intellectual agents, *Standing in Between* will seek to restore the historical role and agency of Afro-Colombians in the digital domain. Connected to Liliana Ángulo's artwork "A case of reparation,"(2) which *liberates* archival sources to reveal historical erasures of the Botanical Expedition, the project is guided by the importance of offering sources to generate analysis with an extensive level of historical detail. Indeed among different local blogs and websites, including *Historia Personajes Afrocolombianos*, *Enamórate del Chocó*, and *República de Colores*, Mosquera Garcés has been included as a historical Afro-Colombian figure. In the form of biographies and informative articles, these private initiatives are rooted in an urgency to present the legacy of Afro-Colombians in order to incorporate these stories as part of the national identity and historical discourse. The University of Vanderbilt has published part of the correspondence of Manuel Zapata Olivella (black novelist) and historical documents of the Pacific Coast, while on a local level the appearance of digital initiatives and archives is still an emerging process.

Standing in Between aims to join these efforts examining Mosquera Garcés's archive, which was preserved by his family but until now it has not been scholarly reviewed, by considering three lenses that influenced his academic and political life: religion, language, and race. Archival material is diverse, and includes photographs, sound archives, bibliographic documents, and correspondence dating from the 1920s to the 1970s. Due to Mosquera's involvement in several periodical publications, as well as his work in the government in different capacities, the privately preserved documents do not offer a complete body of documentation of his political and scholarly life. In order to provide a more comprehensive context, the project has carried archival work in several public archives and libraries, to broadly identify his political agenda and academic interests. The archival work paid special attention to content reflecting his religious thought and conservative partisanship.

The initial work done on the digitization and cataloging of these materials, and the preliminary findings of curating this archive, will be presented in this poster. Additionally, in this early stage of the project, the design of a timeline will be displayed as a way of visualizing the connections between Mosquera Garcés and his generation of peers in his native Chocó poster, all of whom were bridging the gap between the center and the periphery through their participation in the national government. This first visual tool will add references to the collection, other digital projects on Afro-Colombians, and oral histories conducted for this Project.

References

Gallon, K. (2016). Making a Case for the Black Digital Humanities. In her article, Dr. Gallon In Gold M. & Klein L. (Eds.), *Debates in the Digital Humanities 2016* (pp. 42-49). Minneapolis; London: University of Minnesota Press. Retrieved from http://www.jstor.org/stable/10.5749/j.ctt1cn6thb.7

Ángulo, Liliana. (2015). *Un caso de reparación. Un proyecto de reparación histórica y humanidades digitales*. http://uncasodereparacion.altervista.org/?doing_wp_cron=1524636377.0468459129333496093750 (accessed 20 April 2018).

Research Environment for Ancient Documents (READ)

Andrew Glass
asg@uw.edu
Microsoft Corp., University of Washington

Stephen White
stephenawhite57@gmail.com
Stephen White - Italy

Ian McCrabb
ian@prakas.org
University of Sydney, Prakas, Foundation, Australia

The Research Environment for Ancient Documents (READ) is an integrated Open Source web platform for epigraphical and manuscript research. It may be configured as the underlying engine for a text repository or as a complementary research toolset to an existing repository. The defining innovation of this software is the atomization of text into orthographic subunits (as opposed to lines or words). This enables mapping across all layers of textual analysis, from factual data (the location of a character on a surface) through contestable (the transcription of a character) to the purely interpretive (a semantic annotation). This data architecture enables:

- The integration of physical, textual, and interpretive aspects of research
- The transformation of conventional editing practice into optimized workflows
- Granular attribution of components of a text, which allows for alternative interpretations and flexible collaboration

This poster outlines the workflows and outputs supported by version 1.1 (2017) of READ. The first release is optimized for use with Indic languages using ak ara-based writing systems (abugida). We will demonstrate the platform using documents in Gāndhārī language. We will also demonstrate the ability to generalize READ to support other languages and writing systems, e.g., Aramaic, Chinese, English, Italian, and Mayan.

The core workflows of the READ are:

1. **Creating a new item for study and inputting a text transcription**. A researcher creates a new item in READ and adds basic metadata and enters a transcription of the item in free text. Once entered, the researcher can immediately access two types of reports: a wordlist generated from the text; and alternate presentations of the text edition (diplomatic, reconstructed, and hybrid). These reports are available via READ's web interface, as well as the following downloadable export formats: HTML export, RTF, TEI (EpiDoc).
2. **Uploading images of the source text and linking to the text transcription.** READ provides tools to mark segment boundaries around the graphical units of the writing system depicted in the images. These segments are then automatically linked to the transcription entered in step 1. At this point the researcher can view the edition side by side with the image using synchronized scrolling provided by READ's web interface. In addition, the researcher can access a paleographic report generated from the image segments using the linked transcription. The TEI (EpiDoc) export includes the image as Facsimile element. All image segments can be exported as distinct files for paleographic processing using external tools.
3. **Creating a text glossary by adding lexicographical data to the generated wordlist.** The researcher uses tools provided by READ to add lexicographical data to the wordlist that was generated in step 1. At this point, a glossary can be generated and exported (HTML, RTF) or viewed in READ's web interface. Also, the edition viewer in READ's web interface integrates glossary data in flyouts associated with each word in the edition.
4. **Completing the glossary.** The researcher views the glossary created in step 3 in the READ's web interface and adds compound analysis to any compounds occurring in the text. At this point, glossary generation includes cross-reference entries for compound members.
5. **Annotating the edition.** The researcher uses annotation tools provided by READ to add footnotes and tags to the edition. The researcher can add text-structural information as well as textual parallels, translation, and alternate transliteration forms. These annotations can be viewed in the web interface, as footnotes in exported RTF and HTML output.
6. **Cubing the edition.** A researcher can integrate alternate editions of the same text using tools provided by READ. Any alternate editions so integrated, will be linked to the same image added in Step 2. Alternate editions can be viewed side-by-side in READ's web

interface to support comparison between alternate editions of a text.

7. **Sharing the research.** READ has been designed as a collaborative tool from the outset. Researchers can choose to share visibility and editing rights to any of the elements in their work. Work can also be published in mutable and immutable forms via the READ viewer interface, as well full text editions in TEI, exported HTML, and RTF that can be opened in common word processing and desktop publishing software applications.

The READ project began in 2013 and has been funded by Ludwig-Maximilians Universität, Munich, Germany; the University of Washington, Seattle, USA; Université de Lausanne, Switzerland; University of Sydney, Australia; and Prakaś Foundation, Sydney, Australia.

Manifold Scholarship: Hybrid Publishing in a Print/Digital Era

Matthew K. Gold
mgold@gc.cuny.edu
Graduate Center, City University of New York, United States of America

Jojo Karlin
jojo.karlin@gmail.com
Graduate Center,
City University of New York, United States of America

Zach Davis
zach@castironcoding.com
Cast Iron Coding, United States of America

This poster will present the Manifold Scholarship project (http://manifold.umn.org), an open-source scholarly communication and book publishing platform funded by the Andrew W. Mellon Foundation. Created by the University of Minnesota Press, The GC Digital Scholarship Lab, and Cast Iron Coding, Manifold aims to present the scholarly monograph in a new networked and iterative form that still has strong ties to print.

Manifold editions bridge the space between static print and ebook forms and custom web-based projects that are individually designed and programmed to meet the unique and specific needs of a particular scholar. Manifold editions present a multi-dimensional version of the book as we know it—a base text upon which a set of media and user-interaction layers can be added along with an archive space for related research materials. The reading experience offers a set of standard characteristics and constraints so readers who read and interact with one Manifold edition know how to interact with another, no matter the publisher.

Manifold offers a potentially powerful platform for publishers who hope to offer web-based editions of their books at scale. It can ingest ePubs, the format used most often by scholarly presses in their production practices, but it can also ingest Google docs, markdown files, and Microsoft Word docs. It is thus useful not only for scholarly presses, but also for individual DH practitioners who wish to publish their work in an attractive, responsive format with options for annotating and highlighting works. Future development on the platform will enable it to be used in classrooms by groups of students, who might comment together on OER materials that have been published on a Manifold instance.

This poster will explain what Manifold is, how it works, how it integrates into existing university-press publishing workflows, and how others may begun using it on their own for a variety of publishing and pedagogical needs.

Legal Deposit Web Archives and the Digital Humanities: A Universe of Lost Opportunity?

Paul Gooding
p.gooding@uea.ac.uk
University of East Anglia, United Kingdom

Melissa Terras
m.terras@ed.ac.uk
University of Edinburgh, United Kingdom

Linda Berube
l.berube@uea.ac.uk
University of East Anglia, United Kingdom

Introduction

Legal deposit libraries have archived the web for over a decade. Several nations, supported by legal deposit regulations, have introduced comprehensive national domain web crawling, an essential part of the national library remit to collect, preserve and make accessible a nation's intellectual and cultural heritage (Brazier, 2016). Scholars have traditionally been the chief beneficiaries of legal deposit collections: in the case of web archives, the potential for research extends to contemporary materials, and to Digital Humanities text and data mining approaches. To date, however, little work has evaluated whether legal deposit regulations support computational approaches to research using national web archive data (Brügger, 2012; Hockx-Yu, 2014; Black, 2016).

This paper examines the impact of electronic legal deposit (ELD) in the United Kingdom, particularly how the 2013 regulations influence innovative scholarship using the Legal Deposit UK Web Archive. As the first major case

study to analyse the implementation of ELD, it will address the following key research questions:

- Is legal deposit, a concept defined and refined for print materials, the most suitable vehicle for supporting DH research using web archives?
- How does the current framing of ELD affect digital innovation in the UK library sector?
- How does the current information ecology, including not for-profit archives, influence the relationship between DH researchers and legal deposit libraries?

Research Context

The British Library began harvesting the UK web domain under legal deposit in 2013. The UK Web Archive had, by 2017, grown to 500Tb. However, UK legal deposit regulations, based on a centuries-old model of reading room access to deposited materials, affect the archive's significant potential for research: in practice, researchers can only access the full range of UK websites within the walls of selected institutions. DH scholars, though, require access to textual corpora and metadata in addition to interfaces for discovery and reading (Gooding, 2012). Winters argues that "it is the portability of data, its separability from an easy-to-use but necessarily limiting interface, which underpins much of the exciting work in the Digital Humanities" (2017: 246). Restricted deposit library access requires researchers to look elsewhere for portable web data: by undertaking their own web crawls, or by utilising datasets from *Common Crawl* (http://commoncrawl.org/) and the *Internet Archive* (https://archive.org). Both organisations provide vital services to researchers, and both innovate in areas that would traditionally fall under the deposit libraries' purview. They support their mission by exploring the boundaries of copyright, including exceptions for non-commercial text and data mining (Intellectual Property Office, 2014). This contrast between risk-enabled independent organisations and deposit libraries, described by interviewees as risk averse, challenges library/DH collaboration models such as *BL Labs* (http://labs.bl.uk) and Library of Congress Labs (https://labs.loc.gov).

Methodology

This paper analyses the impact of the UK regulatory environment upon DH reuse of the Legal Deposit UK Web Archive. It presents a quantitative analysis of information seeking behaviour, supported by insights from 30 interviews with UK legal deposit library practitioners. Quantitative datasets consisted of Google Analytics reports, and web logs of UK web archive usage, which were analysed in SPSS and Excel. These datasets allowed us to identify broad patterns of information-seeking behaviour.

Practitioner interviews were hand-coded to three levels in Nvivo: initial coding, to provide the foundations for higher level analysis; focused coding, to further refine the data; and axial coding, using the convergence of ideas as a basis for exploring the research questions (Hahn, 2008). This analysis will inform two further research phases: a broader quantitative analysis of UK ELD collections; and qualitative analysis of the ways that the research community, and DH researchers, use ELD collections.

Conclusion

This paper provides a vital case study of how legal deposit regulations can influence library/DH collaboration. It argues that UK ELD regulations use a print-era view of national collections to interpret digital preservation and access. A lack of media specificity, combined with a more cautious approach to text and data mining than allowed under UK copyright, restricts DH research: first, by limiting opportunities for innovative computational research; and second by excluding lab-based library/DH collaborative models. As web preservation activities become concentrated in a small group of key organisations, current regulations disadvantage libraries in comparison to not-for-profits, whose vital work is supported by an ability to take risks denied to legal deposit libraries. The UK's approach to national domain web archiving represents a lost opportunity for computational scholarship, requiring us to rethink legal deposit in light of the differing affordances of born-digital archives.

References

Black, M. L. (2016). The World Wide Web as Complex Data Set: Expanding the Digital Humanities into the Twentieth Century and Beyond through Internet Research. *International Journal of Humanities and Arts Computing*, 10(1): 95–109.

Brazier, C. (2016). Great Libraries? Good Libraries? Digital Collection Development and What it Means for Our Great Research Collections. In Baker, D. and Evans, W. (eds), *Digital Information Strategies: From Applications and Content to Libraries and People*. Waltham, MA: Chandos Publishing, pp. 41–56.

Brügger, N. (2012). Web History and the Web as a Historical Source. *Studies in Contemporary History*, 2 http://www.zeithistorische-forschungen.de/site/40209295/default.aspx (accessed 9 January 2017).

Gooding, P. (2012). Mass Digitization and the Garbage Dump: The Conflicting Needs of Quantitative and Qualitative Methods. *Literary and Linguistic Computing* doi:10.1093/llc/fqs054. http://llc.oxfordjournals.org/content/early/2012/12/22/llc.fqs054.abstract (accessed 30 July 2013).

Hahn, C. (2008). *Doing Qualitative Research Using Your Computer: A Practical Guide*. London: Sage Publications Ltd.

Hockx-Yu, H. (2014). Access and Scholarly Use of Web Archives. *Alexandria*, 25(1/2): 113–27.

Intellectual Property Office (2014). Exceptions to Copyright: Research *UK Government* https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/375954/Research.pdf.
Winters, J. (2017). Coda: Web Archives for Humanities Research - Some Reflections. *The Web as History*. London: UCL Press, pp. 238–48.

Crafting History: Using a Linked Data Approach to Support the Development of Historical Narratives of Critical Events

Karen F. Gracy
kgracy@kent.edu
Kent State University, United States of America

This poster will present a progress report on a project that aims to explore how historians and other humanities scholars can most effectively access and use the data hidden in the silos of digital archival collections to craft narratives about significant developments and critical junctures in historical events, using Linked Data and event-based description. This project has two objectives: 1) to investigate the efficacy of an event-based model of description that will facilitate search across archival inventories and textual documents found in archival collections, and, 2) to develop and test a software tool that will allow scholars to more easily discover and use these hidden nuggets of information about events, and facilitate the construction of explanatory narratives about historical phenomena.

Linked Data and Event-Based Description

In the last two decades, the number of documents, photographs, and other archival material available in open digital archives worldwide has increased dramatically. Yet, these valuable sources of information are often hard to discover, due to long-standing practices in how archival materials are described and cataloged. Archival collections represent a tremendous source of untapped data, which is not discoverable without significant effort on the part of the researcher. Linked Data represents a new approach to information access that goes beyond simple tagging and indexing of documents using a predefined set of topics. Rather, it relies on semantically structured data embedded within the collection inventories, or even in the documents themselves, to interlink related information and make it searchable through semantic queries.

This particular project focuses on the difficulties of finding information on historical events in archival collections. Events are a special form of named entities, as they serve as a nexus point that marks a relationship between specific agents, places, and points in time (Gracy, 2015; Hyvönen, Lindquist, Törnroos, and Mäkelä, 2012). Thus, they act as gathering mechanisms for records of actions and are crucial aspects of archival information systems. To explore the concept of event-based description, the research team for the project has chosen the May 4, 1970 tragedy (during which four students were killed by members of the Ohio National Guard during a Vietnam War demonstration and nine others were injured) as our test case, as it has special resonance for our location at Kent State University. Kent State and other academic institutions have significant archival holdings and other information resources related to this event.

Usefulness of the Event-Based Model for Historical Research

This project employs archival finding aids and selected archival materials to create historical event vocabularies and ontologies, while creating and testing an event-based model that encompasses spatio-temporal dimensions and agents associated with events. The event vocabularies and ontologies are used as the basis for identifying and encoding information about persons, organizations, places, and topics. The event-based description model will be used as the basis for designing an information service that facilitates the linking of historical documents and archival descriptions related to an event, and will also help to link those materials and descriptions to other relevant published and archival sources.

Upon completion of the initial design of the event-based model (which is already underway), the project investigators will develop and test a prototype tool for event information discovery and use which can be used by scholars, students, and others interested in building historical narratives using archival material and related resources. Narrative building, which is the methodological stock in trade for many historians and humanities scholars, relies on the careful accumulation of data via the examination of documents relating to the topic under investigation (Barthes, 1977; White, 1984). This tool will also allow the investigators to test the validity of the event-based model as a suitable approach for facilitating information discovery for archival materials. This project proposes the process of historical research may be aided by a web-based tool designed to help with the discovery, collation, annotation, and sequencing of relevant information, and aims to build a web-based software with that functionality. The investigators propose that this project will have positive outcomes for digital history and humanities work, as it will empower humanities researchers to build complex historical narratives from various primary and secondary sources.

This poster will provide a progress report on the following activities: 1) Testing the event model with semantic metadata drawn from the May 4 Collection, which is an

archival collection from the Kent State University Libraries; 2) Developing and refining a web-based tool to assist historians and cultural heritage scholars in building and testing hypothetical narratives based on the linking of event information from various sources.

References

Barthes, R. (1977). Introduction to the Structural Analysis of Narrative. In Heath, S. (trans), *Image, Music, Text*. New York: Hill & Wang, pp. 79-124.

Gracy, K.F. (2015). Archival Description and Linked Data: A Preliminary Study of Opportunities and Implementation Challenges, *Archival Science*, 15: 239-254. doi: 10.1007/s10502-014-9216-2

Hyvönen E., Lindquist T., Törnroos J., & Mäkelä E. (2012). History on the Semantic Web as Linked Data—An Event Gazetteer and Timeline for the World War I. Proceedings of *CIDOC 2012, Enriching Cultural Heritage*, 10-14 June 2012, Helsinki, Finland. Retrieved from http://www.cidoc2012.fi/en/File/1609/hyvonen.pdf.

White, H. (1984). The Question of Narrative in Contemporary Historical Theory, *History and Theory* 23(1): 1-33.

Prosopografía de la Revolución Mexicana: Actualización de la Obra de Françoise Xavier Guerra

Martha Lucía Granados-Riveros
luciagranadosriveros@gmail.com
Escuela Nacional de Antropología e Historia, Mexico

Diego Montesinos
diegomontesinos@ciencias.unam.mx
Facultad de Ciencias UNAM

En 1985 se publicó el libro *México: del antiguo régimen a la revolución* del historiador Françoise Xavier Guerra, una referencia fundamental para el estudio de la Revolución mexicana. La obra revela las relaciones y tensiones entre la sociedad tradicional, un sistema heredado de la colonia y el Estado moderno proveniente en gran medida de los ideales liberales de la revolución francesa.

El trabajo de Xavier Guerra se inscribe en una amplia tradición de la investigación prosopográfica que se extiende desde el siglo XIX (Verboven, Carlier y Dumolyn, 2007) hasta el relativamente reciente uso de herramientas computacionales para el manejo de bases de datos (Blust, 1989; Keats-Rohan, 2010). El cuerpo biográfico de su investigación está compuesto por más de siete mil actores sociales entre los que figuran individuos y colectividades, con aproximadamente cien mil datos asociados a los movimientos políticos. Para su análisis se construyó una base que sistematizó los datos en más de cincuenta categorías que codifican dos tipos de sucesos; aquellos personales como fecha de nacimiento, muerte y ascendencia familiar y aquellos sucesos relacionados con la vida política y social del actor como participación en batallas o los cargos públicos ocupados. Los sucesos se organizaron en módulos independientes, lo cual permitió enriquecer la base de datos con la captura de nuevos módulos para personajes ya establecidos.

Dicha base de datos fue almacenada originalmente en tres cintas magnéticas de las cuales no se refieren más detalles, realizar nuevos análisis resulta inviable ya que el único medio en que está disponible actualmente es el impreso en los anexos de la obra señalada. El objetivo de este trabajo es la digitalización de la base de datos de Xavier Guerra, que permita la reproducción de los análisis del autor, así como la generación de nuevo conocimiento a partir del cruce de variables.

Con ese fin, se creó un programa en Python que ocupa Tesseract, una biblioteca de reconocimiento de caracteres. Debido a la estructura modular de la base de datos, los renglones, columnas y espacios en blanco son significativos. Por lo tanto, se realizó un pre-procesamiento de las imágenes, para detectar la estructura espacial del texto, de manera que Tesseract procesará pedazos de texto organizados. En esta etapa se ocupó el framework OpenCV y la biblioteca Pytesseract. Posteriormente, el programa organizó la información en un esquema de base de datos dentro de un archivo SQL.

En este póster presentamos el código desarrollado para la recuperación y organización de la base de datos, el funcionamiento de la base mediante algunas réplicas de los análisis que realizó Xavier-Guerra en su obra, así como el resultado de queries inéditos y por último el diseño inicial de la página que permita interactuar con los datos, de modo que los usuarios puedan consultar al sistema en términos de tiempo, geografía, compromisos políticos y relaciones de parentesco o sociales.

References

Blust, N. (1989). Prosopography and the computer: problems and possibilities, en Denley, P. (ed.) *History and computing* . no.2. Manchester, UK: Manchester University Press, pp. 12–18.

Keats-Rohan, K. (2010). Prosopography and Computing: a Marriage Made in Heaven?, *History and Computing*, 12(1), pp. 1–11.

Verboven, K., Carlier, M. y Dumolyn, J. (2007). A Short Manual to the Art of Prosopography, en Keats-Rohan, K. (ed.) *Prosopography Approaches and Applications. A Handbook*. Oxford: University of Oxford, pp. 35–69.

Developing Digital Methods to Map Museum "Soft Power"

Natalia Grincheva
natalia.grincheva@unimelb.edu.au
University of Melbourne, Australia

The project aims to employ Geographical Information Technologies to develop a pilot version of the digital mapping system "Museum Soft Power Map." It explores key factors in the time-space development of museum capacities to contribute to local creative economy by attracting tourism and generating economic activity. In collaboration with the Australian Centre for the Moving Image (ACMI), the project creates a dynamic digital map to visualize a growing in time geographic diversity of the Centre's collections, programming, audiences and partnerships. It reveals what factors affect the development of the ACMI's global brand recognition and influence its capacity to attract larger visitation and revenue.

Contemporary museums, as important actors in the international arena (Sylvester, 2009), increasingly serve as vital economic players helping their cities to compete for talent, tourism, and investment (Towse and Handka, 2013; Vivant, 2011; Werner, 2005). Though Nye's (2004) concept of "soft power" has been recently employed to discuss museum contribution to place branding, urban regeneration and tourism development (Lord and Blankenberg, 2015), there is a significant gap in the academic knowledge on what exact museum resources and activities accumulate "soft power" and how they affect the development of institutional global brand recognition in time and space.

The project tests a theoretical hypothesis that representing, promoting and celebrating cultural diversity help contemporary cultural institutions to attract larger global media attention, increase international visibility and appeal to more diverse audiences and partners (Nye, 2004, La Porte, 2012). The project traces a historic development of the ACMI's global brand that is based on the institutional vision to "be the leading global museum of the moving image" (ACMI, 2016). With diverse collection of hundreds foreign language films, representing a wide variety of cultures across the globe, ACMI runs a dozen of international tours and projects annually to strengthen its "reputation for world class exhibition experiences" (ACMI, 2016). In a close collaboration with ACMI, the project develops a customized Geographic Information System (GIS) that maps a growing international profile and visibility of the ACMI's collections, curatorial expertise and activities through time. The main goal of this digital mapping tool is to explore how attention to diversity on the level of collection acquisitions and a strategic focus on international outreach in its programming help the museum to accumulate institutional "soft power," measured through increase in its audienceship and selfearned income in Melbourne and other hosting cities.

A young, dynamic and ambitious institution, ACMI in 15 years of its existence, managed to develop a large audience reaching in 2016 1.5 million visitors to the Federation Square museum and 500 thousand attendants of its international exhibitions in six countries (ACMI, 2016). With 22% international visitors, ACMI generates $11.5 million through tickets sales and program services annually. As a partner in the project, ACMI is eager to provide its historical institutional records and digital expertise to develop the GIS software which traces and measure the development of its "soft power" in time and space.

The project employs museum records in the last 15 years in collection acquisitions and strategic programming to map and visualise a growing geographic diversity of the museum cultural resources and activities to explore how this international exposure affects audience development. The GIS software operates as a combination of deep mapping layers, each representing a different dimension of museum capitals tied to a specific location on the globe. Resources or Cultural Layer exposes a diversity and scope of museums' collections and main exhibits, highlighting geographic areas of their origins. Outputs or Social Layer maps complex museum "ecosystems" by visualizing museum social resources and telling stories about their engagements with constituencies, partners and audiences on the local and global levels. Impacts or Economic Layer builds on the metric of economic effects, measured through ticket sales at home and abroad, local and international program service revenue, membership dues as well as income received through museum shop, restaurant, and renting. The GIS processes the input data from three dimensions of museum capitals to map, visualize and draw correlations among cultural assets, social outcomes and economic impacts.

Combining and building on recent findings in academic scholarship on deep mapping (Bodenhamer et al., 2010; Gibson et al., 2010; Abrams et al., 2008) and museum evaluations (Jacobsen, 2016), the project designs a GIS system that advances a rapidly developing field of cultural mapping. The major outcome of this project is a research platform that can make a contribution both to applied knowledge and to academic scholarship. On the practical level, this research system can improve ACMI proactive management in global PR and programming. The digital map reveals geographic areas of missed opportunities by exposing locations where ACMI has a low or no cultural affiliations. Also, the system helps to identify "hot spots" of social density in terms of visitation and social activities, as well as to explore if stronger institutional efforts to target specific locations can result in a higher economic return on institutional investments. In academic terms, such a digital mapping tool advances the digital humanities scholarship by developing computational methods to explore cultural institutions and their impacts upon audiences. It combines quantitative and qualitative traditions within cultural mapping to reveal how collections, curatorial expertise and in-

ternational programming strategies can generate museum "soft power."

My poster presentation at the conference will present the first stage of the mapping system development. The first stage is focused on mapping ACMI collections and calculating collection appeal power index to different countries. The demo version of the application is available here: http://victoriasoftware.com/demo.html Integrating content analysis of the multicultural and multilingual collections with cultural analytics data, representing different countries around the globe, the online map shows where ACMI can have a stronger appeal with its offerings and holdings.

ACMI has unbelievably rich and diverse collections. It has 200 thousand original items and more than 40 thousand titles. The majority of the collections are accessible online through the online collection search system which currently allows to search through 41.713 titles. 70% of films are produced outside Australia not only in the US and UK but also in France, Germany, Japan, China or New Zealand. There are movies in around 50 different languages which are spoken in more than 230 countries around the world. For example, extensive collections in English that originate from Australia, New Zealand, Canada, the USA, the UK and other countries provide a potential content access to people from a hundred countries, while films in French could reach people in 38 countries.

To understand the potential appeal power of the ACMI collection to people from different countries, I considered two main types of criteria: collections characteristics and social demographic statistics. Collections criteria indicate how many items where produced in a certain country and how many films in the collection are in the language/s spoken in this country. Social demographic criteria bring to light such nuances as immigration statistics in Melbourne, annual tourism rate, ancestry data and internet penetration rate which affects the collection access and discoverability online. I calculated the collection appeal power index as a weight some of all subsidence's across two key criteria. The demo app available online (http://victoriasoftware.com/demo.html) demonstrates the Appeal Power Index that ranges from 0 to 1 and is visualized by the intensity of the blue color applied to different countries. When you click different countries, the app indicates how many movies from the ACMI collections were produced in this country, how many movies in the collections are in the spoken languages of this country as well as highlights secondary factors like tourism, ancestry and immigration from this country which increases the probability of the collection exposure and visibility among people of this geographic area.

References

Abrams, J. and Hall, P. (2008). *Else/Where: Mapping New Cartographies of Networks and Territories*. Minnesota: University of Minnesota Press.

Australian Centre for the Moving Image (ACMI). 2016. *Annual Report 2015-16*. http://bit.ly/2vEEfNN

Bodenhamer, D., Corrigan, J. and Harris, T. (2010). *The Spatial Humanities: GIS and the Future of Humanities Scholarship*. Bloomington: Indiana University.

Gibson, C., Brennan-Horley, C. and Warren, A. (2010). Geographic Information Technologies for cultural research: cultural mapping and the prospects of colliding epistemologies. *Cultural Trends* 19 (4): 325–348.

Jacobsen, J. (2016). *Measuring Museum Impact and Performance*. Rowman & Littlefield.

La Porte, T. (2012). The Legitimacy and Effectiveness of Non State Actors and the Public Diplomacy Concept. In *Public Diplomacy Theory and Conceptual Issues*, ed. International Studies Association, ISA Annual Convention.

Lord, G. D. and Blankenberg, N. (2015). *Museums, Cities and Soft Power*. Rowman & Littlefield Publishers. AAM Press.

Nye, J. (2004). Soft Power: T*he Means to Success in World Politics*. New York: Public Affairs.

Sylvester, C. (2009). *Art/Museums: International Relations Where We Least Expect It*. London Paradigm Publishers.

Towse, R. and Handka, C. (2013). *Handbook on the Digital Creative Economy*. Routledge.

Vivant, E. (2011). Who brands whom? *Town Planning Review 82* (1): 99-115.

Werner, P. (2005). *Museum, Inc: Inside the Global Art World*. Chicago: Prickly Paradigm Press

Brecht Beats Shakespeare! A Card-Game Intervention Revolving Around the Network Analysis of European Drama

Angelika Hechtl
angelika.hechtl@gmail.com
Vienna University of Economics and Business, Austria

Frank Fischer
ffischer@hse.ru
Higher School of Economics, Russian Federation

Anika Schultz
anika.schultz@hu-berlin.de
Humboldt University of Berlin, Germany

Christopher Kittel
contact@christopherkittel.eu
University of Graz, Austria

Elisa Beshero-Bondar
ebbondar@gmail.com
University of Pittsburgh, United States of America

Steffen Martus
steffen.martus@hu-berlin.de
Humboldt University of Berlin, Germany

Peer Trilcke
trilcke@uni-potsdam.de
University of Potsdam, Germany

Jana Wolf
jana_a_wolf@hotmail.com
University of Potsdam, Germany

Ingo Börner
Ingoboerner86@gmail.com
University of Vienna, Austria

Daniil Skorinkin
dskorinkin@hse.ru
Higher School of Economics, Russian Federation

Tatiana Orlova
taorkon.tootta@gmail.com
Higher School of Economics, Russian Federation

Carsten Milling
cmil@hashtable.de
Higher School of Economics, Russian Federation

Christine Ivanovic
christine.ivanovic@univie.ac.at
University of Vienna, Austria

This poster offers a playful introduction to network analysis as a means to study and compare dramatic texts. Its more serious purpose is a didactic intervention in the now well-established methods of literary network analysis, which are not always applied with sufficient reflection. The calculation of complex network metrics is often not followed by a leap to meaningful interpretation. What does it really mean, for example, that the average path length of the social network extracted from Shakespeare's *Hamlet* is 1.69 and the density of the same network is 0.34? However, when we look at these values in relation to the corresponding values of other dramatic texts, such network statistics become much more meaningful.

In order to cultivate comparative sensitivity in the context of literary network analysis, we build on a gamification approach. Unlike other experiments in this direction – such as the Android and web app *Play(s)* presented at the DHd2016 (Göbel/Meiners 2016), which encouraged the playful correction and enrichment of literary TEI corpora – we produce a true card game that invites players to explore network-analysis data in a new way.

The poster format is applied in two ways: On the one hand, the poster is a data visualisation based on a minimal canon of European drama. On the other hand, it is a card game that playfully acquaints audiences with the meaning of basic network metrics. This approach is not new in the arts and humanities and reaches back to card games like *Plattenbauten. Berliner Betonerzeugnisse* (Mangold et al. 2001), where technical data of different types of prefabricated concrete buildings had to be compared (cf. Richter 2006).

Our drama card game serves to instruct players in literary history, quantitative approaches, and network theory, based on a collection of 32 dramas ranging from the ancient Greeks to the modern age. Instead of a lexicon-like description of such a collection, the descriptive instrument here consists of visual and quantitative values that produce comparability – a type of card game known to English speakers as *Top Trumps* – see https://en.wikipedia.org/wiki/Top_Trumps –, or as *Supertrumpf* in the German context.

Each card presents a visualisation of a social network extracted from one of the 32 plays (very much along the lines of Fischer et al. 2016 and Fischer et al. 2018). Additional information on the cards consists of metadata (author, title, subtitle, year of publication/premiere) and static and dynamic network data (network size, network density, clustering coefficient, average path length, maximum degree incl. the name of the corresponding character, number of scenes). The front card contains an introduction to the project and its background as well as short definitions of network-theoretical terminology.

The poster is generated with the all-in-one drama analysis script *dramavis*, which has received a corresponding function in the new version 0.4 (Kittel/Fischer 2017). The collection of 32 plays used for the conference poster is in no way meant to be definitive or canonical, but is intended to present a diverse collection of plays from the history of European drama that feature comparably interesting social network data. Our collection ranges from antiquity (Aeschylus, Euripides, Sophokles, Aristophanes) to modern times (Marlowe, Shakespeare, Ben Jonson, Calderón de la Barca, Racine, Molière, Aphra Behn, Goldoni, Goethe, Mitford, Victor Hugo, Pushkin, Gogol, Grabbe, Ibsen, Strindberg, Schnitzler, Chekhov, Lasker-Schüler, Shaw, Pirandello, García Lorca, Brecht, and others).

The *dramavis* tool can be fed with a customisable canon file to create your own deck of cards.

References

Fischer, F., Göbel, M., Kampkaspar, D., Kittel, C., Trilcke, P. (2016): *Distant-Reading Showcase. 200 Years of Literary Network Data at a Glance. Proceedings of DHd2016*, Leipzig. DOI: https://dx.doi.org/10.6084/m9.figshare.3101203.v1

Fischer, F., Kittel, C., Milling, C., Schultz, A., Trilcke, P., Wolf, J. (2018): Dramenquartett. Eine didaktische Intervention. *Proceedings of DHd2018*, Cologne. DOI: https://doi.org/10.6084/m9.figshare.5926363.v1

Göbel, M., Meiners, H.-L. (2016): Play(s): Crowdbasierte Anreicherung eines literarischen Volltext-Korpus. *Proceedings of DHd2016*, Leipzig.

Kittel, C., Fischer, F. (2017): dramavis v0.4. On Github, 2017. *Repo*: https://github.com/lehkost/dramavis
Mangold, C. et al. (2001): *Plattenbauten*. Berliner Betonerzeugnisse. Ein Quartettspiel. Berlin.
Richter, P. (2006): *Der Plattenbau als Krisengebiet. Die architektonische und politische Transformation industriell errichteter Wohngebäude aus der DDR am Beispiel der Stadt Leinefelde*. Hamburg, Univ., Diss. URL: http://ediss.sub.uni-hamburg.de/volltexte/2006/3041/

Visualizando una Aproximación Narratológica sobre la Producción y Utilización de los Recursos Online de Museos de Arte.

María Isabel Hidalgo Urbaneja
m.hidalgo-urbaneja.1@research.gla.ac.uk
University of Glasgow, United Kingdom

Publicaciones y exposiciones online, así como otros recursos interactivos, se encuentran entre los recursos online más utilizados por museos de arte en todo el mundo para transmitir historias vinculadas a obras de arte y colecciones. Los formatos tradicionalmente utilizados por museos para contar la historia del arte están siendo reconceptualizados a través de las cualidades y funcionalidades que nos ofrece el medio digital. El proceso experimental que aquí se expone nace con el objetivo de visualizar las particularidades que definen las narrativas generadas en los recursos online de museos de arte. Una selección de seis recursos online representativos de las tipologías más comunes producidos por museos de los Estados Unidos, España y Reino Unido* han sido la base para, por un lado, recabar datos sobre la perspectiva de los productores, y por otro, la de los usuarios especializados—una audiencia de perfil académico/investigador en el área de la historia del arte. Los datos se obtuvieron a través de dos métodos: entrevistas con los productores involucrados en la creación de los recursos online seleccionados, y en el caso de los usuarios, a través del protocolo conocido como "pensamiento en voz alta" (thinking aloud protocol) que ayuda a capturar información relevante a los procesos de navegación de los recursos online. Ambos procedimientos fueron grabados y transcritos para un facilitar el análisis posterior. Estos datos que se codificaron y analizaron desde una perspectiva narratológica permitiendo la observación de elementos configurantes de las narrativas: autoría, recepción como lectoespectador, estructuración, espacialidad, cronología.

Aunque en la investigación doctoral que da origen a los datos utilizados en esta propuesta se siguió una metodología cualitativa de corte más tradicional, en este póster se expone una aproximación experimental basada en la visualización de los códigos extraídos de las transcripciones. La visualización ofrece una visión complementaria al análisis inicial de los datos, orientado a presentar resultados de forma discursiva. De acuerdo con esta premisa, el póster compararía las posibilidades de análisis y presentación de las visualizaciones con el formato discursivo. Un análisis visual de los códigos revela aspectos cuantitativos de los datos, así como las conexiones entre los códigos de forma más explícita. En un cierto sentido, se presenta un resumen visual o vista general. La visualización de datos puede ayudar en la identificación de aspectos que habían sido obviados tras el empleo de la metodología más tradicional, y potencialmente, puede ofrecer nuevas conclusiones en la investigación.

La modalidad de visualización que se emplea en esta propuesta ha sido diseñada partiendo de diferentes metodologías y herramientas de visualización de datos. En primer lugar, toma como punto de partida en el uso de diagramas como herramienta de análisis narratológico (Ryan, 2007). Aunque el diseño de la metodología usada para visualizar datos en este póster emplea específicamente el procedimiento conocido como "map analysis" (Carley, 1993), éste permite la comparación de textos en base a los códigos extraídos y las relaciones entre ellos. Por otro lado, el trabajo de Luther (2017) propone un modelo y herramienta de visualización centrado en la representación de aspectos cualitativos y cuantitativos, éste fue desarrollado con el objetivo de estudiar aspectos de temática socio-histórico artística. No obstante, como herramientas se han elegido Gephi y d3.js ya que permiten representar la frecuencia de los códigos e interrelaciones. Las visualizaciones de este póster representan por separado los datos tanto de productores como de usuarios especializados de los recursos online, permitiendo comparar los seis recursos online. Las visualizaciones permiten estudiar las diferencias y similitudes existentes entre las perspectivas de los creadores, desde un punto de vista referente a la autoría, y las perspectivas de la audiencia especializada, como lectoespectadores de los recursos online. Conclusiones derivadas del proceso de visualización de datos serán argumentadas en el póster. Las visualizaciones se conciben como generadoras de discusiones además de ser una representación de la investigación llevada a cabo, éstas podrán ser consultadas en http://m-hidalgo.com.

Este trabajo es también resultado del proyecto de I+D: „HAR2014-51915-P. Catálogos artísticos: Gnoseología, epistemologías y redes de conocimiento. Análisis crítico y computacional".

*Los estudios de son recursos digitales de las siguientes instituciones: Museo Nacional del Prado, Museo Centro de Arte Contemporáneo Reina Sofía, National Gallery, Londres, National Gallery of Art, Washington DC, Metropolitan Museum of Art y MoMA.

References

Carley, K. M. (1993). Coding choices for textual analysis: A comparison of content analysis and map analysis. *Sociological Methodology*, 23, 75–126. http://www.casos.cs.cmu.edu/publications/papers/carley_1993_codingchoices.PDF

Drucker, J., (2014). *Graphesis. Visual Forms of Knowledge Production*. Cambridge, MA.: Harvard University Press.

Flick, U., (2010). *An Introduction to Qualitative Research*. London: Sage Publications.

Gee, K. (2001). The ergonomics of hypertext narrative: usability testing as a tool for evaluation and redesign. *ACM J. Comput. Doc.* 25, 1 (February 2001), 3-16. DOI=http://dx.doi.org/10.1145/383948.383950

Luther, A. (2017). The Entity Mapper: A Data Visualization Tool for Qualitative Research Methods. *Leonardo*, Volume 50, Issue 3, June 2017. Cambridge: MIT Press, p.268-271. Doi: 10.1162/LEON_a_01148. Abstract available at: http://www.mitpressjournals.org/doi/abs/10.1162/LEON_a_01148

Mann, L. (2016). Online scholarly catalogues: Data and insights from OSCI. *MW2016: Museums and the Web 2016*. Consulted November 26, 2017. http://mw2016.museumsandtheweb.com/paper/online-scholarly-catalogues-data-and-insights-from-osci/

Ryan, M. (2007). Diagramming narratives. *Semiotica*. 165: 1.4, 11-40.

Warwick, C. (2013). Studying users in digital humanities. Terras, Melissa; Nyhan, Julianne, and Vanhoutte, Edward, eds., *Defining Digital Humanities. A Reader*. London: Routledge https://blogs.ucl.ac.uk/dh-in-practice/chapter-1/

Transatlantic knowledge production and conveyance in community-engaged public history: German History in Documents and Images/ Deutsche Geschichte in Dokumenten und Bildern

Matthew Hiebert
hiebert@ghi-dc.org
German Historical Institute Washington DC, United States of America

Simone Lässig
laessigs@ghi-dc.org
German Historical Institute Washington DC, United States of America

This poster presents the technical redesign of the web resource *German History in Documents and Images/ Deutsche Geschichte in Dokumenten und Bildern* (GHDI) as a transatlantic knowledge production and conveyance model for community-engaged public history. It is a multilingual project led and based at the German Historical Institute Washington (GHI) in partnership with DARIAH-DE, the Max Weber Foundation, and the University of Southern California. It was awarded a three-year development grant from the German Research Foundation/ Deutsche Forschungsgemeinschaft (DFG) in 2017. We display the project's theoretical foundations and aims, the resulting technical design, and report on the proof-of-concept phase and first-year of development.

GHDI was first conceived in 2002 by a group of academic historians who sought to make a large collection of German historical documents openly available online in German and English translation. GHDI would consist of ten chronological volumes to cover German history from 1500 to 2009, each of which includes an introduction and a selection of historical documents, images, and maps, accompanied by interpretations. The site currently contains 1,784 German documents (along with an equal number of English translations), 2,374 images, and 55 maps (for a total of 16,068 pages), with content being expanded in the revamp. The project has developed a large and diverse international community of users, registering approximately 100,000 unique visitors a month.

The reconceptualization and revamp of the GHDI includes the encoding of original and new materials in TEI P5, Dublin Core metadata for all content, a site-wide co-created bibliography, and a scholarly annotation system. The integration of, and project development contributions to, *Scalar*—a robust open-source authoring, editing, and publishing platform with support for RDF content—allows users to navigate content in diverse ways and along various critical historiographical paths, challenging "master narrative" approaches to German history. The Scalar adapters developed by the project will link a number of important German archives to English-speaking scholarly communities for the first time, and the GHDI platform will ultimately allow users to use and "mix" this and other content to produce their own and collaborative scholarly outputs.

Data resources of the project are being described using Dublin Core metadata vocabulary. Sources with annotations or other semantic enhancement adhere to TEI (Text Coding Initiative) P5 using the DTA base format. Linked-open data representations are being be stored in RDF-XML. Using Scalar's built-in API, all content will be made available directly via URL-based requests in RDF-XML. This is also the technical basis for user content "remixing" and user publication facilitation being developed within the GHDI environment. Authority control for personal names and other entities, both in consumption and publication, will be assured through GND and similarly broadly accepted standards. Resources suitable for language analyses tools conform to Component MetaData Infrastructure (CDMI) as prescribed by CLARIN-DE data

centers. Geographic data is being encoded in GeoJSON. All data will be published to prioritize permissiveness of use under Creative Commons licensing.

A Tool to Visualize Data on Scientific Performance in the Czech Republic

Radim Hladík
radim.hladik@fulbrightmail.org
Institute of Philosophy of the Czech Academy of Sciences, Czech Republic; National Institute of Informatics, Japan

The poster introduces a project to develop a visualization application for a unique data source on Czech sciences. Information Register of R&D Results (RIV) is the Czech Republic's inventory of the outputs of basic and applied research since 1992. Although it is potentially an important source of data for analyses of various aspects of the intellectual organization and publication culture in Czech sciences, this particular data source has earned itself a pejorative nickname – "a coffee grinder" – for its central role in purely mechanistic science evaluation in the country.

By employing text-mining technique that are standard in the digital humanities and by getting inspiration from visualization platforms such as *Voyant Tools* (Sinclair and Rockwell 2012), the project aims to contribute to the shift in the Czech narrative of science evaluation from the exclusively bibliometric perspective to a more diverse one. For example, the hope is that the visual display of the plethora of topics that are discussed in the research outputs registered in RIV will implicitly criticize the myopic vision in which all disciplines are leveled to the singular measure of the number of publications. The latter system is not only intellectually dubious, but it has had documented adverse effects on the quality of research results. Crucially, it stimulates institutions as well as individuals to prioritize quantity over quality (Good et al. 2015; Grančay, Vveinhardt, and Šumilo 2017).

The ill-fated usage of the RIV data to mold nationwide fiscal policies for scientific research reminds us that data analytics is not necessarily a neutral enterprise. A proper treatment of the data is a matter that confronts a data analyst with questions on the borderline of ethics. Although it is perfectly feasible in technical terms, we wish to discourage users from attempts to track individuals researchers; instead we offer features that display institutional or disciplinary dimensions of the data (see Figure 1). Furthermore, the web application will provide a module to visualize textual information from the register. Textual strings, such as abstracts and keywords, have been part and parcel of the recorded entries, but have only served thus far as mere search terms. Meanwhile, the utility of textual data has been demonstrated in studies that strive to map the intellectual organization and relationships within and between disciplines (Leydesdroff 1989; Moody 2004).

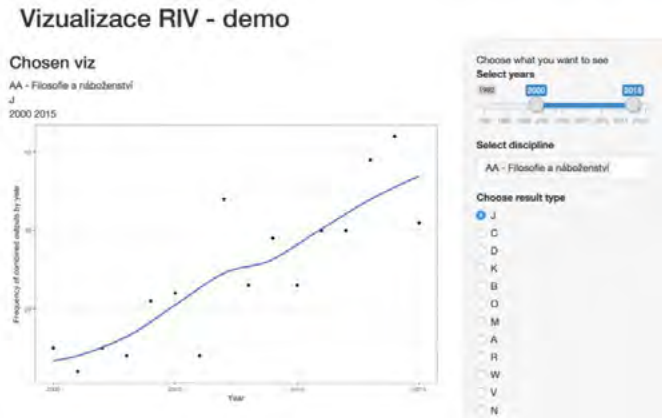


Figure 1. Using RIVVIZ to visualize a trend in the publication frequency of research outputs in the "J" (journal) category of the Information Register of R&D Results for the discipline "Philosophy and Religion" [note: the data are only a sample used in the development version]

The target group of the application are the researchers themselves. Namely, the textual module is intended to serve their needs by providing an overview of the trending topics in research or to identify institutions working on similar problems. The specialist user subgroup is envisaged to come from the fields focusing on social and other studies of science. The accessibility of visualized data and the simplicity of the interface can also attract journalists or other members of the public. The prospective users are also likely to be recruit from among the stakeholders in scientific policy-making and management who may wish to gain quick insights into the quantitatively assessed rates of output per research institutions or funding bodies.

The RIVVIZ application is developed in the R language and deployed on the R Server platform using the standard Shiny library. The data are imported from the publicly available repository of the Czech Research, Development and Innovation Information System. The internal setup is also fairly straightforward, relying predominately on the Tidyverse collection of packages, with ggplot2 library being the primary engine for visualization tasks. The underlying principles of the "grammar of graphics"(Wickham 2009) are particularly suitable for programming a user-oriented environment that allows for a control over a wide range of visualization parameters.

Giving the users more choices should help to make them more engaged with the application, although there is a trade-off between user-friendliness and complexity. Reasonable defaults can partially alleviate this dilemma. The user engagement will be important for the future application development (Galey and Ruecker 2010).

In the case of visualization schemes, locking users in a single – no matter how aesthetically pleasing – perspective is problematic. The apparent self-explanatory style and transparent communication of images may draw attention away from the complex and multifaceted nature of the data by making some of their aspects more easily accessible than others (Drucker 2011).

References

Drucker, J. (2011). Humanities Approaches to Graphical Display. *Digital Humanities Quarterly* (DHQ), 5(1).

Galey A. and Ruecker, S. (2010). How a Prototype Argues. *Literary and Linguistic Computing*, 25 (4): 405-424.

Good, B., Vermeulen, N., Tiefenthaler, B. and Arnold, E. (2015). Counting Quality? The Czech Performance-Based Research Funding System. *Research Evaluation* 24 (2): 91–105.

Grančay, M., Vveinhardt, J. and Šumilo, Ē. (2017). Publish or Perish: How Central and Eastern European Economists Have Dealt with the Ever-Increasing Academic Publishing Requirements 2000–2015. *Scientometrics* 111 (3): 1813– 37.

Leydesdroff, L. (1989). Words and Co-Words as Indicators of Intellectual Organization. *Research Policy* 18 (4): 209–223.

Moody, J. (2004). The Structure of a Social Science Collaboration Network: Disciplinary Cohesion from 1963 to 1999. *American Sociological Review* 69 (2): 213–238.

Sinclair, S., Rockwell, G. and the Voyant Tools Team. (2012). *Voyant Tools* (web application).

Wickham, H. (2009). *Ggplot2: Elegant Graphics for Data Analysis*. Dordrecht: Springer.

Augmenting the University: Using Augmented Reality to Excavate University Spaces

Christian Howard
ch4zs@virginia.edu
University of Virginia, United States of America

Monica Blair
mkb4rf@virginia.edu
University of Virginia, United States of America

Spyros Simotas
ss4ws@virginia.edu
University of Virginia, United States of America

Ankita Chakrabarti
ac4ze@virginia.edu
University of Virginia, United States of America

Torie Clark
vrc7de@virginia.edu
University of Virginia, United States of America

Tanner Greene
tjg6ph@virginia.edu
University of Virginia, United States of America

Project Website: http://reveal.scholarslab.org/

Introduction

Using augmented reality (AR) applications, our project, titled *UVA Reveal: Augmenting the University*, challenges the surface of our perceptions of objects and places. Our project specifically uses the University of Virginia (UVA), a large public state university, as its target. UVA is a southern historic campus with an enrollment of 22,000 students; given its history and recent spotlight in the news, UVA's campus is ripe for the historical inquiry and narrative intervention that our project proposes. In augmenting UVA's campus, we hope to expose the historical, cultural, (inter)national, (trans)sexual, and (dis)ability-related "archeology" of objects, places, and events.

Background of the Project

Augmented reality applications are becoming increasingly prevalent in society (witness Pokémon Go) and in the academy. For instance, a DH project titled *The Whole Story* uses an app that allows users to build AR statues of women and place them in the spatial landscape for others to see. By putting women back in the narrative, the app challenges the unequal statuary landscape and its implication that men are the makers of history. The digital spaces created by AR thus assume an openness and mobility that is lacking in physical space, which may be controlled or limited by socio-economic and political reasons. Nonetheless, these spatial boundaries can seemingly be circumvented in digital spaces,[4] and users can move rapidly across zones that they would be unable to otherwise. *UVA Reveal* is thus designed to explore how real spaces can be experienced through changing, mobile technologies that enable spatial and temporal augmentation.

The objects of our investigation include both buildings and documents at or connected to UVA, especially documents from the special collections library. In particular, we are attempting to renegotiate UVA's narratives about race, gender, and disability. For instance, a prominent mural on our campus depicts troubling scenes, including sexual harassment. We intend to use AR to highlight how women and other minorities are shown in this

4 We recognize that the same can be said about digital spaces, i.e. firewalls, paying services, language barriers, profile/password credentials, profiles set to private, digital literacy, etc. Our project, however, is open-source and freely available to the public.

mural by directing attention to them and challenging the patriarchal gaze.

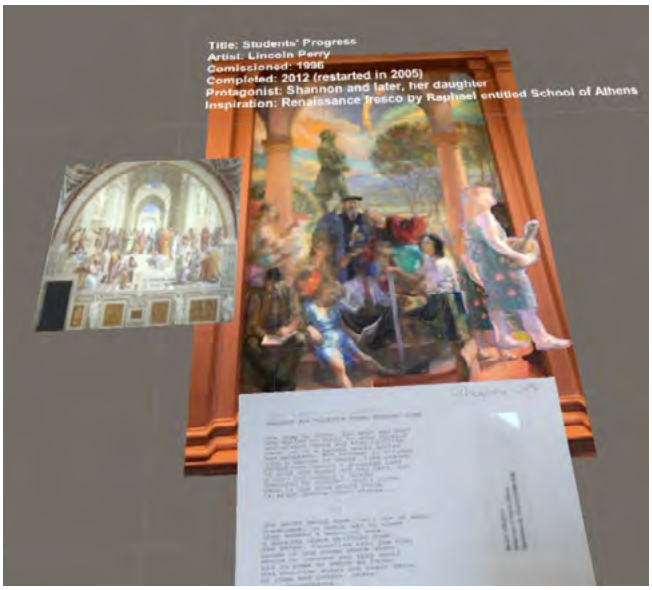


Image 1: Example of the augmentation of a prominent mural on UVA's campus as viewed through the Unity editor.

Theory

The spatial historian Richard White has claimed: "Visualization and spatial history… is a means of doing research: it generates questions that might otherwise go unasked, it reveals historical relations that might otherwise go unnoticed, and it undermines, or substantiates, stories upon which we build our own versions of the past." Similarly, our project neither contests nor reinforces the university's archive; rather, we supplement our archival research with broader research beyond the university's purview. As such, *UVA Reveal* enables viewers to make their own judgments about certain places and objects on UVA's campus by bringing those items to viewers' attention.

Methodology

UVA Reveal will have two primary instantiations: a web-based version and an app. The web version will clearly layout both our research methods and findings. Specifically, as we engaged with Special Collections, we realized that our project could have benefitted from a more directed search experience. To that end, we created a search function using UVA library data. Given a database with a sample of Special Collections holdings, the user may research a topic (narrowly defined for the scope of our project) using multiple keywords that relate to that topic; this cross-search exposes links between thematic data otherwise unavailable. We are using the d3 library to visualize the resulting data. This search function is integrated into our website.

The second version of the project will explicitly draw upon AR technology. In particular, our project uses Unity to layer 3D models on images – including university buildings and physical objects – that will enable the viewer to experience the virtual layering of time upon an object. Unity is easily exportable to Android, iOS, and HoloLens platforms. Our users will thus be able to engage the AR experience through their personal devices.

Our team is committed to open access. Thus, we are using GitHub to manage our content and ensure that our work process is openly accessible.

Conclusion

Through research in Special Collections, we plan to unearth the many historical layers upon which UVA is built. Ultimately, we hope to use AR to allow users to experience these limited-access spaces and objects in new ways that prompt critical reflection on the structure, culture, mission, and history of the university.

References

"Ambient Literature – This Is Your Part of the Story." *Ambient Literature*, UWE Bristol, Bath Spa University, the University of Birmingham, and Calvium Ltd., June 2016. <ambientlit.com/>.

E Silva, Adriana de Souza. "Mobile Narratives: Reading and Writing Urban Space with Location-Based Technologies," *Comparative Textual Media*. Ed. Katherine Hayles and Jessica Pressman. Minneapolis: University of Minnesota Press, 2013. 33-52.

White, Richard. "What Is Spatial History?" *Spatial History Project*, 1 Feb. 2010. <web.stanford.edu/group/spatialhistory/cgi-bin/site/pub.php?id=29>.

"The Whole Story." *The Whole Story*, 2017. <thewholestoryproject.com/>.

An Easy-to-use Data Analysis and Visualization Tool for Studying Chinese Buddhist Literature

Jen-Jou Hung
jenjou.hung@gmail.com
Dharma Drum Institute of Liberal Arts, Taiwan

In the field of Chinese ancient texts digitalization, the digitization of Buddhist scriptures has been regarded as a relatively complete and fruitful collection. The Chinese Buddhist Electronic Text Association (CBETA) has made the Chinese electronic Tripi□aka collection widely available for many years and provided a resourceful platform for the studies on Chinese Buddhist texts. As of the 2016

version(CBETA 2016), more than 210 million Chinese characters are freely and publicly available in digital form through the efforts of the CBETA.

The digital age that we have now entered has provided us with tools which help us in conducting surveys of Buddhist texts at a scale larger than before, and The text analysis techniques has been proofed as useful in many Buddhist literature research studies (Hung 2010, Bingenheimer 2017). It is with this goal in mind, our team made use of these new tools of the digital age to create a digital research environment which tailored to the needs of research in the field of Buddhist studies (and beyond). In order to achieve these goals, we established the CBETA Research Platform (http://cbeta-rp.dila.edu.tw/?lang=en). This research platform provides high-quality digital content from the CBETA corpus, combines with relevant reference materials based on the latest findings. Additionally, we implemented tools for quantitative analysis with the ultimate goal of creating a digital research platform which will assist scholars in their study of Chinese Buddhist texts or the underlying Indian origins.

CBETA Research Platform

The system architecture of CBETA Research Platform is shown in fig. 1. We have integrated the full text of CBETA corpus with Tripitaka catalogue, bibliographic databases, Buddhist dictionaries and authority databases of person and places to form the backend database of CBETA Research Platform. We then create tools to assist researchers in reading, searching and analyzing Buddhist literature.

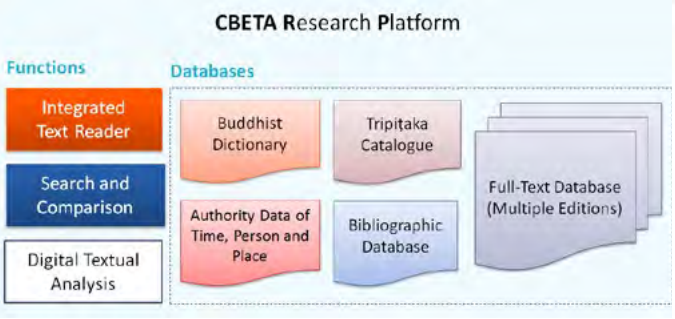


Fig 1. the system architecture of CBETA Research Platform

Concordance Search and Analysis

Concordance Search and Analysis is the first quantitative analysis tool implemented in CBETA Research platform[5]. It is a tool for gaining deeper insight into the search results from CBETA corpus. It allows user to aggregate search results from different dimensions (by Text Category, by Date and Dynasty, by Authors and Translators), and compare the results of multiple search terms.

Start a New Analysis

Concordance Search and Analysis will first require user to enter the keywords they want to compare and specify the search scope.

Fig 2. the start page of Concordance Search and Analysis system

Data

The system retrieves the complete search results and stores the search results for different keywords in the system cache at the same time. On data page, users can examine the complete list of the matches, and delete unwanted records from the result set.

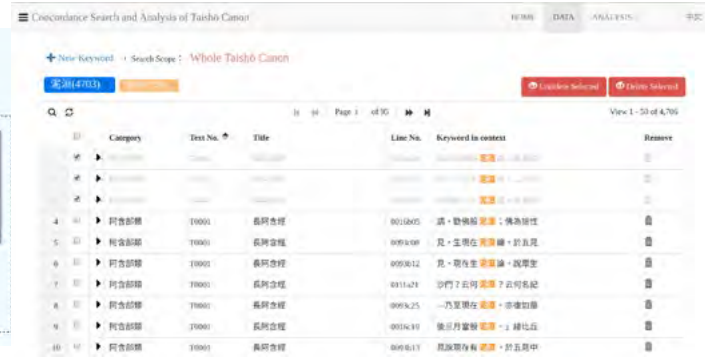


Fig 3. the data page of Concordance Search and Analysis system

Analysis

The System allows user to aggregate search results from different dimensions: by Text Category, by Date and Dynasty, by Authors and Translators, and compare the result of multiple search terms. Fig 4, 5 and 6, show the analysis results of two Synonyms: 泥洹(ní huán)and 涅槃(niè pán) form above-mentioned three different dimensions

5 Besides to Concordance Search and Analysis, CBETA Research Platform has provided an user-friendly reading interface (called as CBETA Online Reader, http://CBETAOnline.dila.edu.tw) for accessing texts and reference materials from backend database.

Fig 4 The statistics keywords in different text categories

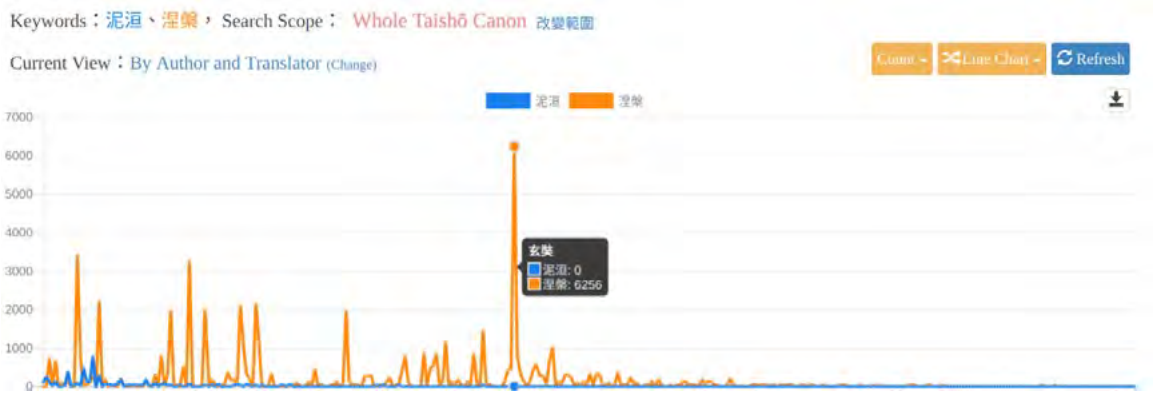


Fig. 5: The statistics of keywords with different translators

Fig. 6 The statistics of keywords in different dynasties

The system offers several statistical range settings. Thus, users are able to observe a wider usage of keyword from large-scale view, and at the same time, to trace a particular phenomenon back to the source text for identification and further research.

Fig. 7. statistics of keywords in different texts from Eastern-Jin Dynasty (C.E. 317 -420)

Fig. 8 statistics of keywords in different fascicles of 長阿含經(Dīrghāgama).

If we click the points represented the fascicle 3 of 長阿含經(Dīrghāgama) in the Fig.8, we will see sentences that actually contain keywords in the text.

Fig 9. sentences that actually contain keywords in fascicle 3 of Dīrghāgama.

In addition, the system also provides the „prefix and suffix analysis“ feature, allowing users to quickly access the statistics of a character before and after the keyword.

Fig 10. prefix and suffix analysis of keywords

In addition, in the spatial analysis function, we use a GIS system to display the location of the text containing the keywords, which allows users to compare the use of keywords geographically.

Fig 11. the spatial analysis of keywords

References

Bingenheimer, M.,Hung, J., and Hsieh, C. (2017) Stylometric Analysis of Chinese Buddhist texts – Do different Chinese translations of the Gandavyūha reflect stylistic features that are typical for their age? *Journal of the Japanese Association for Digital Humanities*, 2(1): 1-30

CBETA. (2016) *CBETA Chinese Electronic Tripitaka Collection*, Available at: http://www.cbeta.org/cbreader/help/index_e.htm (Accessed: 11 July 2017)

Hung, J., Bingenheimer, M., and Wiles, S. (2010) Quantitative Evidence for a Hypothesis regarding the Attribution of early Buddhist Translations *Literary and Linguistic Computing,* 25(1):119-134

'This, reader, is no fiction': Examining the Rhetorical Uses of Direct Address Across the Nineteenth- and Twentieth-Century Novel

Gabrielle Kirilloff
gkirilloff@gmail.com
University of Nebraska-Lincoln, United States of America

Though directly addressing the reader in fiction is often associated with cloying sentimentality, many different forms of direct address are employed across nineteenth- and twentieth-century novels. Upton Sinclair's use of address in *The Jungle* engulfs the reader in a tactile, fictional world, "your knife is slippery, and you are toiling like mad, when somebody happens to speak to you, or you strike a bone" (12). While Harriet Beecher Stowe's *Uncle Tom's Cabin* uses address to implicate the reader in systems of oppression, "And now, men and women of America, is this [slavery] a thing to be trifled with, apologized for, and passed over in silence?" (578). What is fascinating about address, is not only that it can be put to such a variety of purposes, but that these purposes are often antithetical, and have drastically different effects on real readers.

This project seeks to answer questions about the historical usage of address by employing computational methods to detect and extract instances of address from a corpus of 2,000 nineteenth- and early twentieth-century novels.[1] I examine how the frequency of address changes over time and among different groups of authors (such as female authors and African-American authors). In order to detect address I utilize a pattern matching approach that uses regular expressions to match sentences outside of dialogue that contain certain keywords, such as "reader," "you," and "this story." To remove dialogue from the corpus, I developed a pattern matching approach that eliminates quotations. This method accounts for various typographical inconsistencies, including missing quotation marks, embedded quotations, and quotations that extend across paragraphs. In order to learn more about the different types of address that authors have employed, I then used the Stanford Dependency Parser on the sentences extracted from each novel. The Parser is a tool that provides a representation of grammatical relations

1 The corpus is 70% male and 30% female authored; 70% American and 30% British. The texts come from freely available sources. The texts were written between 1800 and 1923.

between words in a sentence. This allowed me to examine the adjectives used to describe the reader or the verbs the reader performs in moments of address. In addition, I performed sentiment analysis on the sentences extracted from each novel using the Syuzhet package in R in order to track the emotional valence of address.

The results from the study indicate the prevalence of address across literary periods. Notably, the mean number of sentences containing address in each novel remains steady over time. Of the 2,000 novels examined, 1,864 contain address, with each novel on average containing 49 instances of address. These results are unexpected given the hypothesis put forward by Garett Stewart in *Dear Reader*: "outlawed in modernism, address went underground [at the beginning of the twentieth-century]" (33). The frequency of address and its prevalence across time push against the critical association (noted by Robyn Warhol in *Gendered Interventions*) of address with mid-nineteenth-century Victorian sentimentality. While the frequency of address remains relatively constant, the form of address radically fluctuates: authors decreasingly use "reader" to address their public in favor of addressing readers as "you."

Address is also correlated with author gender: male authors address their readers more frequently than female authors. Overall, address authored my female writers has a more "positive" emotional valence than address authored by male writers. In addition, male authors are more likely than female authors to use the word "reader" (rather than "you") in moments of address. Although there are notable exceptions, the distribution of "you" and "reader" maintains its correlation with author gender across time and nationality. These results intersect with Robyn Warhol's argument that female authors, more so than male authors, employ the intimate and personal "you" to foster a sense of connection with their readers in order to evoke sympathy for social causes.

References

Stowe, Harriet Beecher. (2009). *Uncle Tom's Cabin or Life Among the Lowly*. Cambridge, MA: Harvard University Press.

Sinclair, Upton. (2005). *The Jungle*. Boston, MA: Bedford/St. Martin's.

Stewart, Garett. (1996). *Dear Reader: The Conscripted Audience in Nineteenth-Century British Fiction.* Baltimore, MD: Johns Hopkins University Press.

Warhol, Robyn. (1989). *Gendered Interventions: Narrative discourse in the Victorian Novel*. New Brunswick, NJ: Rutgers University Press.

Reimagining Elizabeth Palmer Peabody's Lost "Mural Charts"

Alexandra Beall
abeall3@gatech.edu
Georgia Institute of Technology, United States of America

Courtney Allen
callen71@gatech.edu
Georgia Institute of Technology, United States of America

Angela Vujic
av.vujic@gmail.com
MIT, United States of America

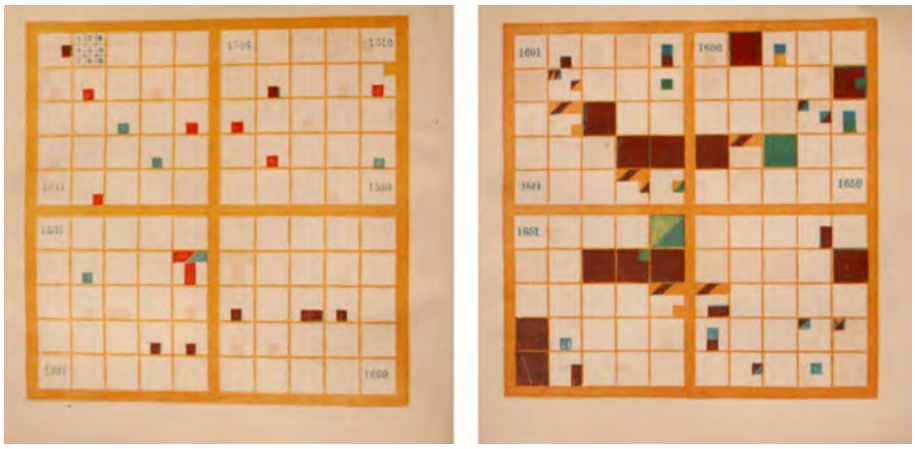
Lauren F. Klein
lauren.klein@lmc.gatech.edu
Georgia Institute of Technology, United States of America

Introduction and Overview

Writing to a friend in 1850, editor and educator Elizabeth Palmer Peabody (1804-1894) complained:

> Just now I am aching from the fatigue of making Charts for the Schools who will take the book… Every school must have a mural chart—& there is but one way of making them (until they can be made by ten thousands) & that is by stencilling… I can do one a day. But I must sell them cheap… To day I worked 15 hours—only sitting down to take my meals—& so I have done all week—so much fatigue stupefies one—but as soon as it is adopted in a few towns I shall be able to hire someone to do this drudgery for me.

In these lines, Peabody provides some of the only extant documentation of her "mural charts"-- large-scale versions of the pedagogical charts that she designed to accompany her U.S. history textbook, *A Chronological History of the United States* (1865). Peabody's textbook promoted data visualization as a pedagogical method. Her visualization scheme involved translating significant historical events into shape and color, and arranging them on a grid (see figures 1). Students could then use the grid as a visual mnemonic, inscribing each century of U.S. history into their memories.



Left: Significant events of the 16th century United States. Right: Significant events of the 17th century United States.

The Mural Chart Project

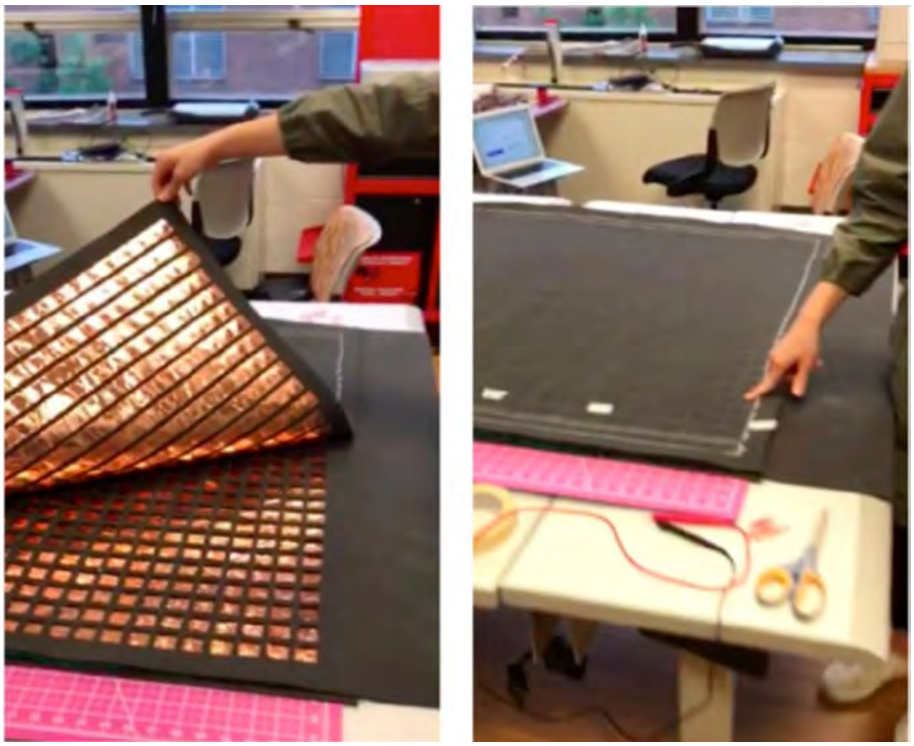
The project team has explored Peabody's visualization scheme in detail (e.g. Klein et al., 2017). But the "mural charts" that she describes in her 1850 letter have not been preserved. Scholarship describes how Peabody would lay the mural charts out on the classroom floor, inviting students to sit around the charts and discuss the colors and shapes that they perceived (Ronda 1999). We were captivated by how, in this particular use, the mural charts seemed to anticipate a form of embodied, experiential learning. We were also taken with the experiential aspects of making the charts-- the "fatigue" and the "drudgery"-- that Peabody describes in her letter. We thus embarked upon a project to recreate Peabody's lost mural charts using physical computing materials, amplifying the embodied and interactive aspects of interpreting the charts that are documented in these archival fragments, and attending to the additional experiential aspects of our own chart-making process. In doing so, we bring together historical fabrication work (e.g. Sayers 2015) with feminist making (e.g. Losh and Wernimont 2014).

Chart Design and Implementation

The reimagined mural chart consists of three layers: a fabric layer that approximates Peabody's original canvas (figure 2, left); a grid of 900 individually-addressable LEDs (figure 2, right); and a soft-button touch interface for toggling each LED off and on (figure 3). The result is an illuminated touch interface that conveys the abstraction of the original grid and the embodied nature of the learning experience, enhanced by contemporary technologies.

Strips of conductive copper tape, arranged in a 30 x 30 matrix and positioned on soft neoprene, are used to register the location of each button press. Two Arudino Megas, daisy-chained together, determine the column and row of the touch. A third Mega, also daisy-chained, takes the location of the button press and illuminates the corresponding LED.

Left: Fabric layer before assembly. Right: LED layer before assembly.



Left: The conductive layers of the touch interface. Right: The assembled touch interface.

Next Steps

Currently, the chart allows the user to touch any square to turn on the corresponding LED. The next steps are to design and implement the interaction that will allow the user to create and input their own events; and to design and implement a color picker, perhaps employing a digital interface. The goal for this phase of the project is to complete a start-to-finish interaction from selecting a historical event, choosing its color and position, and then visualizing it on the mural chart.

References

Klein, L., Foster, C., Hayward, A., Pramer, E., and Negi, S. (2017). The Shape of History: Elizabeth Palmer Peabody's Feminist Visualization Work. *Feminist Media Histories* 3 (3): 149-153.

Ronda, B. (1999). *Elizabeth Palmer Peabody: A Reformer on Her Own Terms*. Cambridge: Harvard University Press.

Sayers, J. (2015). Prototyping the Past. *Visible Language* 49 (3): 156-177.

Wernimont, J. and Losh, E. (2014). Feminist Digital Humanities: Theoretical, Social, and Material Engagements around Making and Breaking Computational Media. https://jwernimont.com/2014/06/02/feminist-digital-humanities-theoretical-social-and-material-engagements-around-making-and-breaking-computational-media/ (accessed 24 April 2018).

TOME: A Topic Modeling Tool for Document Discovery and Exploration

Adam Hayward
adam.hayward@gatech.edu
Georgia Institute of Technology, United States of America

Nikita Bawa
nbawa3@gatech.edu
Georgia Institute of Technology, United States of America

Morgan Orangi
moorangi@gatech.edu
Georgia Institute of Technology, United States of America

Caroline Foster
cfoster2@gatech.edu
Georgia Institute of Technology, United States of America

Lauren F. Klein
lauren.klein@lmc.gatech.edu
Georgia Institute of Technology, United States of America

Introduction and Overview

In the past several years, the utility of topic modeling for the humanities has been clearly established. Scholars can now point to projects that convincingly employ topic modeling to explore the figurative language employed in ekphrastic poetry (Rhody 2012), to trace the "quiet transformations" of literary studies (Goldstone and Underwood

2014), and to distill the epistemic dimensions of novels (Erlin 2017), among others. And yet, broader applications of the technique remain limited by the computational and statistical expertise required to implement a topic model and interpret its results. While there has been some work to develop topic model "browsers" (e.g. Goldstone 2014, Murdock and Allen 2015), these projects are designed to facilitate the exploration of the model itself, rather than to leverage the affordances of topic modeling for humanities scholars. By contrast, our interface was conceived so that non-technical humanities scholars can employ a topic model of their corpus in order to discover the documents most salient to their research (Klein et al. 2015).[1]

Corpus, Model, and Database

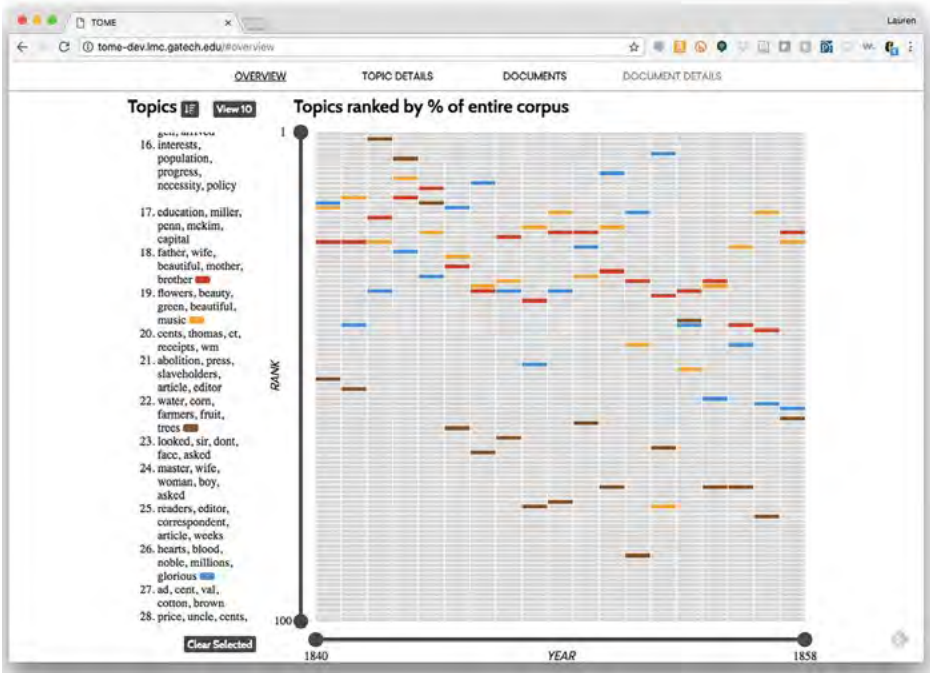
Our corpus consists of nearly 300,000 documents drawn from a collection of nineteenth-century abolitionist newspapers. The documents were scraped from the Accessible Archives website, as per an agreement with Accessible. Additional cleaning of the data, as well as metadata creation, was performed through custom Python scripts.

The topic model of our corpus was created using gensim, the vector space and topic modeling library (Rehurek and Sojka 2010). We employed gensim's wrapper for Latent Dirichlet Allocation (LDA) from MALLET (McCallum 2002). We generated 100 topics after 100 iterations, filtering the 100 most common words. We printed the topics and topical composition of each document to CSV files. We then ingested the data into a MySQL database using Django's ORM framework.[2]

Interface and Sample Interaction

Our interface is the result of a several-month design process during which we considered a variety of user scenarios. Our goal was to scaffold the process of document discovery so that the user could draw new insights as they moved through each section of the interface: Topic Overview, Topic Details, Document Overview, and Document Details.[3]

The user begins with the Topic Overview section (Figure 1), which employs a custom visualization in order to display each of the 100 topics according to its change in rank over time. The user can also filter the topics by keyword or sort according to overall prevalence.



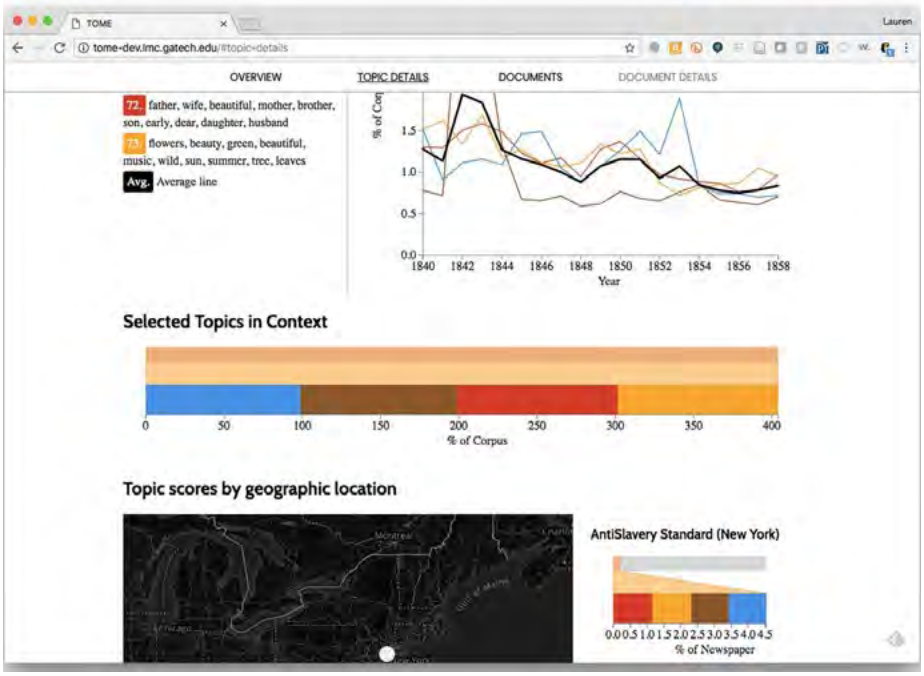
Topic Overview

When the user has selected their topics of interest, they scroll to see details about those topics: change in percentage of the corpus over time; distribution in each newspaper over time; and geographic distribution (Figure 2). These visualizations work together to show which topics were most prevalent at which times; which sources were reporting on which topics at particular times; and where each topic was being reported on. From there, the user can either return to the Topic Overview to further refine the topic set (Gelman 2004), or scroll down to the Document Overview section.

1 The first round of research on TOME was conducted between 2013 and 2015 in collaboration with Jacob Eisenstein, Assistant Professor of Interactive Computing at Georgia Tech, funded by NEH Office of Digital Humanities Startup Grant HD-51705-13. See Klein et al. 2015.

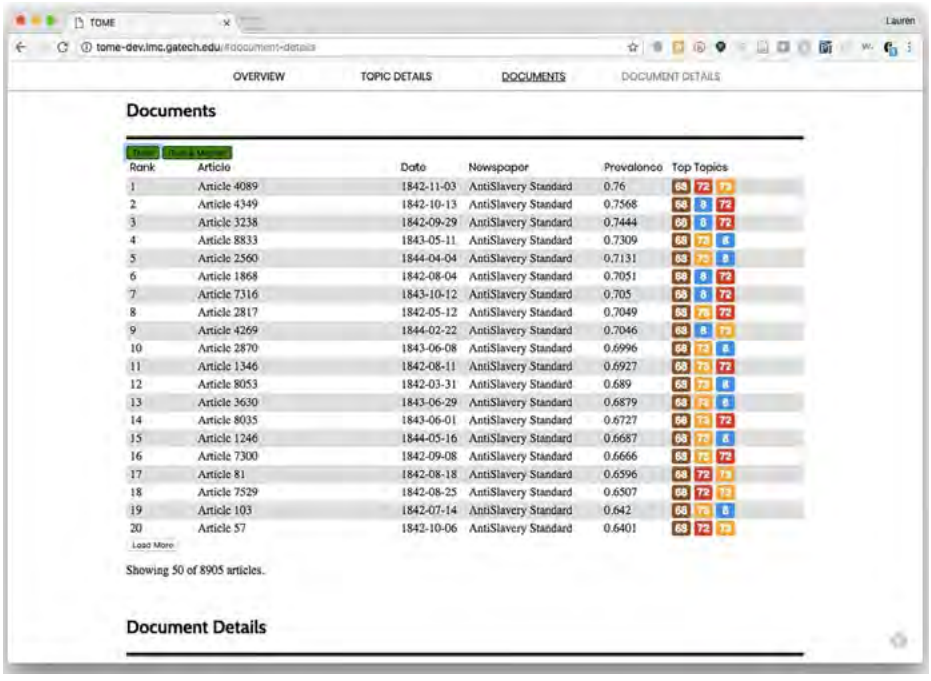
2 The topic model and related processing scripts can be found at: https://github.com/GeorgiaTechDHLab/TOME/.

3 A live version of this interface can be found at: http://tome.lmc.gatech.edu/.



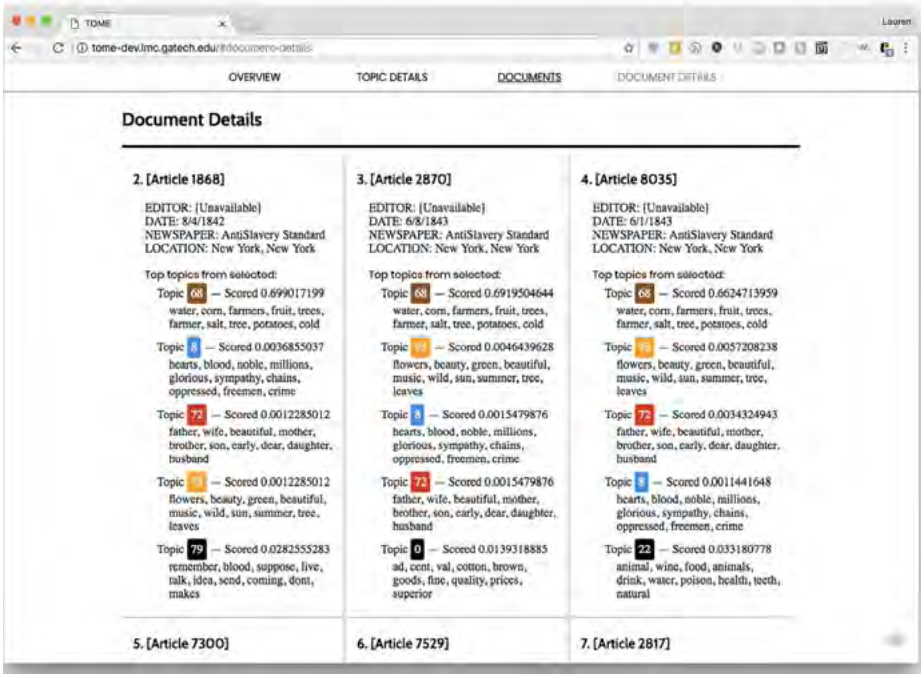
Topic Details

The Document Overview (figure 3) section allows the user to further refine the set of documents they will eventually read. They can toggle between a standard list view of all the documents, ranked in terms of what percentage of the selected topics they contain, and a dust-and-magnets view (Yi et al. 2005).



Document Overview

From there, they move to Document Details (figure 4), which displays the metadata associated with each article in the corpus, ordered according to the percentage of the selected topics they contain. This allows the user to click through to the articles themselves, having narrowed down a set of articles relevant to their research.



Document Details

The interface is implemented using HTML and JavaScript, including D3.js, the JavaScript-based visualization library, and AJAX for client-side data retrieval.

Initial research on TOME was conducted from 2013 to 2015 in collaboration with Jacob Eisenstein, School of Interactive Computing, Georgia Institute of Technology, funded by NEH Office of Digital Humanities Startup Grant HD-51705-13.

References

Erlin, M. (2017). Topic Modeling, Epistemology, and the English and German Novel. *Cultural Analytics*.

Gelman, A. (2004). Exploratory Data Analysis for Complex Models. *Journal of Computational and Graphical Statistics* 13 (4): 755–779.

Goldstone, A., and Underwood T. (2014). The Quiet Transformations of Literary Studies: What Thirteen Thousand Scholars Could Tell Us. *New Literary History* 45 (3): 359-384.

Goldstone, A. (n.d.). DfR Browser. https://agoldst.github.io/dfr-browser/ (accessed 25 April 2018).

Klein, L., Eisenstein, J., and Sun, I. (2015). Exploratory Thematic Analysis for Digitized Archival Collections. *Digital Scholarship in the Humanities* 30 (Supp. 1): i130-i141.

McCallum, A. (2002). MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu (accessed 25 April 2018).

Murdock, J. and Allen, C. (2015). Visualization Techniques for Topic Model Checking. *AAAI Conference on Artificial Intelligence*, Austin, TX, January 2015.

Rehurek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valetta, Malta, May 2010.

Rhody, L. M. (2012). "Topic Modeling and Figurative Language." *Journal of Digital Humanities* 2 (1).

Yi, J.S. (2005). Dust & Magnet: Multivariate Information Visualization Using a Magnet Metaphor. *Information Visualization* 4 (4): 239-256.

Bridging Digital Humanities Internal and Open Source Software Projects through Reusable Building Blocks

Rebecca Sutton Koeser
rebecca.s.koeser@princeton.edu
Center for Digital Humanities,
Princeton University, United States of America

Benjamin W Hicks
bhicks@princeton.edu
Center for Digital Humanities,
Princeton University, United States of America

Software development is often an integral aspect of Digital Humanities projects. By working to generalize and build small modules or utilities targeting specific needs rather than large-scale systems, DH software developers have the capacity to generate tools with greater potential for scholarly reuse, which should enable more rapid development on future projects, and allow developers to focus on innovative work. This poster demonstrates a case study of modular software developed as part of ongoing DH projects.

There is a tendency among some institutions, particularly libraries, to adopt existing large-scale Open Source Software solutions and adapt them for local needs; but as Hector Correa points out, this approach results in skipping the work of thinking carefully about users and local needs (Correa, 2017). If large-scale software solutions developed by coalitions of libraries are problematic (Princeton University Library Systems, 2017) where needs are at least similar, even where content structures or workflows differ, this problem is redoubled for research software, which is much more likely bespoke to a particular problem. As Correa argues, single-purpose software is less complex and easier to understand and manage; and understanding the logic of code is crucial for research that is based on or otherwise makes use of software (Koeser, 2015).

Applying best practices from software development such as modular design can mitigate these problems through an emphasis on delivering working components of software and focusing on simplicity of purpose—a single, well-honed and balanced knife rather than a multi-tool with every imaginable attachment. This approach is consistent with the design philosophy from one of the greatest success stories of modern open-source software, UNIX and its derivatives (Raymond, 2003).

There are certainly possible drawbacks and concerns about this approach. It may require more effort, and perhaps different skills, to create, release, and manage independent software packages or modules. According to Glass' *Facts and Fallacies of Software Engineering*, it is "three times as difficult to build reusable components as single use components" (Glass, 2003: 49). In our case, when new software modules were being developed and extended in tandem with an existing software project, finalizing a new release of that project involved releasing and publishing multiple software modules. There is also a danger of generalizing too soon; another familiar rule of thumb in software is that you have to do something three times before you know how to generalize it properly (Glass, 2003).

As a case study, our poster will present an overview of the software written for two annotation projects that were developed at the same time. "Derrida's Margins" analyzes the work of Jacques Derrida through references in *De la grammatologie* and corresponding annotations in the books he cited. "The Winthrop Family on the Page" examines a community of readers connected through books over time via annotations. This software ecosystem includes two project codebases (Koeser et al., 2018; Koeser and Hicks, 2018a) that make use of four new reusable components (Koeser and Hicks, 2018b; Koeser, 2018b), two of which (Koeser, 2018a; Koeser and Hicks, 2018c) were adapted from the "Readux" codebase (Koeser et al., 2017), which was previous developed at Emory University. In the process, we also used and made minor updates to a related, pre-existing module (Koeser, 2018c).

For each of these tools, a use case emerged in one project which could be generalized for other projects, with potential for broader reuse. As an example, "viapy"—a Python module for searching and providing VIAF data to a web framework—was adapted from previous work, and first existed as code for one of the annotation projects, but it proved generalizable. In fact, it proved easier to extract as a reusable component rather than duplicate; one project team discovered a bug that had previously gone undetected, and creating a reusable package allowed us to correct the problem once for both projects. Likewise, code for storing and displaying annotations from the Readux project was ripe for repackaging as a general module because of its relatively direct purpose despite the different intellectual aims of these projects. However, these codebases also contain similar, potentially reusable functionality that is not yet ready for generalization.

These projects provide a view into the ongoing process of balancing customized solutions to DH projects with generalizing focused portions of functionality. Modular design aimed at 'doing one thing and doing it well' offers the possibility of creating an ecosystem of reusable packages that are widely useful and applicable, and can participate in a larger community of open source and other DH software research.

References

Correa, H. (2017). Build your own software *Hector Correa* http://hectorcorrea.com/blog/build-your-own-software/70 (accessed 28 November 2017).

Glass, R. L. (2003). *Facts and Fallacies of Software Engineering*. Addison-Wesley Professional.

Koeser, R. S. (2015). Trusting Others to 'Do the Math'. *Interdisciplinary Science Reviews*, 40(4): 376–92 doi:10.1080/03080188.2016.1165454. http://dx.doi.org/10.1080/03080188.2016.1165454 (accessed 29 June 2016).

Koeser, R. S. (2018a). *Django-Annotator-Store: Django Application to Act as an Annotator.js 2.x Annotator-Store Backend*. Python Center for Digital Humanities at Princeton https://github.com/Princeton-CDH/django-annotator-store.

Koeser, R. S. (2018b). *Viapy: VIAF via Python*. Python Center for Digital Humanities at Princeton https://github.com/Princeton-CDH/viapy.

Koeser, R. S., Glover, K., Li, Y., Varner, J. and Thomas, A. (2017). *Readux: Django Web Application to Display, Annotate, and Export Digitized Books in a Fedora Commons Repository*. JavaScript Emory Center for Digital Scholarship https://github.com/ecds/readux.

Koeser, R. S. and Hicks, B. W. (2018a). *Winthrop-Django: Django Web Application for the Winthrop Family on the Page Project*. Python Center for Digital Humanities at Princeton https://github.com/Princeton-CDH/winthrop-django.

Koeser, R. S. and Hicks, B. W. (2018b). *Django-Pucas: Django App to Streamline CAS Auth and Populate User Attributes from LDAP*. Python Center for Digital

Humanities at Princeton https://github.com/Princeton-CDH/django-pucas.
Koeser, R. S. and Hicks, B. W. (2018c). *Djiffy: Django Application to Index and Display IIIF Manifests for Books*. Python Center for Digital Humanities at Princeton https://github.com/Princeton-CDH/djiffy.
Koeser, R. S., Hicks, B. W., Glover, K. and Budak, N. (2018). *Derrida-Django: Django Web Application for Derrida's Margins*. Python Center for Digital Humanities at Princeton https://github.com/Princeton-CDH/derrida-django.
Koeser, R. S. (2018). *Piffle: Python Library for Generating and Parsing IIIF Image API URLs*. Python Center for Digital Humanities at Princeton https://github.com/Princeton-CDH/piffle.
Princeton University Library Systems (2017). Valkyrie *Princeton University Library Systems by Pulibrary* https://pulibrary.github.io/2017-07-06-valkyrie (accessed 28 November 2017).
Raymond, E. S. (2003). *Art of Unix Programming, The*. Addison-Wesley Professional http://proquest.safaribooksonline.com/book/operating-systems-and-server-administration/unix/0131429019.

Building Bridges Across Heritage Silos

Kalliopi Kontiza
kalliopi.kontiza@ng-london.org.uk
The National Gallery, United Kingdom

Catherine Jones
catherine.jones@uni.lu
University of Luxembourg, Luxembourg

Joseph Padfield
joseph.padfield@ng-london.org.uk
The National Gallery, United Kingdom

Ioanna Lykourentzou
ioanna.lykourentzou@list.lu
Luxembourg Institute of Science and Technology, Luxembourg

Building Bridges aCROSS CULTural Heritage Silos

This research considers how best to cross the divides that exist between: (1) disparate practices between research fields (2) disparate interpretations of shared cultural heritage by the public and (3) disparate cultural heritage objects.

Figure 1The CrossCult Consortium and Partners

Building bridges across disciplines

Within the field of heritage research there still remain, to this day, many silos between researchers in sciences or the humanities, professionals, practitioners and information technologists. In this poster we consider how best to bridge these gaps between the disciplines. We present, as a demonstrator, an H2020 project named CrossCult (http://www.crosscult.eu). The project brings together inter-disciplinary researchers including: Social scientists, Data and Information scientists, Heritage and Digital Heritage Scientists, Engineering, Humanities and Digital Humanities (Archaeologists and Digital Archaeologists, Linguists, Museum Professionals), Practitioners (Conservators, Curators), and Information Technologies (Backend and Front end and app Developers, Programmers,

Semantic Web specialists, Gamers). We achieve collaboration and discussion through shared common goals and research objectives, and we support dialogue through tools. When possible, we use open source technology to support us for Communication, Programming, Data Structuring/Editing, Visualisation, Conceptual Mapping. We follow standards to be compatible with other people's work, produce reusable research outputs and collaborate with other European projects towards the same goal.

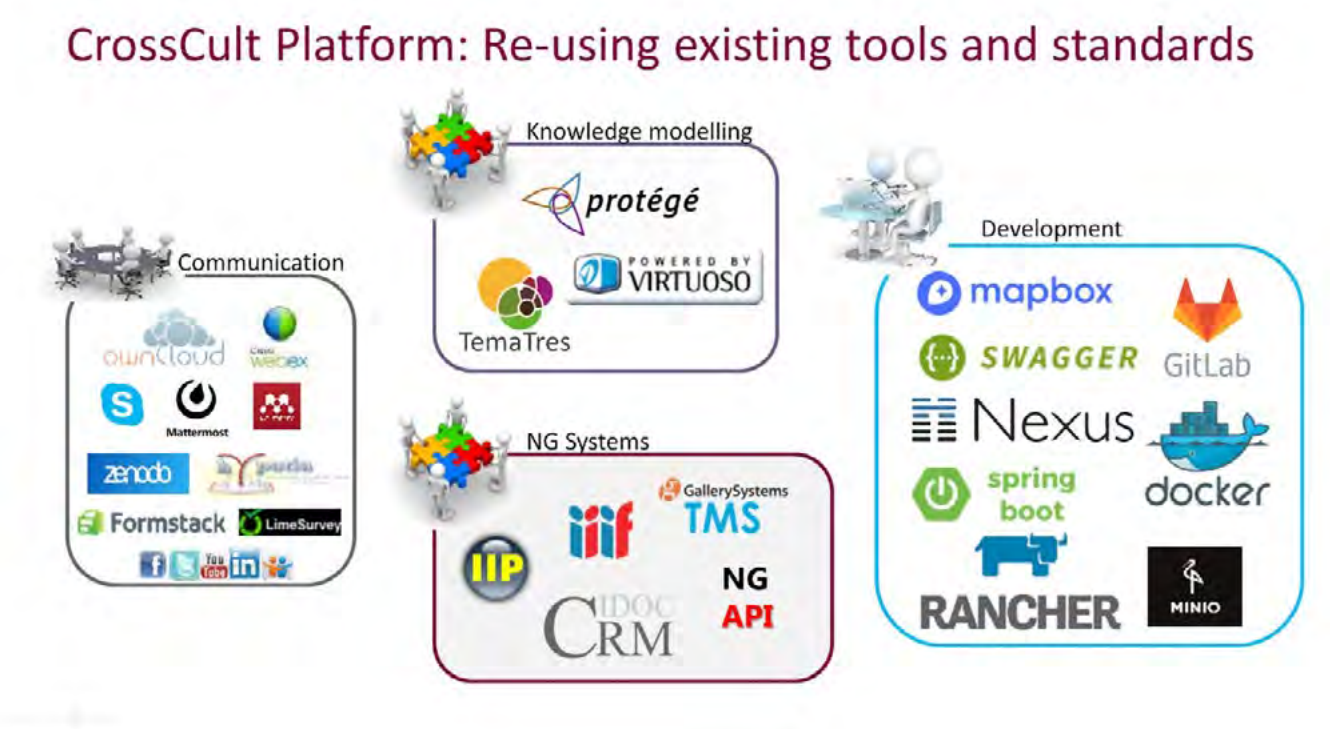


Figure 2 Reusing existing tools and Standards

Building bridges across members of the public

The challenge also extends beyond researchers and continues into the lived experience of our shared Heritage. It raises the challenge of how CROSSCultures can challenge siloed opinions and interpretations of Cultural Heritage (Lykourentzou et al., 2016). At the heart of this project is the desire to build bridges between disciplines to explore innovative practices that can present historic knowledge to non-specialist audiences in an engaging way. European history is an exciting mesh of interrelated facts and events, interpretations and narratives that cross countries and cultures. However, public history is a challenging practice that must be mindful of the audiences, their interests and goals; in this research we are concerned with the museum or the city visitor (Vasilakis et al., 2016).

Building bridges across cultural objects

The final challenge we explore is how to build bridges between disparate objects of our common Heritage. We use heritage objects and historical resources to trigger reflection, individually and collectively, on European history and to showcase the importance to bridge the past and its connection to the present (ERCIM News, 2017).

Using the CROSSCULT project we demonstrate how we are can address the three challenges by developing around four use cases: from large museums to small ones, and from indoors to outdoors. In this presentation, we discuss two of the project's four pilots (Pilot 1 and Pilot 4), which highlight the comparison between the *Indoor and Outdoor Exhibition*; in the first case with the museum/gallery and its paintings and in the second case with the city and its geo-located Places of Interest (POIs). The exhibits (both POIs and paintings) are represented as semantically structured data, linked through our Knowledge Base (Vlachidis et al., 2017). They are our stepping stones to create stories that connect one item to the other, and invite the user (gallery visitor or city traveller) to discover them. The POIs are either discovered outside (in Pilot 4) and can lead to the museum/gallery or vice versa (Pilot 1), eventually bridging the outdoors with the indoors and creating a seamless cultural discovery experience.

Figure 3 CrossCult H2020 project – Overview of the four pilots and their supporting technologies

Pilot 1: Large multi-thematic venue - The National Gallery London- Building narratives through personalisation

We use the gallery's large collection to offer the visitors personalised stories that highlight the connections among people, places and events across European history, through art. Semantic reasoning, recommender systems and path routing optimisation are employed to ensure that each visitor will be navigated through the conceptually linked exhibits that interest them the most, while avoiding congested spaces as much as possible. The experience combines technologies, balancing in a unique way individual visitor needs with museum-wide objectives, can be extended and customised to serve the needs of various other large venues across Europe.

Pilot 4: Multiple cities - City of Valletta in Malta and City of Luxembourg. Building narratives through location based gaming and crowdsourcing.

Pilot 4 takes place outdoors in the two cities to trigger reflection through urban discovery. Focusing on the topic of migration, past for Malta and present for Luxembourg, and using the technologies of location-based services, urban informatics and crowdsourcing, it invites people to walk the two cities, discover and share stories. Visitors and residents engage in comparative reflection that challenges their perception on topics touched by migration such as identity, quality of life, traditions, integration and sense of belonging (Jones et al., 2017).

Acknowledgements:

The work described in this presentation has received funding support from European Union's Horizon 2020 research and innovation programme under grant agreement no 693150.

References

ERCIM News. (2017, October) Reinterpreting European History Through Technology: The CrossCult Project. Retrieved from https://ercim-news.ercim.eu/en111/special/reinterpreting-european-history-through-technology-the-crosscult-project (accessed 02 May 2018)

Jones, C. E., Liapis, A., Lykourentzou, I., Guido, D. (2017). Board Game Prototyping to Co-Design a Better Location-Based Digital Game. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems*. ACM Press, New York, USA, pp. 1055-64. Available from: https://doi.org/10.1145/3027063.3053348

Lykourentzou, I., Naudet Y., Vandenabeele, L. (2016). Reflecting on European History with the Help of Technology: The CrossCult Project. *Proceedings of the Eurographics Workshop on Graphics and Cultural Heritage*. The Eurographics Association, pp. 67-70. Available from: https://doi.org/10.2312/gch.20161384

Vassilakis, C., Antoniou, A., Lepouras, G., Wallace, M., Lykourentzou, I., Naudet, Y., 2016. Interconnecting Objects, Visitors, Sites and (Hi)Stories Across Cultural and Historical Concepts: The CrossCult Project, in: Ioannides, M., Fink, E., Moropoulou, A., Hagedorn-Saupe, M., Fresa, A., Liestøl, G., Rajcic, V., Grussenmeyer, P. (eds.), *Digital Heritage. Progress in Cultural Heritage: Documentation, Preservation, and Protection*. Springer International Publishing, Cham, pp. 501–10. Available from: https://doi.org/10.1007/978-3-319-48496-9_39

Vlachidis, A., Bikakis, A., Kyriaki-Manessi, D., Triantafyllou, I., Padfield, J., Kontiza, K. (2017). Semantic Representation and Enrichment of Cultural Heritage Information for Fostering Reinterpretation and Reflection on the European History. *Paper presented to the final Conference of the Marie Skodowska-Curie Initial Training Network for Digital Cultural Heritage, ITN-DCH 2017*. Olimje, Slovenia, 23–25 May.

Voces y Caras: Hispanic Communities of North Florida

Constanza M. López Baquero
constanza.lopez@unf.edu
University of North Florida, United States of America

Voces y Caras: Hispanic Communities of North Florida is an ongoing project that explores the power of digital *testimonio* (Benmayor, 2012) to make visible hidden communities and enable processes of self-discovery by students of Latinx origin in the U.S. The project engages heritage speakers of Spanish in the process of developing questions and recording interviews with members of the Hispanic/Latinx community in North Florida, a population

that has been, according to many, deliberately made invisible.

Since the inception of the project in 2012, 109 interviews have been conducted, recorded, transcribed and archived. The project serves at least four purposes: (1) It recognizes immigrants as an indispensable part of our society in a political environment increasingly hostile to them, (2) it puts students who are heritage speakers of Spanish in contact with their cultural and historical backgrounds, (3) it gives these students the opportunity to recognize themselves in the stories of others, and (4) it serves as a pedagogical tool that creates communities in and outside of the classroom.

Digital *testimonio* provides an important tool for teaching bicultural students who are searching for their own identities, particularly those who live in an area, like North Florida, where they feel pressure to assimilate or avoid the stereotypes that surround being Latinx in the United States. In many cases, these students are largely disconnected from their own histories, as the Hispanic roots of much of the United States, as well as the history of Latin America, are barely present, if represented at all, in mainstream curriculum. As the Latinx community in the U.S. gains visibility, in part through the negative ramifications of the current political climate, these students are increasingly interested in understanding how they fit into a larger Latinx identity, as well as in vindicating the misperceptions or distortions of Latinx people that they witness in the media.

Since our students live in a large geographical area without a center for immigrants, or a specifically Latinx neighborhood like you would find in Orlando or Miami, many feel lost because they are not fully accepted into the mainstream culture. Furthermore, Latinx make up a small percentage of the university's population and this furthers their feeling of alienation. When they come to my class, they learn about the value of community and history. Voces y Caras is a collection of stories that are testimonial and as Rina Benmayor has stated, "*Testimonio*, thus, expresses the central values of situated knowledge production, embodied theorizing, and community engagement, and thus can be considered a signature pedagogy," which can be "grounded in liberatory values and methods." By learning about other Latinx and what they are doing to influence our city, students discover their own stories. The sacrifices and traumas of other immigrants help them shape their own identities and claim their rights to belong to the U.S., and also to the culture where they, or their parents came from. Benmayor highlights the benefit of this type of projects because it "engages students first hand in reproducing the processes of (1) situated knowledge production, (2) embodied theorizing, and (3) collective practice that are foundational to the field. These processes constitute core epistemologies for Latin[x] Studies, ones that we hope all of our students learn to perform in their lives as well as in their professional futures" (2012: 509).

As I ponder upon my project, I believe that its value resides largely with the opportunities it offers for engagement with local communities. As Will Fenton argued in a recent opinion piece in the *Chronicle of Higher Education*, such use of scholarship to connect with the public is sorely lacking in the Digital Humanities today. I believe, furthermore, that this project demonstrates how digital approaches can be deployed in ways that are truly transformational for students from a variety of disciplinary backgrounds.

There is an organic connection between oral history projects and digital humanities. Listening to the stories of others make us more empathetic. These stories arouse feelings of love and compassion because we can recognize our stories in others. In this line, Voces y Caras highlights the achievements of the community. This is particularly relevant in our present political environment where immigrants have been perceived as a problem rather than what they are; an indispensable part of our society that contributes greatly to its growth. The recordings, excerpts of the interviews, and pictures of the interviewees are available online at vocesycaras.weebly.com

References

Benmayor R. (2012). Digital Testimonio as a Signature Pedagogy for Latin@ Studies. *Equity and Excellence in Education.* 45, 507-524.

Benmayor, R. (2008). Digital Storytelling as a Signature Pedagogy for the New Humanities. *Arts and Humanities in Higher Education.* 7, 188-204.

Fenton, W. (2018). Literary scholars should use digital humanities to reach the oft-ignored 'public' (opinion). *Technology and Learning Blog: Inside Higher Ed.* 2018-01.

Empatía Digital: en los pixeles del otro

Carolina Laverde
ca.la1412@gmail.com
Biblioteca Nacional de Colombia, Colombia

Vivimos en la sociedad de la imagen y la información (Manuel Castells, 1996) que se caracteriza por la hiperproducción de conocimiento. Esto representa un verdadero desafío estructural a la hora de formular proyectos relevantes en los que el desarrollo y la investigación no sean los únicos enfoques de un humanista digital: se requiere aplicar la empatía digital como puente que equilibre el desarrollo de productos digitales.

"La empatía digital es un proceso en el cual una persona puede analizar > reflexionar > proyectar > predecir > sentir mediante la comunicación con lo digital" (Friesem, 2105). La empatía es un proceso subvalorado como herramienta en las primeras etapas de la creación

de un proyecto, cuando realmente la empatía aplicada a los contextos digitales es crucial para poder formular y formar productos responsables, sostenibles y cohesivos desde un contexto de creación multidisciplinar.

Asimismo, para desarrollar proyectos en Humanidades Digitales resulta necesario estructurarlos a partir de tres preguntas: ¿qué se quiere generar?, ¿cómo se quiere construir? y ¿cómo se va a presentar?, y de esta manera encontrar **insights** que generen empatía con el usuario, que den cuenta de sus motivaciones y gustos para lograr proyectar un tono de comunicación, línea de pensamiento e interacción, entendiendo desde un plano mucho más profundo las necesidades de del usuario para poder crear un resultado y producto más efectivo (McDonag y Tomas J, 2010) al establecer una conexión emocional que se convierte en una oportunidad creativa.

> Es la habilidad cognitiva y emocional de ser reflexivo y socialmente responsable mientras se utilizan estratégicamente medios digitales (Friesem, 2015).

En otras palabras, enviar el mensaje en el formato adecuado apelando a la sensibilidad del público objetivo y a direccionar una estrategia de valor a través de la emoción lleva a "humanizar los productos digitales", al mismo tiempo que permite observar y analizar más allá de la superficialidad comercial de algunas herramientas como *focus group* o encuestas, al identificarse con estados emocionales, cognitivos y con actitudes de otras personas por medio de la experiencia indirecta, es decir, "ponerse en los zapatos del otro".

En este orden de ideas, mi propuesta es un poster que permita a los participantes del congreso encontrar **insights** de una manera sencilla a través de una caja de herramientas que funcione como base de un proyecto acertado y sostenible. Por lo tanto, este poster permitirá al usuario llevarse algo práctico de él con claves rápidas y pasos simples para empezar a fortalecer la habilidad der ser empático y así utilizarlo como una herramienta para conectar a un nivel emocional como valor agregado a los proyectos digitales.

Sobre el contenido del poster se plantean 3 formatos con ejercicios y técnicas básicas como primer acercamiento al concepto de empatía digital divido en tres secciones a partir de tres preguntas básicas que son ¿qué? ¿cómo? ¿por qué? Con la finalidad de **Sentir + compartir+ reaccionar = experiencia de usuario.**

Finalmente se quiere evidenciar los procesos creativos de la Biblioteca Nacional de Colombia en el área de Humanidades y Desarrollos Digitales y como han se han transformado utilizando este puente como herramienta que serán aplicados en la creación de este poster ya que después de un proceso de conceptualización por medio de metodologías como design thinking entre otras herramientas y por su puesto desde la empatía digital se utilizarán unos colores específicos por conceptos que se van a comunicar basados en la teoría del color para aplicarlos mediante técnicas como: ilustración digital y tipografías que hacen alusión a la estética del mundo digital como los pixeles o el código.

La interacción será análoga en la medida que el usuario pueda entender los insights para ponerlos en práctica en sus procesos al poder revelar el contenido del poster con ayuda de un elemento externo para poder filtrar el contenido de cada color, esto es posible al hacer uso de un recurso visual como la adición de colores primarios RGB. En ese orden de ideas al tener todos los contenidos impresos al mismo tiempo cada tipo de contenido en una tinta (verde, rojo o azul), se genera una recarga o confusión visual resultando complicado para el usuario entender la información en la primera impresión. Al ayudarse con los elementos de filtrado de color pueden obtener la información por medio de filtrado por lo tanto se propone es que haya un cambio de visión y perspectiva como lo requiere la habilidad de ser empático para su posterior aplicación a proyectos de humanidades digitales.

References

Dave M Berry. (2010). *The Computational Turn: Thinking About the Digital Humanities*.

Yonty Friesem. (2016). *Chapter 2 - Empathy for the Digital Age: Using Video Production to Enhance Social, Emotional, and Cognitive Skills*.

IDEO. (2016). *What is Human-Centered Design?*

Jon Kolko. (2010). Abductive Thinking and Sense making: *The Drivers of Design Synthesis*. Vol. 26, No. 1 (Winter, 2010), pp. 15-28.

Jon Kolko. (2015). *Design Thinking Comes of Age*.

Hasso Plattner (2013). *Empathy field guide. Institute of Design at Stanford.*

Mark Considine (2012). *Thinking Outside the Box? Applying Design Theory to Public Policy: Applying Design Theory to Public Policy.*

McDonagh y Deana. (2010). *Rethinking Design Thinking: Empathy Supporting Innovation.*

Sanhueza y Camila Holven. (2012). *Design Empathy in Service Design Methodology*.

Elena González García (2016). *Un nuevo camino hacia las humanidades digitales: el laboratorio de innovación en humanidades digitales de la uned (linhd)*.

Atlas de la narrativa mexicana del siglo XX y la representación visualizada de México en su literatura. Avance de proyecto

Nora Marisa León-Real Méndez
nora.marisa@itesm.mx
Tecnológico de Monterrey, Mexico

En esta presentación se busca mostrar el avance obtenido en un año de trabajo en el proyecto de creación de un mapa literario de México en el que se representen gráficamente las obras y los espacios en las que éstas se desarrollan. El *Atlas de la Narrativa Mexicana del siglo XX* compila y presenta visualmente información geográfica proveniente de las obras más representativas de la literatura mexicana contemporánea con un propósito educativo. Además, el proyecto busca servir como base para la realización de conexiones sociohistóricas que los estudiantes pueden realizar, pues la visualización de las distintas versiones de México presentes en la literatura es un paso importante para la evolución de la identidad cultural del país, así como una manera innovadora de reconocer nuestra realidad dentro de los textos. Por otra parte, este proyecto requiere analizar la narrativa sobre México utilizando herramientas de análisis literario, historia y geografía, a través de medios digitales. Esta naturaleza interdisciplinaria vuelve al proyecto pertinente dentro del marco de las Humanidades Digitales y arroja ya resultados que contribuyen a la metodología de su aplicación en clase.

El primer paso del proyecto (aprobado por la Convocatoria de Experimentación en Innovación Educativa NOVUS en agosto de 2017) ha sido recopilar la información necesaria, creando un corpus de las novelas más representativas de la literatura mexicana del siglo XX (de inicio, por medio de un compilado de Novelas de la Revolución Mexicana: *Los de abajo*, de Mariano Azuela; *El águila y la serpiente*, de Martín Luis Guzmán; *Cartucho*, de Nellie Campobello; y *Los relámpagos de agosto*, de Jorge Ibargüengoitia), considerando la representación narrativa que hacen del espacio mexicano. Luego, se asignó la lectura de los primeros textos a los participantes del proyecto para realizar las anotaciones y capturar los datos. Con esta información se crearon categorías espaciales que puedan ser marcadas en un mapa de México, de acuerdo con el estado, región o población mencionados en las obras. Por otra parte, estos espacios han sido también clasificados en dos categorías narrativas: aquellos en los que sucede la acción de la novela y los que son mencionados como referentes de eventos fuera de la trama. Esta información se ha vertido en un primer borrador del *Atlas*, un mapa digital realizado con herramientas de acceso abierto propias de las HD, en el que se proyecta visualmente la información de forma que se pueda interactuar con ella: conocer qué porciones del territorio mexicano aparecen con mayor frecuencia en las obras, u observar la predominancia de los espacios rurales o urbanos, por ejemplo. Esta información cartográfica nos permite sacar ya algunas conclusiones con respecto a la representación de la Revolución Mexicana en la literatura, considerando los espacios de acción de las obras en su proporción con la extensión geográfica del país y de los hechos sucedidos en la historia de México. Pero, sobre todo, este proceso ha servido como práctica para proponer el método de creación del *Atlas* así como las áreas en las que hay oportunidad de mejora.

Eventualmente, se busca que el *Atlas* pueda ser utilizado como herramienta de enseñanza de la literatura mexicana en cursos de preparatoria y profesional, permitiendo a los estudiantes contribuir en su crecimiento, aportando nuevos datos según sus lecturas. La información recopilada de manera gráfica permitirá continuar encontrando conexiones entre distintas obras y movimientos literarios, que luego podrían ser analizados por estudiantes e investigadores de la literatura mexicana contemporánea.

Al final del proyecto, se espera contar con un producto demostrable y perfectible (el *Atlas de la Narrativa Mexicana del siglo XX*), así como con grupos de estudiantes que han pasado por el proceso de contribuir a su creación y que, a través de ello, han aumentado su interés y desempeño en las clases de literatura mexicana. De manera tangible, los alumnos serán capaces de mostrar en un mapa de México los espacios detectados dentro de las obras literarias leídas, así como de explicar distintas relaciones entre el espacio y la obra.

HuViz: From _Orlando_ to CWRC… And Beyond!

Kim Martin
kmarti20@uoguelph.ca
University of Guelph, Canada

Abi Lemak
alemak@uoguelph.ca
University of Guelph, Canada

Susan Brown
sbrown@uoguelph.ca
University of Guelph, Canada

Chelsea Miya
cmiya@ualberta.ca
University of Guelph, Canada

Jana Smith-Elford
smithelf@ualberta.ca
University of Guelph, Canada

The Orlando Visualizer (OViz) was originally conceived in 2010 as a tool that would display extracts from The Orlando Project's textbase as a series of interconnected nodes in a graph. Since then, the project has grown to address digital humanities research more generically. Now called HuViz (fig. 1), the Humanities Visualizer is a browser-based, interactive interface that allows for the exploration of semantic relationships and ontologies represented using Linked Open Data (LOD). LOD is a practice of creating, sha-

ring, and interlinking bits of information on the Semantic Web (linkeddata.org). At its core, LOD is a way of structurally representing data as connected. More broadly, it challenges how information networks are built within digital environments and calls attention to the importance of making these networks and the data they house open and accessible. In the spirit of LOD, HuViz came together as a tool designed to make available the contents, along with the contexts, of portions of the Semantic Web to experts and lay-users alike in ways that are open, editable, and transferable. This poster will provide an overview of HuViz's development, shaped by the results of user-testing, the demands of Orlando's complex data, as well as the growing ontology of the Canadian Writing Research Collaboratory (CWRC), which is building out from the Orlando data. Future possibilities include use by other projects housed by CWRC's infrastructure (see beta.cwrc.ca).

The CWRC ontology team has been using HuViz to visualize the Orlando datasets, translating the textbase's XML-encoded entries on women writers into RDF assertions (also referred to as triples) (Simpson and Brown, 2013). The test extractions made from the Orlando textbase range from Virginia Woolf to Margaret Atwood, encompassing everything from the schools they attended, to the places they lived, the writers they influenced, and the overlapping and often contradictory cultural forms that contribute to an author's social identity. Given the scope of the Orlando data, which contains millions of connections, as well as the immeasurably bigger Semantic Web itself, the task of visualizing massive hoards of data in meaningful ways remains a central question in developing HuViz. This problem of visualizing large-scale datasets is by no means new for digital humanities scholars (Duke, 2005; Sherratt, 2011). With increased attention paid to the value of LOD for humanities scholarship in the past decade, the question of how to make these graphs both interactive and legible has arisen as a major concern (Katifori, 2007; Ghorbel et al., 2016). Beyond interface design, questions of tool mediation and the "avenues of interpretation" (Warwick, 2012) made available to the user are central to discussions surrounding the false neutrality of technology (McPherson, 2012; Nakamura, 2002; Chun, 2005). With these concerns in mind, the design of HuViz incorporates some aspects of D'Ignazio and Klein's "feminist data visualization" principles (2016). The ability to visualize data along with the structure of the ontology that governs it, for instance, aims to enable interrogation, such as Jacqueline Wernimont's, of "how and where we might locate feminist ideology and politics within digital archives" (2013).

In the latest iteration of HuViz, these concerns have materialized in features supporting:

Context awareness (provision of source snippets; ability to visualize ontologies as well as data; support for web annotation data model)

1. Collaboration (HuViz code available on Github; forthcoming edit button)
2. Transparency (users may import their own data and ontology; CWRC ontology extensively documented and published in HTML)

This poster and tool demonstration will show the growth of this tool over the past several years. The poster will provide an overview of feature development and indicate how a growing body of user tests have shaped that process, highlighting a number of enhanced features. These include:

- Enhanced control over shape, colour, size and weight of edges, nodes, and background both for user preference and to aid accessibility
- Visualization of LOD ontologies
- Loading a dataset or an ontology from a URL (eg. GitHub)
- And perhaps most excitingly, the chance for users to upload and explore their own datasets.

The tool demonstration will introduce attendees to basics of HuViz and invite them to play with it. We will have multiple datasets and ontologies for users to explore, and will provide a link to detailed instructions on how to upload their own datasets or ontologies.

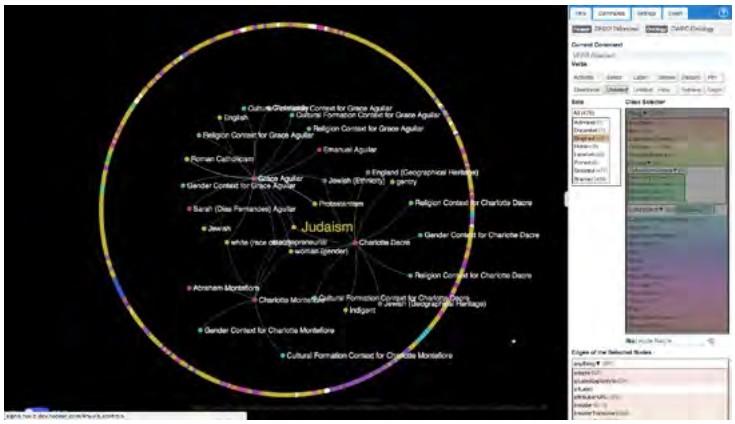


Figure 1. HuViz visualizing Orlando data via CWRC Ontology.

References

Chun, Wendy Hui Kyong. (2005). "On software, or the persistence of visual knowledge." *Grey Room* 18: 26-51.

Canadian Writing Research Collaboratory. beta.cwrc.ca (accessed 25 Nov. 2017).

D'Ignazio, Catherine, and Lauren F. Klein. (2016). "Feminist data visualization." Paper presented at the 2016 IEEE VIS Conference, Baltimore, October 23–28.

Duke, David J., Ken W. Brodie, David. A. Duce, and Ivan Herman. (2005). "Do you see what I mean? [Data visualization]." *IEEE Computer Graphics and Applications* 25.3: 6-9.

Ghorbel, Fatma, Nebrasse Ellouze, Elisabeth Métais, Fayçal Hamdi, Faiez Gargouri, and Noura Herradi. (2016). "MEMO GRAPH: An ontology visualization tool for everyone." *Procedia Computer Science* 96: 265-274.

Humanities Visualizer. http://alpha.huviz.dev.nooron.com/ (accessed 25 Apr. 2018).

Katifori, Akrivi, Constantin Halatsis, George Lepouras, Costas Vassilakis, and Eugenia Giannopoulou. (2007). "Ontology visualization methods—a survey." *ACM Computing Surveys (CSUR)* 39.4: 10.

Linked Data. http://linkeddata.org/ (accessed 25 Nov. 2017).

McPherson, Tara. (2012). "Why are the Digital Humanities so white? Or thinking the histories of race and computation." In M. Gold (ed). *Debates in the Digital Humanities*. Minnesota: University of Minnesota Press, pp. 139-160.

Nakamura, Lisa. (2002). *Cybertypes: Race, Ethnicity, and Identity On the Internet*. London: Routledge.

Sherratt, Tim. (2011). "It's all about stuff: Collections interfaces, power and people." *Journal of Digital Humanities* 1: 1-1.

Simpson, John, and Susan Brown. (2013). "From XML to RDF in The Orlando Project." *Culture and Computing (Culture Computing), 2013 International Conference*. IEEE.

Warwick, Claire. (2012). "Studying users in digital humanities." In *Digital Humanities in Practice*, edited by Claire Warwick, Melissa Terras, and Julianne Nyhan, 1–21. London: Facet Publishing.

Wernimont, Jacqueline. (2013). "Whence Feminism? Assessing feminist interventions in digital literary archives." *DHQ: Digital Humanities Quarterly* 7.1.

Endangered Data Week: Digital Humanities and Civic Data Literacy

Brandon T. Locke
blocke@msu.edu
Michigan State University, United States of America

Endangered Data Week (http://endangereddataweek.org) emerged in the early months of 2017 as an effort to encourage conversations about government-produced, open data and the ways in which it may become endangered due to political, technical, and social factors.

The 2016 US election set off a wave of activism surrounding government data, particularly in the collection and mirroring of environment and climate change data. While much of this attention has been focused on the United States, similar conditions have affected and continue to threaten governments around the world. Endangered Data Week presented an opportunity to funnel even more attention to the issue of potential federal data loss, while also providing opportunities to include lessons on data literacy, civic issues and policy advocacy, data management and curation, technical skills for data capture, and open access and open data in scholarship.

The inaugural Endangered Data Week (April 17-21, 2017) was comprised of 57 formally registered events from 30 institutions and organizations, including virtual participation from hundreds of participants from around the world. The second annual Endangered Data Week will be February 26 - March 2, 2018.

One particularly interesting strain of events in Endangered Data Week is civic data literacy. While so many other projects, including DataRescue, the Preservation of Electronic Government Information (PEGI) project and Environmental Data and Governance Initiative (EDGI) are focused on capturing and preserving government data, Endangered Data Week data literacy events focus on the capacity of the user communities. They seek to enable broader communities to use, interpret, and analyze open data.

The required knowledge and tools for working with civic data overlap significantly with much of the work digital humanists do with data. The creation of datasets often requires scraping information off of the web in flat HTML or confusing databases. Data in both contexts is often irregularly formatted or melded together from multiple sources, requiring the cleaning and reorganization. Meaningful research often requires an iterative process of researching the contexts in which the data was created and the data itself to resolve undocumented meaning in the data. Both contexts also require interpretation for both specialized and non-specialized audiences.

This poster will include a brief overview of Endangered Data Week and will focus on the existing efforts to teach civic data literacy, including an exploratory framework for the most essential skills, knowledges, and tools that are required for diverse communities to use civic data, and the relationship between these events and the broader role of digital humanities faculty, librarians, and staff within our institutions and the communities in which we live.

Herramienta web para la identificación de la técnica de manufactura en fotografías históricas

Gustavo Lozano San Juan
gustavolsj@gmail.com
Instituto de Investigaciones Estéticas
Universidad Nacional Autónoma de México, Mexico

Introducción

Este proyecto consiste de una metodología para identificar el proceso fotográfico en fotografías históricas, está inspirado en el concepto de árbol de decisiones utilizado en las ciencias de datos para clasificar entidades con base en sus diferentes atributos y valores, la implementación ha sido realizada por medio de una herramienta web en idioma español.

Este recurso está dirigido a archivistas historiadores y otros profesionales de archivos históricos en Latinoamérica y les permite identificar el proceso fotográfico entre una gama de 29 alternativas utilizadas a los largo de los siglos XIX y XX, para lo cual los usuarios son guiados paso a paso a través de la metodología por medio de preguntas sobre las características de la fotografía que buscan catalogar.

Figura 1. Diferentes procesos fotográficos históricos

Antecedentes

La identificación del proceso fotográfico es una de las tareas fundamentales que realizan los archivos históricos en el ámbito de la catalogación de fotografías ya que brinda a los investigadores información sobre su temporalidad y características físicas, como el color, el tipo de soporte, el formato, entre otras.

Comúnmente la identificación del proceso fotográfico es una habilidad visual especializada que se transmite de persona a persona de manera empírica mediante la observación detallada de cientos de fotografías y el estudio de su evolución tecnológica. Esta forma de aprendizaje limita la diseminación de este conocimiento entre los profesionales de los archivos y como resultado de ello un gran número de fotografías se encuentran incorrectamente clasificadas dentro de los catálogos.

En la bibliografía sobre conservación de fotografías se han propuesto varios esquemas de clasificación, Lavedrine propone la división inicial de las fotografías por polaridad, posteriormente por soporte y finalmente por tono, aunque este es un modelo útil pone el énfasis en la conservación y no en la identificación (Lavedrine, 2009: 15). Reilly abordan específicamente el tema de la identificación, aunque enfocado únicamente a impresiones del siglo XIX, y no contempla negativos o impresiones a color del siglo XX (Reilly, 1986: 40). El Graphics Atlas (IPI, 2017) es una página que brinda una vasta información que ilustra y describe las características físicas de las fotografías y ayuda al usuario a identificarlas, sin embargo, al igual que las fuentes anteriores se ocupa principalmente de los procesos fotográficos más comunes en los archivos de Estados Unidos y en Europa y su contenido se encuentran en idioma inglés, lo que limita su utilidad y aplicación en archivos de Latinoamérica.

Desarrollo

Una revisión bibliográfica de la literatura en español permitió definir la terminología y los conceptos más adecuados para nombrar cada uno de los procesos fotográficos las características físicas y sus valores (Barra y Gutiérrez, 2000: 19; Boadas et al. 2001: 211; SE, 2016: 20), con esta información posteriormente se elaboró una tabla de datos común para todos los procesos y sus atributos.

Clasificaciones		Atributos comunes				Atributos particulares								
Proceso	Tipología	Soporte primario	Iluminación	Polaridad	Tono	Fechas	Estratigrafía	Magnificación	Tonalidad	Brillo	Superficie	Particularidades Objeto	Texto	Deterioro
Daguerrotipo	Imagen de cámara	Metal	Reflexión	Positivo	Monocromático	1839 - 1860			Neutro	Muy brillante		Positivo-negativo		Delaminación de plata, corrosión de cobre
Ambrotipo	Imagen de cámara	Vidrio	Reflexión	Positivo	Monocromático	1851 - 1865			Café	Brillante		Luces fechosas		
Ferrotipo	Imagen de cámara	Metal	Reflexión	Positivo	Monocromático	1855 - 1890			Café	Semi mate		Magnetico		Corrosión, faltantes, craqueladuras
Cianotipo	Impresión	Papel	Reflexión	Positivo	Monocromático	1840 - 1920	Una	Fibras visibles	Cian	Mate				
Albúmina	Impresión	Papel	Reflexión	Positivo	Monocromático	1851 - 1890	Dos	Fibras visibles	Amarillo, Café rojizo	Semi mat	Textura del pap	Soporte primario delgado, Soporte secundario grueso		Craqueladuras amarillamiento, perdida de densidad en las luces y sombras
Carbón	Impresión	Papel	Reflexión	Positivo	Monocromático	1860 - 1940	Dos		Otro		Relieve en sombras			
Colodión de impresión directa	Impresión	Papel	Reflexión	Positivo	Monocromático	1885 - 1910	Tres	No se ven fibras	Purpura_Rojo	Brillante	Textura lisa	Soporte primario grueso, Soporte secundario grueso, Iridiscencia		Abrasiones, perdida de densidad en las luces
Plata gelatina de impresión directa	Impresión	Papel	Reflexión	Positivo	Monocromático	1885 - 1910	Tres	No se ven fibras	Amarillo	Brillante	Textura lisa	Soporte primario grueso, Soporte secundario grueso, Iridiscencia		perdida de densidad en las luces
Goma bicromatada	Impresión	Papel	Reflexión	Positivo	Monocromático	1890 - 1930	Dos		Otro					
Platinotipo	Impresión	Papel	Reflexión	Positivo	Monocromático	1890 - 1930	Una	Fibras visibles	Neutro					Ghosting
Plata gelatina	Impresión	Papel	Reflexión	Positivo	Monocromático	1890 - 2018	Tres	No se ven fibras	Neutro					
Colodión mate de impresión directa	Impresión	Papel	Reflexión	Positivo	Monocromático	1895 - 1910	Tres	No se ven fibras	Neutro, Café (platino), purpura (oro)	Semi mate	Texturizado	Soporte primario grueso, Soporte secundario grueso y de color,		Ghosting, No hay perdida de densidad
Difusión de plata	Impresión	Papel	Reflexión	Positivo	Monocromático	1942 - 2018	Tres		Neutro		Lisa	Pestañas, borde irregular o perforado, superficie con restos de adhesivo,	Polaroid	Revelado irregular, amarillamiento y desvanecimiento por recubrimiento irregular Craqueladuras
Cromógeno	Impresión	Papel	Reflexión	Positivo	Policromático	1940 - 2018	Tres			Brillante	Lisa o texturizada	Papel de fibra, RC, acetato pigmentado		<1980 amarillamiento, perdida de balance de color y desvanecimiento de colorantes
Difusión de colorantes por transferencia	Impresión	Papel	Reflexión	Positivo	Policromático	1963 - 2018	Tres			Brillante	Lisa	Pestañas, borde irregular o perforado, superficie con restos de adhesivo,	Polaroid	Revelado irregular.
Blanqueo de colorantes	Impresión	Papel	Reflexión	Positivo	Policromático	1963 - 2018	Tres							
Difusión de colorantes por transferencia inferior	Impresión	Papel	Reflexión	Positivo	Policromático	1972 - 2018	Tres			Brillante	Lisa,	Marco blanco con contenedor de químicos de procesado,	Polaroid	Revelado irregular, craqueladuras, migración de colorantes en areas blancas
Colodión húmedo	Negativo	Vidrio	Transmisión	Negativo	Monocromático	1851 - 1885			Café		Aglutinante irregular, Barniz	Vidrio grueso e irregular		Abrasión
Gelatina seca	Negativo	Vidrio	Transmisión	Negativo	Monocromático	1880 - 1925			Neutro		Aglutinante irregular	Vidrio delgado y recto		Espejo de plata
Plata gelatina Nitrato de celulosa	Negativo	Plástico	Transmisión	Negativo	Monocromático	1889 - 1950						Muescas "V"	Nitrate	Amarillamiento del soporte, deformación,
Plata gelatina Acetato de celulosa	Negativo	Plástico	Transmisión	Negativo	Monocromático	1935 - 2018						Muescas "U"	Safety	Canales, burbujas, deformación, olor a vinagre
Plata gelatina Poliester	Negativo	Plástico	Transmisión	Negativo	Monocromático	1955 - 2018						Birrefringencia	Safety	
Cromógeno Acetato de celulosa	Negativo	Plástico	Transmisión	Negativo	Policromático	1942 - 2018							Safety	
Cromógeno Poliester	Negativo	Plástico	Transmisión	Negativo	Policromático	1955 - 2018						Birrefringencia	Safety	
Plata gelatina Vidrio	Transparencia	Vidrio	Transmisión	Positivo	Monocromático	1880 - 1940								
Plata gelatina Acetato de celulosa	Transparencia	Plástico	Transmisión	Positivo	Monocromático	1935 - 2018							Safety	
Plata gelatina poliester	Transparencia	Plástico	Transmisión	Positivo	Monocromático	1955 - 2018						Birrefringencia	Safety	
Procesos aditivos	Transparencia	Vidrio	Transmisión	Positivo	Policromático	1907 - 1935		Retícula			Lineas rectas paralelas	Bajo brillo y saturación		Delaminación, puntos verdes
Cromógeno Acetato de celulosa	Transparencia	Plástico	Transmisión	Positivo	Policromático	1935 - 2018							Safety	
Cromógeno Poliester	Transparencia	Plástico	Transmisión	Positivo	Policromática	1955 - 2018						Birrefringencia	Safety	

Figura 2. Tabla de datos de procesos, características físicas y valores.

Posteriormente la tabla de datos se tradujo en un árbol de decisiones, las características físicas se convirtieron en nodos de decisión, sus valores en ramas y los procesos fotográficos en hojas. Por un lado, este esquema proporciona la estructura de navegación a la página web y por otro le permite al usuario visualizar el panorama del universo posible y comprender las distintas combinaciones de atributos que caracterizan a los procesos fotográficos.

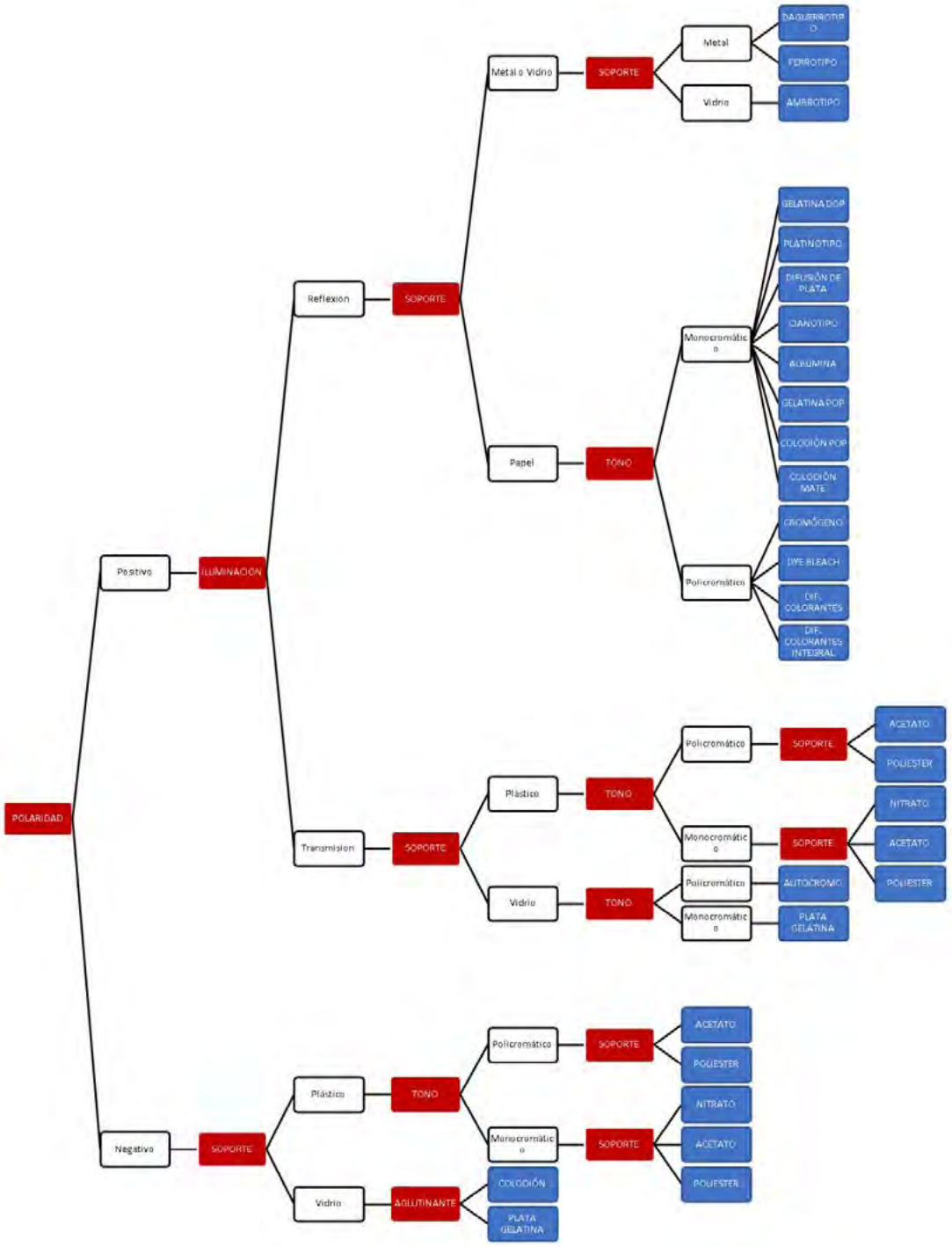


Figura 3. Árbol de decisión para la identificación de procesos fotográficos

Mediante el diseño y programación de la página web se recreó la estructura del árbol de decisión y utilizando preguntas con un lenguaje claro y sencillo se guía al usuario a través de la metodología, las respuestas se ilustran con galerías de imágenes que permiten comparar la fotografía que se busca identificar y encontrar similitudes. El objetivo principal de esta fase del proyecto fue hacer accesibles conceptos que son difíciles de comprender sólo verbalmente pero que son fáciles de reconocer de manera visual.

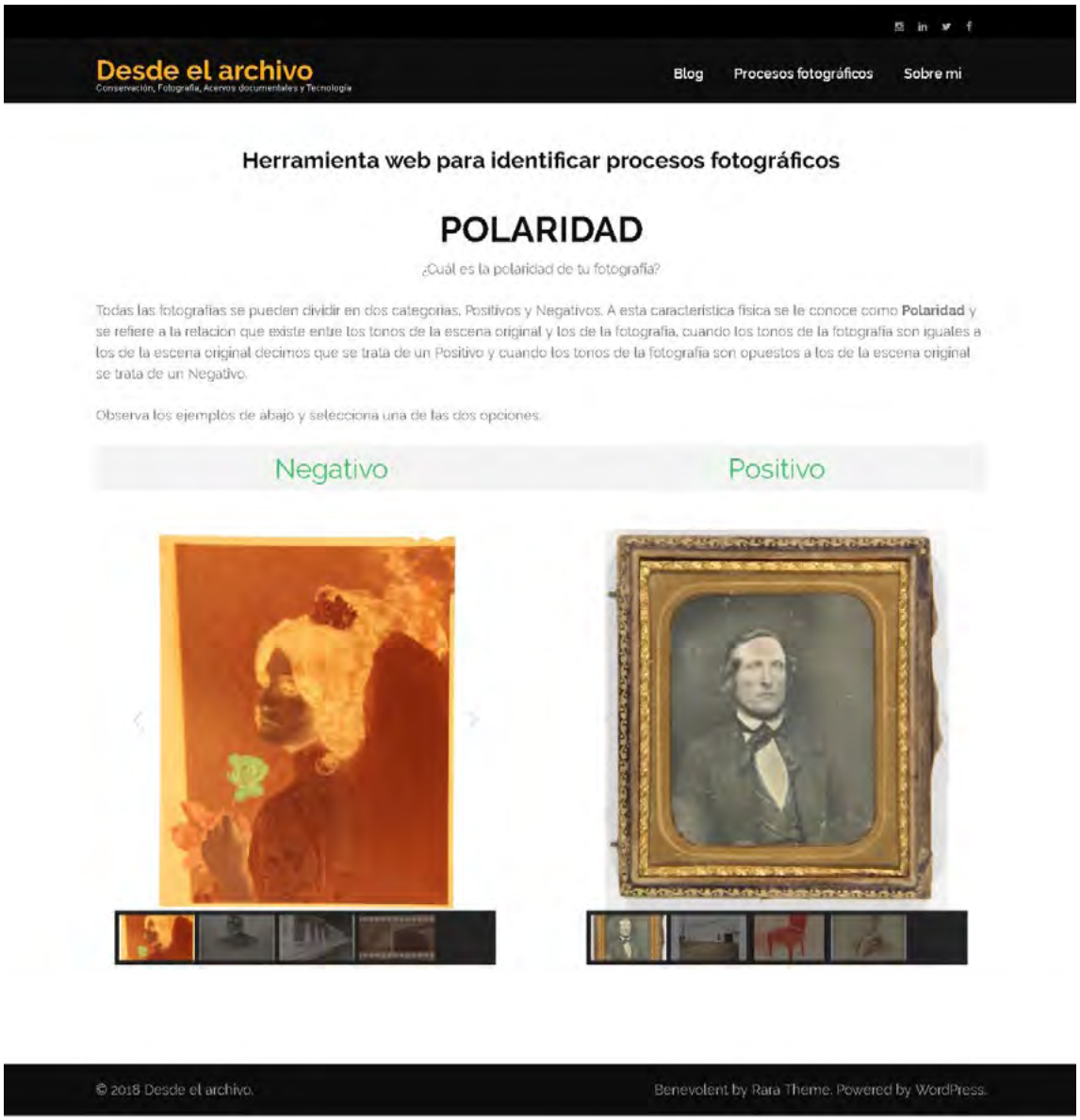


Figura 4. Interfaz de usuario de la herramienta web

Conclusiones

Gracias a las posibilidades comunicativas de la tecnología web, herramientas como esta pueden contribuir a diseminar conocimientos especializados que son poco accesibles, lo cual a su vez permite a un mayor número de personas comprender y valorar la materialidad de las fotografías analógicas resguardadas en los archivos históricos.

Dirección web. http://www.desdeelarchivo.com/procesos-fotograficos/

References

Barra, P., y Gutiérrez, I. (2000). *Normas Catalográficas del Sistema Nacional de Fototecas del INAH*, México: INAH/CONACULTA.

Boadas, J., Casellas, L., y Suqyet, M. (2001). *Manual para la Gestión de Fondos y Colecciones Fotográficas*. Girona: CCG ediciones.

IPI. Image Permanence Institute. (2017). *Graphic Atlas*. http://www.graphicsatlas.org (recuperado el 16 de noviembre de 2017).

Lavedrine, B. (2009). *Photographs of the Past: Process and Preservation*. Los Ángeles: Getty Conservation Institute.

Reilly, J. (2009). *Care and Identification of 19th-century Photographic Prints*. Rochester: Eastman Kodak Co.

SE. Secretaría de Economía. (2016). *Norma Mexicana NMX-R-069-SCFI-2016. Documentos fotográficos. Lineamientos para su Catalogación*. México: Secretaría de Economía.

Propuesta interdisciplinaria de un juego serio para la divulgación de conocimiento histórico. Caso de estudio: la divulgación del saber histórico sobre la vida conventual de los carmelitas descalzos del ex-Convento del Desierto de los Leones

Leticia Luna Tlatelpa
letyludigital@gmail.com
Universidad Autónoma Metropolitana Cuajimalpa, Mexico

Fabián Gutiérrez Gómez
fabian.gutierrez.gomez@gmail.com
Universidad Autónoma Metropolitana Cuajimalpa, Mexico

Edné Balmori
ednebalmori@gmail.com
Universidad Autónoma Metropolitana Cuajimalpa, Mexico

Feliciano García García
felicianogarcia.9@gmail.com
Universidad Autónoma Metropolitana Cuajimalpa, Mexico

Luis Rodriguez Morales
luis.rodriguezm12@gmail.com
Universidad Autónoma Metropolitana Cuajimalpa, Mexico

Resumen

El surgimiento de las narrativas hipermedia y transmedia así como de los productos culturales propios de la era digital como los *Juegos Serios*, expanden el espectro de los medios de comunicación tradicionalmente usados para divulgar la historia. Con estos nuevos medios es factible realizar divulgación bajo el modelo contextual, el cual, al contrario del modelo de déficit, considera las respuestas del público y la complejidad del fenómeno comunicativo de la divulgación (Leewenstein, 2003). En este sentido, se propone un *Juego Serio* sobre la vida conventual de los Carmelitas Descalzos que habitaron el Desierto de los Leones que estimule una experiencia memorable en jóvenes.

Divulgación de la historia

Se retoma el concepto de comunicación de la ciencia propuesto por Burns *et al.* (2003: 1991) para aplicarlo a la divulgación de la historia. Ellos plantean que la divulgación debe provocar alguna de las siguientes respuestas en la audiencia con relación a la ciencia: sensibilización, disfrute, interés, construción de opinión y/o análisis de contenido.

Se siguió la metodología de interpretación temática propuesta por Ham (2013) para responder a la pregunta de cómo comunicar el conocimiento histórico acerca de los carmelitas a un público joven, este último definido según Feixa (2003), a través de un videojuego y generar lazos de identidad fincados en valores culturales.

Esta metodología resalta la importancia de conocer la audiencia y sus intereses para provocar en ella reflexiones, cuestionamientos o la generación de nuevos significados, de modo que recuerde la experiencia divulgativa. Gándara añade, además, la importancia que tiene la narrativa para crear divulgación significativa (Gándara citado por Sánchez, 2016). Es de destacar también, que las emociones son relevantes en el divulgación, ya que pueden lograr que esta sea más recordada (Bonfil, 2003).

Si bien hay antecedentes del uso de videojuegos para tratar temas históricos, (Caldera Estudios, 2010), (Mulaka, 2018), (Nomdedeu, 2015), (Rodríguez et al, 2017), (Salinas et al, 2017), no se encontró uno cuyo diseño emplee la interpretación temática.

Los Carmelitas Descalzos

Esta orden fue una de las últimas que llegaron a la Nueva España. Aún cuando el cometido de estos frailes fue evangelizar a la población indígena, decidieron que su misión sería continuar su vida contemplativa que, según su cosmovisión, los acercaba a Dios (Ramírez, 2015). Formaron la Provincia de San Alberto con conventos en las principales ciudades de la Nueva España. Así, la Corona les permitió edificar el Desierto en el Monte de Santa Fe. Este yermo fue un espacio donde renunciaban a todo placer sensorial, como el disfrute de la comida, la prohibición de mirar a otra persona a los ojos, y el cumplimiento de la regla del silencio y del claustro para dedicarse a la contemplación (Báez, 1981).

Videojuegos

Se escogió el juego serio ya que "buscan cumplir un propósito más allá del propósito autocontenido de los juegos de entretenimiento" (Mitgush y Alvarado, 2012: 121, citado por González, 2017). Las emociones son observadas desde el enfoque de los videojuegos, como lo propone Lazzaro (2004). Finalmente, para el diseño del videojuego se siguieron los lineamientos propuestos por Shell (2005) y la metodología de interpretación temática. En esta, es primordial escoger un Tema (Ham, 2003) que guíe el contenido del objeto comunicativo.

Después de analizar los documentos históricos y desde la perspectiva de una audiencia jóven, se definió el tema como **los espacios donde se ejerce control sobre las personas anulan la individualidad y fomentan el desarraigo**. A partir de allí, se modeló la vida de los frailes en términos de tentaciones, recuerdos de la vida pasada, castigos, recompensas, el diablo, reglas, obediencia. El tema también determinó la estética gráfica, la animaciones y la música.

El videojuego se titula *Tentación en el Desierto* y es de tipo *Click and Point*. El jugador ayuda a un fraile Carmelita recién llegado al Yermo de Santa Fe a luchar contra tentaciones para que lo acepten como ermitaño mientras explora elementos de diferentes espacios del convento. Hay cinemáticas que muestran los recuerdos del personaje antes de convertirse en fraile; las tentaciones, castigos y retos del diablo son representados como minijuegos que el jugador debe resolver; la obediencia a las reglas es la energía del fraile; hay un diario escrito en primera persona el cual brinda información acerca del contexto histórico.

Conclusiones

Las pruebas con el prototipo analógico del juego mostraron que los jugadores problematizaron la vida carmelita y manifestaron emociones e interés. Se recomienda evaluar con la versión digital para determinar si la interpretación temática aplicada al diseño de un juego serio genera experiencia memorable al divulgar conocimiento histórico.

References

Báez M., E. (1981). *El Santo Desierto. Jardín de Contemplación de los Carmelitas Descalzos en la Nueva España*, México: Universidad Nacional Autónoma de México.

Bonfil, O., M. (2003). Una estrategia de guerrilla para la divulgación: Difusión cultural de la ciencia. *Congreso Latinoamericano Ciencia, comunicación y sociedad, Costa Rica.*

Burns, T.W., O'Connor, D. J., y Stockmayer, S. M. (2003). *Science communication: a contemporary definition. Public understanding of science*, 12(2), 183-202.

Caldera Estudios. (2010). Peluconas. [en linea] disponible en: http://caldera-estudio.com/proyectos/asi-se-veia-mexico-hace-250-anos/ [consultado 20 abril 2018].

Feixa, C. (2003). Del reloj de arena al reloj digital. Sobre las temporalidades juveniles. *Jóvenes, Revista de Estudios sobre la Juventud.* 7 (19), 6-27.

González, A. (2017). *Comunicación de la ciencia en videojuegos: evolución y juegos serios.* Tesis de Maestría. Universidad Nacional Autónoma de México.

Ham, S. H. (2013). *Interpretation: making a difference on purpose*. Fulcrum publishing

Lewenstein, B. (2003). Models of public communication of science and technology. *http://communityrisks. cornell. edu/Background-Materials/Lewenstein2003. pdf*.

Lazzaro, N. (2004). *Why we play games: Four keys to more emotion without story.*

Mulaka. (2018). Mulaka. [en linea] disponible en: https://www.lienzo.mx/mulaka/?lang=es [consultado 20 abril 2018].

Nomdedeu, L. (2015). *RAÍCES, un juego serio social para revalorizar las culturas originarias.* Tesina de Licenciatura. Universidad Nacional de la Plata.

Ramirez, J. (2015). *Los Carmelitas Descalzos en la Nueva España. Del activismo misional al apostolado urbano. 1585 - 1614.*, México, INAH.

Rodríguez, F. C., Palacios D., E., Marín G., P., Ortiz M., B. y Romero Q., G. (2017).Wirikuta. [en linea] disponible en: https://leiva2017.wordpress.com/proyectos/wirikuta/ [consultado 20 abril 2018].

Salinas, I., Hernández, E., Rodríguez, S. (2015). El desarrollo social a través de la valoración del sistema estético-comunicativo de los pueblos nativos de Baja California. [en linea] disponible en: http://www.redalyc.org/pdf/4981/498150319057.pdf [consultado 20 abril 2018].

Sánchez, M. (2016). *La Museología como herramienta de vinculación entre el profesor y el patrimonio. Propuesta de un curso de capacitación a profesores que imparten la asignatura estatal Patrimonio Cultural y Natural del Distrito Federal.* Tesis de Maestría. Escuela Nacional de Conservación, Restauración y Museografía.

Schell, J. (2015). *The art of game design: a book of lenses*. CRC Press.

Digital 3D modelling in the humanities

Sander Münster
sander.muenster@tu-dresden.de
TU Dresden, Germany

For more than 30 years, digital 3D modelling and in particular reconstruction methods have been widely used to support research and education in the digital humanities, especially but not exclusively on historical architecture. While technological backgrounds, project opportunities, and methodological considerations for the application of digital 3D modeling techniques are widely discussed in literature (e.g. Arnold and Geser, 2008, European Commission, 2011, Frischer, 2008, Bentkowska-Kafel et al., 2012, Bentkowska-Kafel, 2013, Kohle, 2013), my interest is to investigate digital 3D modeling in the humanities as a scholarly area and to derive implications for further organizational and methodical development. Against this background, my research investigates the following research questions:

-
- What marks a scholarly culture of 3D modelling in the humanities?
- What are technical and designal implications and workflows for model creation and presentation?
- How can digital 3D modelling techniques be learned and taught?

The research presented is part of an ongoing post doc thesis work dedicated to draw a "big picture" on digital 3D modelling techniques as research tools in the humanities. Against this background, my own and my department's activities include to investigate (1) a scholarly community, (2) usage practices occurring within single projects and to gain implications for further methodical development. We develop (4) technologies and workflows to enhance both, the creation of 3D models and user-centered interfaces and investigate how 3D models are (5) perceived and how 3D modeling techniques can be used in (6) education (Table 1). Research has been carried out since 2010 in 12 projects on local, national and EU level so far.

Area	Research Interest	Investigation
Scholarly community	Who are main authors? What are academic structures?	[A] Social network (c.f. Wellman, 1988) and bibliometric analysis (c.f. De Solla Price, 1963) of publications from major conferences in the field of digital cultural heritage 1990-2015 (n=3917)
	What are topics?	[B] Automated topic mining of 3917 articles, manual classification of 452 articles plus 26 project reports via qualitative content analysis (c.f. Mayring, 2000)
	Who funds projects?	[C] Qualitative content analysis of 518 project activities in the field of digital cultural heritage including
	What marks a disciplinary culture?	[D] Three stage investigation including a questionnaire-based survey during three workshops with 44 participants to gain a general overview; 15 guideline based interviews with researchers to investigate research culture in depth (Mieg and Näf, 2005, Gläser and Laudel, 2009); online survey with 988 participant to quantify findings
Usage practices	What are phenomena and strategies for cooperation?	[E] 4 case studies: Data collection via expert interviews (c.f. Gläser and Laudel, 2009) and observation (c.f. Lamnek, 2005). Data analysis via heuristic frameworks (c.f. Kubicek, 1977) and grounded theory (c.f. Bryant and Charmaz, 2010)
	How to support cooperation in 3D modelling projects?	[F] Employment and evaluation of SCRUM as agile project management approach (Schwaber, 2004) in 2 educational project seminars so far with 13 student teams
Methodological development	What are current challenges?	[G] Three group discussions (c.f. Lamnek, 2005) on workshops at national/ international conferences (~60 participants) to examine; online survey with 650 participants
	How to systematize?	[H] Classification scheme developed and applied for 8 projects yet
Technologies	How to create 3D models?	[I] Development of workflows and toolsets to automatically create 3D models from historic photos [removed for reviewing], and semi-automatic creation from GIS data [removed for reviewing]
	How to improve user interaction with 3D models?	[J] Development and testing of 4D geo browsers; browser-based augmented and virtual reality interfaces for mobile devices
Perception	What factors are influencing perception of 3D models?	[K] 2 expert workshops and literature survey yet to identify influencing factors
	How are virtually represented structures perceived?	[L] Two studies to investigate how virtually represented structures and proportions are perceived, involving 21 persons and using usability testing (c.f. Nielsen, 1993)
Education	How to use 3D modelling techniques in education?	[M] 3 student seminars to develop and test team project-based learning approaches via formative & summative assessment (c.f. Dumit, 2012)

Table 1 - Investigational parts

Some results at a glance

What are some results at a glance? Considering a scholarly community on digital 3D reconstruction and modeling, discourses on major conferences during the last 25 years were mainly led by institutions from European Mediterranean countries, covering primarily technological topics. Especially statues and buildings in Mediterranean countries dating from all periods Anno Domini deliver rich content for such reconstruction. Due to the high complexity and team-based workflows, aspects and usage practices for communication, cooperation, and quality management are of high relevance within 3D reconstruction projects. Especially if people with different disciplinary backgrounds are involved, visual media are intensively used to foster communication and quality negotiations, for example by comparing source images and renderings of the created virtual reconstruction. Furthermore, several projects successfully adopted highly standardized conventions from architectural drawings for interdisciplinary exchange. To support a methodological development I ran five workshops to identify prospects and demands for further development, involving around 60 researchers and an online survey to verify findings from these workshops. Costs and training were named most frequently as currently pressuring issues. With regards to technologies, a big hurdle to overcome in order to use augmented and virtual reality is the current need to download and install additional software. Since current browser generations allow the visualization of 3D content natively, our focus is on user-friendly interaction concepts to access both, visualizations and underlaying informations. Regarding the perception of virtual 3D models relatively little visual information is needed to allow observers to distinguish buildings from each other or to identify a single building and to gain information about its spatial relation and shape [removed for reviewing]. Moreover, we adopted and evaluated team project-based learning approaches to support student education in digital 3D reconstruction. As observed in two courses so far, a development of procedures and strategies for cooperation within student project teams for creating virtual representations evolves slowly, and mostly as reaction of upcoming problems and demands. Related competencies are based highly on implicit knowledge and experience. As consequence, a teaching of best practices prior to a project work is less effective than coaching during the project work.

Next steps

What are next steps? Since 3D models in the humanities are primarily accessed via visualizations, a toolset for assessing visualization and interactivity of 3D models and presentations is currently missing and will be in focus of a next research stage. Many of the already completed investigations are of qualitative nature or focus on particular aspects. Consequently, a further validation for adjacent aspects as well as a verification of findings are alltime tasks. To proceed, further investigations on the scholarly use of 3D models and historical photographs or the design of interfaces for virtual museums are under development as well as a survey to further investigate challenges and perspectives of 3D modeling. Since the research is intended to enhance the validation and dissemination of 3D modeling technologies in the humanities both, education and organizational development are key issues. Beside the further development and establishment of teaching concepts and university courses, especially strategies for self-driven and scalable learning as MOOCs or open educational resources seems promising. Finally, beneficial and methodologically grounded best practice examples, an institutionalization of chairs and institutes as well as an increased awareness seem to be crucial for a further organizational establishment.

References

ARNOLD, D. & GESER, G. 2008. *EPOCH Research Agenda – Final Report,* Brighton.

BENTKOWSKA-KAFEL, A. 2013. Mapping Digital Art History. Available: https://bentkowska.files.wordpress.com/2013/05/annabentkowska-kafel__gettyd-ah-lab_2013.pdf.

BENTKOWSKA-KAFEL, A., DENARD, H. & BAKER, D. 2012. *Paradata and Transparency in Virtual Heritage,* Burlington, Ashgate.

BRYANT, A. & CHARMAZ, K. 2010. *The SAGE Handbook of Grounded Theory,* Thousand Oaks, SAGE.

DE SOLLA PRICE, D. 1963. *Little Science - Big Science,* New York, Columbia Univ. Press.

DUMIT, N. Y. 2012. *Diagnostic/Formative/Summative Assessment,* n.n.

EUROPEAN COMMISSION 2011. *Survey and outcomes of cultural heritage research projects supported in the context of EU environmental research programmes. From 5th to 7th Framework Programme,* Brussels, European Commision.

FRISCHER, B. 2008. *Beyond illustration : 2D and 3D digital technologies as tools for discovery in archaeology,* Oxford, Tempus Reparatum.

GLÄSER, J. & LAUDEL, G. 2009. *Experteninterviews und qualitative Inhaltsanalyse als Instrumente rekonstruierender Untersuchungen,* Wiesbaden, VS Verlag für Sozialwissenschaften.

KOHLE, H. 2013. *Digitale Bildwissenschaft,* Glückstadt.

KUBICEK, H. 1977. Heuristische Bezugsrahmen und heuristisch angelegte Forschungsdesigns als Element einer Konstruktionsstrategie empirischer Forschung. *In:* KÖHLER, R. (ed.) *Empirische und handlungstheoretische Forschungskonzeptionen in der Betriebswirtschaftslehre.* Stuttgart.

LAMNEK, S. 2005. *Qualitative Sozialforschung. Lehrbuch,* Weinheim.

MAYRING, P. 2000. Qualitative Content Analysis. *Forum Qualitative Sozialforschung,* 1**,** Art. 20.
MIEG, H. A. & NÄF, M. 2005. *Experteninterviews,* Zürich.
NIELSEN, J. 1993. *Usability Engineering,* Salt Lake City, Academic Press.
SCHWABER, K. 2004. *Agile Project Management with Scrum,* Redmond.
WELLMAN, B. 1988. Structural Analysis. From Method and Metaphor to Theory and Substance. *In:* WELLMAN, B. & BERKOWITZ, S. D. (eds.) *Social Structures: A Network Approach.* Princeton: Princeton University Press.

Question, Create, Reflect: A Holistic and Critical Approach to Teaching Digital Humanities

Kristen Mapes
kmapes@msu.edu
Michigan State University, United States of America

Matthew Handelman
handelm@msu.edu
Michigan State University, United States of America

Teaching digital humanities at the undergraduate level is as much about issues of critical theory, inclusion, and diversity as it is about teaching digital tools and methods. Examining DH methods such as topic modeling introduces students to the concept of algorithmic bias, pointing to the algorithms that shape our daily lives. Working with DH tools such as Palladio enables students to confront and reveal the layers of representation (and inequality) that structure the virtual and physical spaces that we inhabit. And creating digital archives with platforms such as Omeka challenges students to question the purpose and limits of digital tools, offering opportunities to reflect on the ethics of (digital) representation. The dialectics of teaching new DH tools and questions of critique, the archive, and representation central to the humanities form the basis of the undergraduate Digital Humanities Minor at our institution, in which students take two sequential, required courses: "Introduction to Digital Humanities" and the "Seminar in the Digital Humanities". Our talk will explore how we weave together these courses to create a holistic and critical approach to the foundations of digital humanities at the undergraduatelevel.

In "Introduction to Digital Humanities," students examine a range of DH methods and activities and create a final project of their own choosing. We explore DH approaches to humanities questions by evaluating digital projects that engage with the Harlem Renaissance and its context. By centering students' exposure to DH on one broad but unifying topic, we can avoid the trap of the carousel of tools into which an Intro DH class could fall. The Harlem Renaissance centers the course because it touches on cultural areas of critical interest spanning disciplines – art, music, literature, economic history, social history, political history, and urban planning – and has several DH projects either directly on the Harlem Renaissance or on related topics. By rooting the course in a historical cultural period, students are introduced to structural trends and issues that reverberate today.

In analyzing digital projects as a class, we critique the data behind the project, its presentation - in terms of style, effectiveness, and accessibility - and the structures in which it was made. We discuss what role grant funding plays in promoting certain types of projects, how crowdsourcing relates to labor ethics and the digital, who the project's users may be, and what its long term preservation prospects are. We then apply this critical framework to projects ranging from a digital edition (such as Claude McKay's Early Poetry) to large scale image analysis (such as On Broadway) to linked data and network analysis (such as Linked Jazz). We also talk to project leaders (from Virtual Harlem, Umbra Search, and the Mapping the Second Ku Klux Klan projects) to get a behind-the-scenes perspective on project management, origins, and goals. The bulk of the second half of the semester is spent on student projects. Students choose any topic they like and develop a critical research question. It is then is up to each student to choose a DH method and to find, gather, and clean their data. Class time is built in for one-on-one assistance from the professor and the embedded librarian to guide the students through the frustration and joy of the iterative DH project. By the end of the semester, the same digital project evaluation framework is used to analyze the students' projects.

The second semester in this year-long sequence, "Seminar in Digital Humanities," deepens students' skills with DH tools and methods, applies these skills in a semester-long DH project, and combines students' DH knowledge with the reflective practices of critical theory. As both "Text, Technology, and the Body" (spring 2016) and "Digital Humanities and Critical Theory" (spring 2017), students participate in a collaborative DH project, in which they design and build an online collection using archival materials from our institution's Special Collections as well as analyze and reflect on their digital work and the content of our archive. Whether it is digitizing criminology broadsheets from 17th Century Europe or early-twentieth century comics, this course frames DH as a continuation of - instead of a break with - critical debates over media, technology, and culture - from classics such as Walter Benjamin to current critical voices in DH such as Alan Liu and Laura Klein. The goal of these projects is not only to enable students to conceptualize and execute a student-led DH project, but also to develop their ability to read and critique digital tools and recognize their affordances, limitations, and political implications.

Exploring and employing a variety of digital techniques, "Seminar in the Digital Humanities" adapts and expands on the "read, play, build" approach to teaching DH proposed by Joanna Swafford at DH 2016. The semester is divided into seven units, the first two of which position DH within contemporary (and critical) debates in the humanities and introduce students to the historical and disciplinary context pertaining to our subject matter. For each unit, students read theoretical texts and articles that contextualize the tool under consideration as part of a larger historical-critical discourse within media studies, critical theory, and the history of DH. These readings provide the background in which students then learn how to implement these tools and explor examples aided by guest DH specialists from around our institution. The final phase of each unit provides a collaborative space for class members to create a working plan to apply this technique to our project - in order, for example, to clean our metadata, digitize our selected archival materials, and set up the Omeka site. Finally, students execute this plan as their individual project and compose a reflective essay that positions their work in the critical debates and comments on the technological, epistemological, and ethical choices that went into their digital work. These individual projects and critical reflections provide a self-reflective context for our digital collection, while allowing the students to cultivate their identities as critical thinkers and digital humanists.

Taken together, these two undergraduate courses expose students to a range of digital tools and methods for humanistic inquiry, providing them with experience overseeing their own DH project from conception to completion as well as participating in a semester-long team project. In different ways, the courses introduce students to critical frameworks for asking humanities questions of the digital and for using the digital to ask humanities questions. Teaching DH and critical thought as two sides of the same coin, this DH sequence provides students with tools to not only understand, but also intervene in a world increasingly mediated by digital processes.

"Smog poem". Example of data dramatization

Piotr Marecki
piotr.marecki@ha.art.pl
Jagiellonian University, Poland

Leszek Onak
leszek.onak@gmail.com
Jagiellonian University, Poland

The proposed poster is a visual presentation of the literary experiment "Smog Poem" (2018) by a Polish poet Leszek Onak developed in the UBU lab at the Jagiellonian University run by dr Piotr Marecki. Drawing on the terminology of expressive processing developed by Noah Wardrip-Fruin, and platform studies by Nick Montfort and Ian Bogost the authors of the poster present the process of the creation of the work.

According to World Health Organization ambient particle pollution kills about 6.5 million people annually affecting all regions of the world. In Poland, it amounts to nearly 50 thousand deaths each year. Krakow, the former capitol of Poland, is ranked third among the European cities with the highest levels of particulate matter (PM 10).

"Smog Poem" is a text and graphics generator that uses the data on the environmental pollution to change the tissue of the text, its graphic elements, and other components depending on the pollution's intensity. The algorhythm has a form of an internet browser plugin; after its installation, the users browsing through the internet will experience the air pollution in front of their own eyes through the glitches appearing on the websites they use, the replacement of the photos and text modification. Some articles will be replaced by a separate generated text based on the syntactic mechanisms and by using the rules of the "Game of Life" by John Conway.

The piece consists of two main engines. One mechanism is pulling data on the actual air pollution with Particulate Matter (PM 10 and PM 2.5), Nitrogen Dioxine NO2, Sulfur Dioxide SO2 and Carbon Monoxide CO. Each of those indicators is responsible for a different element distorting the content. The second mechanism is responsible for the upload of data from the websites and its modification depending on the particle pollution. If the air quality does not exceed the norms, the content of websites remains unchanged.

The algorhythm is representative of the growing trend of digital art based on resources and presenting them in a way to influence the user's consciousness. It refers to the concept of 'data dramatization' by Liam Young, who once said: 'Data Dramatization, as opposed to data visualization presents a data set with not only legibility or clarity but in such a way as to provoke an empathetic or emotive response in its audience.'

„Smog poem" is one of the of the digital works developed in the UBU lab at the Jagiellonian University. The lab primarily produces digital works that can function in a few fields of the demoscene, electronic literature, video games and media art. The research conducted in the lab focuses on, among other things, local phenomena in the digital media field, e.g. strategies for cloning platforms in Central and Eastern Europe, as well as the digital genres and their specific features in Central and Eastern Europe. The artists, programmers and scholars affiliated with the lab develop new genres and communication practices (technical reports, open notebook science) to describe the creative process in its widest definition in the era of digital textuality. The project has been made possible through the support of the Polish Ministry of Science and Higher Education "National Programme for the Development of Humanities".

ANJA, ¿dónde están los encabalgamientos?

Clara Martinez-Canton
cimartinez@flog.uned.es
LINHD, UNED, Spain

Pablo Ruiz-Fabo
pablo.ruiz@linhd.uned.es
LINHD, UNED, Spains

Elena González-Blanco
egonzalezblanco@flog.uned.es
LINHD, UNED, Spain

Introducción

Encabalgamiento es el desajuste entre la pausa métrica y la sintáctica (Domínguez Caparrós, 2000: 103) que ocurre cuando una unidad de sentido se rompe entre dos versos. Este fenómeno, desde siempre utilizado con distintos fines expresivos (énfasis, ambigüedad, etc.) es difícil de delimitar formalmente.

El estudio más sistemático realizado para su caracterización en español sigue siendo el realizado en su tesis por Quilis (1964). El estudioso experimentó con lecturas de prosa, buscando demostrar qué unidades sintácticas no permiten pausa de sentido en su interior. Basándose en los resultados definió una serie de categorías gramaticales y sintácticas cuya separación en versos distintos produce encabalgamiento. La tipología allí establecida se considera ya clásica. El estudio de Quilis proporciona una definición formal y empírica del fenómeno. Con base en sus reglas se ha creado una herramienta capaz de detectar el encabalgamiento y sus tipos.

Este póster presenta la interfaz ANJA para el análisis automático del encabalgamiento desde una sencilla aplicación web: http://prf1.org/anja/index/, desarrollada dentro del proyecto ERC POSTDATA GA- 679528[1].

Estado del arte

La naturaleza formal del análisis métrico lo hace un campo propicio para su tratamiento computacional (Birnbaum and Thorsen, 2015; Delente and Renault, 2015). El procesamiento del lenguaje natural (PLN) ofrece muchas posibilidades para la métrica, pues las reglas de definición lingüística permiten llevar a cabo análisis y extracción automática de grandes cantidades de información de corpus textuales.

Para la automatización del análisis métrico en español destacamos los estudios de escansión silábica y acentual de Navarro-Colorado (2017), Agirrezabal (2017) y Gervás (2000). También los trabajos de generación automática de poesía con patrones métricos (Gervás, 2000b) y (Gervás, 2015).

En el campo de las interfaces cabe distinguir entre aquellas que exploran datos de textos ya analizados, recogidos en una base de datos, y aquellas que permiten la entrada y análisis de cualquier poema. Del primer tipo destacamos For Better For Verse[2] (Tucker, 2011) y Database of Czech Metre[3] (Plecháč and Kolár, 2015). Entre las que permiten introducir textos destacamos, en español, la ligada a la herramienta de Navarro-Colorado[4], que analiza versos endecasílabos. Otros sitios con interfaz de entrada para análisis métrico son Separarensílabas[5] o Lexiquetos[6]. En otras lenguas destacamos Metricalizer[7] (Bobenhausen and Hammerich, 2015) para alemán, Aoidos[8] (Mittmann, 2016) para portugués y español, y RhymeDesign[9] (McCurdy et al., 2015) especializado en rima en inglés.

Una interfaz para el análisis del encabalgamiento representa, sin embargo, una novedad en el campo.

Herramienta y resultados

El programa de detección del encabalgamiento en español, basado en PLN, se desarrolló en 2016-2017 y fue evaluado sobre dos corpus de test de distintos periodos (Ruiz et al., 2017). ANJA proporciona una interfaz web simple para este programa. El sistema consta de tres componentes: módulo de preprocesado para uniformar el formato de los poemas, pipeline de PLN (basada en IXA Pipes (Agerri et al., 2014) para POS-tagging, constituyentes y dependencias sintácticas) y módulo de detección de encabalgamiento (basado en reglas y diccionarios) y ampliamente documentado en el sitio web[10]. Se ha utilizado esta herramienta para etiquetar un corpus de más de 4000 sonetos alojado y documentado en https://github.com/postdataproject/disco.

El código de la herramienta de detección de encabalgamientos está disponible en https://bitbucket.org/pruizf/anja_public/.

Interfaz gráfica de usuario

ANJA es una interfaz pública y gratuita, alojada en: http://prf1.org/anja/index/. Permite cargar los poemas que el

1 Este trabajo se enmarca dentro del proyecto de investigación Starting Grant Poetry Standardization and Linked Open Data: POSTDATA (ERC-2015-STG-679528), financiado por el European Research Council (ERC) bajo el programa: European Union´s Horizon 2020 research and innovation programme, dirigido como Investigador Principal por la profesora Elena González-Blanco, LINHD UNED (http://postdata.linhd.es/).

2 http://prosody.lib.virginia.edu/
3 http://versologie.cz/v2/web_content/
4 http://adso.gplsi.es/index.php/es/demostracion/
5 http://www.separarensilabas.com/index.php
6 http://lexiquetos.org/silio/
7 https://metricalizer.de/en/metrikanalyse/poem
8 http://aoidos.ufsc.br/
9 http://www.sci.utah.edu/~nmccurdy/rhymeDesign/
10 https://sites.google.com/site/spanishenjambment/

usuario decida y analizarlos en el momento. También ofrece la carga de archivos ZIP que contengan archivos en texto plano.

La interfaz de usuario está construida con el framework Django (Python), con las plantillas de Bootstrap 3. Las vistas de Django se llaman con AJAX para poblar los elementos de la UI. Para el análisis de PLN, Django accede a servicios web Java (IXA Pipes) implementados en nuestro servidor.

ANJA presenta dos ventanas de navegación (Fig. 1), la principal, para introducir poemas, a la derecha y, a la izquierda, una mínima guía de uso que explica su funcionamiento y enlaza a la web del proyecto.

Captura de ANJA

Los resultados se ofrecen dos formatos: *Standoff* (tipo de encabalgamiento y línea), e *Inline* (etiquetado gramatical y tipo de encabalgamiento por línea, ver Fig. 2 para *Inline*). Las anotaciones PLN en que se basa en sistema se ofrecen en las pestañas *PosTags* (etiquetas gramaticales) y *FullNLP* (pipeline completa).

El enlace **legend**[11] da acceso a la leyenda que explica los tipos de encabalgamiento, las etiquetas gramaticales y otras convenciones de representación:

Standoff | Inline | PosTags | FullNLP [legend]

#	Text	Position	Enjambment Type
1	{Ya A} {se Q} {doran V} {las D} {hojas N} {, O} {ya A} {las D} {tardes N}	B	ex_subj_verb
2	{Son V} {lo D} {mejor G} {del P} {día N}	I	ex_subj_verb
3	{Bajo P} {su D} {luz N} {madura G} {de P} {sosiego N} {. O}	O	
4	{Ya A} {comienza V} {el D} {curso N} {. O}	O	
5	{Aprender V} {, O} {enseñar V} {, O} {lecciones N} {, O} {aulas N} {. O}	O	
6	{Siempre A} {en P} {octubre O} {comenzaba V} {el D} {año N} {. O}	O	
7	{¡ O} {Y C} {cuántas Q} {veces N} {esa D} {luz N} {de P} {otoño N}	O	

Anotaciones de encabalgamiento en formato *Inline*

La existencia de una aplicación web simple para la utilización esta herramienta la hace accesible para una gama mucho más amplia de usuarios.

References

Agerri, R., Bermudez, J. and Rigau, G. (2014). IXA pipeline: Efficient and Ready to Use Multilingual NLP tools. *Proceedings of LREC 2014, the 9th International Language Resources and Evaluation Conference*, vol. 2014. Reykjavik,Iceland, pp. 3823–3828

11 https://sites.google.com/site/spanishenjambment/legend

http://www.lrec-conf.org/proceedings/lrec2014/pdf/775_Paper.pdf (accessed 20 April 2017).
Agirrezabal, M. (2017). Automatic Scansion of Poetry San Sebastián/Donosti: Universidad del País Vasco.
Birnbaum, D. J. and Thorsen, E. (2015). Markup and meter: Using XML tools to teach a computer to think about versification. *Balisage: The Markup Conference* http://www.balisage.net/Proceedings/vol15/print/Birnbaum01/BalisageVol15-Birnbaum01.html (accessed 22 April 2017).
Bobenhausen, K. and Hammerich, B. (2015). Métrique littéraire, métrique linguistique et métrique algorithmique de l'allemand mises en jeu dans le programme Metricalizer2. *Langages*(3): 67–88.
Delente, É. and Renault, R. (2015). Outils et métrique: un tour d'horizon. *Langages*(3): 5–22.
Domínguez Caparrós, J. (2000). *Métrica Española*. Madrid: Síntesis.
Gervás, P. (2000a). A Logic Programming Application for the Analysis of Spanish Verse. *Computational Logic—CL 2000*. Berlin: Springer Berlin Heidelberg, pp. 1330–44.
Gervás, P. (2000b). Wasp: Evaluation of different strategies for the automatic generation of spanish verse. *Proceedings of the AISB-00 Symposium on Creative & Cultural Aspects of AI*. pp. 93–100 https://www.researchgate.net/profile/Pablo_Gervas/publication/228609235_Wasp_Evaluation_of_different_strategies_for_the_automatic_generation_of_spanish_verse/links/00b4952aada6407047000000.pdf (accessed 22 April 2017).
Gervás, P. (2015). Tightening the Constraints on Form and Content for an Existing Computer Poet. *AISB Convention 2015* http://eprints.sim.ucm.es/37000/ (accessed 22 April 2017).
McCurdy, N., Srikumar, V. and Meyer, M. (2015). Rhymedesign: A tool for analyzing sonic devices in poetry. *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*. pp. 12–22.
Mittmann, A. (2016). Escansão automática de versos em português. https://repositorio.ufsc.br/handle/123456789/175819.
Navarro-Colorado, B. (2017). A metrical scansion system for fixed-metre Spanish poetry. *Digital Scholarship in the Humanities* doi:10.1093/llc/fqx009. https://academic.oup.com/dsh/article-abstract/doi/10.1093/llc/fqx009/3064339/A-metrical-scansion-system-for-fixed-metre-Spanish (accessed 19 April 2017).
Plecháč, P. and Kolár, R. (2015). The Corpus of Czech Verse. *Studia Metrica et Poetica*, 2(1): 107–118.
Quilis, A. (1964). *Estructura Del Encabalgamiento En La Métrica Española*. Consejo Superior de Investigaciones Científicas, patronato' Menéndez y Pelayo,' Instituto' Miguel de Cervantes,'.
Ruiz Fabo, P., Bermúdez Sabel, H., Martínez Cantón, C. I., González-Blanco, E. and Navarro-Colorado, B. (2018). The Diachronic Spanish Sonnet Corpus (DISCO): TEI and Linked Open Data Encoding, Data Distribution and Metrical Findings. Humanidades Digitales 2018 (DH 2018). Ciudad de México, México.
Ruiz Fabo, P., Bermúdez-Sabel, H., Martínez Cantón, C. I. and Calvo Tello, J. (2017). *Diachronic Spanish Sonnet Corpus (DISCO). Madrid: UNED.* Madrid https://doi.org/10.5281/zenodo.1012567.
Ruiz, P., Martínez Cantón, C., Poibeau, T. and González-Blanco, E. (2017). Enjambment Detection in a Large Diachronic Corpus of Spanish Sonnets. *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Association for Computational Linguistics, pp. 27–32.
Tucker, H. F. (2011). Poetic data and the news from poems: A for better for verse memoir. *Victorian Poetry*, 49(2): 267–281.

Combining String Matching and Cost Minimization Algorithms for Automatically Geocoding Tabular Itineraries

Rui Santos
ruipds@gmail.com
IST and INESC-ID,
University of Lisbon, Portugal

Bruno Emanuel Martins
bruno.g.martins@ist.utl.pt
IST and INESC-ID,
University of Lisbon, Portugal

Patricia Murrieta-Flores
p.murrietaflores@chester.ac.uk
Digital Humanities Research Center,
University of Chester, United Kingdom

Historical itineraries, often accessible as tables or as sequential lists of names for the places visited in the context of a particular journey, are abundant resources and also important objects of study for Humanities scholars, providing 'snapshots' of particular socio-cultural events, insights into the development of human mobility, and invaluable information related to the establishment of road networks. Well-known examples include the 3rd century *Itinerarium Antonini Augusti* or the *Itinerarium Burdigalense,* written between the 8th and 10th centuries, among others. Many historical manuscripts and/or transcriptions containing information on itineraries, dating from the medieval period to the 20th century, are nowadays available in digital formats, through initiatives such as Europeana or the Internet Archive, or in the context of Digital Humanities projects like Pelagios.

Few historical tabular itineraries are nonetheless directly associated with map-based representations and, in many cases, there is little information on the actual routes

taken in between locales. As such, there are many interesting questions related to early traveling routes, in need of further study. We believe that the analysis of historical itineraries (e.g., for consistency checking, or enabling new inquires/inferences about the routes) can be facilitated through the analytical tools of Geographical Information Systems (GIS) and/or through map-based representations. The research reported in this poster concerns with automatically geocoding historical itineraries, leveraging innovative methods that explore the idea that travelers tend to choose the most efficient routes (e.g., itineraries will likely minimize the distance between locations).

In brief, we propose an automated method for geocoding tabular itineraries based on a sequence of four stages, combining string similarity search and well-known optimization procedures (Santos et al., 2017b). On the first stage, we use string similarity to look for candidate disambiguations in a large-coverage gazetteer. State-of-the-art string matching methods (Santos et al., 2017a, 2018), leveraging supervised learning, can then optionally be used to further filter/restrict the set of disambiguation candidates. A least-cost path between pairs of candidates, visited in sequence over the itinerary, is afterwards estimated on the third stage. We tested geodesic paths over the Earth's surface, or least-cost path calculations (Douglas, 1994) leveraging terrain slope and land-coverage for estimating movement costs. Finally, Step 4 leverages the distance associated to each of the paths between candidate pairs, computed in Stage 3, to find an overall best path for the entire itinerary, also disambiguating each of the toponyms to the most likely candidate. A dynamic programming algorithm similar to Viterbi decoding (Forney, 1973) is used at this stage to efficiently compute the global path that minimizes the traveled distance.

The proposed method was tested with manually geocoded itineraries (e.g., measuring the distance between the estimated disambiguation and ground-truth geo-spatial coordinates for the places in each itinerary). We relied on a dataset of well-known European historical itineraries (see http://www.peterrobins.co.uk/itineraries/list.html), containing 24 instances corresponding to sequences of varied lengths. We also used the GeoNames gazetteer for supporting the disambiguation of toponyms into geo-spatial coordinates, i.e. a resource which focuses on the modern administrative geography that nonetheless lists many historical variants as alternative place names. Our experiments showed that while approximate string matching can already achieve very low median errors (e.g., many of the itinerary toponyms match exactly with entries in GeoNames, and thus the median distance towards the correct disambiguations is quite low), the combination with cost optimization can significantly improve results in terms of the average distance. Moreover, using Least-Cost Paths (LCPs) for reconstructing the most likely routes can enable new inquiries and inferences. Although LCP analysis is commonly used within computational archaeology (Murrieta-Flores, 2012), the application that is reported through this poster is particularly innovative.

Our work shows that methods leveraging the intuition that travelers tend to choose the least-costly routes, in combination with approximate string matching for finding gazetteer entries that corresponding to historical toponyms, are indeed effective for automatic geocoding. We focused on the validation of the automated method but we believe that, if implemented within plugins for popular GIS environments, the proposed ideas can effectively help Humanities scholars in the analysis of data pertaining to historical itineraries.

Figure 1 - Ground-truth trajectory for the pilgrimage of Jehan de Tournay from Valenciennes to Venice (left), compared to the estimated trajectory for the same itinerary (right).

Acknowledgements

This research was supported by the Trans-Atlantic Platform for the Social Sciences and Humanities, through the Digging into Data project with reference HJ-253525. The researchers from INESC-ID also had financial support from Fundação para a Ciência e Tecnologia (FCT), through the INESC-ID multi-annual funding from the PIDDAC program, (UID/CEC/50021/2013)

References

Douglas, D. H. (1994). Least-cost Path in GIS Using an Accumulated Cost Surface and Slopelines. Cartographica: *The International Journal for Geographic Information and Geovisualization*, 31(3).

Forney, G. D. (1973). *The Viterbi Algorithm. Proceedings of the IEEE*, 61(3).

Murrieta-Flores, P. (2012). *Traveling through past landscapes - Analyzing the dynamics of movement during Late Prehistory in Southern Iberia with spatial technologies*. Ph.D. Dissertation, University of Southampton.

Santos, R., Murrieta-Flores, P. and Martins, B. (2017a). Learning to combine multiple string similarity metrics for effective toponym matching. *International Journal of Digital Earth*.

Santos, R., Murrieta-Flores, P. and Martins, B. (2017b). *An Automated Approach for Geocoding Tabular Itineraries. Proceedings of the ACM Workshop on Geographic Information Retrieval*. New York: ACM Press.

Santos, R., Murrieta-Flores, P., Calado, P. and Martins, M. (2018). Toponym Matching Through Deep Neural Networks. *International Journal of Geographical Information Science*, 32(2)

How We Became Digital? Recent History of Digital Humanities in Poland

Maciej Maryl
maciej.maryl@ibl.waw.pl
Institute of Literary Research of the Polish Academy of Sciences, Poland

Digital humanities suddenly erupted in Poland in the second decade of the 21st Century: first digital humanities centres were established (2013-2015); Poland joined important European networks and consortia like CLARIN (2013), NeDiMAH (2014), DARIAH (2015), or OPERAS (2017) while establishing national consortia CLARIN-PL (2013), and DARIAH-PL (2015); finally, it hosted important international conferences: CLARIN 2015 in Wrocław and ADHO's Digital Humanities 2016 in Kraków. Yet, this sudden eruption by no means marks the beginning of DH in Poland. The first digital projects in the humanities could be traced backed to early 2000s as the data collected in the survey by Werla & Maryl (2014) suggest. Those events should then be understood as landmarks in the process of the institutionalization of digital humanities in Polish scholarship.

This paper explores the specificity of digital humanities in Poland through the analysis of the events and projects which lead to this institutionalization. As O'Sullivan et al. 2015 point out "Tracing the emergence of academic disciplines in a national context is a useful undertaking, as it goes beyond the definition of a field to an assessment of its evolution within a more specific cultural context." They also claim that the emergence of the field is closely connected to the social as well as economic trends. It is true for Poland, where humanities computing evolved slowly due to technological deficiencies and budgeting problems. Moreover, Polish humanities in the 1990s (especially in the field of literature, culture and history) were also preoccupied with removing the "white spots", i.e. conducting research on topics that could not have been accounted for before 1989 for political reasons. On the other hand, when discussing the development of DH in a country which was hardly a forerunner of digital methods, but rather its late adopter, heavily influenced by the experiences of foreign institutions, it is extremely difficult to pinpoint the regional specificity of digital research practices (cf. Schreibman 2012). Is there any local flavour of the practices, materials, or tools selected? Does it go beyond mere linguistic differences? Are region-specific research questions being asked?

The discussion will be based on selected projects (Werla & Maryl 2014), conferences, as well as on the observations of the forming phase of DARIAH-PL consortium (2013-2015), which would serve as a case-study. The issue of national specificity of DH in Poland in comparison to other European countries will be addressed in the light of the results of DARIAH VCC2 survey on digital methods (Dallas et al. 2017), conducted in 2014-15 by the Digital Methods and Practices Observatory (DiMPO) Working Group of DARIAH-EU. The discussion will be informed by Roopika Risam's concept of "DH accent" which allows to account for "both local specificity and global coherence in DH" (2017:378).

Although the authors of *Digital_Humanities* claim that "The mere use of digital tools for the purpose of humanistic research and communication does not qualify as Digital Humanities" (ibid.) The results of DARIAH VCC2 survey on digital methods and tools in the humanities show that the application of digital methods in the humanities is gradual. The tools like word processors, web search engines and various online resources (digital libraries, archives, journals) are widely adopted. Yet, a bit more advanced tools (e.g. bibliography managers or specialized note-taking applications) are relatively less popular. And there are still some types of applications (e.g. databases,

Content-Management-Systems, or use of social media in scholarly practices) which are used only by a small group of scholars. Therefore being a digital humanist means placing oneself on the scale ranging from the basic tools nearly all of us use to the most advanced stage on which new methods and software capacities enable us to pose completely new research questions (or to answer the old ones in a fundamentally different manner).

This process of *becoming* digital, i.e. adopting digital methods and practices by scholars in the humanities, will be analysed through the conceptual framework of "three waves" of digital humanities: (1) early remediation of traditional methods of scholarly inquiry (cf. Svensson 2009); (2) taking the advantage of the new medium in creating new methods and genres (Pressner 2011; Davidson 2008; Svensson 2010) (3) critical scrutiny of the epistemic constraints of the medium (Berry 2011, Rogers 2015). Those waves, although sometimes understood chronologically, are here considered as co-occuring in a DH community.

Polish sample of the DARIAH survey does not differ greatly in terms of the digital tools applied by scholars in comparison to the European sample. They use less often bibliography managers or personal databases, but Polish results seem to be rather consistent with European sample, what – in turn – shows that Polish DH, although developed beyond the existing networks, show similar patterns of growth. There are however important differences in terms of disciplinary background, career status and perceived needs of the Polish scholars, who were more interested in enhancing their existing research practices (improved access to the sources or software, networking), and are less open to new methods and tools (advice, courses, support options).

By means of such comparative perspective this paper engages with the conference topic, discussing how digital approaches may be instrumental in building 'bridges' between various research communities, which in turn may contribute to levelling the differences with regards to centres and peripheries of contemporary DH. Understanding the tension between local and transnational initiatives is important to capture the specificity of Polish DH, which could be viewed also as a heavily institution-related. Poland participates in CLARIN and DARIAH, yet Polish scholars are not that active in ADHO (there is no Polish Association of DH). Given the emerging national and international DH initiatives in Eastern Europe, as well as the plans to establish DARIAH Hub for the region, it may be a good moment to reflect on the interplay of regional and external factors of this process. A better understanding of how we have become digital humanists, offered here on the example of Poland, may inform those initiatives.

References

Berry, D.M. (2011). The Computational Turn: Thinking About the Digital Humanities. *Culture Machine,* vol. 12 , https://www.culturemachine.net/index.php/cm/article/view/440/470.

Dallas, C., Chatzidiakou, N., Benardou, A., Bender, M., Berra, A., Clivaz, C., Cunningham, J., et al. (2017). *European Survey on Scholarly Practices and Digital Needs in the Arts and Humanities - Highlights Report*. Zenodo. doi:10.5281/zenodo.260101.

Davidson, C. N. (2008) "Humanities 2.0: Promise, Perils, Predictions". *Publications of the Modern Language Association of America (PMLA)* 123(3), 707-717.

O'Sullivan, J., Murphy, O. and Day, S. (2015). The Emergence of the Digital Humanities in Ireland. *Breac: A Digital Journal of Irish Studies*, https://breac.nd.edu/articles/the-emergence-of-the-digital-humanities-in-ireland/

Presner, T. (2010). Digital Humanities 2.0: A Report on Knowledge. *OpenStax CNX. 8* http://cnx.org/contents/2742bb37-7c47-4bee-bb34-0f35bda760f3@6

Risam, R. (2017). Other worlds, other DHs: Notes towards a DH accent. *Digital Scholarship in the Humanities*, 32(2), 377-384.

Rogers, R. (2015). *Digital methods*. Cambridge: MIT press.

Schreibman, S. (2012). Controversies around the Digital Humanities. *Historical Social Research / Historische Sozialforschung,* 37(3):141, 46-58.

Svensson, P. (2009). Humanities Computing as Digital Humanities. *Digital Humanities Quarterly* 3: 3.

Svensson, P. (2010). "The Landscape of Digital Humanities" *Digital Humanities Quarterly* 4:1.

Werla, M., and Maryl, M. (2014). *Humanistyczne projekty cyfrowe w Polsce*. Poznań-Warszawa, http://lib.psnc.pl/publication/831.

Hacia la traducción automática de las lenguas indígenas de México

Jesús Manuel Mager Hois
mmager@turing.iimas.unam.mx
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Mexico

Ivan Vladimir Meza Ruiz
ivanvladimir@turing.iimas.unam.mx
Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Mexico

Introducción

En México existen 68 lenguas indígenas oficialmente reconocidas (Diario oficial, 2013). Esta riqueza lingüística forma parte del mosaico multicultural que define la identidad de nuestro país. Sin embargo, la predominancia cultural del español y el rezago generalizado del acceso a las tecnologías de información (Sandoval-Forero, 2013) por parte de los hablantes de estas lenguas crea barreras

culturales que dificultan la transferencia del conocimiento entre los pueblos indígenas.

En los últimos años se ha consolidado el campo de traducción automática. Parte de la consolidación de la traducción automática se debe a la traducción estadística (SMT) (Koehn, 2009; Lopez, 2008). Ésta metodología usa ejemplos de oraciones en ambas lenguas (corpus paralelos) para determinar los parámetros de un modelo estadístico que permite tal traducción. Adicionalmente, en los últimos años se han abierto paso a los modelos de traducción automática basados en redes neuronales (NMT) (LeCun *et al.*, 2015), los cuales permiten traducción multilingüe, en donde se crea un modelo de traducción común entre múltiple lenguas, el cual se utiliza posteriormente para mejorar la traducción entre pares de lenguas (Cho *et al., 2014*).

Metodología y resultados

En este proyecto presentamos nuestros avances en la creación de traductores automáticos para cinco lenguas indígenas al español: wixarika, náhuatl, yorem nokki, purépecha y mexicanero. Para obtener una visión completa sobre el campo decidimos hacer una comparación entre SMT y NMT. En ambos casos entrenamos los modelos usando segmentación morfológica que ha mostrado mejores resultados para lenguas polisintéticas (Mager, et. Al, 2016).

Para SMT fue utilizado el traductor por frases MOSES (Kohen, *et. al*., 2007) junto con el alineador GIZA++ (Och y Ney, 2003). Para los experimentos de NMT fue utilizado el modelo neuronal Codificador-Decodificador (Seq2Seq) con Redes Neuronales Recurrentes Bidireccionales (BiRNN) y con celdas de Unidades Recurrentes con Compuertas (GRU) (Cho., *et. al.*, 2014). Las pruebas fueron llevadas a cabo con OpenNMT (Klein, *et. al.*, 2017) con un corpus que consta de 985 frases traducidas a los 5 idiomas y que incluyen notación morfológica (Gómez y López, 1999; Chamoreau, 2003; Freeze, 1989; Lastra, 1980). Cada modelo ha sido evaluado de manera automática usando Bilingual Evaluation Understudy (BLEU) (Papineni, *et. al.*, 2002), y su salida fue valorada de manera manual, de tal manera que ha sido posible identificar los retos y limitaciones de las propuestas.

	NMT	SMT
Mexicanero-Español	2.95	23.47
Náhuatl-Español	3.04	10.14
Purépecha-Español	0	5.38
Wixarika-Español	0	0
Yorem Nokki-Español	0	2.44

Tabla 1: BLEU de los resultados experimentales de traducción de los cinco pares de idiomas con NMT y SMT

Como podemos ver en la tabla 1, los resultados de SMT superan los de NMT debido al corpus tan reducido con que se entrenaron. Mexicanero y náhuatl tuvieron un mejor desempeño que el wixarika, dado que el wixarika es una lengua con morfología con mayor cantidad de morfemas por palabra que el náhuatl (Kann, et. al. 2018).

Discusión

Si bien, se lograron mejorar las traducciones de manera importante, estos no son suficientes para ser usadas en la práctica cotidiana de manera autónoma o para asistencia humana. A través del desarrollo de estos traductores que hemos identificado los siguientes retos:

- **Escasez de los recursos.** Para poder generar un traductor automático es necesario contar con cientos de miles de pares de oraciones entre las dos lenguas; sin embargo, el poco uso de tecnologías de las comunidades nativo hablantes hace difícil la construcción de este corpus.
- **Complejidad morfológica**. Dada la naturaleza polisintética de estas lenguas, se necesita mejorar la segmentación morfológica automática para evitar la dispersión de datos (Kann, et. al. 2018).
- **El español es una lengua distante a los idiomas indígenas** que, en su gran mayoría tienen una topología morfológica polisintética, a diferencia del español que es fusionante y con orden Sujeto-Verbo-Objeto.
- La **falta de estandarización de la ortografía de las lenguas y el amplio espectro dialectal interno en las lenguas**.

Conclusiones

El presente trabajo expone primeros avances en traducción automática de cinco lenguas indígenas al español con SMT y NMT, identificando retos y limitaciones. Para trabajos futuros planteamos; mejorar el análisis y la segmentación morfológica de las lenguas indígenas, dada la fuerte correlación entre traducción y calidad de segmentación; la generación de corpus paralelos sintéticos a partir de modelos de aumento de datos; y la recopilación de más datos paralelos escritos para todos los idiomas indígenas trabajados, además de incorporar más idiomas.

References

Bahdanau, D., Cho, K., y Bengio, Y. (2014). 'Neural machine translation by jointly learning to align and translate'. *arXiv preprint arXiv:1409.0473*.

Canger, U. (2001). *Mexicanero de la sierra madre occidental*. El Colegio de México.

Chamoreau, C. (2003). *Purépecha de Jarácuaro* (p. 162). El Colegio de México.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., y Bengio, Y. (2014). Learning

Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724-1734.
Diario Oficial. (2014) Programa Especial de los Pueblos Indígenas 2014-2018, *Diario Oficial de la Federación*, México, Distrito Federal, 20 de abril.
Freeze, R. A. (1989). *Mayo de Los Capomos, Sinaloa (Mayo of Los Capomos, Sinaloa)*.
Gómez, P., & López, P. G. (1999). *Huichol de San Andrés Cohamiata, Jalisco* (Vol. 22). El Colegio de México.
Kann, K., Mager, M., Meza, I. Schütze, H. (2018) Fortification of Neural Morphological Segmentation Models for Polysynthetic Minimal-Resource Languages *16th Annual Conference of NAACL-HLT 2018*, New Orleans, Louisiana, US.
Klein, G., Kim, Y., Deng, Y., Senellart, J., y Rush, A. (2017). OpenNMT: Open-Source Toolkit for Neural Machine Translation.*En Proceedings of ACL 2017, System Demonstrations*, pp. 67-72.
Koehn, P. (2009). *Statistical machine translation*. Cambridge: Cambridge University Press.
Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., ..., y Dyer, C. (2007) Moses: Open source toolkit for statistical machine translation. *En Proceedings of the 45th annual meeting of the ACL*. Association for Computational Linguistics, pp. 177-180.
Lastra de Suárez, Y. (1980). Náhuatl de Acaxochitlán (Hidalgo). *Archivos de lenguas indígenas de México. DF: Colegio de México*.
LeCun, Y., Bengio, Y., y Hinton, G. (2015). Deep learning. *Nature*, 521(7553): 436.
Lopez, A. (2008). Statistical machine translation. *ACM Computing Surveys (CSUR)*, 40(3): 8.
Mager Hois, J. M., Barrón Romero, C., y Meza Ruiz, I. V. (2016). Traductor estadístico wixarika-español usando descomposición morfológica. *Memorias de COMTEL*. Lima, Perú,
Och, F. J., y Ney, H. (2003). A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1): 19-51.
Papineni, K., Roukos, S., Ward, T., y Zhu, W. J. (2002). BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pp. 311-318.
Sandoval-Forero, E. A. (2013). Los indígenas en el ciberespacio. *Agricultura, sociedad y desarrollo*, 10(2): 235-256.

Towards a Digital History of the Spanish Invasion of Indigenous Peru

Jeremy M. Mikecz
mikecz@usc.edu
University of Southern California, United States of America

What role did indigenous activity play in shaping the events of 'conquest'? How can digital tools aid in the reconstruction of this activity? These are the core questions driving my research on the Spanish invasion of Peru.

My research experiments with the use of digital methods to assist in the rewriting of indigenous history during the early period of European invasion. In this poster, I will introduce some of these digital methods – particularly the use of data and geo-visualizations to a) identify gaps and silences in colonial sources, and b) to fill in some of those gaps with information recovered from indigenous sources.

These methods draw on diverse inspirations. In reconstructing the hidden geography or spatiality of historical texts, it follows literary geographers' recent innovative mapping of fictional sources.(Cooper et al., 2016; Cooper and Gregory, 2011; Moretti, 1998) In visualizing and recognizing patterns within these sources through the creation of a wide variety of data visualizations, it draws on work ranging from nineteenth-century information graphics to twenty-first-century data science. Finally, it also finds inspiration in pioneering work in Historical GIS, spatial history, and qualitative and even indigenous cartography. (Eltis and Richardson, 2015; Knowles et al., 2014; For indigenous cartography, see the work of Margaret Wickens Pearce, including: Pearce and Hermann, 2010)

The role of geography and indigenous activity in European invasions of the indigenous Americas – first elided or erased by colonial authors – has remained largely overlooked by modern scholars. In Inka Peru, Spanish conquistadors encountered a complex imperial infrastructure and labor system that mitigated much of the geographic challenges of an invasion of the Andes. Native guides showed them the way, native informants advised them on potential obstacles to their journey, native allies offered military and political support, native messengers relayed information between the Spanish and their allies, native porters carried their supplies, and native villagers provided them with lodging and support.

While recent work – especially increased use of indigenous sources – has begun to reconstruct some of this activity(Matthew and Oudijk, 2007), I propose a new methodology to more fully reconstruct indigenous geography and activity and to present an alternate vision of the invasion of Peru. This is accomplished in two steps. First, I use digital text analysis methods to examine how colonial sources hid or erased indigenous activity. Second, I use geovisualizations to reconstruct indigenous activity in conquest-era events as it played out across space and time. This reconstruction of indigenous activity draws on a diverse range of indigenous sources. These include: 1) indigenous polities' petitions to the Crown documenting the service they provided the conquistadors during the invasion, 2) *cacicazgo* cases which document an indigenous group's history (for disputes over hereditary succession to leadership positions) and often include some references to the conquest era, and 3) the trial testimony of indigenous witnesses describing their experiences during the period.

This reconstruction and mapping of indigenous activity will be the focus of this poster. I will provide examples of four types of data and geo-visualizations I use to reconstruct this indigenous activity. These include:

1. **Geographic Knowledge Maps:** Mapping the geographic extents (and limits) of European knowledge of the Americas– places known and unknown – makes clear just how limited their knowledge and, by extension, their power was.
2. **Mood Maps:** First created by literary geographers, mood maps allow the mapping of an author's subjective experiences of a landscape.
3. **Density Plot of Events:** Graphing the density and range of events described in historical literature allows the comparison and contrast of how the story of the conquest of Peru has changed over time.
4. **Indigenous Activity Maps,** which trace the often hidden role of indigenous actors in conquest events.

References

Cooper, D., Donaldson, C., Murrieta-Flores, P. (Eds.), 2016. *Literary Mapping in the Digital Age, New edition edition*. ed. Routledge, Farnham, Surrey, England ; Burlington, VT.

Cooper, D., Gregory, I.N., 2011. Mapping the English Lake district: A literary GIS. *Trans. Inst. Br. Geogr.* 36, 89–108.

Eltis, D., Richardson, D., 2015. *Atlas of the transatlantic slave trade*. Yale University Press, New Haven, CT.

Knowles, A.K., Cole, T., Giordano, A., 2014. *Geographies of the Holocaust*. Indiana University Press.

Matthew, L.E., Oudijk, M.R., 2007. *Indian conquistadors: indigenous allies in the conquest of Mesoamerica*. University of Oklahoma Press, Norman.

Moretti, F., 1998. *Atlas of the European novel, 1800-1900*. Verso, London; New York.

Pearce, M.W., Hermann, M.J., 2010. Mapping Champlain's Travels: Restorative Techniques for Historical Cartography. *Cartogr. Int. J. Geogr. Inf. Geovisualization*. https://doi.org/10.3138/carto.45.1.32

Style Revolution: Journal des Dames et des Modes

Jodi Ann Mikesell
jm4470@tc.columbia.edu
Columbia University, United States of America

Avery Schroeder
abschroeder4@gmail.com
City University of New York, The Bard Graduate Center, United States of America

Anne Higonnet
ahigonne@barnard.edu
Columbia University, United States of America

Alex Gil
agil@columbia.edu
Columbia University, United States of America

AnaKaren Aguero
agueroak@gmail.com
Columbia University, United States of America

Sarah Bigler
scb2180@columbia.edu
Columbia University, United States of America

Meghan Collins
mmc2267@columbia.edu
City University of New York, The Bard Graduate Center, United States of America

Emily Cormack
emily.cormack@bgc.bard.edu
Columbia University, United States of America

Zoë Dostal
azd2103@columbia.edu
Columbia University, United States of America

Barthelemy Glama
bg2601@columbia.edu
Columbia University, United States of America

Brontë Hebdon
bah416@nyu.edu
New York University, Institute of Fine Arts, United States of America

Recently rediscovered at The Morgan Library, fashion plates from the *Journal Des Dames et Des Modes*, taught all Europeans how to look, read, and entertain themselves as modern individuals. Dating from 1797-1804, they represent the most radical changes in all of clothing history. This revolution in consumer culture signals the birth of fashion as we know it and transformed conceptions of identity, gender, and power. Their revolutionary representations of fashion generates cult followings within both academic and hobbyist circles; among whom are art historians, antiquarian bibliophiles, and historical fashionistas. However, the plates lack circulation and few digital sources present research that is both academically rigorous and accessible to learners of all levels. Our work seeks to remedy this issue and bridge the accessibility gap by creating a digital exhibit of the most rare and stylistically revolutionary plates. In doing so, we have produced our exhibit using minimal computing approaches developed at Columbia University Library and the Group for Experimental Methods in the Humanities.

Our website serves as a resource for viewing the *Journal Des Dames et Des Modes* color plates themselves, but also includes resources which contribute to furthering the observer's contextual understanding. We've done this by providing concise and easily digestible academic essays,

translation glossaries for both terms and color, a historical timeline, and an interactive map which visually situates the fashion plate figures within 18th century Paris. Our conference poster reflects the importance of our topic's historically democratic roots, describes our use of Wax (a suite of tools for minimal exhibitions), and collaboration structures; and directly links our undertaking to the democratic production and dissemination of knowledge through the aesthetics of minimal computing. By creating an accessible public-facing entry into a collection of art historical objects we create a channel to information without which *Journal Des Dames et Des Modes* scholarship would remain siloed in an institution's basement.

Ten graduate students—whose diverse institutional affiliations range from Columbia University, NYU, and The Bard Graduate Center—collaborated under the direction of Professors Anne Higonnet and Alex Gil to accomplish an unprecedented digital archive and scholarly online resource. The course, "Style Revolution," was a hybrid between traditional Art History seminar and an innovative Digital Humanities seminar. Students enrolled in the course had had no prior knowledge of coding in any of the languages used (HTML, CSS, Markdown, Bash, YAML, etc) nor familiarity with any of the additional software tools that were employed to create our final site. A wide range of literacies were taught, practiced, shared and acquired, from multiple lenses and disciplines, through multi-directional pedagogy, where all became teachers for one another at some point.

The site's main functionality was built using an early version of Jekyll Wax, which creates iiif compliant tiles and manifests, and generates pages with complete sets of YAML metadata converted from a spreadsheet. The iiif in turn allows our use of Open Sea Dragon for interacting with high resolution images, without burdening the browser with front-loaded data. The spreadsheet made it possible for all graduate students, regardless of technical inclination, to contribute metadata to each plate in the archive without the need for a database or forms. Additionally, because the resulting data is in CSV format and the complete site lives on GitHub, we share all data with the public directly. Leveraging the power of markdown and Jekyll, each student was able to contribute unique multimodal 'essays' to the project, from mapping exercises to digital art, based on original research.

By providing an online resource for the *Journal Des Dames et Des Modes* we are engaging the public in creating a greater understanding of current fashion phenomena, but one for which we lack a historical framework. The *Journal Des Dames et Des Modes* helps to create this framework and guides the viewer to a deeper, more meaningful understanding of how a seemingly inconsequential fad within fashion can create a paradigm shift in societal conceptions of consumer culture and its importance in material representations of our modern day identity. Simultaneously, we are modeling how collaborative work in the beginner digital humanities classroom can achieve almost complete control of an online exhibit of public import. This work will act as the foundation for an ongoing, larger project— and has already begun to be added upon. We look forward to the constant evolution of new projects, as we believe the increased attention our site provides will generate a response of scholarship, with which, we will continue to expand our project.

The Two Moby Dicks: The Split Signatures of Melville's Novel

Chelsea Miya
cmiya@ualberta.ca
University of Alberta, Canada

There has been a longstanding debate over the cetology sections in Herman Melville's Moby Dick. These chapters, which are interwoven into the mid-section of the novel, are curiously devoid of characters or plot development and instead describe whaling biology and behavior. Some Melville scholars, including Charles Olson and Lawrence Buell, have suggested that the novel might have been written as two separate texts that were spliced together in the final stages. As the original manuscripts have been lost, this has never been confirmed. However, I hope to show that the way in which the chapters cluster together reveals that the novel does indeed have two unique stylistic signatures. This is perhaps compelling evidence in favor of the "two Moby Dicks," a phenomenon that has been much speculated upon but never proven.

References

Bastian M., Heymann S., Jacomy M. (2009). *Gephi: an open source software for exploring and manipulating networks*. International AAAI Conference on Weblogs and Social Media.

Eder, Maciej. Kestemont, Mike and Rybicki, Jan. (2015). *'Stylo': a package for stylometric analyses*.

devochdelia: el Diccionario Etimolójico de las Voces Chilenas Derivadas de Lenguas Indíjenas Americanas de Rodolfo Lenz en versión digital

Francisco Mondaca
f.mondaca@uni-koeln.de
Universität zu Köln, Germany

devochdelia es la versión digital y en línea[1] del *Diccionario Etimolójico de las Voces Chilenas Derivadas de Lenguas*

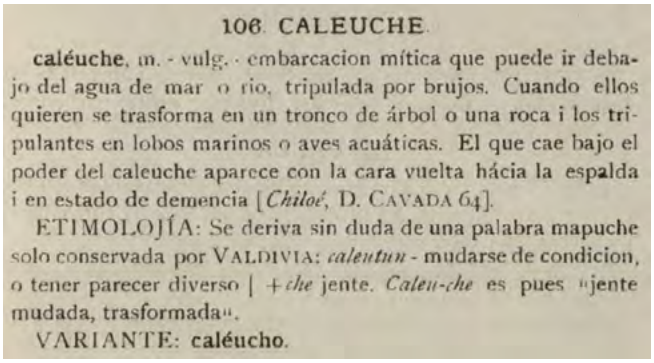
1 http://devochdelia.cl

Indíjenas Americanas (1905-1910) (Diccionario) compilado por el lingüista alemán-chileno Rodolfo Lenz. Esta obra ha sido fundamental en el desarrollo de la lexicografía chilena e hispanoamericana por su innovador y minucioso método de compilación. La digitalización de textos antiguos y valiosos como lo es el Diccionario presenta problemas engorrosos que dificultan el proceso en sí y el acceso a los datos obtenidos. En este proyecto se pueden apreciar soluciones accesibles a este tipo de dificultades facilitando tanto la digitalización de diccionarios impresos como su consulta en línea.

Acerca del diccionario impreso

La relevancia del Diccionario para lexicografía chilena radica en su enfoque descriptivo[2], que lo distingue de los diccionarios publicados en Chile hasta ese entonces. Si bien ya se habían publicado obras de americanismos con esta perspectiva, tanto en España (De Alcedo 1789) como en Cuba (Pichardo 1836), el Diccionario presenta innovaciones que lo destacan a nivel mundial. Entre ellas cabe mencionar la clara y detallada descripción del método de compilación empleado y de la teoría subyacente; la coherencia en la estructura y tipografía de los artículos, así como en la clasificación geográfica del área de empleo de los vocablos (Lenz 1905-1910[1980]:16).

Como nunca antes en la lexicografía chilena, un autor realiza un trabajo tan exhaustivo al comparar la información recabada con diccionarios publicados en Chile e Hispanoamérica. Pero no se limita a eso, también organiza conferencias con colegas, estudiantes e interesados en el tema para verificar la información reunida y añadir a su manuscrito nuevas palabras de origen indígena (Lenz 1905-1910:22ff).



106. CALEUCHE.

caléuche, m. - vulg. - embarcacion mítica que puede ir debajo del agua de mar o rio, tripulada por brujos. Cuando ellos quieren se trasforma en un tronco de árbol o una roca i los tripulantes en lobos marinos o aves acuáticas. El que cae bajo el poder del caleuche aparece con la cara vuelta hácia la espalda i en estado de demencia [*Chiloé*, D. CAVADA 64].

ETIMOLOJÍA: Se deriva sin duda de una palabra mapuche solo conservada por VALDIVIA: *caleutun* - mudarse de condicion, o tener parecer diverso | + *che* jente. *Caleu-che* es pues "jente mudada, trasformada".

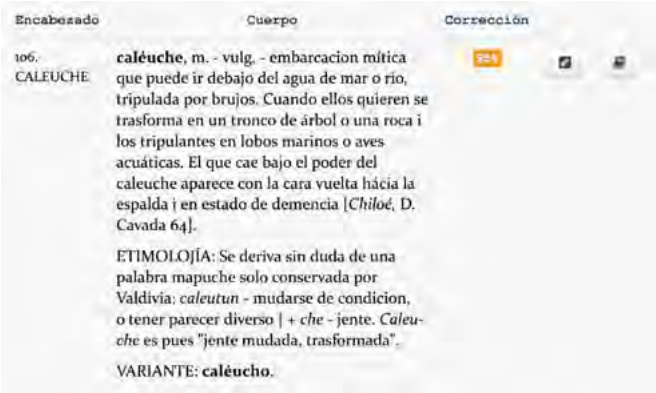
VARIANTE: **caléucho**.

La entrada ‚caleuche' en el Diccionario (Lenz 1905-1910:163)

El Diccionario cuenta con 1665 entradas que se dividen en encabezado y cuerpo. En el primero se aprecia la voz indígena propiamente tal y en el segundo se tratan las palabras chilenas derivadas de ella. Como suele ser tradición en los diccionarios semasiológicos, luego del lema se aprecia la categoría gramatical y el significado. Siempre se encuentra la „etimología", pudiendo no estar presentes secciones como „variantes" o „derivados".

Acerca del diccionario digital[3]

Un diccionario es un objeto cultural cuya función es aclarar dudas de carácter lingüístico. Por otra parte, el proceso de extracción de texto desde imágenes (OCR), es propenso a generar errores, lo que no se espera encontrar en ningún texto, menos en diccionarios. Las decisiones técnicas en este proyecto se tomaron bajo la premisa de poner en línea una versión digital del Diccionario con la menor cantidad posible de errores y, al mismo tiempo, acceder a todas las entradas del mismo. El formato elegido para la generación de texto en OCR fue Hypertext Markup Language (HTML), porque permite mantener cursivas y negritas, además de presentarse en un navegador de Internet sin problemas. Corregir todos los encabezados de las entradas, permitió la extracción de las 1665 entradas dentro de sus límites, e hizo posible buscar y encontrar las entradas mediante el número que Lenz les asignó o por el texto del encabezado. De los 1665 cuerpos, 1000 han sido corregidos.



La entrada ‚caleuche' en *devochdelia*

Una vez extraídas las entradas, se creó una aplicación web donde se pueden buscar y corregir las entradas, la cual está hecha con el *framework* Maalr (Neuefeind y Schwiebert 2013). En su versión básica, Maalr permite trabajar con entradas de diccionario en formato de texto simple. Como el fin de *devochdelia* es permitir que los usuarios ayuden a corregir las entradas, hubo que hacer dos modificaciones a Maalr:

a) que se pueda mostrar y editar texto en formato HTML,
b) que se puedan mostrar las imágenes correspondientes a cada entrada para que los usuarios vean la fuente impresa, y también editar las entradas de manera adecuada.

Cada entrada puede ser corregida y estas modificaciones ser vistas sin la necesidad de registrarse o iniciar sesión. Asimismo, cada corrección tiene que ser autoeva-

2 „I la ciencia exije que no escluyamos nada, que no dejemos de apuntar ninguna palabra" (Lenz 1905-1910:20)

3 Para más detalles, ver: http://www.devochdelia.cl/about

luada por el corrector, comunicando el nivel de la corrección a otros usuarios y a los editores.

Este proyecto muestra que, con pocos recursos, es posible digitalizar obras lexicográficas complejas haciendo partícipes en el proceso a quienes se interesan por ellas. Asimismo sirve de base para digitalizar diccionarios a otra escala.

References

De Alcedo, A. (1789). *Diccionario geográfico-histórico de las Indias Occidentales ó América*. Tomo V. Madrid: Imprenta de Manuel González.

Lenz, R. (1905-1910). *Diccionario Etimolójico de las Voces Chilenas Derivadas de Lenguas Indíjenas Americanas*. Santiago: Imprenta Cervantes.

Lenz, R. ([1905-1910] 1980). *Diccionario Etimológico de las Voces Chilenas Derivadas de Lenguas Indígenas Americanas.* Edición dirigida por Mario Ferreccio Podestá. Santiago: Universidad de Chile.

Neuefeind, C. y Schwiebert S. (2013). Introducing Maalr: A Modern Approach to Aggregate Lexical Resources. *Language Processing and Knowledge in the Web, the proceedings of the 25th Conference of the German Society for Computational Linguistics (GSCL 2013)*, Darmstadt, Alemania, 25-27 febrero 2013. https://gscl2013.ukp.informatik.tu-darmstadt.de/fileadmin/user_upload/Group_UKP/conferences/gscl2013/demo_maalr-gscl2013.pdf (consultado el 25 de abril de 2018)

Pichardo, E. (1836). *Diccionario Provincial de Voces Cubanas*. Matanzas: Imprenta de la Real Marina.

Unsustainable Digital Cultural Collections

Jo Ana Morfin
jo.morfin@gmail.com
Universidad Nacional Autónoma de México, Mexico

This paper analyzes how in the context of Mexican museums, the lack of policies, frameworks and strategic planning has led to the creation of unsustainable cultural digital collections. It focuses on the challenges in rescuing the digital collection "Bienal Internacional de Poesía Visual y Experimental" [Biennale of Visual and Experimental Poetry], held at the Mediateque of the Museo Universitario del Chopo.

The Mexican artists Araceli Zúñiga y César Espinosa organized the International Biennales of Visual and Experimental Poetry between 1985 and 2009. These events brought together practitioners from all over the world whose work is placed at the intersections of the fields of contemporary visual writing, copy art, concrete music, mail art and performance.

Throughout the years, Zúñiga and Espinosa became interested in creating a "memory" of these events. Therefore, they started to gather videos, mail art works, photography, artistic electrography, from each event. The collection was stored at their house and classified and organized by the artists themselves. Through the years, the collection became a key source for researching and tracing the development of alternative and experimental art practices in Mexico.

Given the significance of this collection and with the aim of preserving and providing greater access to its contents, Zúñiga and Espinosa agreed with the Museo Universitario del Chopo in digitizing the materials and donating a digital version to be included in the collection of the museum. Over 2,000 artworks were digitized. In 2015 the museum received a grant to put these contents online. However, during the development of the project we realized that most of the digital objects were unstable. Given this situation, the project focused on rescuing this digital collection from the oblivion.

The project brought to light several concerns, such as the lack of a digital preservation planning, the deficient use of metadata standards, the shortage of expertise, and more importantly, the lack of institutional policies to create sustainable digital collections. The museum´s team did not follow clear guidelines, standards and best practices for the creation of digital objects and their subsequence management. Thus affecting the ability to read, access and understand the digital materials.

This poster describes the rescuing process and the guidelines we create in order to prevent the creation of unsustainable digital collections within cultural memory institutions.

La automatización y "digitalización" del Centro de Documentación Histórica "Lic. Rafael Montejano y Aguiñaga" de la Universidad Autónoma de San Luis Potosí, mediante la autogestión y software libre

José Antonio Motilla
jamotilla@gmail.com
Universidad Autónoma de San Luis Potosí, Mexico

Ismael Huerta
ismaelhuerta.ten@gmail.com
Universidad Autónoma de San Luis Potosí, Mexico

La presente ponencia tiene como objetivo presentar el estudio de caso del proceso de modernización del Centro de Documentación Histórica "Lic. Rafael Montejano y Aguiñaga" de la Universidad Autónoma de San Luis Potosí (CDHRMA-UASLP), México, constituido por una colección de aproximadamente 100 mil volúmenes bibliográ-

ficos, una amplia sección de manuscritos, publicaciones periódicas, e impresos , y un gran acervo documental que incluye el Archivo Histórico de la UASLP.

Hacia el año 2014, el CDHRMA-UASLP trabajaba con un sistema fundamentalmente análogo, al no contar con un catálogo electrónico ni de herramientas tecnológicas que le permitieran preservar y difundir sus materiales. En ese contexto, se emprendió un profundo diagnóstico del Centro, que buscaba detectar sus carencias con el fin de hacer más eficientes sus procesos. El análisis arrojó como resultado la necesidad de fortalecer cuatro áreas fundamentales: la implementación de un Sistema Integral de Automatización de Bibliotecas (SIAB); el manejo de los inventarios y catálogos mediante bases de datos eficientes; la digitalización de los materiales de alta demanda para garantizar su conservación; y la investigación académica de sus fondos y colecciones.

Ante la falta de presupuesto institucional, el equipo encargado del desarrollo del proyecto tomó la decisión de desarrollar el proyecto mediante el empleo de Software Libre y desarrollar estrategias para reducir al máximo los costos; así, para el desarrollo del SIAB se recurrió a la plataforma de acceso libre Koha; se migraron y sistematizaron las bases de datos en la plataforma File Maker (único software de paga que fue utilizado); para la digitalización de materiales se adquirió una cámara de alta resolución y se creó un soporte con iluminación no profesional para digitalizar documentos; y se implementó un equipo de investigación, coordinado por el departamento de investigación del Centro, con el apoyo de becarios, para crear bases de datos y analizarlas bajo el paradigma de las humanidades digitales.

Como resultado, al día de hoy se cuenta con un catálogo electrónico con más de 7 mil registros, dos periódicos del siglo XIX completamente digitalizados y en consulta, un inventario general de la biblioteca, la descripción detallada y digitalización de algunos fondos del archivo histórico, y herramientas de investigación realizadas mediante minería de texto.

La experiencia y reflexión planteada en ésta ponencia, busca poner sobre la mesa la importancia que herramientas como el software libre, y el desarrollo de aplicaciones tecnológicas e informáticas, puede impactar de manera favorable en la conservación y difusión de acervos bibliográficos y documentales de alto valor patrimonial, y garantizar el acceso a ellas por parte de los investigadores y público interesado tanto del presente como de generaciones futuras.

A Comprehensive Image-Based Digital Edition Using CEX: A fragment of the Gospel of Matthew

Janey Capers Newland
janeycapers.newland@furman.edu
Furman University, United States of America

Emmett Baumgarten
emmett.baumgarten@furman.edu
Furman University, United States of America

De'sean Markley
desean.markley@furman.edu
Furman University, United States of America

Jeffrey Rein
jeffrey.rein@furman.edu
Furman University, United States of America

Brienna Dipietro
brienna.dipietro@furman.edu
Furman University, United States of America

Anna Sylvester
anna.sylvester@furman.edu
Furman University, United States of America

Brandon Elmy
brandon.elmy@furman.edu
Furman University, United States of America

Summey Hedden
summey.hedden@furman.edu
Furman University, United States of America

This poster (with accompanying downloadable dataset and application) will demonstrate as a proof-of-concept a comprehensive image-based publication and analysis of a text bearing artifact, [catalog number redacted for anonymous review], a hitherto unpublished 10th Century palimpsest fragment of the Gospel According to Matthew. The fragment contains most of the "Parable of the Sower".

In editing this text, we sought to be as comprehensive as possible, capturing:

- Natural light and UV images, both overview images and details
- A diplomatic transcription of the overwritten text of Matthew and any legible characters from the undertext
- A word tokenization of the diplomatic transcription, mapping to each token:
  - a normalization
  - editorial status
  - lexical status
  - morphology, part of speech, and syntactic relations
  - alignment to the image data
- A character tokenization, aligned to the image data
- An edition of the whole Gospel according to Matthew from a critical edition, for comparison and context
- Translations aligned to the text
- Editors' comments

In publishing it, we sought simplicity, longevity, and clarity. While we use TEI XML as a format for capturing an initial transcription, the overlapping analytical categories, many-to-many alignments of text and image, and open ended possibilities for commentary precluded implementing a coherent data model fully in XML. At the same time, we wanted a concise and integrated publication.

By using the CEX format[1], a plain text, self-documenting format based on the CITE/CTS architecture, we are able to bring together these many levels of analysis in a form that is at once disaggregated, with each scholarly primitive explicitly and unambiguously citable, while still united in a single file. CEX allows us to distribute a fully integrated dataset in the form of a single plain text file and a single directory of images.

Our publication, a CEX file and a directory of images, is technology-agnostic readable by humans, but also able to serve as the data for an end-user application. We will describe, and have available for download and on USB thumbdrives, a lightweight, zero-configuration single page web application (SPA), fully usable offline, that integrates the data and images for this publication for end-users.[2]

Finally, we will outline the low cost, low technology, collaborative work behind the digitization and editing of this manuscript fragment: off-the-shelf cameras, simple handheld UV lighting, readily available FOSS software.

We believe that this work will be of interest to the international Digital Humanities research community both as a new publication of a Byzantine Greek text and as a demonstration of a replicable and sustainable combination of technology and workflow. We think this approach provides a compelling alternative to XML or RDF editions and complex database-driven end user applications, offering advantages both on the back end (a flexible, scalable, and self-documenting format for implementing diverse data models), and on the front end (lightweight and portable presentation for readers). At the same time the data we present as CEX and images is easily transferrable to other standard formats.[3]

All project data is under version control in a public GitHub repository, and licensed under a CC-BY license. All source code is under either a GPL or MIT public license.

1 CEX (CITE Exchange Format) is a plain text format for capturing data about texts and collections, based on the CITE/CTS architecture and developed by C. Blackwell (*Homer Multitext*), T. Köntges (University of Leipzig), and N. Smith (*Homer Multitext*). For implementations and projects using CEX, see: T. Köentges, (Meletē)ToPān (topic modelling environment): https://thomask81.github.io/ToPan/; C. Blackwell, N. Smith, CEX Library (Scala): https://github.com/cite-architecture/cex; C. Blackwell, N. Smith, CEX Dataset Repository: https://github.com/cite-architecture/citedx

2 This application is based on the ScalaJS implementation of "CITE App" by C. Blackwell and N. Smith: https://github.com/cite-architecture/CITE-App

3 Existing code libraries for working with CEX include a microservice framework that delivers textual and other data from CEX files as JSON objects, via HTTP requests (see https://github.com/cite-architecture/scs-akka) and libraries that export CEX data into other formats, such as 2-column tabular data or 82XF (see https://github.com/cite-architecture/scm).

Using Zenodo as a Discovery and Publishing Platform

Daniel Paul O'Donnell
daniel.odonnell@uleth.ca
University of Lethbridge, Canada

Natalia Manola
natalia@di.uoa.gr
OpenAIRE

Paolo Manghi
paolo.manghi@isti.cnr.it
Zenodo, Switzerland; CNR, Italy

Dot Porter
dot.porter@gmail.com
University of Pennsylvania

Paul Esau
paul.esau@gmail.com
University of Lethbridge, Canada

Carey Viejou
c.viejou@uleth.ca
University of Lethbridge, Canada

Roberto Rosselli Del Turco
robertorossellidelturco@gmail.com
University of Pisa, Italy; University of Turin, Italy

We are 25 years into the World Wide Web revolution. While Humanities researchers have been at the forefront of many uses of networked communication to disseminate their research, they have lagged other disciplines in their adoption of formal discovery and organisational tools (Spiro, 2016; Borgman, 2009; Anderson et al., 2012). Some of the core tools that characterise current best practice in other disciplines—ORCID, DOIs, discipline-wide repositories, mega and overlay journals—have seen slow or limited adoption in the case of Humanities researchers. Data Management and Citation practices tend to be less well-developed and widely practised in the Humanities than in other areas. Humanities publishing, too, especially scholar-led publishing, still commonly involves less than optimal practice—custom, project-held URLs, storage on private/commercial data servers, a lack of formal attention to versioning, backups, and long-term preservation (Copland et al., 2016).

This poster shows how two projects at the University of Lethbridge are addressing these long-standing problems through the use of OpenAIRE/Zenodo (the final form of the poster is O'Donnell et al., 2018). In one case, the project is looking for an open and FAIR (Findable, Accessible, Interoperable, and Reusable) method of publishing project data—a small (by cross-disciplinary standards) set of 2D and 3D images and point clouds, annotations, and textual transcriptions involving medieval

cultural and textual heritage. The goal here is to establish an expansible repository that will allow for non-negotiated additions and reuse by external projects and survive and remain citable long after the originating project has concluded and funding has run out.

The second is the publication platform for a graduate-student run journal. In this case, the students needed a platform that would provide their early career authors with some guarantee of permanent archiving and discoverability while recognising and accommodating the inherently unstable nature of a graduate-student run editorial board: while this year's board is enthusiastic about the project, we have no way of guaranteeing that this will be true of future generations of graduate students.

Although other options exist to solve both these problems, our poster demonstrates the degree to which OpenAIRE/Zenodo provides an extremely simple and durable platform for ensuring the long-term discoverability and preservation of Humanities research in these common use cases.

References

Anderson, D. E., Dwyer, G. and Leahy, S. (2012). Fine-Tuning the Institutional Repository: Evaluating the Self-Archiving Behavior of Researchers in Music. *The Serials Librarian*, 63(3-4). Routledge: 277–87 doi:10.1080/0361526X.2012.722594. https://doi.org/10.1080/0361526X.2012.722594.

Borgman, C. L. (2009). DHQ: Digital Humanities Quarterly: The Digital Future is Now: A Call to Action for the Humanities http://www.digitalhumanities.org/dhq/vol/3/4/000077/000077.html (accessed 20 June 2017).

Copland, C., Carrell, S., Davidson, G., Grandfield, V. and O'Donnell, D. P. (2016). Kiernan, Kevin S. 2015. Electronic Beowulf - Fourth Edition. *Digital Medievalist*, 10 http://www.digitalmedievalist.org/journal/10/copland/ (accessed 25 April 2017).

O'Donnell, D. P., Manola, N., Manghi, P., Porter, D., Esau, P., Viejou, C., Del Tuco, R. R. and Singh, G. (2018). Using Zenodo as a Discovery and Publishing Platform Paper presented at the DH 2018, Mexico doi:10.5281/zenodo.1234474. https://zenodo.org/record/1234474.

Spiro, L. (2016). Studying how digital humanists use GitHub *Digital Scholarship in the Humanities* https://digitalscholarship.wordpress.com/category/open-access/ (accessed 27 November 2017).

SpatioScholar: Annotating Photogrammetric Models

Burcak Ozludil Altin
bozludil@njit.edu
New Jersey Institute of Technology, United States of America

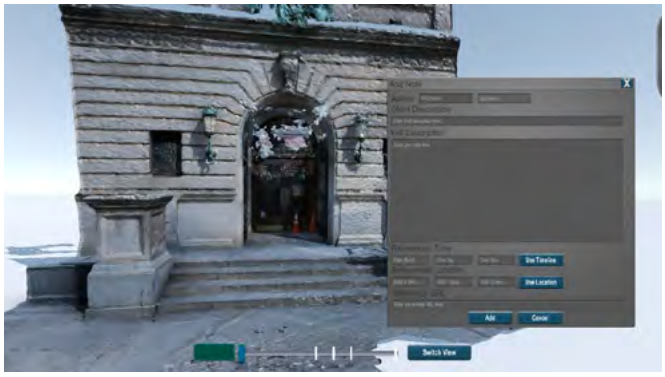
Augustus Wendell
wendell@njit.edu
New Jersey Institute of Technology, United States of America

This poster presents a new phase in the development of *SpatioScholar* which is a platform for studies that require spatial and temporal processing, visualization, and analysis, including art/architectural history and urban research (Wendell et al., 2016). The platform is a scholarly application created in Unity3D synthesizing 3D models with textual information and research documents. (Figure 1) *SpatioScholar* provides a computational close reading system for spatial and temporal data sets by using the following functionalities: (1) through a timeline slider, it demonstrates the phases through which a certain building or location passed; (2) through a simulation, it provides the viewer with the ability to experience the space in first person, and to track any desired aspect of life inside buildings or locations; (3) through a reconnection of the primary materials and the conclusions derived from them, viewers can browse and review the relevant information (photographs, drawings, textual primary documents etc.) that are cross-referenced with the "scene;" and (4) through a "Shared Scholarship" feature, viewers and users can leave notes, comments or browse others' notes.

SpatioScholar interface displaying time slider, mini-map, primary source panel and an enlarged historic photograph that shows the same scene in the 1910s (Photograph source: Sihhat Almanaki, 1933.)

At this phase of the *SpatioScholar* development, we are testing the platform with photogrammetry models. (Figure 2) Photogrammetry is a computational process coordinating optical data recorded in a series of photographic images, solving matched data points for a 3D point cloud, and outputting a 3D model with applied photographic textures. The benefits of photogrammetry to digital art/architectural history and digital heritage in comparison to traditional 3D model building are well debated (Allen et al., 2003; El-Hakim et al., 2007; Webb, 2016). While other scholars have included photogrammetry data in spatial simulations (Ozer and Nagakura, 2016), we are extending this line of work by integrating a shar-

able spatial annotation feature within a single distributed application.



A photogrammetric model imported into SpatioScholar. The "Add Note" interface element shown is used for the sharable spatial annotation

Incorporating photogrammetry in *SpatioScholar* presents multiple advantages: first, it eliminates the need to create a 3D model from scratch for projects that are not previously modelled. Second, the use of simple photographic data allows non-technical or non-traditional users to capture, research, and create accurate 3D representations of space, even with smart phones (6). Adapting *SpatioScholar* to photogrammetry will widen the user base as this technology becomes more readily available and accessible in the field.

SpatioScholar implements a custom developed space based annotation toolset that allows notation of the photogrammetric 3D model through a web accessible database. This feature, combined with a WebGL delivery mode allows a research project to be delivered via the web in the same interface for input, comments, and collaborations without the need to transfer or use another medium. This single interface in *SpatioScholar* combines the **research phase** inherent to scholarly production and its **sharing** with the outside world.

The components that create the *SpatioScholar* functionality within Unity3D are programmed elements that actively manage models, database interaction, user interface, and primary source document coordination. As it stands now, the user imports an FBX format version of photogrammetry model into a Unity3D enabling all the functionalities of the platform by using a previously created "SpatioScholar Unity3D Template Project." By dropping their imported FBX file onto the coordinating *SpatioScholar* component, Unity3D creates temporal, primary document and annotation associations based on existing metadata mapping within the FBX file.

SpatioScholar was conceptualized first and foremost as a platform to facilitate and share research, not as a tool to merely navigate the virtual reconstruction of a building or site. The incorporation of photogrammetry as a fairly accessible technology into the platform paves the path to opening of the platform to a wider user-base that can employ its functionalities to foster research. This poster demonstrates the potentials in bringing spatial data into *SpatioScholar* to create a web-distributable spatial research project, by enlisting temporal awareness, trajectory tracking, primary document coordination, and shared annotation features.

References

Allen, P. K., Troccoli, A., Smith, B., Murray, S., Stamos, I., Leordeanu, M. (2003). New methods for digital modelling of historic sites. In *IEEE Computer Graphics and Applications*, 23(6), 2003, pp. 32–41.

El-Hakim, S, Gonzo, L., Voltolini, F., Girardi, S., Rizzi, A., Remondino, F., Whiting, E. (2007). Detailed 3D Modelling of Castles. In *International Journal of Architectural Computing* 5(2): 200-220.

Fassi, F. (2012). Complex architecture in 3D from survey to web. *International Journal of Heritage in the Digital Era*, 1(3): 379-398.

Osman, M. (1933). *Sihhat Almanaki*, Kader Matbaasi, Istanbul.

Ozer, D.G. and Nagakura, T. (2016). Simplifying architectural heritage visualization – *AUGMENTEDparion.* In Herneoja, A., Österlund T. and Markkanen, P. (eds.), *Complexity & Simplicity - Proceedings of the 34th eCAADe Conference* - Volume 2, University of Oulu, Oulu, Finland, 22-26 August 2016, pp. 521-528.

Webb, N., Buchanan, A. and Peterson, J.R. (2016). Modelling medieval vaults: comparing digital surveying techniques to enhance our understanding of gothic architecture. In Herneoja, A., Österlund T. and Markkanen, P. (eds.), *Complexity & Simplicity - Proceedings of the 34th eCAADe Conference* - Volume 2, University of Oulu, Oulu, Finland, 22-26 August 2016, pp. 493-502.

Wendell, A., Ozludil Altin, B. and Thompson, U. (2016). Prototyping a temporospatial simulation framework:case of an ottoman insane asylum. In Herneoja, A., Österlund T. and Markkanen, P. (eds.), *Complexity & Simplicity - Proceedings of the 34th eCAADe Conference* - Volume 2, University of Oulu, Oulu, Finland, 22-26 August 2016, pp. 485-491.

Decolonising Collections Information – Disrupting Settler Colonial Power In Information Management in response to Canada's Truth & Reconciliation Commission and the United Nations Declaration on the Rights of Indigenous Peoples

Laura Phillips
laura.phillips@queensu.ca
Queens University, Canada

Standard collections information management principles in use by settler colonial cultural institutions derive from the foundation of museums as repositories to showcase the extent of empire and, as with all 'Euro-Western' disciplines, are not capable of objectivity in approach or in reflecting the multiplicity of identities in non-Western world views (Garneau, 2016). As a reflection of contemporary society, cultural institutions must be at the forefront of the decolonisation movement, and not simply initiate projects that perpetuate the museum as the authority to further (consciously or unconsciously) settler colonialist aspirations as one of the "…lasting effects of European colonialism on the multiple stagings and worldings of nations and societies across the globe" (Byrd, 2017: 176). Decolonisation in Canada means critically reflecting on the colonial bias for accepted 'truths' projected by the actions and ethos of cultural institutions, especially museums, to analysis bias in information management from the point of ingestion to management and re-presentation.

Having participated in efforts to build museums based on non-Western world views in both Qatar (Taylor, 2014) and the Cree Nation in Eeyou Istchee (Pashagumskum, 2016), my research continues my progression in deconstructing professional museum practice by exploring these questions:

- How can contemporary museology incorporate Indigenous perspectives to address the power imbalance that perpetuates colonial mythology and the related presumption of ownership rights?
- What practical methods can reframe methods of engagement between Indigenous communities and museums?
- How can critiques of museum practice by Indigenous knowledge keepers be presented to museums to change established procedures?
- How can Indigenous values and traditional knowledge be shared with museums to centre the Indigenous perspective, while respecting unique traditions for each community and safeguarding their intellectual property rights?
- How can museums and curators identify the settler colonial bias in their work?
- Is any of this even possible given that museums are founded on 'scopophilia' (Garneau, 2016)?

My innovative, community-centric research approach will improve the efficacy demonstrated in existing case studies of community based research (Smith, 1999; Tuck, 2009; Tuck and Wang, 2012; Tuck and Wang, 2014). The focused application of Indigenous knowledge to museology, including collections information management, will generate guidance required for imperative revisions in museum policies and procedures to become consistent with Canada's Truth & Reconciliation Commission (Truth and Reconciliation Commission, 2015) and United Nations Declaration on the Rights of Indigenous Peoples (United Nations, 2008).

My poster will present ideas for shifts in information management as perceived during my Ph.D. research in Cultural Studies, including case studies from Indigenous institutions in Canada to demonstrate ways to disrupt the current colonial power structures. The ideas presented will provoke discussion that will ultimately help to create self-empowering principles to engage international, national and provincial cultural institutions to form the basis of new standards of decolonised cultural information management. My poster will include examples of decolonisation efforts taking place in Canada to de-centre the settler colonial hegemony, an overview of theoretical approaches used as the foundation for this shift, and explain how militant research principles (Colectivo Situaciones, 2003; Brown, 2013) can be applied to day to day cultural information management on an individual level to disrupt the current paradigm.

References

Brown, N. (2013). *Militant Research Handbook*. New York: New York University.

Byrd, J. (2017). American Indian Transnationalisms. In Goyal, Y. (ed), *The Cambridge Companion to Transnational American Literature*. Cambridge: Cambridge University Press, pp. 174–89.

Colectivo Situaciones (2003). On the Researcher-Militant *European Institute for Progressive Cultural Policies* http://eipcp.net/transversal/0406/colectivo-situaciones/en.

Garneau, D. (2016). Imaginary Spaces of Conciliation and Reconciliation: Art, Curation, and Healing. In Robinson, D. and Martin, K. (eds), *Arts of Engagement: Taking Aesthetic Action In and Beyond the Truth and Reconciliation Commission of Canada*. Waterloo: Wilfred Laurier University Press, pp. 21–41.

Pashagumskum, S., Menarick, P., Phillips, L., Laurendeau, G. and Scott, K. (2016). Seeing Ourselves: The Path to Self-curation, Cultural Sovereignty and Self-Representation in Eeyou Istchee. In Hele, K. (ed), *Survivance and Reconciliation: 7 Forward / 7 Back: 2015 Canadian Indigenous Native Studies Association Conference Proceedings*. Manitoba: Aboriginal Issues Press, pp. 60–87.

Smith, L. (1999). *Decolonizing Methodologies: Research and Indigenous Peoples*. 2nd ed. London: Zed Books Ltd.

Taylor, D., Phillips, L., Al Malek, N. and Alathbah, N. (2014). Collective Opportunities: Collections Management in Qatar. In Erskine-Loftus, P. (ed), *Museums and the Material World: Collecting the Arabian Peninsula*. Edinburgh: Museums Etc, pp. 412–52.

Truth and Reconciliation Commision of Canada (2015). Honouring the Truth, Reconciling for the Future: Summary of the Final Report of the Truth and Reconciliation Commission of Canada http://www.trc.

ca/websites/trcinstitution/File/2015/Findings / Exec_Summary_2015_05_31_web_o.pdfhttp://www.trc.ca/websites/trcinstitution/File/2015/Findings/Exec_Summary_2015_05_31_web_o.pdf (accessed 1 June 2017).
Tuck, E. (2009). Re-visioning Action: Participatory Action Research and Indigenous Theories of Change. *Urban Review*, 40(11): 47–65.
Tuck, E. and Ree, C. (2013). A Glossary of Haunting. In Jones, S., Adams, T. and Ellis, C. (eds), *Handbook of Autoethnography*. London: Routledge, pp. 639–58.
Tuck, E. and Wang, K. W. (2012). Decolonization is not a metaphor. *Decolonization: Indigeneity, Education & Society*, 1(2): 1–40.
Tuck, E. and Wang, K. W. (2014). R-Words: Refusing Research. In Paris, D. and Winn, M. (eds), *Humanizing Research: Decolonizing Qualitative Inquiry with Youth and Communities*. Thousand Oakes: Sage Publications, pp. 223–47.
United Nations (2008). United Nations Declaration on the Rights of Indigenous Peoples , http://www.un-.org/esa/socdev/unpfii/documents/DRIPS_en.pdf.
Wilson, J. (2016). Gathered Together: Listening to Musqueam Lived Experiences. *Biography*, 39(3): 469–94.

An Ontological Model for Inferring Psychological Profiles and Narrative Roles of Characters

Mattia Egloff
mattia.egloff@unil.ch
University of Lausanne, Switzerland

Antonio Lieto
lieto@di.unito.it
University of Turin, CAR-CNR, Italy

Davide Picca
davide.picca@unil.ch
University of Lausanne, Switzerland

Introduction

The modelling of the inner world of narrative characters and the ability to capture and formally shape their deep psychological characteristics are at the center of the reflection of a part of literary criticism and remains, today, an open challenge in the Digital Humanities. In this paper, we present an ongoing work of a preliminary version of the Ontology of Literary Characters (OLC), that allows to capture and inference psychological characters' traits starting from their linguistic descriptions as they appear in literary texts.

The ontology of literary characters

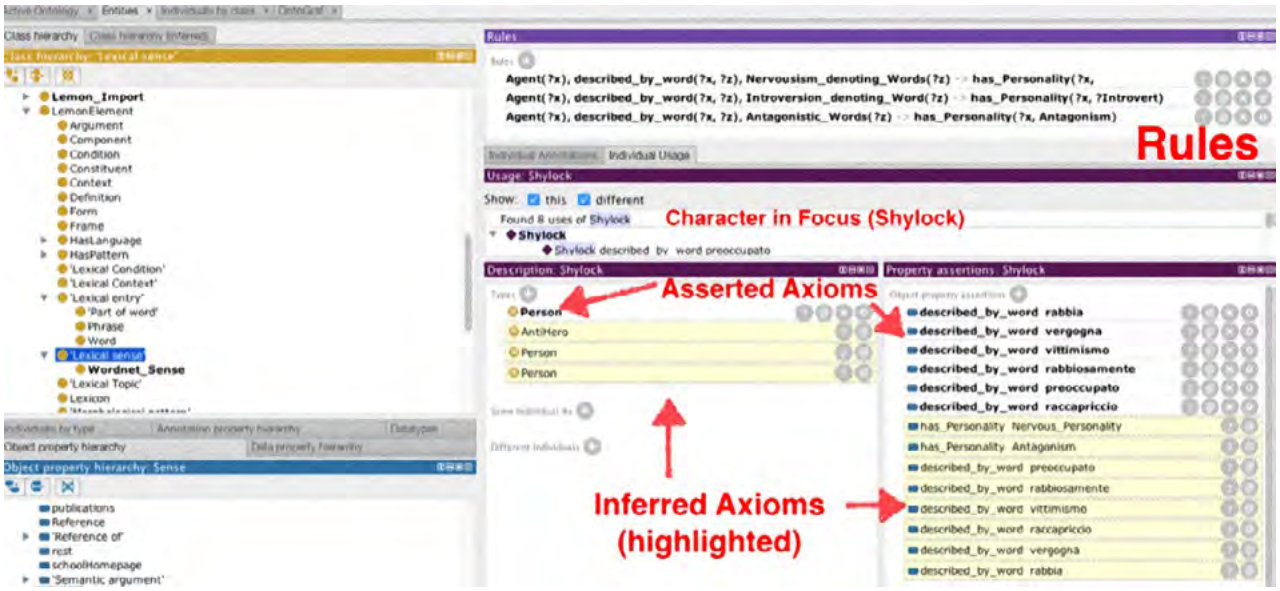
The ontology of literary characters (OLC) integrates different ontological models already available in conceptual models literature. In particular, it integrates the ontology framework LEMON (The Lexicon Model for Ontologies, (McCrae et al., 2011)) and the Ontology of Emotion (OE) (Patti et al., 2015) (encoding affective knowledge in emotional categories based on both Plutchik's (Plutchik, 1997)) and Hourglass's models in (Cambria et al., 2012)) with two additional models:

- a preliminary ontology of narrative roles
- a model of psychological profiles relying on the model of the Big 5 personality traits (Digman, 1990).

In our ontology, the word level is encoded in our model as Lexical Entry in the LEMON module. Lexical Entries are linked to their corresponding Emotion through the property *describes emotion*. The different set of Emotions is represented with the OE model that currently includes 32 emotional concepts. Each of such concept, as specified above, is connected to the word level and, in addition, is connected with specific concepts represented the micro-ontology of the Big Five Personalty Traits. The latter integrated model allows to categorize the psychological profiles of the characters along the axes of Openness to experience Conscientiousness, Extraversion, Agreeableness and Neuroticism. Finally, the concepts of Big Five micro-ontology are connected with those represented in an additional module that allows to represent the narrative roles played by the characters in a given story. Such integrated micro-ontology of narrative roles has been based on the archetypes of HERO, ANTI-HERO and VILLAIN which are commonly used in the narrative realm (Lieto and Damiano, 2014). Regarding the HERO class is represented with the following relevant narrative features: e.g. the fact that it is characterized by his/her fights against the VILLAIN of a story, the fact that his/her actions are necessarily guided by general goals to be achieved in the interest of the collectivity, the fact that they fight against the VILLAIN in a fair way and so on. The ANTI-HERO, on the other hand, is described as characterized by the fact of sharing most of its typical traits with the HERO (e.g. the fact that it is the protagonist of a plot fighting against the VILLAIN of the story); however, his/her moves are not guided by a general spirit of sacrifice for the collectivity but, rather, they are usually based on some personal motivations that, incidentally and/or indirectly, coincide with the needs of the collectivity. Furthermore the ANTI-HERO may also act in a not fair way in order to achieve the desired goal. A classical example of such archetype is Shylock which is described with the words "rabbia"/"anger", „vergogna"/"shame", etc (See Figure 1) . Each of these words is associated with a specific emotion of the OE ontology. In addition, each emotion is linked in the ontology to a

particular Psychological Profile from the Big Five Model. Finally, each Personality of the Big Five Model is semantically connected with a particular narrative role. Finally the VILLAIN is represented as a classic negative role in a plot and is characterized as the main opponent of the protagonist/HERO.

The overall integrated ontological model allowed us to show how a given character (e.g. Shylock in figure 1) described in the text with some particular psychological-denoting words (e.g. described by the words "rabbia"/"anger" …) can be automatically associated to one of the 5 classes of the personality traits of the Big Five and, as as consequence, also to the corresponding narrative role played in a story. Such semantic association is performed by using the ontological connections between the lexical level and the Emotional Concepts and an additional layer of SWRL rules connecting specific types of Words to specific Personality Traits, (See Figure 1).



Example Shylock.

Conclusion

In this paper, we presented an ongoing work on a first version of the Ontology of Literary Characters (OLC),As already observed by (Egloff et al., 2016) this ontology highlights the close relationship between character and language. In particular, where words play a significant role is crafting what we would now call the "personalities" in literature. As a result of these semantic connections it is possible to infer, starting from the natural language description of a given character, which is his/her psychological profile and his/her role played in the plot. In the case of Shylock, the system automatically infer that this character plays the role of ANTI-HERO in the plot. This ontological approach offers a new mean to scholar in order to isolate and analyze these verbal features of character going from natural language description of literary characters to the automatic assignment of their narrative role.

References

Cambria, E., Livingstone, A. and Hussain, A. (2012). The hourglass of emotions. *Cognitive Behavioural Systems*: 144–157.

Digman, J. M. (1990). Personality structure: Emergence of the five-factor model. *Annual Review of Psychology*, 41(1): 417–440.

Greenwade, G. D. (1993). The Comprehensive Tex Archive Network (CTAN). *TUGBoat*, 14(3): 342–351.

Lieto, A. and Damiano, R. (2014). A hybrid representational proposal for narrative concepts: A case study on character roles. *OASIcs-OpenAccess Series in Informatics*, vol. 41. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Egloff, M., Picca, D. and Curran, K. (2016). How IBM Watson Can Help Us Understand Character in Shakespeare: A Cognitive Computing Approach to the Plays. *In Digital Humanities 2016: Conference Abstracts*. Jagiellonian University and Pedagogical University, Kraków, pp. 488–92.

McCrae, J., Spohr, D. and Cimiano, P. (2011). Linking lexical resources and ontologies on the semantic web with lemon. *Extended Semantic Web Conference*. Springer, pp. 245–259.

Patti, V., Bertola, F. and Lieto, A. (2015). ArsEmotica for arsmeteo. org: Emotion-Driven Exploration of Online Art Collections. *FLAIRS Conference*. pp. 288–293.

Plutchik, R. (1997). The circumplex as a general model of the structure of emotions and personality.

A Graphical User Interface for LDA Topic Modeling

Steffen Pielström
pielstroem@biozentrum.uni-wuerzburg.de
University of Würzburg, Germany

Severin Simmler
severin.simmler@stud-mail.uni-wuerzburg.de
University of Würzburg, Germany

Thorsten Vitt
thorsten.vitt@uni-wuerzburg.de
University of Würzburg, Germany

Fotis Jannidis
fotis.jannidis@uni-wuerzburg.de
University of Würzburg, Germany

Using LDA (Latent Dirichlet Allocation) for analyzing the content structure of digital text collections is a possibility, that aroused the interest of many digital humanists in the recent years. The method allows to generate a so called 'topic model' from a text corpus, each 'topic' in the model being represented by a probability distribution over the words in the corpus. In each of these topics, another group of semantically related words appears with high probability scores. By labeling topics with their most probable words and then calculating the relative contributions of the topics to each text or text segment, researchers can use LDA as an unsupervised method to survey the contents of a text corpus (Blei 2012, Steyvers and Griffiths 2006).

However, to actually use LDA, technical skills lacked by the majority of humanities scholars is necessary. There is a number of accessible implementations of the LDA algorithm, the most popular being in MALLET (McCallum 2002), a Java program that has to be run and controlled from the command line and Gensim (Rehurek und Sojka 2010), a text analysis library for the Python programming language. Basically, most existing implementations of the algorithm require programming skills to be used efficiently, and for most use cases one has to switch between systems, tools and programming languages to complete the entire workflow from preprocessing to the analysis of results.

With the aim of lowering the threshold to use LDA for humanities scholars, we developed a programming library in Python that significantly reduces the complications to control the whole process of topic modeling from preprocessing to the visualization of results with a single Python script. The library, developed with funding from the European infrastructure project DARIAH (https://de.dariah.eu/), allows to choose from three different LDA implementations (MALLET, Gensim, and the 'LDA' package by Allan Riddell; https://pypi.python.org/pypi/lda). It provides a number of interactive, extensively annotated jupyter notebooks (http://jupyter.org/) that can be used as tutorials for beginners and template workflows that can be adjusted to individual needs.

Many potential users are not yet familiar with programming at all, but interested in the method and eager to experiment with it a little before deciding if it is worth learning a new set of skills to use it to its full extent. For them the learning curve of a jupyter notebook is still too steep. That at least was the feedback we received in our workshops which we organized to get feedback from scholars: the wish for a GUI to access at least the basic functionalities was expressed frequently. To meet this demand, we started the development of a 'GUI Demonstrator' that mirrors the working steps and explanations in the notebooks, and allows users to analyse their own texts using LDA with a limited set of options.

The current version, that is implemented in the FLASK microframework (http://flask.pocoo.org/) and runs within a browser window (Fig 1.), includes all steps necessary to get from a number of raw text files (txt and xml file formats are supported) to a visualized output, currently an interactive heat map showing the distribution of topics over texts (Fig. 2). As the quality of results depends on removing frequent words that appear in all texts, users can decide on the number of most frequent words to remove, or provide their own stopword list. They can control the number of topics to be generated, and the number of iterations the algorithm should run. The latter is important, because a large number of iterations will produce more stable results, but the algorithm will take longer to finish the task.

The next working steps include the implementation of standalone graphics in the Qt library (https://www1.qt.io/), and in allowing for flexibility in the choice and use of the results and outputs users are specifically interested in. The possibility to include metadata and evaluation results is another focus for upcoming developments, e.g. to sort text in the output heatmap according to different categories, or topics according their quality indicated by evaluation metrics.

Both the library and the Demonstrator as a standalone executable for Windows and OSX are open source and available on Github (https://github.com/DARIAH-DE/Topics).

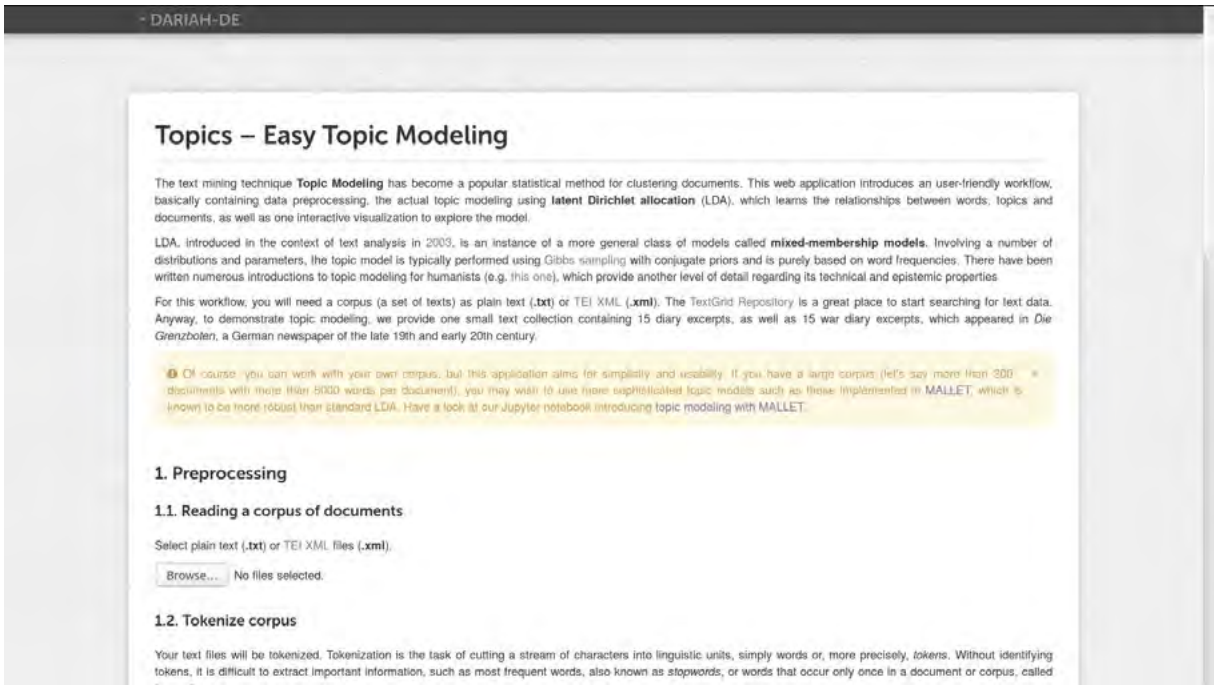


Figure 1: Screenshot of the upper end of the input screen in the current version of the GUI Demonstrator.

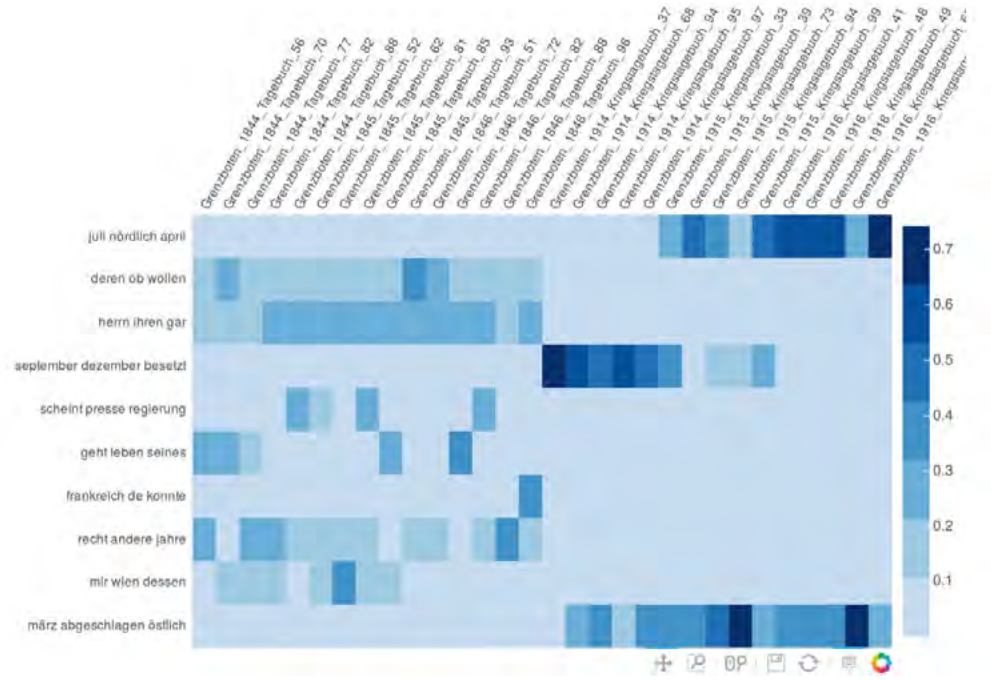


Figure 2: Example for an interactive heatmap output in the current version of the GUI Demonstrator.

References

Blei, David M. (2012): „Probabilistic Topic Models“, in *Communication of the ACM* 55, Nr. 4 (2012): 77–84. doi:10.1145/2133806.2133826.

McCallum, Andrew K. (2002): *MALLET : A Machine Learning for Language Toolkit*. http://mallet.cs.umass.edu.

Rehurek, Radim/ Sojka, Petr (2010): "Software framework for topic modelling with large corpora." *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*.

Steyvers, Mark/ Griffiths, Tom (2006): „Probabilistic Topic Models“, in *Latent Semantic Analysis: A Road to Meaning*, herausgegeben von T. Landauer, D. McNamara, S. Dennis, und W. Kintsch. Laurence Erlbaum.

Eliminar barreras para construir puentes a través de la Web semántica: Isidore, un buscador trilingüe para las Ciencias Humanas y Sociales

Sthephane Pouyllau
stephane.pouyllau@cnrs.fr
CNRS, Huma-Num, France

Laurent Capelli
laurent.capelli@huma-num.fr
CNRS, Huma-Num, France

Adeline Joffres
adeline.joffres@huma-num.fr
CNRS, Huma-Num, France

Desseigne Adrien
adrien.desseigne@huma-num.fr
CNRS, Huma-Num, France

Gautier Hélène
helene.gautier@huma-num.fr
CNRS, Huma-Num, France

"ISIDORE" es un buscador creado por una infraestructura francesa de investigación: la TGIR Huma-Num. No solamente ofrece una plataforma de búsqueda, sino que también normaliza y enriquece los datos y metadatos que cosecha, integrándolos en la Web semántica.

Desde hace dos años, la plataforma "ISIDORE" lanzada en diciembre de 2010 puede enriquecer e indizar metadatos y recursos digitales en Ciencias Humanas y Sociales (CHS) en 3 idiomas: francés, inglés y español. Esta posibilidad es un gran avance para "ISIDORE" y abre perspectivas de colaboración científica en distintos continentes. Conforme a los principios de ciencia abierta y respetuosa de los principios "FAIR", este enfoque permite el intercambio cada vez más estrecho de numerosos datos integrados en la Web de datos.

De hecho, ahora cuenta con más de 5 millones de recursos digitales (libros, revistas científicas, artículos científicos, anuncios y programas de eventos, convocatorias, blogs, mapas, archivos, documentos audiovisuales, etc.) interconectados mediante referenciales, indizados por un motor de búsqueda. Estos datos enriquecidos son accesibles en tres formas : un portal web (http://www.rechercheisidore.fr/), una API (http://www.rechercheisidore.fr/api) y un acceso unificado (http://www.rechercheisidore.fr/sparql) en una óptica de metadatos abiertos según el formalismo RDF. De hecho "Isidore" promueve el uso de estándares interoperables.

Así, "ISIDORE" es capaz de cosechar corpus y bases de datos en español y en inglés, pero ofrece también enriquecimientos multilingües enlazados entre sí mediante las posibilidades ofrecidas por el linked data. Para lograrlo, "ISIDORE" utiliza las alineaciones de los conceptos entre tesauros y vocabularios disponibles en la web semántica como los Registros de Autoridad y Referencia de Materia de la Biblioteca Nacional de España (http://datos.bne.es/temas) para los datos en español, o bien los encabezamientos de materias del referencial de la Biblioteca del Congreso de EEUU (Library of Congress Subject Headings – LCSH, http://id.loc.gov/authorities/subjects.html) para los datos en inglès. De esta manera, los conceptos de estos dos referenciales mayores están alineados en parte con los conceptos del referencial francès Rameau de la BnF (Biblioteca Nacional Francesa).

Junto con tesauros multilingües ya integrados en "ISIDORE" (como Pactols, Lexvo, GeoEthno, GEMET, etc.), y con el sistema de categorización/clasificación también multilingüe (categorías del sistema francés de archivos abiertos HAL-SHS y del sistema "Calenda" de anuncios de eventos científicos y convocatorias del CLEO-CNRS), "ISIDORE" es capaz de proponer un sistema de enriquecimientos/clasificación en 3 idiomas con la posibilidad de cambiar de idioma durante la búsqueda en la interfaz del portal www.rechercheisidore.fr y de la interfaz para tableta/smartphone (http://m.rechercheisidore.fr/?lang=es).

Esta característica permite al investigador no-francófono de tener acceso a nuevos datos con enriquecimientos, enlaces y clasificaciones en tres idiomas, permitiéndole medir, por ejemplo, el interés de fuentes en idioma francés sugeridas por "ISIDORE" en la interfaz (bien sea en inglés o en español).

De momento, casi 220 000 documentos en español se encuentran en "ISIDORE" y la plataforma contempla cosechar aún más en el futuro.

En paralelo, otros desarrollos que hacen de "ISIDORE" una herramienta cada vez más personalizada, han venido completando sus funcionalidades y abriendo perspectivas. Es el caso del widget "IMoCO", ISIDORE Motor Constructor que permite crear sólo en unos clics, una interfaz de consulta personalizada de los recursos disponibles en la plataforma "ISIDORE" (por ejemplo recursos específicos sobre un tema). Así, "IMoCO" está diseñado para los usuarios que deseen incluir en su sitio Web el buscador "ISIDORE" haciendo una simple copia/pega de un código HTML. Simple y neutral, se adapta a la mayoría de los sitios Web. Además, IMoCO puede ser totalmente personalizado con sus estilos CSS. También el widget multilingue WordPress "ISIDORE suggestions" (https://fr.wordpress.

org/plugins/isidore-suggestions/) permite al usuario de blogs WordPress conseguir sugerencias de documentos presentes en "ISIDORE". Estas sugerencias se hacen basadas en palabras claves asociadas al artículo que el usuario esté leyendo. Es posible afinar su búsqueda, subrayando el contenido del artículo consultado o seleccionando una o varias disciplinas.

Con este poster, quisiéramos mostrar todas las posibilidades que ofrece actualmente "ISIDORE" para el mundo hispánico en CHS, con lo que nos permite también contemplar colaboraciones fructuosas que contribuirán sin duda a alimentar esta plataforma y, al final, a enriquecer las búsquedas de investigadores o estudiantes francófonos de "ISIDORE" que tendrían acceso a más recursos en español, así como las investigaciones de usuarios hispanohablantes y angloparlantes. También tener la oportunidad de presentar este póster en el cuadro del congreso DH en México permitiría intercambiar con usuarios potenciales sobre sus necesidades, y alrededor de los futuros desarrollos de la plataforma.

SSK by example. Make your Arts and Humanities research go standard

Marie Puren
marie.puren@inria.fr
INRIA, France

Laurent Romary
laurent.romary@inria.fr
INRIA, France; Centre Marc Bloch, Germany

Lionel Tadjou
lionel.tadonfouet@inria.fr
INRIA, France

Charles Riondet
charles.riondet@inria.fr
INRIA, France

Dorian Seillier
dorian.seillier@inria.fr
INRIA, France

Arts and Humanities research has to address new challenges raised by the increasing amount of digital sources, contents and tools. New digital practices and protocols, new digital methodologies and services, new software and databases, offer a completely renewed framework for research, and encourage the emergence of a next generation of digitally-aware scholars.

Digital infrastructures, such as PARTHENOS, aim at supporting and accompanying the rise of this new generation of scholars by offering innovative solutions to connect digital tools and contents to Arts and Humanities researchers' needs. PARTHENOS has thus acknowledged the growing importance to develop a data-centered strategy for the management of scientific data (European Commission, 2010), and is currently developing the Standardization Survival Kit ("SSK") to help Arts and Humanities scholars understand the crucial role that proper data modelling and standards have to play in making digital contents sustainable, interoperable and reusable.

Accompanied by a live demo of the website[1], the poster will be composed of three parts: introducing the Standardization Survival Kit or "SSK", using the SSK, customizing the SSK.

Even if it is not obvious that the Arts and Humanities would be well-suited to taking up the technological prerequisites of standardization, it is yet essential that standardization takes a crucial role in the management of Arts and Humanities data. In this framework, this poster will present the Standardization Survival Kit, an overlay platform dedicated to promote a wider use of standards within Arts and Humanities. This comprehensive interface aims at providing documentation and resources concerning standards (especially authoritative references for each standard such as sources, Standard Development Organizations), and at covering three types of activities related to the deployment and use of standards in the Arts and Humanities scholarship: documenting existing standards by providing reference materials, supporting the adoption of standards, and communicating with all Arts and Humanities research communities.

The SSK is designed as a comprehensive interface for guiding Arts and Humanities scholars through all available resources (collected within a dedicated Zotero library[2]), on the basis of reference scenarios identified since the beginning of the project (PARTHENOS, 2016). The interface intends to provide a single entry point for both novice and advanced scholars in the domain of digital methods, so that they can have quick access to the information needed for managing digital content, or applying the appropriate method in a scholarly context. Users will be able to explore the platform according to their needs, thanks to precise research criteria: disciplines, standards, research activities and research objects. The poster will show how an Arts and Humanities scholar can navigate the Standardization Survival Kit website, by taking the example of an actual reference scenario. A live demo of the interface will also accompany the presentation, so that those interested in the poster will be able to search the website according to their needs.

To stress the importance of standards for Arts and Humanities scholarly work, let us take the example of a sociologist who is a novice in digital methods, but who wants to disseminate a collection of field survey data online, so that they could be used by other researchers in the long-term. By browsing in the SSK, she or he will find

1 The beta-version of the website can be found here: https://ssk-application.parthenos.d4science.org/ssk/#/scenarios
2 https://www.zotero.org/groups/427927/parthenos-wp4

a standardized scenario that could be perfectly suited to her or his needs: "Encode and modelize field surveys for their online dissemination". The poster will follow this researcher exploring this reference scenario, and going through its nine steps[3] with the associated resources. Let us take some of the scenario's steps as examples:

- the fourth step "Anonymize" offers a curated and up-to-date list of resources to help the researcher respect ethical practices and adopt proven techniques for anonymizing the collected data.
- the second and sixth steps stress on the importance of using tested standard - such as EAD to "Collect and classify" the data, and TEI to "Transcribe the interviews" -, highlight the importance of proper data modelling before disseminating them, and give access to appropriate resources on the subject.

More advanced users will also be able to edit the scenarios themselves, by submitting new resources or adding new steps. They can also create new scenarios. The SSK scenarios and steps can be easily extended, reused and customized, thanks to their flexible data model in TEI[4]. A dedicated interface in the Standardization Survival Kit will enable users to make suggestions, automatically converted in TEI according to the appropriate schema. The poster will present this interface and the associated functionalities. And for those who will be eager to test it, a live demo will be provided.

References

Romary, L., Banski, P., Bowers, J., Degl'Innocenti, E., Ďurčo, M., Giacomi, R., Illmayer, K., et al. (2017). *Report on Standardization (Draft)*. Technical Report Inria https://hal.inria.fr/hal-01560563 (accessed 27 April 2018).

Romary, L., Degl'Innocenti, E., Illmayer, K., Joffres, A., Kraikamp, E., Larrousse, N., Ogrodniczuk, M., Puren, M., Riondet, C. and Seillier, D. (2016). *Standardization Survival Kit (Draft)*. Research Report Inria https://hal.inria.fr/hal-01513531 (accessed 27 April 2018).

(2018). *SSK: Development of the Standardization Survival Kit*. XSLT ParthenosWP4 https://github.com/ParthenosWP4/SSK (accessed 26 April 2018).

Riding the Wave. How Europe can gain from the rising tide of scientific data, *FOSTER FACILITATE OPEN SCIENCE TRAINING FOR EUROPEAN RESEARCH* https://www.fosteropenscience.eu/content/riding-wave-how-europe-can-gain-rising-tide-scientific-data (accessed 26 April 2018a).

Standard Survival Kit https://ssk-application.parthenos.d4science.org/ssk/#/ (accessed 26 April 2018b).

3 1. Obtain the informed consent of the participants, 2. Collect and Classify, 3. Select and digitize, 4. Anonymize, 5. Convert into sustainable formats, 6. Transcrive the interviews, 7. Add metadata, 8. Contextualize the research, 9. Disseminate and archive.
4 https://github.com/ParthenosWP4/SSK/spec

Monroe Work Today: Unearthing the Geography of US Lynching Violence

RJ Ramey
rj@findauut.com
Auut Studio, United States of America

MonroeWorkToday.org, launched in November 2016, is a digital history project that synthesizes current historical research on the scope of American lynchings. The website was updated again in October 2017 with additional content digitized from Tuskegee University Archives.

Lynchings in the United States were perpetrated as homegrown acts, not orchestrated regionally in any way. This exhibit focuses on people of color murdered over 100 years in this fashion under the pretext of white supremacy. Yet unlike most academic studies, the project does not compartmentalize by region (e.g. the South or West) or by group (e.g. Mexican-Americans). By contrast, *Monroe Work Today* is the first of its kind to use web technologies to visualize the entirety of these documented events, connecting scholarship about African Americans, Native Peoples, Mexicans, Sicilians and Chinese immigrants across the United States (Carrigan and Webb, 2013) (Frazier, 2015) (Pfaelzer, 2007) (Pfeifer, 2013) (and others). Through four years of work, Auut Studio meticulously created a database and directory in the form of a map, compiling all modern academic research with century-old archives of the Tuskegee Institute. This national map carries the names of 4166 victims of lynchings and nearly 600 other victims of racialized mob violence. The project gives clarity to the sheer extent of the murders.

Previous inquiries into the lynching record have relied on tabulations and statistics, enumerating one tally for each state or county – such as 531 lynchings in Georgia vs. 205 in Kentucky, etc. (Tolnay and Beck, 1995) (Guzman, c.1960) (Pfeifer, 2013). This project, however, transforms the public's interaction with **each** lynching using maps and extensive contextual narrative. Its goal is to spawn a public discussion about the logic of white supremacy.

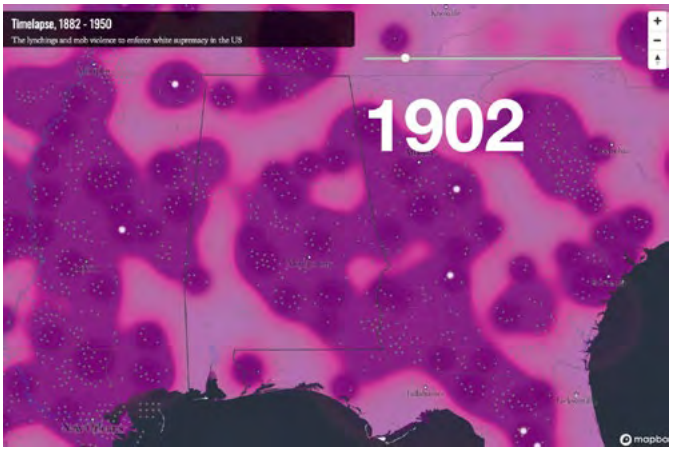
As a second phase to the project, the author now proposes a novel approach to using GIS to understand these murders. Acts of lynching are better examined like other crime data: not as tallies, but rather as incidents with a geographic location. As the commission of overt intimidation over people of color, they were in fact perpetrated with a specific geography in mind. The terrorizing effect was intended to carry over the nearby locale: it enforced the racial order "around **here**." In this context, maps of smaller areas may better recreate the historical truth about lynching, and a geospatial visualization of the regional landscape may better illuminate the original effect of these individual violent acts.

A new computer model created by the author animates regional maps of the USA, weighting the nearby radius

around a murder but also the persistence of its memory over a span of many years. The model makes certain blanket assumptions about the duration of trauma and fear–how long does the grotesque murder of a neighbor dissuade one's actions? These are starting assumptions which the author readily admits may be **wrong**, but they are coded as parameters. This allows different scholars for the first time to test their various interpretations of historical trauma and compare the visual output of competing viewpoints in the model.

This geo-temporal-visual model has the potential to drastically reframe the academic interpretation of lynching by unearthing multiple evolving shapes of the pockets of terror in the historical United States. As a stepping point for future research, this model for the broad reach of real, acute fear could be laid upon a map with other major events in the history of the Jim Crow South and brave acts of popular resistance.

In this poster session, the author will demonstrate the software model to attendees, exchange ideas and suggestions, as well as interrogate on-screen with them several new maps created with the model.



References

Berg, M. (2011). *Popular Justice: A History of Lynching in America*. Chicago: Ivan R. Dee.

Carrigan, W. (2004). *The Making of a Lynching Culture: Violence and Vigilantism in Central Texas 1836-1916*. Urbana: University of Illinois Press.

Carrigan, W. and Webb, C. (2013). *Forgotten Dead: Mob Violence against Mexicans in the United States, 1848-1928*. New York: Oxford University Press.

Frazier, H. (2015). *Lynchings in Kansas, 1850s-1932*. Jefferson, NC: McFarland Publishers.

Frazier, H. (2009). *Lynchings in Missouri, 1803-1981*. Jefferson, NC: McFarland Publishers.

Gonzales-Day, K. (2006). *Lynchings in the West, 1850-1935*. Durham, NC: Duke University Press.

Guzman, J (ed.). (c.1960). Lynching records of Tuskegee Institute as a database typewritten on paper. Tuskegee, AL: Tuskegee University Archives.

Leonard, S. (2002). *Lynching in Colorado, 1859-1919.* Boulder: University Press of Colorado.

Loewen, J. (2005). *Sundown Towns: A Hidden Dimension of American Racism*. New York: New Press.

Newkirk, V. (2009). *Lynching in North Carolina: A History, 1865-1941*. Jefferson, NC: McFarland & Company Inc.

Pfaelzer, J. (2007). *Driven Out: The Forgotten War Against Chinese Americans*. New York: Random House.

Pfeifer, M (ed.). (2013). *Lynching Beyond Dixie: American Mob Violence Outside the South*. University of Illinois Press.

Phillips, P. (2016). *Blood at the Root: A Racial Cleansing in America.* W.W. Norton & Company.

Rushdy, A. (2012). *American Lynching.* New Haven: Yale University Press.

Tolnay, S. and Beck, E.M. (1995). *A Festival of Violence: An Analysis of Southern Lynchings, 1882-1930*. Urbana: University of Illinois Press.

Thompson, V. (2014). *Clinton, Louisiana: Society, Politics, and Race Relations in a Nineteenth-Century Southern Small Town*. Lafayette: University of Louisiana at Lafayette Press.

Educational Bridges: Understanding Conservation Dynamics in the Amazon through The Calha Norte Portal

Hannah Mabel Reardon
hannahmreardon@gmail.com
McGill University, Canada

Calha Norte is the northernmost region of the Brazilian Amazon, and the largest mosaic of protected areas in the world, encompassing nearly 14 million hectares. Given the vastness of this area, government enforcement of parks and conservation zones can be poor, and scarce resources prevent authorities from providing much-needed support to the inhabitants of protected areas. This poster focuses on the Calha Norte Portal, a digital project that constitutes a personal initiative to encourage awareness of conservation efforts in the region. The portal is an educational tool intended to demonstrate the power of digital technologies for fostering greater transparency in conservation management. It also aims to provide a clearer understanding of the social, political, economic and historical dynamics which have shaped the challenges to protecting the Amazon forest today.

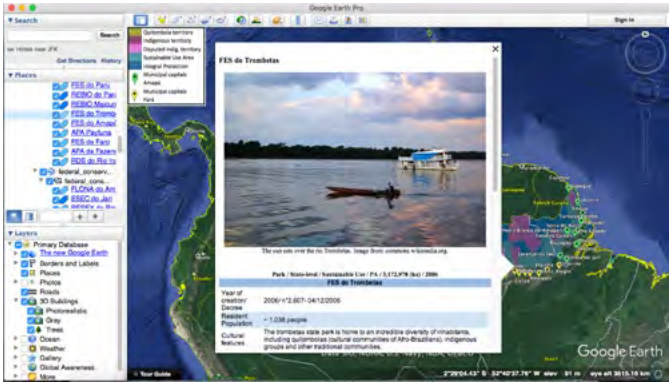
The data for the Calha Norte Portal was gathered during my work with the Social Policy department of the Amazonian Institute for Man and the Environment (Imazon), an environmental NGO based in Belém. In accordance with the department's focus on communities in

Ex. 1&2: A bauxite mine in the Saraca-Taquera national park. Top, a satellite image of the mine in 1986, bottom, the same mine in 2017.

the Calha Norte region, I compiled information from various sources about the region's history, cultural diversity, transportation networks, governing bodies, development indices, demographics, economic activities, protected area implementation, and accessibility. This data was then used to create the Calha Norte Portal, a website and blog with an interactive Google map of the municipal capitals and protected areas in the region. The Google Earth application allows the user to navigate through protected areas, indigenous territories, maroon communities, and municipalities. At a click, each area on the map displays a pop-up window with historical information, demographic statistics, economic and political data, photos, deforestation figures and an implementation index for protected areas. Furthermore, users can look back in time at satellite images from 1960 to the present and visualize patterns of deforestation, and urban sprawl over time.

The project focuses mainly on political, economic, historical, cultural and social data for populations in protected areas and the surrounding municipalities. As an anthropologist, I am particularly interested in dispelling the myth of Amazonia as an uninhabited biological entity, and exposing the important historical dynamics which have shaped the Amazon region as it is today. Understanding the human forces which have pushed the economic development of the region is a crucial first step for conservation policy which can protect both human livelihoods and biodiversity, in line with current sustainable development benchmarks. I also hope to draw attention to the power of digital technologies for overcoming communication barriers between isolated regions and institutional bodies, a major issue in developing informed and tailored conservation policy.

My hope is that, in breaking down the collected data in a visual, interactive format, the uninitiated user will be able to play with the information and learn about the region in any way that suits their interests. The user's guide and tutorials available on the portal offer a guided introduction, but the stand-alone map itself is meant to be played with, manipulated and explored, in ways that dismantle a traditional historical narrative. This poster presentation will elaborate on the features of the Calha Norte Portal and its contribution to greater awareness of regional conservation efforts. The overarching aim is to convey the importance of transparency in the institutionalization of protected areas and to encourage a more thorough understanding of the cultural fabric of the Northern Amazon region, so that research and conservation initiatives might be better tailored to the realities of local communities and their involvement in the protection of the natural resources upon which their livelihoods depend.



Ex.3: Calha Norte in Google Earth. The portal offers users the opportunity to navigate through the online version of the map, or the option to download Google Earth and the Calha Norte KMZ file, for a more complete user experience.

References

Reardon, H. (2018). *Calha Norte Portal.* [Online] Available at: calhanorteportal.com

Building a Community Driven Corpus of Historical Newspapers

Claudia Resch
claudia.resch@oeaw.ac.at
Austrian Academy of Sciences, Austria

Dario Kampkaspar
dario.kampkaspar@oeaw.ac.at
Austrian Academy of Sciences, Austria

Daniela Fasching
daniela.fasching@oeaw.ac.at
Austrian Academy of Sciences, Austria

Vanessa Hannesschläger
vanessa.hannesschlaeger@oeaw.ac.at
Austrian Academy of Sciences, Austria

Daniel Schopper
daniel.schopper@oeaw.ac.at
Austrian Academy of Sciences, Austria

Faced with the challenge of organizing the digital processing and publication of a large collection of historical newspaper data from the 18th century publication known as the *Wien[n]erisches Diarium*, a small project located at the Austrian Centre for Digital Humanities (ACDH) in Vienna has opted for a user-centred, participatory approach and employs methods of community involvement to tackle the specific challenges that arise from the particular qualities of the historical source material.

Founded in 1703, the newspaper under investigation is among the oldest periodical publications still being published today, and was regarded as the most important newspaper of the Habsburg Monarchy for a considerable time span during the 18th century. The value and significance of the newspaper as a source is undeniable, not only due to the density of the information it contains, but also because of the virtually gapless preservation of its run from its foundation in 1703 up until today and the full availability of these original sources. So far, no computer-based processing of this historical data cache has been undertaken. The ACDH project aims at facilitating the use of the source in a digital environment and creating a cornerstone resource, making the *Diarium* freely and easily available to researchers everywhere.

The more than 10.000 issues from the 18th century constitute a mass of text and data. As resources are limited, a number of issues manageable within the project's run had to be selected. For now, the project will thoroughly edit a corpus of approximately 500 issues from all decades of the 18th century. The priority is the quality of the data and the creation of a reliable HTR model that will improve automatic processing and pave the way for expanding or completing the existing corpus at a later point.

As not all queries and research questions that may be posed to the sources can be anticipated, it is the project's primary aim to secure and process the full text of the newspaper in a way that does not disregard or omit any of the relevant information – regardless of the querying researcher's field or discipline. In order to determine which aspects are of particular relevance, where the interests of different disciplinary fields overlap, and how the issues should be prepared and presented to make them useful for the largest number of (academic) users, the digitization project has devised a way to work closely with researchers from various backgrounds.

The project's **community-driven approach** invites and relies on participation on several levels, effectively allowing future users to follow, accompany and shape the project throughout the course of its duration. The following three methods of user involvement were or are being employed in the course of the digitization and annotation process:

1) In spring 2017, a **call for nominations** promoted via digital channels and the print version of the newspaper provided an opportunity for prospective users to nominate specific issues or sets of issues for digitization.
2) While the text recognition process does not involve users, the project team nevertheless upholds the principle of transparency by allowing users to track the progress of the procedure: A **reporting tool** developed for this purpose is accessible via the project website, provides a current list of the issues selected for processing and allows users to track the daily progress in real time.
3) A series of community-driven **annotate-a-thons** allow the project team to survey and adapt to the user community's needs. Consulted as experts and prospective users, (peer) researchers are involved in the annotation process early on and contribute specialised knowledge to the enrichment of the data.

To ensure users' ongoing engagement with the texts even beyond the initial phase and to provide a way to preserve and publicize the results, the platform has been designed with continuous annotation activities in mind. Any user shall be able to make annotations and contribute to the encoding source via the web-app, which will support four basic types of annotations: 1) full text, 2) named entity identification, 3) text or layout corrections, and 4) semantic or structural annotations.

In pioneering a user-centred approach in the development of a digital newspaper resource, the *Diarium* project generates new insights in the potential of community involvement for similar projects. It roadtests methods for motivating both digital and 'traditional' humanities researchers to contribute to a collaborative resource and for creating highly sustainable and re-usable resources

that will meet the needs of diverse user communities, and encourage ongoing engagement.

Expanding Communities of Practice: The Digital Humanities Research Institute Model

Lisa Rhody
lrhody@gc.cuny.edu
CUNY Graduate Center, United States of America

Hannah Aizenmann
haizenmann@gc.cuny.edu
CUNY Graduate Center, United States of America

Kelsey Chatlosh
kchatlosh@gradcenter.cuny.edu
CUNY Graduate Center, United States of America

Kristen Hackett
khackett@gradcenter.cuny.edu
CUNY Graduate Center, United States of America

Jojo Karlin
jojo.karlin@gmail.com
CUNY Graduate Center, United States of America

Javier Otero Peña
javo01@gmail.com
CUNY Graduate Center, United States of America

Rachel Rakov
rrakov@gradcenter.cuny.edu
CUNY Graduate Center, United States of America

Patrick Smyth
patricksmyth001@gmail.com
CUNY Graduate Center, United States of America

Patrick Sweeney
pswee001@gmail.com
CUNY Graduate Center, United States of America

Stephen Zweibel
szweibel@gc.cuny.edu
CUNY Graduate Center, United States of America

In his preface to *Doing Digital Humanities: Practice, Training, Research* (2016), Ray Siemens points out that imagining digital humanities as a community of practice wherein participants come into conversation with one another over shared approaches to craft establishes a "methodological commons" where fields intersect by sharing their work processes. Presenting a taxonomy of approaches to training that span from the informal to the formal within the methodological commons, Siemens suggests that the variety of possible approaches builds an infrastructure for "self-determination" in humanists' approach to learning useful skills. Somewhere between informal consultations and formal degree programs, short courses and "bootcamps" offer professional and research skill development opportunities that scholars can choose from based on their most pressing needs.

Digital humanities skill development cannot be automated; it is resource intensive. It depends upon a limited number of people to deliver highly personalized training to relatively small cohorts of scholars--a model that is difficult to fund and harder to scale. As interest in and demand for training in digital humanities research methods continues to increase, overall capacity to reach the needs and interests of diverse populations of scholars in the wide range of institutional contexts where they do their work has not kept pace.

Committed to building a vibrant community of scholars who deploy a critical use of digital technologies in their teaching and research, the CUNY Graduate Center will run its fourth week-long digital research institute in January 2018. Between 2016 and 2017, GC Digital Initiatives offered a combined 100 hours of instruction on digital research methods to more than 100 students, faculty, staff, and librarians across the CUNY system.[1] Our institute model has focused on reducing the time required to develop new curricula through sharing and versioning, expanding the number of participants per institute through collaborative learning environments, and supporting participants through community-building. The success of our model is demonstrated by continued, growing interest from students, faculty, and staff each year.

As interest in digital humanities at universities, museums, libraries, and archives increases, so too does the demand for faculty, administrative staff, librarians, postdocs and graduate students who are tasked with expanding DH research and teaching capacity with relatively few resources. With funding from the National Endowment for the Humanities, we will be expanding our model to create a sustainable, reproducible model for digital methods training that can be adapted and used in a variety of institutional contexts. Our institute model is designed to integrate feedback so that it can be replicated, modified, and reproduced in new contexts, lowering the barrier to entry for digital humanities scholars by meeting scholars where they are rather than requiring participants to travel to receive training.

In June 2018, 15 individual participants will participate in the first Digital Humanities Research Institute. The DHRI emphasizes foundational technical skills, such as the command line, git, Python, and databases, that provide a flexible technology "stack" and that better enable DH researchers to become more confident autodidacts and mentors in their own right. While participants develop

1 GC Digital Research Institute http://cuny.is/gcdri

familiarity with useful tools, they learn more importantly how to navigate a computer's information architecture, read technical documentation, and reason through simple systems, leading to a greater conceptual vocabulary and increased confidence approaching technology with a critical eye. As participants learn skills to support their individual research goals and professional growth, they will also learn how to lead similar digital humanities institutes in their local communities over the following academic year. Through the process of iterating, refining, and building the institute model, we intend to share the lessons learned to increasingly wider communities of learners and build a network of curricular models and support.

Our poster will feature curricula, pedagogical materials such as datasets, and resources developed for the ten-day residential institute, where participants will explore interdisciplinary digital humanities research and teaching with leading DH scholars, develop core computational research skills through hands-on workshops, and begin developing versions of the DHRI for their own communities. We will share lessons learned and provide information about forthcoming institutes. Short video clips will feature our unique approach to digital humanities pedagogy and interviews with previous institute instructors and participants.

References

Crompton, Constance, Richard J. Lane, and Ray Siemens. *Doing Digital Humanities: Practice, Training, Research*. Routledge, 2016.

Hispanic 18th Connect: una nueva plataforma para la investigación digital en español

Rubria Rocha
rubria@tamu.edu
Texas A&M University, United States of America

Laura Mandell
mandell@tamu.edu
Texas A&M University, United States of America

18thConnect.org es una comunidad en línea de académicos que realizan revisión por pares de materiales digitales obteniendo metadatos de los mismos para colocarlos en nuestro buscador que está disponible de forma gratuita. Los materiales patentados, tales como Early English Books Online (EEBO) y Eighteenth-Century Collections Online (ECCO) también se pueden buscar a través de nuestro asistente de búsqueda. Además, los libros y documentos de las colecciones EEBO y ECCO de la literatura moderna temprana están disponibles en 18thConnect para que los usuarios corrijan sus transcripciones mecánicas, a través de nuestra herramienta TypeWright. Cualquier persona que corrige un documento puede, entonces, tenerlo tanto en formato de texto plano como en XSLT. Exhortamos a los especialistas a corregir textos, crear ediciones digitales en GitHub o enviarlas al TEI Archiving and Publishing Access Service (TAPAS), así como a enviar sus ediciones a 18thConnect para su revisión por pares y para publicarlas en acceso abierto.

Mientras que 18thConnect ha estado en línea desde el 2009, la idea de crear Hispanic 18th Connect, resultó de la necesidad de ayudar en el proyecto *Primeros Libros* de la Texas A & M University, financiado por la NEH, para desarrollar OCR para documentos históricos escritos en español. Dados los resultados positivos en el proyecto *de Primeros Libros* creemos que es momento de extender nuestros recursos para su uso en otras bibliotecas hispanas y para hacer disponibles sus colecciones en nuestro sitio.

Para comenzar este proceso, elegimos traducir y adaptar nuestra interfaz al idioma español y a la cultura hispana. Nuestra justificación es que el español es la segunda lengua materna más hablada en el mundo, así mismo, el 18% de los habitantes en los Estados Unidos habla español y se estima que para el 2060, E.U. sea el segundo país con mayor número de hispanohablantes después de México (Llorente, 2017). Además, la cultura hispana permea en múltiples países, donde también se puede observar una especial motivación por conocer más de esta cultura. Esto último, se ve reflejado en el creciente interés por hacer investigación y desarrollar proyectos relacionados a la lengua y cultura hispanas desde las humanidades digitales (Gutiérrez y Ortega, 2014 y AtlasCS-HD, 2015).

El proyecto de Hispanic 18th Connect consiste en 6 etapas: 1) traducción al español; 2) revisión del funcionamiento de la interfaz; 3) prueba piloto con colegas humanistas cuyos intereses sean en estudios hispánicos con y sin experiencia previa en la interfaz de 18th Connect en inglés; 4) análisis de los resultados de la prueba piloto; 5) adición o modificación de contenidos de acuerdo a las respuestas y comentarios de la prueba piloto; 6) presentación oficial de la interfaz de Hispanic18th Connect.

La presentación del póster de Hispanic 18th Connect tiene varios aspectos a cubrir: por un lado, dar a conocer que la plataforma 18thConnect será más accesible para la comunidad hispanohablante por tener la opción de navegar en su sitio en español; presentar los retos que implicó la traducción de esta plataforma tanto en cuestión de términos, como en relación a los aspectos culturales que creemos pueden impactar (resultados de la etapa 4) y es dónde se pudiera visualizar cómo las características de la comunidad podrían modificar las instrucciones y/o las herramientas con las que cuenta 18thConnect para que pueda ser relevante en el estudio del siglo 18 hispano.

El principal objetivo para este primer momento, es ofrecer esta plataforma traducida al español, y darla

a conocer con el material de las colecciones existentes, y en un segundo momento, que es nuestro objetivo a mediano plazo, incluiremos nuevas colecciones y buscaremos identificar las posibles necesidades de nuevas herramientas que sirvan al estudio del siglo 18 hispano.

Nuestro póster mostrará el trabajo realizado en las 6 etapas, los retos, los hallazgos, los cambios y las diferencias respecto a la interfaz en inglés. Así mismo, contendrá la información más importante de la interfaz y ejemplos que demuestren la manera en que se realizarán las búsquedas de documentos en español (provenientes de EEBO y ECCO), la forma en que se corrigen documentos con TypeWright, y cómo se crean las ediciones digitales para que puedan ser sometidas a revisión por pares en Hispanic 18th Connect.

References

18thConnect Eighteenth-century Scholarship Online (2009). Available at: http://www.18thConnect.org (Accessed 17 November 2017).

AtlasCSHD (2015). Atlas de Ciencias Sociales y Humanidades Digitales. Available at: http://medialab.ugr.es/proyectos/atlas-de-ciencias-sociales-y-humanidades-digitales/ Mapa: http://grinugr.org/mapa/#

Llorente, A. (2017). ¿En qué países se habla español fuera de España y América Latina? *BBC Mundo*. Available at: http://www.bbc.com/mundo/noticias-america-latina-38021392 (Accessed 17 November 2017).

Ortega, É. y Gutiérrez, S. (2014). MapaHD. Una exploración de las Humanidades Digitales en español y portugués. In Romero, E. y Sánchez M. (eds), *Ciencias Sociales y Humanidades Digitales Técnicas, herramientas y experiencias de e-Research e investigación en colaboración*. CAC, Cuadernos Artesanos de Comunicación, pp. 101-128.

TAPAS, TEI Archiving and Publishing Access Service (2014). Available at: http://tapasproject.org/ (Accessed 30 April 2018).

Lorenzetti Digital

Elvis Andrés Rojas Rodríguez
elarojasrod@unal.edu.co
Universidad Nacional de Colombia, Colombia

Jose Nicolas Jaramillo Liévano
jonjaramilloli@unal.edu.co
Universidad Nacional de Colombia, Colombia

Lorenzetti Digital es un proyecto de historia digital e historia pública que pretende mostrar perspectivas de la Edad Media desde Latinoamérica a un público especializado, escolar y no especializado. Esto se hace buscando conexiones en el mundo medieval desde los frescos de Ambrogio Lorenzetti *Le Allegorie del Buono e Cattivo Governo e dei loro Effetti*, pintados en la ciudad de Siena, Italia, en el siglo XIV. A través de los personajes alegóricos del fresco se relacionan fuentes pictóricas y fuentes primarias textuales para reconstruir el contexto histórico de cada personaje.

El proyecto nace como un ejercicio académico estudiantil para las materias **Historia Digital** e **Historia Medieval**, a través de la plataforma wix. En principio solo se proyectaba como una herramienta digital de difusión del conocimiento histórico desde los estudiantes. Sin embargo, ahora las ambiciones del proyecto son más grandes. Lorenzetti Digital se proyecta como una herramienta de comunicación de la historia medieval y, por otro lado, como un repositorio de fuentes primarias medievales. Esto significa que el sitio web resolverá las necesidades de los usuarios brindándoles un primer acercamiento interactivo a la historia medieval, para luego profundizar en diferentes niveles de investigación a través de las fuentes primarias del repositorio. Para esto, implementamos herramientas como el HTML5 en vez de la plataforma wix. Para finales de 2018, se espera que el proyecto tenga un sitio web con dominio propio y que estén consolidados tanto los aspectos didácticos, visuales e investigativos como el repositorio.

Se trata, también, de dar una perspectiva de la historia europea medieval desde Latinoamérica. Consideramos que hay un **mercado** para la historia medieval en Colombia y Latinoamérica en general, propiciado por la industria del entretenimiento y desaprovechado por los historiadores. Por tanto, Lorenzetti Digital es una herramienta para los curiosos, los estudiantes y los investigadores por igual.

Lorenzetti Digital ha participado en las dos últimas conferencias de la International Federation for Public History, donde recibió críticas y comentarios útiles para el proyecto. El vínculo entre las humanidades digitales y la historia digital en este caso radica en la necesidad de aplicar herramientas informáticas como el desarrollo de sitios web para la divulgación del conocimiento histórico. Esto es, enseñar un período de la historia que recibe mucho interés por parte del público. Se trata de un ejercicio de estudiantes de historia para responder a una necesidad de la sociedad, que debe ser tratada de una forma no convencional para el historiador. Es decir, desde formas digitales con contenidos con potencial hipertextual.

Por supuesto, esto presenta muchos desafíos, preguntas y problemas de entrada. Primero, el hecho de que algunos estudiantes y profesores de historia se enfrenten al desarrollo y diseño web es algo para resaltar. Esto radica en un reto de interdisciplinariedad para lograr un desarrollo multimedia equilibrado entre historia, estética y funcionalidad.

Por otro lado, está el problema de la investigación. El ejercicio curatorial y de investigación que hay detrás

es un entramado de conexiones y redes complejas entre el fresco de Lorenzetti, la tradición iconográfica medieval, renacentista y antigua y las fuentes textuales como *La Divina Comedia*, *El Decamerón*, los mitos grecorromanos y los mitos judeocristianos. Lorenzetti Digital también se trata de un ejercicio investigativo nativo digital, donde la mayoría de esas fuentes están disponibles en línea. Sumado a esto nos enfrentamos a un reto de carácter epistemológico y temático. Sitios web sobre historia medieval existen en grandes cantidades con contenidos precisos y de alta rigurosidad, esto implica reconfigurar la estructura y narrativa del sitio web, manteniendo el nivel de asertividad sobre el pasado que se quiere comunicar. Esto a través de una lectura **por capas** o **hipertextual** de los personajes del fresco.

Por último, está el problema de hacer la plataforma más participativa, no solo proveyendo información, datos, interpretaciones y fuentes, sino también recibiendo ideas, comentarios, nuevos trabajos y retroalimentación en general. Así, no intentamos desarrollar un sitio web plano y estático, sino más bien uno dinámico y participativo, digno de la Web 2.0.

Hasta la fecha, no tenemos conocimiento de proyectos similares desarrollados desde Latinoamérica. Sin embargo, sí hay referencias de otras obras de arte que a través de sus personajes narran o explican ciertas ideas. Por eso mismo, el proyecto plantea más retos que soluciones, se trata de explorar el campo de las humanidades digitales, el diálogo con otros saberes fuera de las ciencias sociales y humanas, y de entablar una conversación virtual con el público del proyecto.

References

"An Empirical Framework For Learning (Not a Methodology)". Consultado el 11 de marzo de 2018. http://scrummethodology.com/.

Gentile, Gianni. Luigi Rogna y Anna Rossi. *Multistoria 1. La civilità medievale*. Vicenzo Bona: Editrice La Scuola, 2013

"Exposition Monet 2010 - RMN - Grand Palais - Paris". Consultado el 12 de marzo de 2018. http://www.monet2010.com/.

"Jheronimus Bosch - de Tuin der Lusten". Consultado el 12 de marzo de 2018. https://tuinderlusten-jheronimusbosch.ntr.nl/.

Skinner, Quentin. *El artista y la Filosofía Política.* Madrid: Cambridge University Press: 2009

Traditional Humanities Research and Interactive Mapping: Towards a User-Friendly Story of Two Worlds Collide

Vasileios Routsis
v.routsis@ucl.ac.uk
University College London, United Kingdom

Background

Historians have been using printed maps to illustrate movements of people, trends or any other kind of information for a long time. However, only recently the technological advances made it possible to produce digital interactive environments. Digital Humanities is born out of the need to use computational methods to facilitate humanities research, and data visualisation and digital cartography are two important areas within the Digital Humanities spectrum of research fields.

Objectives

This poster draws on the conclusion of the first phase of the *Mapping the Enlightenment: Intellectual Networks and the Making of Knowledge in the European Periphery* [1] (MtE) project funded by the Research Centre for Humanities in Greece[2]. The major deliverable was the creation of an online interactive mapping tool capable of indexing and visualising data of movements of Greek-speaking scholars during the Enlightenment Era. The first public version of the tool was released in late December 2017.

The project's goal is to enhance users' understanding of the emergence of modern science and technology as the expression of a dynamical geography. Addressing the spatiality of knowledge, it focuses on associating particular cultural traits with specific points on a map, and work on tracking down the various paths and encounters through which such cultural traits and the respective knowledge practices evolved.

By digitising and mapping the original data in a user-friendly way and using the latest modern technology available, the team behind this project hopes to re-emerge existing knowledge out of obscurity and ideally cultivate the ground that can lead to the development of new knowledge around this topic. Two of the major benefits of creating the digital tool include: i) availability/access to information: It is easier to access a website than a printed copy and ii) understanding of information: Interactive visualisation helps users explore and retrieve the information they want easier and in ways that may engage them further.

1 https://mapping-the-enlightenment.org/
2 https://www.rchumanities.gr/en/

The mapping tool

The tool uses a holistic approach to deliver the data with a unified all-in-one interface. Within this framework, there are no separate web pages, and the entirety of the available information is accessible via the tool's dashboard. Communication between the server and the clients is asynchronous. A considerable effort has been put to enrich the user experience by providing flexibility of the interface to improve data comprehension and to accommodate users' diverse navigational preferences and different screen resolutions (see Figure 1: The tool interface with its dashboard sidebar collapsed, and different windows opened at the same time.Figure 1,Figure 2: The tool interface with the sidebar open, the timeline placed at the bottom of the screen and an informational window opened. Figure 2 anFigure 3: In this screenshot, the timeline is contained within the sidebar with various data graphs open at the same time on top of the map.d Figure 3 in Appendix).

The tool is custom-built, and its technical infrastructure supports open-source software. On the server side, Apache, PostgreSQL, PHP, and GeoServer with PostGIS library is used. On the client side, the latest versions of the web standards model HTML5, CSS3, and JavaScript provide a modern and user-friendly user interface. Leaflet.js and D3.js are the main libraries that drive the mapping and visualisation system core. The combination of these technologies gives life to the historical data of the project by combining powerful visualisation components and a data-driven approach to DOM manipulation.

Discussion

Stemming from our own experience developing MtE, the poster intends to discuss and exchange ideas on how modern geohumanities projects can be designed and delivered successfully from their early to final stages. As it is known amongst Digital Humanities scholars, digitisation of information is far from being straight-forward and often involves highly complicated techniques to extract and transform the data to the desired format. Furthermore, as the digital age expands and the underlying technologies change, the problem of digital obsolescence lurks, the situation when a digital resource is no longer supported and readable. There may also be challenges in keeping the necessary balance between offering a simple and user-friendly environment without at the same time compromising the integrity and richness of the original data. In addition, each project may have different needs, peculiarities, and objectives. The discussions that are hoped to be made through this poster aim to lead to an exchange of knowledge from both technical and theoretical perspectives that will help to build better similar digital humanities projects in the future.

Finally, instead of a conclusion, it is worth mentioning that such projects and tools are especially valuable if they contribute in raising the academic and public interest in historical, cultural and societal matters - especially if these engage within a critical discourse. In this context, digital technology is used as a tool and means for these purposes and not as a self-referencing end.

Appendix

Figure 1: The tool interface with its dashboard sidebar collapsed, and different windows opened at the same time.

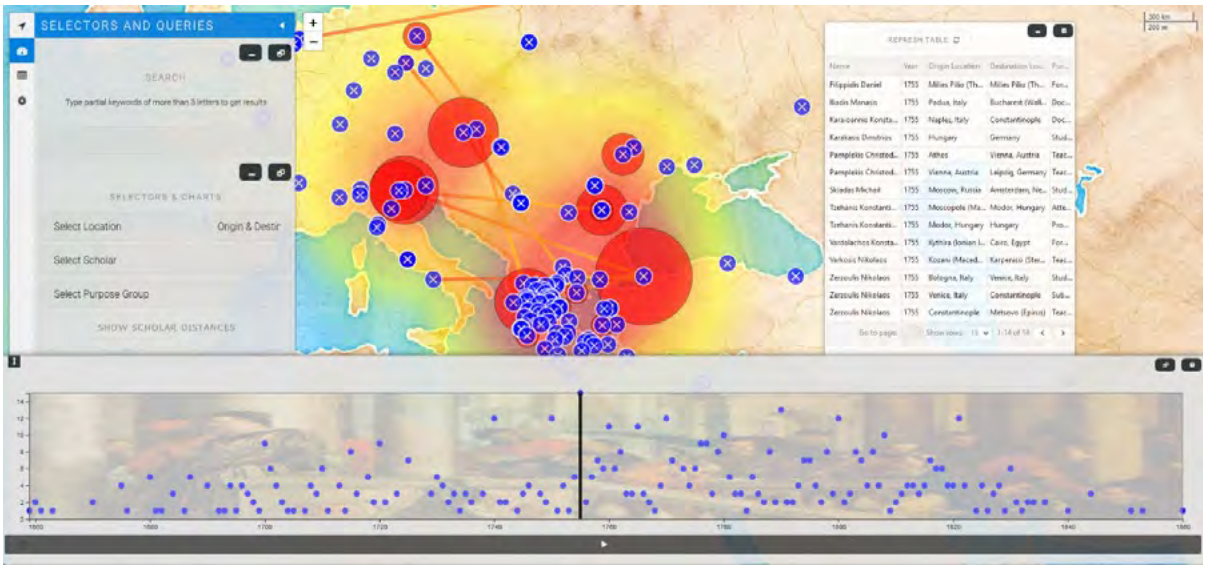


Figure 2: The tool interface with the sidebar open, the timeline placed at the bottom of the screen and an informational window opened.

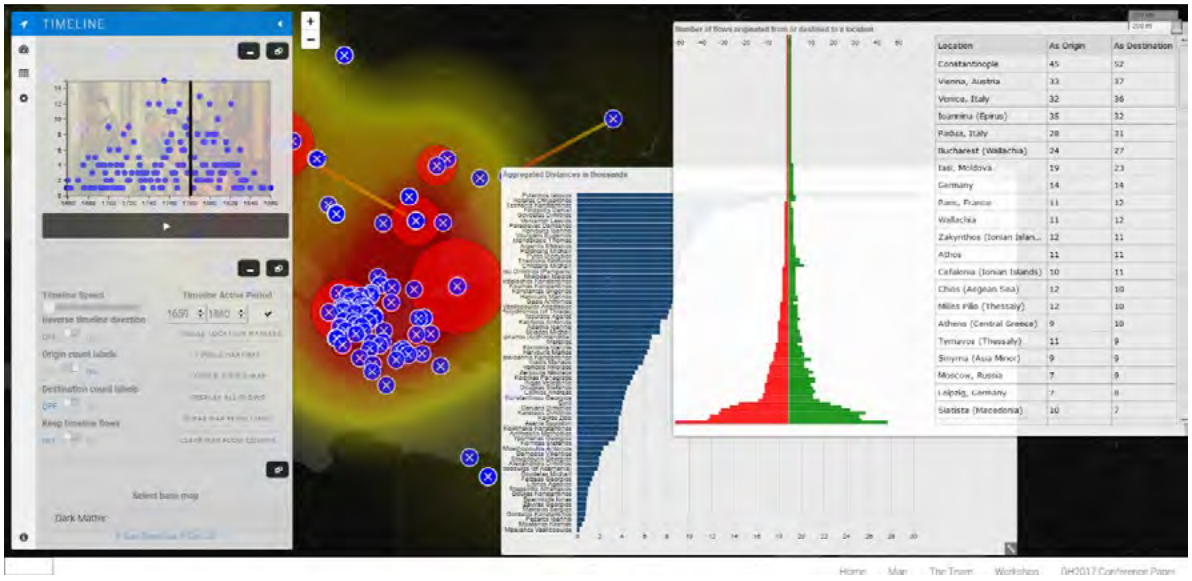


Figure 3: In this screenshot, the timeline is contained within the sidebar with various data graphs open at the same time on top of the map.

Digital Humanities Storytelling Heritage Lab

Mariana Ruiz Gonzalez Renteria
mruizgo1@asu.edu
Arizona State University, United States of America

Angélica Amezcua
aamezcu1@asu.edu
Arizona State University, United States of America

We are proposing to develop a storytelling tool that integrates multimodal mapping for use in language classrooms. Through a Digital Humanities approach on Storytelling Labs, we will be integrating the App StoryMapJS in order to create a storymap of their cultural heritages. This DH tool is very accessible and it will allow the student to engage mapping narrative through images, videos, music, writing and maps; so the heritage learners will interpret space by their personal print, and it will let to other readers from the course or outside the course, to confront other sociopolitical contexts.

The DH Storytelling Heritage Lab will reforge the spatial, and emotional relation from our heritage learners as individuals that can create their own mapping. In a pedagogical perspective the heritage learner will improve their writing, oral, listening and reading skills in Spanish. In a

linguistic research approach we will analyze the outcome of the students, a qualitative discourse analysis.

The workshop will be divided in two sections: the narrative without the DH tool: the student will engaged their narratives through family albums, objects, drawings, and recordings. The second part is to transform the storytelling into a digital narrative with the StoryMapJS. At the end of the Lab the student will have the opportunity to exhibit their narratives maps. The final stories will be compiled in a single web page for their distribution in different areas.

The idea to expand the personal stories and experience in the US of the heritage learners is essential for the course; so the learners engaged Spanish in the sociopolitical context of bilingualism of their own families and community. Their narratives, our narratives, will enrich the course.

Digital Humanities Under Your Fingertips: Tone Perfect as a Pedagogical Tool in Mandarin Chinese Second Language Studies and an Adaptable

Catherine Youngkyung Ryu
ryuc@msu.edu
Michigan State University, United States of America

Learning Chinese, now one of the most widely studied foreign languages in the United States and worldwide, can be challenging, especially for those without any prior exposure to the Chinese tonal system. Mandarin Chinese has four main tones, and one sound carries four different meanings, each tied to a particular tone. For example, "ma" in tone 1 means a "mother"; in tone 2, "hemp"; in tone 3, a "horse"; and in tone 4, a verb to "yell." Chinese as a tonal language thus differs fundamentally from how English speakers often use tone, pitch, and volume to add personal texture to communication. Novice Chinese learners are in great need of sustained and rigorous tonal training with multiple native speakers to develop and sharpen their tonal perception. However, it is usually not feasible to receive such training through in-class or online instruction constrained by time. Digital resources or tools designed for self-guided tone training can help remove such barriers and make tone learning more widely accessible to novice learners in particular.

How does Tone Perfect as a multimodal database render Mandarin Chinese (MC) tone learning accessible?

To create an optimal digital space of learning for each user with different backgrounds, skill sets, and learning styles, Tone Perfect includes by design multiple channels through which the users can synergistically integrate "seeing" and "hearing" into tone learning. Such multiple channels include: (1) a novel color-coded tone visualization (tone 1-yellow; tone 2-green; tone 3-blue; tone 4- red) to enable users to associate the tones with specific colors; (2) a waveform accompanying each sound file to enable the users to see how each of six native speakers produces the same target sound with a particular tone differently, which is also inflected by gender; (3) an additional conventional method of visualizing the tonal information with numbers, so as to aid users with color blindness and to reinforce what the user may have learned through formal instruction; and (4), both simplified and traditional Chinese characters together with a Romanization system (*pinyin*) to enable users to learn tone, sound, and character simultaneously.

How does Tone Perfect maximize its potential as a digital open source?

Tone Perfect is comprised of 9,864 audio assets representing an exhaustive set of monosyllabic sounds in Mandarin Chinese produced by six native speakers (3 female; 3 male). These audio files were produced at MSU to develop a Mandarin tone learning app game, Picky Birds (scheduled to be released in summer 2018). This app game, a digital tool for self-guided tone learning, is an outshoot of a 100% web-based experiment on the efficacy of different methods of visualizing the Mandarin tonal information (i.e., tone-number, tone-pitch contour, tone-color). The app itself was also subsequently utilized as an innovative experiment instrument for another Mandarin tone perception empirical experiment. That is to say, Tone Perfect now serve as an active digital repository that can be accessed by users from various backgrounds for different purposes, for example, as the audio resources for Mandarin Chinese sound tables, computer musical compositions, acoustic analysis, Mandarin linguistics experiments, etc. All audio files can be downloaded directly from the website to enable a wide range of applications of this resource.

Overview

This poster presentation features a multidisciplinary project, Tone Perfect—an interactive audio database—as an example of a multimodal approach to optimizing accessible learning in second language acquisition, specifically for Mandarin tone learning. Tone Perfect also serves as an example of an adaptable multipurpose database that simultaneously functions as an active repository, maximizing the preservation of existing digital materials and amplifying their full potential as digital resources.

Through a hands-on demonstration of how to navigate this database, as well as its metadata structure, this presentation aims to solicit feedback from the audiences

from various backgrounds attending the digital humanities conference. This will enable our team to further enhance the usability of Tone Perfect so as to build an inclusive and accessible space of optimal learning.

References

Godfroid, A., Lin, C., and Ryu, C. (2017). Hearing and seeing tone through color: an efficacy study of web-based, multimodal Chinese tone perception training. *Language Learning*, 76: 819-857.

Grimes, Ryan (2016). With colors and tones, MSU researcher's game gives your brain the tools to learn Mandarin. *The Next Idea*. Michigan Radio. April 14, 2016. http://michiganradio.org/post/colors-and-tones-msu-researcher-s-game-gives-your-brain-tools-learn-mandarin (accessed April 27, 2018).

Ryu, C. and Michigan State University Libraries (2017). Tone Perfect: Multimodal Database for Mandarin Chinese. Michigan State University. East Lansing, Michigan https://tone.lib.msu.edu/ (accessed April 27, 2018).

Codicological Study of pre High Tang Documents from Dunhuang : An Approach using Scientific Analysis Data

Shouji Sakamoto
sakamoto@mac.com
Ryukoku University, Japan / Centre de Recherche sur la Conservation des Collections (CRCC), France

Léon-Bavi Vilmont
leon-bavi.vilmont@mnhn.fr
Centre de Recherche sur la Conservation des Collections (CRCC), France

Yasuhiko Watanabe
watanabe@rins.ryukoku.ac.jp
Ryukoku University, Japan / Centre de Recherche sur la Conservation des Collections (CRCC), France

Dunhuang documents consist of about 40 thousand documents from 5th to 11th century. The documents were discovered in Mogao Cave 17, in Dunhuang, China, by the Daoist monk Wang Yuanlu in 1900. At that time, many foreign explorers visited Central Asia and especially Dunhuang: the Hungarian-British archaeologist Aurel Stein in 1907, followed by the French Sinologist Paul Pelliot in 1908. Both brought back to Europe thousands of documents that they bought from the monk. Although scattered in many countries, the documents are available on the International Dunhuang Project website and the Gallica website of the French National Library; however, except the digital images and bibliographic data, there is no further information on the constituent materials (paper, ink, dyes etc.)

This priceless treasure represents an invaluable resource that led to the creation of a new research field called Dunhuang studies, in order to contribute to a better knowledge of the evolution of paper over 6 centuries. To date the manuscripts, in the 1990s, Prof. Akira Fujieda, a Japanese scholar adopted codicological analysis, focusing both on paper and morphology of the manuscripts, and on the shape of characters written (e.g. Clerical script (□書), Regular script (楷書), etc.). Only preeminent documents were deliberately taken into account by him and thus ignoring many manuscripts. As a result, 5 classes were determined according to the historical periods, that is Northern dynasty (386-581 CE), Sui dynasty (581-618), Early and High Tang dynasty (618-765), 765-786 and Tibetan Empire and Guiyi Circuit (786-1036) (Fujieda, 1999).

In addition to Prof. Fujieda's study, we investigate more details of paper using nondestructive scientific analysis on more than 400 Chinese Dunhuang manuscripts from the Pelliot collection and the Stein collection, and collected various data using a high-resolution digital microscope (Keyence VHX-1000) together with visual checks. Information contained in colophons (date, title of manuscripts) was also collected. A manuscript title is useful for categorization of the manuscripts. Analysis results show differences that can be criteria for differentiate paper. We developed the database (http://www.afc.ryukoku.ac.jp/pelliot/index.html), including scientific analysis data such that microscopic images from the Pelliot collection, as part of new digital archives for old documents.

As we obtained new data by scientific analysis and visual check, we can define new classes, A2, A3, B1 and C1, besides Fujieda's classes, A1, B2 and C2, as follows; A1: Fujieda's class from Northern Wei (北魏 (386-534)) and Western Wei (西魏 (535-556)). Paper is Ma-shi (麻紙) including hemp or ramie with 4~6 laid lines/cm, and with clerical like script of northern dynasty style. On the other hand, A2: paper from southern dynasty, Liang (梁 (502-557)) and Chen (陳 (557-589)), is high quality Ma-shi, and have finer laid lines, 8~9 laid lines/cm, paper width is 49~50 cm and well dyed, and is written sutra with clerical like script of southern dynasty style. Moreover, A3: new class paper from Northern Zhou (北周 (556-581)) is not Ma-shi but Cho-shi (楮紙) including mulberry paper (B. papyrifera, M. alba, etc.), and they have around 6 laid lines/cm, and with clerical like script. B1: new defined class paper from Sui (隋 (581-618)) is Cho-shi with 6~8 laid lines/cm and paper width is narrow, 41~43 cm. Few paper include rice starch. But B2: Fujieda's class from Sui. Paper is Cho-shi with about 6~7 laid lines/cm, paper width is wide, 50~53 cm, and well dyed, and is written sutra with clerical like or regular script of southern dynasty style. C1: paper in this new small class from early and high Tang (初

唐 (618-712), 盛唐 (712-765)) is similar to the ones in B1, that is Cho-shi with 6~8 laid lines/cm and paper width is 37~44 cm. Some paper include rice or millet starch. C2 is also Fujieda's class from early and high Tang. Paper in C2 is Cho-shi and high quality Ma-shi with fine laid line, about 8~10 laid lines/cm, paper width is wider than the ones in C1, 45~51 cm, and well dyed, and is written sutra with regular script.

As mentioned above, scientific analysis data is very useful for Dunhuang studies, for example, the data improved Fujieda's classification. We developed the database, Scientific Analysis of Pelliot Collection, digital archives, including such data

References

Fujieda, A. (1999). *Dunhuang Study and Related Topics*. *Brain Center*, pp.24-56. (in Japanese)

Connecting Gaming Communities and Corporations to their History: The Gen Con Program Database

Matt Shoemaker
mshoemaker@temple.edu
Temple University Libraries, United States of America

2017 saw the 50th anniversary of the Gen Con gaming convention, the oldest and largest continuously running gaming convention in the United States. Started in 1967 as a wargaming convention, Gen Con faced exponential growth following the 1974 creation of Dungeons & Dragons by one of its founders, Gary Gygax. Since then, Gen Con has seen a wealth of change. Evolving from a wargaming convention to a roleplaying game convention, growing to encompass video games and board games and finally reaching its current state of a gaming convention with close ties to popular culture. Aside from the content Gen Con has covered, the convention has also seen fluctuations in the populations that attend the event. All of these factors make Gen Con a prime target for scholarly study in areas of popular culture, games, gender in games studies, and the impact of Dungeons & Dragons. Scholars in media studies, history, material culture and gender studies, to name a few, would all be interested in data related to Gen Con.

Though Gen Con offers a wealth of possibility for scholarship, the information about the convention has largely remained inaccessible to scholars. As a corporate entity, Gen Con LLC, the company that currently runs Gen Con, keeps the majority of their records confidential. One resource that is publically available, however, are the programs from each year of the convention. The quality of the data within the programs varies from year to year, but they generally contain information pertaining to events that were run, who ran them, and descriptions of those events along with other information. Another barrier regarding these programs is that the vast majority of them exist only in physical form, with no digital counterparts. Many of these paper programs are also quite rare, particularly from the conventions that took place in the

1960s and 1970s. An additional resource that is dwindling is those who attended and organized the convention during its early years. Gen Con's most famous founder, Gary Gygax, passed away in 2008. Many of the others involved with the convention from its inception are approaching an advanced age and part of an insular group within gaming culture that few outside of it have approached. These barriers to access have, thus far, limited the scholarship that could be conducted on the Gen Con game convention.

With the above in mind and the 50th anniversary of the convention quickly approaching, we took the opportunity to undertake a project to make resources related to Gen Con more accessible to scholars. The primary work for the first phase of this project took place during 2016 and the first 3 quarters of 2017. We set out to first collect digital and physical copies of all 50 years of Gen Con which we were successful in doing. Second, we converted all event data from these programs into a database of more than 150,000 records which scholars and members of the gaming and Gen Con communities could access online via a Black Light discovery layer. Third, we conducted oral history interviews of several people involved in the history of Gen Con's past and present and transcribed them. Fourth, we conducted some preliminary research using textual analysis and data analysis methods to showcase some of the research that could be conducted using this data and other resources. Finally, we created an Omeka instance and Neatline timeline to both house these resources and make them available for others to use. All of this information can be found at http://best50yearsingaming.com/

We are continuing to conduct research with this dataset and are creating workshops that utilize the dataset in order to educate students in how to use large datasets. We also would like to increase awareness of this open dataset in order to connect more scholars to the resource so they can utilize it in their own research. This project has been able to connect the gaming community, the Gen Con community, and the Gen Con LLC community over a dataset they all have interest in, and we would like to see them connected with more scholars as well. The work we conducted for this project and knowledge of the availability of this dataset is something that attendees of DH2018 would be interested in, particularly those looking for a 20th and 21st century data set suitable for textual and other forms of data analysis, and we hope you will allow us to present it to them.

References

Best 50 Years in Gaming Project Website. http://best-50yearsingaming.com/

Resolving South Asian Orthographic Indeterminacy In Colonial-Era Archives

Amardeep Singh
amsp@lehigh.edu
Lehigh University, United States of America

One of the challenges of doing archival research with respect to colonial-era Indian print archives is orthography. A substantial number of Indian newspapers produced under have now been digitized, and are accessible through services such as Readex's "South Asian Newspapers" archive, the Digital Library of India, the Panjab Digital Library, and others.

Within the English-language archive, the searchability of these archives is limited, in large part due to idiosyncratic choices made by editors and authors in rendering words from South Asian languages in Roman script. Thus, the pioneering feminist doctor whose name is usually rendered as "Rukhmabai" by present-day scholars was quite often represented as "Rukmabai," "Rukmibai" and "Rukhmibai" in English-language newspapers from the British colonial era. The Roman rendering of Bengali-language names such as "Chatterjee" and "Tagore" also have similar indeterminacy (Chatterjee could be rendered in Indian print archives as "Chatterji," "Chaterjee," or "Chattopadhyay"; "Tagore" could be "Thakur").

The orthographic indeterminacies also proliferate beyond how authors' names are rendered; indeed, we see the issue occurring with reference especially to the representation of South Asian vowel forms ("i" vs "ee"; "u" vs. "oo"), aspirated consonants ("d vs" "dh"; "t" vs "th"; "b" vs. "bh"), and labials ("b" vs. "v"). Given that these archives tend to have simple search features that do not feature intelligent spelling correction, searching for topics of historical interest ("sati" or "satee" or "suttee"?) can lead to highly incomplete results.

Finally, orthographic indeterminacy can be an issue within and across South Asian languages themselves. "V" sounds in the Punjabi language, for instance, are frequently pronounced and spelled with "b" or "bh" in Hindi. The "ā" vowel sound common in many north Indian languages is rendered as "p" (that is to say, a soft "o" sound) in Bengali.

A possible solution to the South Asian orthographic indeterminacy problem might be found by appropriating tools developed by digital humanists in Early Modern studies. A team at Newcastle University, led by pioneering DH scholar Hugh Craig, has developed a tool called Corella, which is designed to help resolve orthographic indeterminacies in early modern English corpora (Craig 2010). Here, we propose to use a limited corpus from an existing archive of texts by British authors in India (the Kipling family) as well as a series of Indian authors (the afore-mentioned Rukhmabai as well as several others). We will aim to train Craig's Corella tool to work with Indian languages rather than with early modern orthography. This will allow us to address linguistic indeterminacies in the Roman rendering of Indian languages along the lines of those mentioned above. Can the searchability of these archives be improved via the use of such tools? What are the prospects of training tools such as Corella to work with larger corpora?

References

Hugh Craig, R. Whipp, "Old spellings, new methods: automated procedures for indeterminate linguistic data." *Literary and Linguistic Computing*, Volume 25, Issue 1, 1 April 2010, Pages 37–52

Brâncuși's Metadata: Turning a Graduate Humanities Course Curriculum Digital

Stephen Craig Sturgeon
stephen.sturgeon@bc.edu
Boston College, United States of America

This poster outlines the planning stages for introducing a substantial digital assignment to a paper-based graduate Humanities course and describes techniques for making metadata interesting to graduate students who have never had occasion to give much thought to it. It also details the experience of a librarian co-teaching a graduate seminar, and may provide a basis for reflection on where the particular types of bridges that get built in these activities lead: are they bridges that well-prepared students will take into a competitive job market? Bridges that subject librarians and faculty members will use to traverse a new collaborative environment? Bridges that students will send their scholarly ideas and projects across to a web-based public? Or bridges for university administrators to point to for the comparison of their respective bridges?

A Style Comparative Study of Japanese Pictorial Manuscripts by "Cut, Paste and Share" on IIIF Curation Viewer

Chikahiko Suzuki

ch_suzuki@nii.ac.jp
Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems, Japan

Akira Takagishi
taka@l.u-tokyo.ac.jp
University of Tokyo, Japan

Asanobu Kitamoto
kitamoto@nii.ac.jp
Center for Open Data in the Humanities, Joint Support-Center for Data Science Research, Research Organization of Information and Systems; National Institute of informatics, Japan

Introduction

Today, many institutions provide digital image data for their collections. Easy access to high-quality images not only improves efficiency in art history research but also changes how research is conducted. Our approach to a style-comparative study makes use of this trend with a web-based tool called the "IIIF Curation Viewer," built using IIIF (International Image Interoperability Framework), to change the input and output of research.

We studied pictorial manuscripts called "Emaki," "Eiribon," or "Nara Ehon" (illustrated scrolls and books with calligraphy) from the Edo period in Japan through the IIIF Curation Viewer, then discussed the efficiency and shareability of this approach.

Tools and materials

Composing lists of notable elements from target materials is a fundamental step in style comparison in art history research. The IIIF Curation viewer, developed by the Center for Open Data in the Humanities (CODH), is a useful tool for IIIF-compliant image resources. It has a function called "curation" that creates a list of interesting canvases with metadata. It reduces the effort of using cut and paste for the target material [Figure 1]. The result of cutting and pasting can easily be saved and shared in a JSON format.

Figure 1. Selecting element by mouse drag operation

The "selected thumbnails" function shows a list of 20 curated elements at a time. This function is useful for comparing small details [Figure 2].

Figure 2. Example of the "Selected thumbnail" mode and list of facial expressions

Analysis with the IIIF Curation Viewer

We picked up all facial expressions from four Eiribon and compared lists of the facial expressions using the IIIF Curation Viewer. Comparison suggests that pictures in each Eiribon were painted by different painters, but the same calligrapher wrote the texts. It also suggests that these Eiribon were created by a workgroup of artists.

We further analyzed using the IIIF Curation Viewer by comparing the above-mentioned curation with other Eiribon created by anonymous painters and calligraphers. We found that two anonymous works have the same drawing style as pictures in Asakura's Eiribon [Figure 3].

Figure 3. Comparing facial expressions in Eiribon: Asakura's text (above) and an anonymous work (below)

We found that our approach was useful for both the style comparing process and the sharing process. It is helpful to share pictures as evidence that supports the conclusion of a paper, but many journals did not allow it because of space limitations. Sharing the curation and citing it from the paper can solve this problem. For example, the evidence used in this paper is accessible at the CODH web site, as shown in the reference, so that other researchers can easily verify the results. Curated data can be increasingly promoted, due to its shareablity and reusability, by publishing them in repositories with persistent identifiers.

Conclusion

The IIIF Curation Viewer is important not only for making the entrance process easy through its cut and paste function, but also for making the output process useful through its sharing function. Both insertion and output is useful to art history research, in particular, in Eiribon research. There are many remaining unexamined Eiribon ; each Eiribon has many facial expressions. An easy cut, paste and share tool has been long awaited, and we hope it will enable the creation of a comprehensive facial expression database of Eiribon and Emaki.

We focused on Japanese art in this paper, but we can use this tool for any artwork as long as the images are served in IIIF. For example, we picked up facial expressions from portraits in the Yale Center for British Art and grouped them by century. The increased reusability of research extends possibilities for art historians terms of education and machine learning. Curated data can be reused as training data for machine learning.

Two issues remain for futures study. First, we need to increase IIIF-compliant image services. Especially in Japan, few institutions provide digital images in IIIF. Second, we need an ecosystem for sharing the results of curation, such as correcting metadata, identifier, and a repository for sharing and editing.

References

Center for Open Data in the Humanities (CODH). (2017). IIIF Curation Viewer, http://codh.rois.ac.jp/software/iiif-curation-viewer/ (accessed 20 April 2018a).

Center for Open Data in the Humanities (CODH). (2017). IIIF Global Curation : Facial expression data: British Portraits, http://codh.rois.ac.jp/curation/exhibition/2/index.html.en (accessed 20 April 2018b).

Center for Open Data in the Humanities (CODH). (2017). "Curation" used in this paper (accessed 20 April 2018c). *Daikoku-mai* (Original version provided by National Institute of Japanese Literature, DOI: 10.20730/200006198) http://codh.rois.ac.jp/software/iiif-curation-viewer/demo/?curation=http://codh.rois.ac.jp/curation/exhibition/3/json/daikoku.json. *Rashomon* (Original version provided by National Institute of Japanese Literature, DOI: 10.20730/200003096) http://codh.rois.ac.jp/software/iiif-curation-viewer/demo/?curation=http://codh.rois.ac.jp/curation/exhibition/3/json/rashou.json. *Tomonaga1/2* (Original version provided by Digital Collection of Keio University Library, ID: 132X@56@2@1) http://codh.rois.ac.jp/software/iiif-curation-viewer/demo/?curation=http://codh.rois.ac.jp/curation/exhibition/3/json/tomo.json. *Tomonaga2/2* (Original version provided by Digital Collection of Keio University Library, ID: 132X@56@2@2) http://codh.rois.ac.jp/software/iiif-curation-viewer/demo/?curation=http://codh.rois.ac.jp/curation/exhibition/3/json/tomo2.json. *Story of Kumano-Gongen* (Original version provided by Digital Collection of Keio University Library, ID: 11X@31@1) http://codh.rois.ac.jp/software/iiif-curation-viewer/demo/?curation=http://codh.rois.ac.jp/curation/exhibition/3/json/kumano.json

Complex Networks of Desire: Fireweed, Fuse, Border/Lines

Felicity Tayler
felicity.tayler@utoronto.ca
University of Toronto, Canada

Tomasz Neugebauer
tomasz.neugebauer@concordia.ca
Concordia University Library

We present ongoing research using data visualization and complex network analysis to historicize the production of three periodicals: *Fireweed, Fuse,* and *Border/Lines.* Computational methods allow for the visualization of metadata describing these magazine issues as a complex network – but what do these visualizations reveal about real social relations involved in the production and circulation of these magazines?

Fireweed, Fuse, and *Border/Lines* emerged between 1976 and 1986 in Toronto, Canada, from a hotbed of lesbian and gay liberation, feminist and cultural race politics, thereby circulating in relation to transnational social, political and cultural movements (Butling and Rudy, Gonosko and Marcellus, Monk, Robertson). Whereas digital art historical scholarship often applies computational methods to the analysis of visual images (Zorich, Manovich), this paper instead applies complex network analysis to bibliographic metadata describing artist-led magazine publishing. We propose that there is a correlation between the magazine as a site of imagined community (as a discursive site where artistic scenes and poetic community are formed) (Allen, 12-17); and the complex networks visualized from metadata describing production teams and content of each printed issue (Knight, Long, Lincoln, Liu).

At this time, we have completed the data gathering stage. Prior to our initiative, *Fireweed* and *Fuse* were not digitized, nor were they comprehensively indexed on digital platforms. A complete data set was created using human cataloguers and a pre-existing metadata schema developed for the e-artexte open repository of publications on contemporary art. *Border/Lines* was previously digitized, and housed in an open journal repository. However, this online collection is not complete, further, it was not possible to extract the metadata from this platform in a consistent format. Contributor names and roles were indexed for each magazine issue (editor, author, translator, etc.). Many of the contributor names and roles

already exist within the e-artexte authority files, and standard indexing protocols were expanded to include roles that are not usually recorded in the metadata (members of editorial committees, designers, typesetters, etc.).

Once indexed in e-artexte, the data became publicly accessible and exportable into various formats, including EPrints XML. A conversion to Graph GML files used Apache Pig Latin scripts (Neugebauer). The resulting Graph GML data was imported into Gephi.

To borrow an expression from Hoyt Long's mapping of literary community, resulting graphs encourage a "sliding back and forth" between the macroscale of the generated graphs and the microscale of the discourse of the artistic and poetic communities represented (316).

A Multi Modal graph (Figure 1) maps relationships between individual magazine issues, contributors (writers, editors, and designers, etc.), artists as subjects of articles, and publishers. Edges were assigned a colour according to magazine title. Node size has been mapped to betweenness centrality, with a filter applied to a range higher than .01.

Lisa Steele and Clive Robertson feature prominently as contributors to *Fuse* magazine, with a high degree of betweenness centrality. This is not a surprise, as both authored multiple articles in the magazine, are founding editors and key figures in the Toronto artistic and activist scenes bridged through the magazine's content (Robertson, Monk). More remarkable is the prominence of Lynne Fernie in the network, best known for later success as the director of documentary films addressing LGBTQ histories. Fernie's high degree of betweenness centrality and position as a connector between the cluster of nodes surrounding *Fuse* magazine and *Fireweed,* provides a bridge between these two magazines as spaces that shared an impulse towards lesbian and feminist liberation. Poet and activist Dionne Brand, who works at the intersection of race and gender, also bridges *Fuse* and *Fireweed*. Cultural policy analyst Jody Berland, and gay activist and environmentalist Alexander Wilson bridge *Fuse* and *Border/Lines*. Feminist cultural historian Rosemary Donegan bridges all three discursive spaces.

A second graph, a Single Mode Contributor Projection will map relationships between individual contributors through their frequency of co-occurence in magazine issues. The graph will be filtered through edge weight, which represents co-occurrence in a minimal number of journal issues. We will colour the graph through community detection on this network of contributor relations using the modularity functionality in Gephi (Blondel et al.).

We anticipate that contributors with a high betweenness centrality will emerge as catalysts for artistic community as it is represented by the discursive spaces of these magazines. Although some of these names may be iconic, "famous" artists and writers, other careers may not have had the same trajectory of visibility. Additional graphs will be generated by publication year to illustrate how the network structure and centrality measures changed over time.

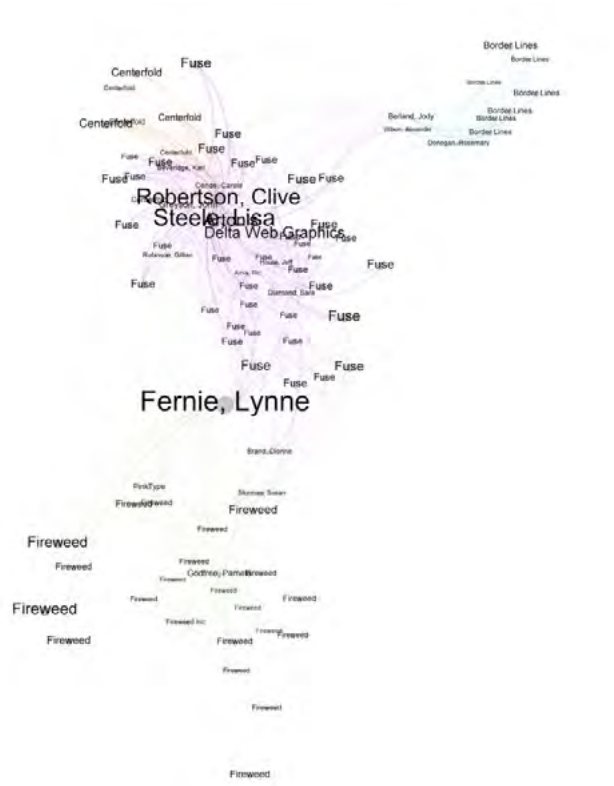


Figure 1. A Multi Modal graph (Figure 1) maps relationships between contributors and issues of magazines with a high degree of betweenness centrality: *Fuse* (purple) previously titled, Centerfold (orange); Fireweed (green), and *Border/Lines* (blue).

References

Allen, G. (2016). *The Magazine.* Cambridge, Mass.: MIT Press.

Butling, P. and Rudy, S. (2005). *Writing in Our Time: Canada's Radical Poetries in English (1957-2003).* Waterloo, Ont.: Wilfred Laurier University Press.

Gonosko, G. and Marcellus, K. (2005). Dead Downtown: Writing the Cultural Obituary of the Alternative Press. *Topia,* 14: 23-35.

Knight, A. R. (2017). Putting them on the map: Mapping the Agents of the Colored Co-operative Publishing Company. https://www.arcgis.com/apps/MapSeries/index.html?appid=665eb933117f4ed-68f0535b4560b5744

Lincoln, M. (2016). "Social Network Centralization Dynamics in Print Production in the Low Countries, 1550-1750" *International Journal of Digital Art History* 2: 134-157.

Long, H. (2015). "Fog and Steel: Mapping Communities of Literary Translation in an Information Age" *The*

Journal of Japanese Studies, 41(2): 281-316. DOI 10.1353/jjs.2015.0062
Liu, A. (2012). "Friending the Humanities Knowledge Base: Exploring Bibliography as Social Network in RoSE." NEH Office of Digital Humanities White Paper. https://rosedocumentation.files.wordpress.com/2012/07/rose-white-paper-as-submitted-to-neh.pdf
Manovich, L. (2015). "Data Science and Digital Art History" *International Journal of Digital Art History* 1:12-34. DOI: 10.11588/dah.2015.1.21631
Neugebauer, T. (2017). "EPrintsData2GML" Eprints Interest Group, 2017 International Conference on Open Repositories. https://github.com/photomedia/EPrintsData2GML
Monk, P. (2016). *Is Toronto Burning*? *Three Years in the Making (and Unmaking) of the Toronto Art Scene.* Toronto: AGYU.
Robertson, C. (2006). *Policy Matters: Administrations of Art and Culture.* Toronto: YYZ Books.
Blondel, V.D. et al. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* 10: 10008-10020. DOI 10.1088/1742-5468/2008/10/P10008
Zorich, D. (2012). "Transitioning to a Digital World: Art History, Its Research Centers, and Digital Scholarship," *Kress Foundation.* http://www.kressfoundation.org/news/Article.aspx?id=35338

Locating Place Names at Scale: Using Natural Language Processing to Identify Geographical Information in Text

Lauren Tilton
ltilton@richmond.edu
University of Richmond, United States of America

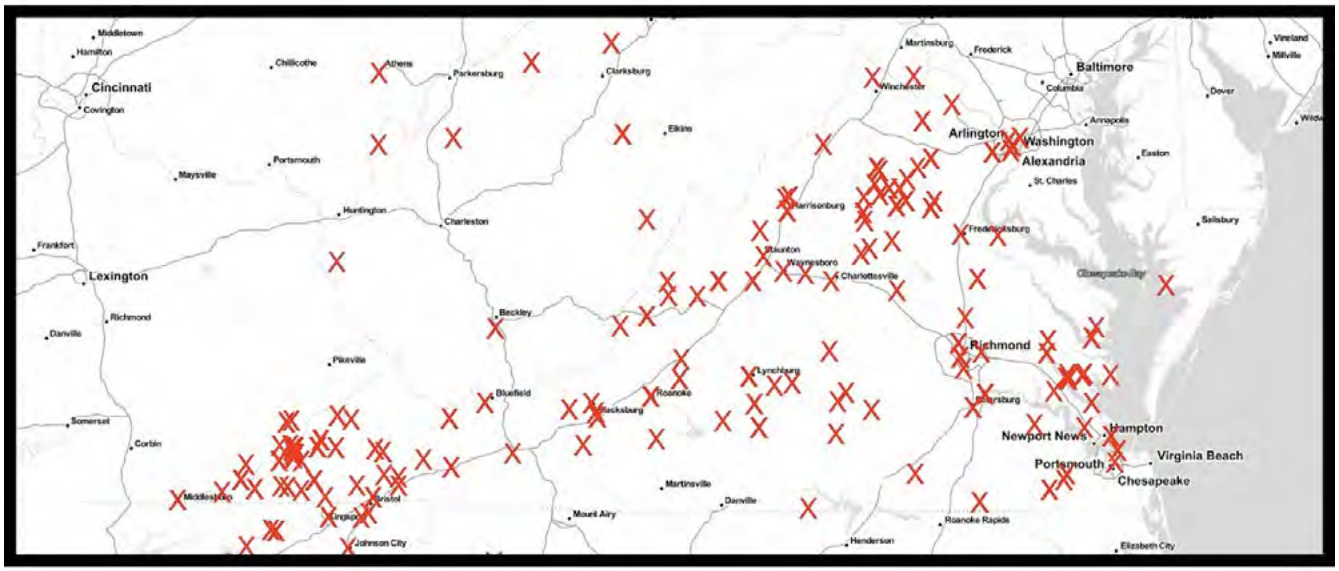
Taylor Arnold
tarnold2@richmond.edu
University of Richmond, United States of America

Courtney Rivard
crivard@email.unc.edu
University of North Carolina - Chapel Hill, University States of America

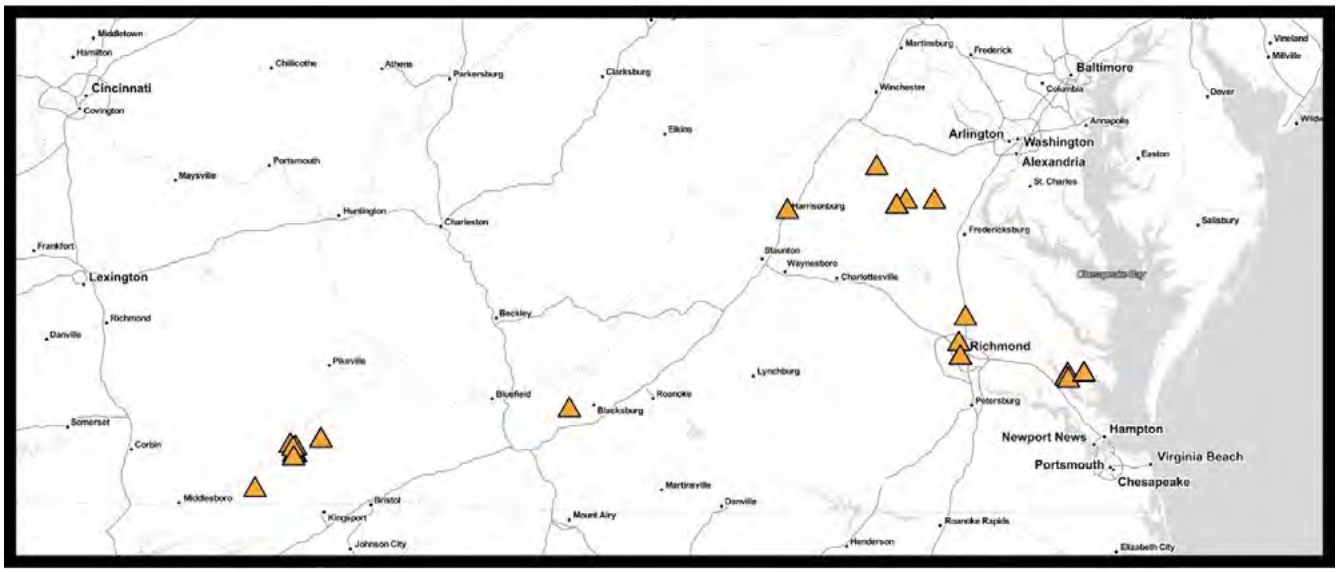
Historical sources are often tagged with metadata about place such as where the object was created, acquired, or stored. Rich latent geographical information is often also mentioned throughout textual documents. A challenge though is how to extract this spatial information at scale. For example, when a text mentions Paris, does the writer mean Paris, Texas, USA or Paris, France? Out of context, most would assume the reference is to more populous capital of France, but it could also be the city in Texas. While close reading would provide an answer, this becomes a challenge when working with hundreds and thousands of documents. How might we be able to more accurately predict the exact location using the broader context?

Our poster „Locating Place Names at Scale: Using Natural Language Processing to Identify Geographical Information in Text" addresses how computational methods can be used to identify and geolocate place-based data. We show how Named Entity Recognition (NER), a natural language processing (NLP) technique, can locate place names using the document's context. We then discuss how to geolocate those places names using a series of computational techniques. Specifically, we start by finding references to specific political divisions (countries, states, and cities), georeferencing them through the Google API. Any political divisions that are uniquely determined become reference points. The reference points are then used to disambiguate terms with multiple results, such as Paris, France and Paris, Texas. Disambiguation is done by appending the political division to the name of the place in order of specificity. If this fails to uniquely determine locations, distances to the closest reference points in the text are used to break ties. This strategy increases proper place name identification and can be applied automatically over a large corpus of digitized texts.

Finally, we turn to an example from our collaborative project on the United States Federal Writers' Project (FWP) entitled Voice of a Nation: Mapping Documentary Expression in New Deal America. During the New Deal, thousands of life histories were written to capture the American experience. While the location of the interviews provides insight into the geographic expanse of the collection (Figure 1), the interviewees consistently spoke about places beyond the location of the physical interview. We apply NER and NLP to identify the place names in the interviews. We are then able to identify and map the many different locations that interviewees mentioned (Figure 2). Across over a thousand interview, what we see is that many of those interviewed spoke of migration - whether their own or their kin - generating a more complex understanding of movement and place during the early 20th century in the United States.



Triangles represent where the metadata identified the interview location in Virginia.



Red "X"s represent locations identified by the use of our algorithm, based on named entity recognition, to the text of the interview referenced in Figure 1

4 Ríos: una construcción transmedia de memoria histórica sobre el conflicto armado en Colombia

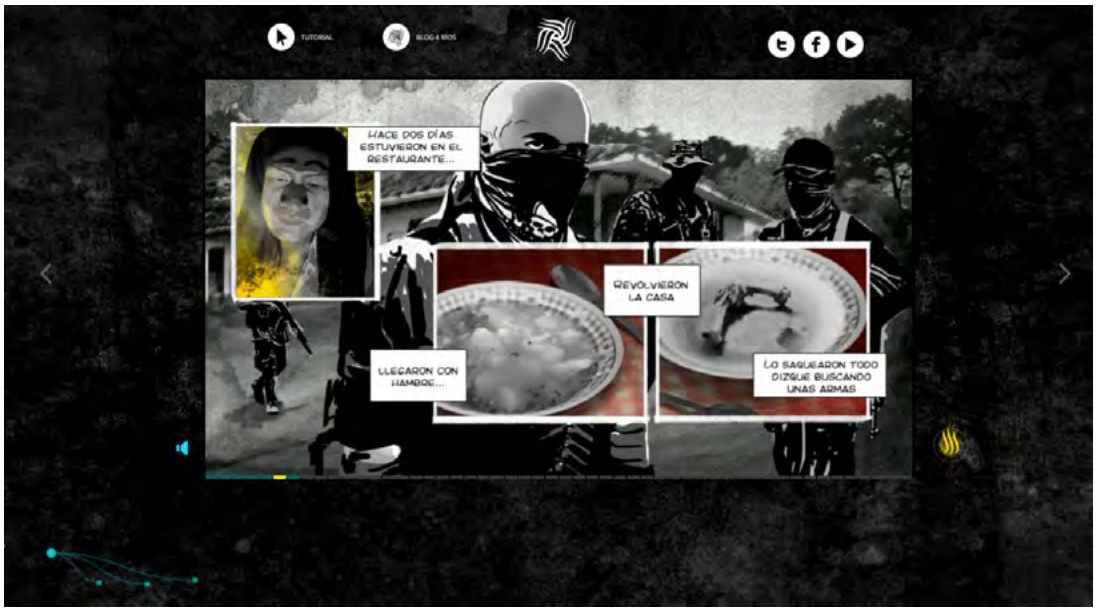
Elder Manuel Tobar Panchoaga
eldertobar@gmail.com
Orgánica Digital; Universidad de los Andes, Colombia

Colombia ha vivido las secuelas del conflicto armado durante más de cinco décadas; el Centro Nacional de Memoria Histórica (CNMH) calcula que existen más de seis millones de desplazados por la violencia, principalmente población civil, campesinos e indígenas.

Desde el inicio del conflicto las víctimas han manifestado de múltiples formas sus sentires, vivencias, recuerdo, testimonios acerca del conflicto y sus consecuencias. Prueba de ello son las más de 177 iniciativas no gubernamentales llevadas a cabo por población víctima de la violencia armada, las cuales giran alrededor de la memoria histórica del conflicto y que fueron registradas en una investigación a fondo realizada por el CNMH (Centro de Memoria Histórica, 2013)


abuelo...
tenemos que irnos
Armas...
...creen que somos Guerrilleros


A
ABUELO...
B1
TENEMOS QUE IRNOS
B2
B3
ARMAS...
(REEN QUE SOMOS GUERRILLEROS


TUTORIAL
BLOG 4 MOS
HACE DOS DÍAS ESTUVIERON EN EL RESTAURANTE...
REVOLVIERON LA CASA
LLEGARON CON HAMBRE...
LO SAQUEARON TODO DIZQUE BUSCANDO UNAS ARMAS

FOGRAFÍA 01

DATOS OFICIALES DE LA MASACRE

Asesinatos y cifras desde la perspectiva de sus protagonistas.

"En los mismos expedientes judiciales se habla de que no hay certeza del número, debido a que varios de los cadáveres fueron arrojados a abismos."[1]

SEGÚN HH[2]

24 o 25
personas asesinadas entre ellas 2 menores de edad y 3 mujeres.

SEGÚN "AUTORIDADES JUDICIALES"[3]

220 Paramilitares

32 Personas asesinadas

10 Personas desaparecidas

SEGÚN JUSTICIA Y PAZ[4]

24 Asesinatos

10 Desapariciones forzadas

67 paramilitares del Bloque Calima fueron condenados a 40 años de cárcel en febrero de 2005. En octubre de 2008, el Consejo de Estado condenó a la Nación por omisión y falla en el servicio, y ordenó indemnizar con seis mil millones de pesos a varias víctimas por el delito de desplazamiento.

500 Hombres del Bloque Calima (Aproximadamente)[5]

PARTE DE VICTORIA DE CARLOS CASTAÑO[6]

"después de combatir durante 72 horas lograron incursionar en el Alto Naya y dar de baja 42 narcoterroristas del ELN y las FARC"

DATOS DE VÍCTIMAS (KITEK KIWE, ONIC, ACNUR, ETC)[7]

400 Paramilitares (Aproximadamente)

más de 40 indígenas, afrodescendientes y campesinos fueron asesinados

60 Personas siguen desaparecidas

más de 1800 Desplazados

COMANDANTES DE LAS AUC

Que participaron en el operativo o la masacre

Elkin Casarrubia, **alias "El Cura"**
Jair Alexander Muñoz Borja, **alias 'Sisas'**
Armando Lugo, **alias 'El Cabezón'**
Luis Fernando Arce Martínez, **alias "Chilapo"**

KITEK KIWE[8]

más de 500 Hombres de las AUC.

3.500 Desplazados o más

Operativo inicia el 6 de abril, el 8 de abril se arman 2 retenes e inicia la masacre con el asesinato de una persona; los asesinatos continúan hasta el 17 de abril.

"... la Fiscalía General de la Nación registró en Abril de 2001 el levantamiento de veintisiete cuerpos en el Alto Naya, y reconoció la existencia de catorce cuerpos más que yacen en fosas comunes en San Antonio, bajo Naya.

...El Cabildo Kitek Kiwe denuncia que en el contexto de la masacre del Naya se han presentado más de cien muertes ocasionadas por los grupos armados ilegales."

Invitación a compartir el contenido para así vivir la experiencia 4 ríos y expanderla a varios públicos. Está en nosotros no olvidar la historia.

COMPARTE ESTE CONTENIDO

1. http://terranova.uniandes.edu.co/noticias/octubre/justicieverdadyreparacionoctubre.pdf
2. http://www.youtube.com/watch?v=qbDy-0BkHDc&list=PL0CB88CC756A4D40A
3. http://www.verdadabierta.com/masacres-seccion/3157-la-fuerza-publica-y-la-masacre-del-naya
4. http://www.verdadabierta.com/masacres-seccion/3187-las-deudas-con-la-comunidad-de-el-naya
5. http://www.wdh.ed.cr/BibliotecaWeb/Varios/Documentos/BD_438003671/Resoluciones-Colombia/Resolucion%20009.htm
6. http://ibs.org.co:81/biblioteca/dwnlds/taq/taq11/taq11-01.pdf (p 33)
7. Rios Aparicio, L., Floresmiro (2001). Caracterización del desplazamiento indígena en el departamento del Cauca. Popayán CNUR, ONIC, RSS. http://servindi.org/pdf/ElCasodelNaya.pdf
8. http://www.humanas.unal.edu.co/colantropos/documentos/Cartilla%20Kitek%20Kiwe%20FINAL%20version%20digital.pdf

El CNMH ha identificado por lo menos tres usos de la memoria dentro de estas acciones comunitarias y sociales, en la primera la memoria es expuesta en búsqueda del esclarecimiento de los hechos sucedidos para exigir justicia por parte del Estado y las instituciones encargadas. En la segunda, la memoria sirve como elemento pedagógico de lo acontecido en búsqueda de la no repetición de estos hechos; mientras que en su tercer uso, la memoria apunta al duelo, a la dimensión reparadora, a proponer 'una oportunidad para restablecer los vínculos sociales y un horizonte para la reconstrucción de lo que se perdió' (Centro de Memoria Histórica, 2013).

El cambio de percepción acerca del conflicto armado a partir de sucesos como la firma del acuerdo de Paz con

la guerrilla de las Farc, la creación del Centro de Memoria Histórica Nacional o la promulgación de la Ley de Víctimas y Restitución de Tierras o Ley 1448 de 2011, ha revitalizado el interés en la construcción y recuperación de las memorias de la violencia, lo que ha repercutido en la producción de múltiples productos artísticos, periodísticos y comunicativos relacionados con este tema.

En ese contexto surge '4 Ríos', un proyecto transmedia que narra historias del conflicto armado en Colombia a través de distintos medios, entre ellos un cómic interactivo, un aplicativo web de memoria además de una exposición interactiva compuesta de maquetas con realidad aumentada.

La primera historia producida está basada en la masacre del Naya, perpetrada durante la Semana Santa del año 2001 en la región del Naya donde alrededor de 300 paramilitares asesinaron a más de 30 personas, lo que provocó el desplazamiento de miles de habitantes de la región.

En el inicio del proceso investigativo, el proyecto se puso en contacto con la población desplazada de la masacre, sin embargo, luego de meses de charlas y reuniones telefónicas, la comunidad manifestó que no estaba interesada en trabajar en nuevos procesos alrededor del tema lo que impidió realizar un trabajo de campo con las víctimas, en cambio las autoridades del cabildo autorizaron el uso de fuentes de archivo donde habían contribuido de forma activa. De esta forma el proceso investigativo se enfocó en la búsqueda, clasificación y curaduría de diversos archivos, investigaciones, tesis, fotografías y noticias de medios públicos.

Una vez establecida la orientación curatorial de la información, se consolidó un equipo de trabajo multidisciplinario integrado por diseñadores gráficos, artistas plásticos, dibujantes de cómic, desarrolladores de software, programadores y animadores que trabajaron en la producción total de todas las plataformas: un cómic interactivo que mezcla una narración ficcional basada en la masacre (disponible en www.4rios.co), acompañado de un aplicativo web que permite a los usuarios dejar mensajes en forma de texto, gráfico o un archivo de audio, además de un corto animado que se complementa con una exposición interactiva compuesta de 3 maquetas que fusionan elementos materiales con animaciones en Realidad Aumentada.

Posterior a su lanzamiento en el año 2016, el proyecto ha interactuado con más de 5.000 usuarios a través de sus distintas plataformas, explorando temas como la narración del conflicto armado a través de Realidad Aumentada. En Internet, el cómic y el Flujo de Memoria han permitido la visualización de historias y documentación además de recibir mensajes, dibujos y audios que reflexionan sobre temas relacionados a las consecuencias del conflicto armado en el territorio nacional.

Así, 4 Ríos busca aportar a la construcción de memoria histórica alrededor del conflicto armado en Colombia, citando a Paloma Aguilar, a través de "una 'memoria prestada' que el sujeto no ha experimentado personalmente, y a la que llega por medio de documentos de diverso tipo" (Aguilar,1996) en donde el trabajo interdisciplinario busca proponer nuevas experiencias que reúnan otras formas de narrar, investigar, crear y construir a través de arte y tecnología.

References

Aguilar Fernández, P. (2008). *Políticas de La Memoria y Memorias de La Política: El Caso Español En Perspectiva Comparada*. Madrid: Alianza Editorial.

Comisión Nacional de Reparación y Reconciliación (Colombia) (ed). (2013). ¡*Basta Ya! Colombia, Memorias de Guerra y Dignidad: Informe General*. Segunda edición corregida. Bogotá: Centro Nacional de Memoria Histórica.

Building a Bridge to Next Generation DH Services in Libraries with a Campus Needs Assessment

Harriett Green
green19@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Eleanor Dickson
dicksone@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Daniel G. Tracy
dtracy@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Sarah Christensen
schrstn@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

Melanie Emerson
memerson@illinois.edu
University of Illinois at Urbana-Champaign, United States of America

JoAnn Jacoby
jjacoby@ColoradoCollege.edu
Colorado College, United States of America

This poster reports on a needs assessment for digital humanities library services undertaken at large research university in order to provide a basis for transition to a

next phase of Digital Humanities (DH) support at a library supporting a growing amount of DH work on campus. It reports key findings and how the library services will evolve to meet needs identified on campus.

A recent survey tallied over ninety research centers and initiatives around the world that support DH research, and the majority are associated with university campuses. The recent ARL report *Supporting Digital Scholarship* (Mulligan 2016) observed the trend for digital scholarship support to be centered in a single department, sometimes in a dedicated digital scholarship center, but with support for digital scholarship extending throughout the library. Despite the growing number of DH initiatives and support models for digital scholarship at institutions of higher education around the U.S. and world, few have conducted formal needs assessments on their campuses to ascertain the needs of researchers and other stakeholders. The professional literature that provides a strong guiding framework for this study includes the report on the University of Colorado's recent digital humanities needs assessment (Lindquist et al., 2013) and the Ithaka S+R Sustaining Digital Humanities study and Implementation Toolkit (Maron and Pickle, 2014).

The members of the working group conducting this needs assessment sought to use the study to provide a bridge to the next generation of DH services in the library. The timing was opportune, as there were several features of the library and campus environment at that moment made this a good time to asses DH needs. First, the library's digital scholarship public service space had entered its fifth year and had begun planning to move to a new, larger, and more visible space. Researchers' and instructors' interest in DH collaborations with the library had steadily grown since the foundation of DH services in 2010. The library had grown support for digital scholarship and communication in recent years and, like many peer institutions, sought to increase capacity by involving more librarians in DH services. All of this planning required updated knowledge of campus DH activity in order to evolve services appropriately.

For the first phase of the study, two members of the team conducted a total of 15 interviews with faculty, administrators, academic professionals, and graduate students from multiple colleges and campus units with interest or active involvement in digital humanities research and teaching. The group also reviewed recent dissertations across a range of arts, humanities, and humanistic social science fields to identify recent DH related work and the advisors for those projects.

From the interview responses, the working group developed a survey protocol for the second stage of the study. The group administered a survey that was sent to a random sample of 5% of faculty and graduate students from the colleges and units of Liberal Arts and Sciences, Fine and Applied Arts, College of Media, and School of Information Sciences; as well as targeted sampling of known practitioners of digital scholarship on campus. The survey was open for two months from November 2016 through early January 2017, and gathered 55 responses.

The group identified several areas of need expressed by researchers. These included access to collections and data, funding, networks of research and community, education, and infrastructure and research support. The study showed some differences between needs of graduate students and researchers. For example, graduate students saw a greater urgency around library support for tools and software. Faculty and staff saw greater urgency across all other areas including access to library expertise, assistance with access to digital content, and data storage. Access to digital collections as data appeared as a key barrier to researchers pursuing projects.

Based on these needs, the group developed six broad recommendations for library services: (1) provide opportunities for in-depth training; (2) connect the library's role in research data curation to digital scholarship creation; (3) expand the library's strengths in discovery and access to digital collections; (4) build space and opportunities for people to form communities of practice, (5) act as a key node in the network of digital scholarship research initiatives, and (6) build library personnel capacity for digital scholarship services. Each of these recommendations had specific associated action items.

This poster will provide an opportunity to discuss these findings, the steps being taken by the library to accomplish the goals identified, and the general landscape of next generation DH services in libraries.

References

Lindquist, T., et al. (2013). *dh+CU: Future Directions for Digital Humanities at CU Boulder*. Boulder, CO: University of Colorado, University Libraries Digital Humanities Task Force. http://scholar.colorado.edu/libr_facpapers/32/

Maron, N., and Pickle, S. (2014). *Sustaining the Digital Humanities: Host Institution Support beyond the Start-Up Phase*. Ithaka S+R. https://doi.org/10.18665/sr.22548

Mulligan, R. (2016). *Supporting Digital Scholarship*. SPEC Kit 350. Washington, DC: Association of Research Libraries, May 2016. https://doi.org/10.29242/spec.350

Chromatic Structure and Family Resemblance in Large Art Collections — Exemplary Quantification and Visualizations

Loan T Tran
lxt110930@utdallas.edu
The University of Texas at Dallas, Richardson, TX, United States of America

Kelly Park
kelly.park@utdallas.edu
The University of Texas at Dallas, Richardson, TX, United States of America

Poshen Lee
sephonlee@gmail.com
The University of Washington, Seattle, WA, United States of America

Jevin West
jevinw@uw.edu
The University of Washington, Seattle, WA, United States of America

Maximilian Schich
maximilian.schich@utdallas.edu
The University of Texas at Dallas, Richardson, TX, United States of America

Computational pattern recognition has made ground-breaking progress in recent years by combining advanced methods of machine learning with ever increasing amounts of visual data. Algorithms that learn to learn, combined with massive parallel computation in so-called GPU clusters, and billions of images a day acquired via sensors, or uploaded by Web users, have led to a situation where computers are able to recognize faces, spot cats in any body-configuration, and even drive cars without human interaction. In Art History such advanced methods of pattern recognition increasingly aim to compete with human connoisseurship. Relevant studies, for example, successfully identify duplicate photos in image archives (Resig, 2013), quickly find artworks given a certain object (Crowley and Zisserman, 2014), quantify the innovativeness of paintings (Elgammal and Saleh, 2015), convincingly discern and date architectural styles at a mega-city scale (Lee et al., 2015), and track the evolution of color contrast in Western Art from *chiaroscuro* to landscape painting (Kim et al., 2014 and Lee et al., 2017). What is missing is a rigorous reconciliation between state-of-the-art computer science techniques and established art historical standards based on trained observation and hermeneutic interpretation. Such a reconciliation is hard due to both the so-called "curse of dimensionality" in machine learning, and the cognitive limit of individual researchers confronted with potentially millions of images.

Our project aims to work towards a reconciliation of the computational and hermeneutic perspectives via two pathways. First, through visualizing the chromatic structure of paintings up to entire collections by consistently sorting color pixels, we uncover hidden color patterns of individual paintings, artist oeuvres, periods, and museum collections. Here, we also deal with a well-known multidimensional phenomenon, i.e. color, which could be a starting point to deal with hidden dimensions in machine learning using a traditional hermeneutic approach. Second, using cutting-edge deep learning algorithms and dimension reduction techniques that reduce the high dimensions of the machine learning results to a human-digestible level, we calculate visual family resemblance, generate a variety of clustering possibilities, and produce different visualizations. Combining both pathways, while performing these analyses on three different art collections, we will be able to evaluate the machine learning results, from both an art historian's and a computer scientist's perspective, in a manner that is understandable by a broad audience.

We work with three datasets: the Dallas Museum of Art, a "universal" art collection, circa 18,000 artworks; the Barrett collection, a comprehensive private collection of Swiss art, circa 400 paintings; and WikiArt, an encyclopedic online collection of circa 75,000 paintings. The DMA data is particularly strong in its six-thousand-year coverage, well in line with the exponential growth of world population and cultural production. The Swiss art collection, including high resolution images taken under controlled lighting conditions, is strong in its topical coherence. The WikiArt dataset, though subjects to shortcomings in lighting conditions and temporal coverage, is widely used as a de facto benchmark among machine learning community, and is therefore used for comparative analysis with the other collections.

(**A**) Color palette of Claude Monet at DMA

(**B**) Color palette of Piet Mondrian at DMA

(**C**) Color palette of Georgia O'Keefe at DMA

(**D**) Color palette of Mark Rothko at DMA

(**E**) Color palette of Gerhard Richter at DMA

Fig. 1. Colors in the oeuvre of individual artists. Colors are consistently sorted by luminosity, indicating color frequency (number of pixels), equivalent to area coverage. The strips for (**A**) Monet, (**B**) Mondrian, (**C**) O'Keefe, (**D**) Rothko, and (**E**) Richter reveal striking individual differences between artists. Similar sense-making comparison can be used to differentiate collection coverage as well as canonicity of artists, departments, and other sub-selections of works across museums.

References

Crowley, Elliot J., and Andrew Zisserman. (2014). In Search of Art. *Workshop at the European Conference on Computer Vision*. Springer International Publishing, pp. 54-70. https://www.robots.ox.ac.uk/~vgg/publications/2014/Crowley14a/crowley14a.pdf

Elgammal, Ahmed, and Babak Saleh. (2015). Quantifying Creativity in Art Networks. *arXiv preprint arXiv:1506.00711*. http://arxiv.org/abs/1506.00711

Kim, Daniel, Seung-Woo Son, and Hawoong Jeong. (2014) Large Scale Quantitative Analysis of Painting Arts. *Scientific Reports* 4: 7370. https://www.nature.com/articles/srep07370

Lee, Byunghwee, Daniel Kim, Hawoong Jeong, Seunghye Sun, and Juyong Park. (2017). Understanding the Historic Emergence of Diversity in Painting via Color Contrast. *arXiv preprint arXiv:1701.07164*. https://arxiv.org/pdf/1701.07164.pdf

Lee, Stefan, Nicolas Maisonneuve, David Crandall, Alexei A. Efros, and Josef Sivic. (2015). Linking Past to Present: Discovering Style in Two Centuries of Architecture. *IEEE International Conference on Computational Photography*. http://dx.doi.org/10.1109/ICCPHOT.2015.7168368

Resig, John. (2013). Using Computer Vision to Increase the Research Potential of Photo Archives. *Journal of Digital Humanities* 3: 3-2. http://journalofdigitalhumanities.org/wp-content/uploads/2014/07/Using-Computer-Vision-to-Increase-the-Rese-John-Resig.pdf

Ethical Constraints in Digital Humanities and Computational Social Science

Anagha Uppal
auppal@vols.utk.edu
University of Tennessee, United States of America

As it developed, the field of Digital Humanities has had a particular set of advantages in making advancements and gaining approval among the scientific community, allowing it to serve as a "means to revitalize the humanities" in the face of decreased funding and appreciation for its contributions (Reid 2011, pp. 352-353). Both for Digital Humanities and Computational Social Science, principal among these advantages are:

- Easy and fast access, via the Internet, to data resources and databases.
- Inexpensive computational power, including large amounts of inexpensive memory and physical storage.
- New forms of data (especially text) that can be easily obtained from many sources, particularly social media and blogs.

- Open-source software and a culture of code-sharing
- Modern advocacy and acceptance of interdisciplinary and multidisciplinary research (Alvarez 2016, pp. 3-4)

Watts (2013, p. 7) adds to this list a shorter timescale and lower cost for experiments in theory.

But alongside these advantages come challenges in the use of such data and methods that, if ignored, have the capacity to harm the public and the advancement of knowledge. From the perspective of the researcher, the necessary combination of tools and applications required, often from "multiple research traditions," are not all familiar to any individual researcher (Watts, 2013, pp. 5-6). Data acquisition is becoming more and more difficult, with much proprietary big data (such as the Social Security Administration database or IRS database that would be useful for the study of job networks and the economy) locked away and expensive. Data, once made available, is also messy, unreliable and easily falsified. In order to be usable, it must be grounded with offline findings or other web data. When decentralized online data is found to be false, there is no system of institutional accountability, further increasing uncertainty and eroding trust in the use of the web to crowdsource the production of data and knowledge (Conte et. al, 2012, p. 336). Additionally, now that the use of social network sites is becoming more common, users become more adept at toggling privacy controls and choosing which content to share publicly and which to keep hidden, and the availability of social media data decreases (Giglietto & Rossi, 2012, p. 25).

For study participants, the concerns of weight particularly relate to data acquisition, and its privacy and confidentiality, security and reliability. As social media data is extensively used in DH studies, we demarcate the line at which it is appropriate to use such information without users' consent by confronting extant questions of public/private arenas of publishing and accountholder motivation. Although it is important to retain the approval of users and collect private data ethically, failure to do so has its most damaging consequences when those who have access once it is collected are able to identify users and withdraw participants' privacy, and therefore, we discuss individual-level data and ways to retain people's confidentiality.

We also review ways of benefiting from data that comes from online sources, despite its inherent exclusion of those of low income and low socioeconomic status throughout much of the world, including the U.S. Also excluded are independent researchers, students and those associated with small organizations – especially interdisciplinarians – conducting this work often requires special supercomputers, and many humanities researchers do not have access to such resources or the skillset to use them. A number of papers have been written about data use ethics in other fields of research. This paper attempts to review and combine these needs for the specific purposes of Digital Humanities and Computational Social Science. Through an extended literature review, it collects ethical questions surrounding data use, and applies them to two infamous case studies: that of AOL's release of search data in 2006 and of Facebook's emotional contagion study published in 2014.

It is feasible to imagine that computational advantages, and the promise of DH and CSS, lead to a world of the analysis of not only text, but also sound, images and video, of richly-visualized data so that a maximum number of people can overcome confirmation bias and understand complex research results and contribute, and large-scale undertaking of crowd-sourced data and sophisticated citizen science is commonplace enough to allow us to solve high-impact questions. As we move towards such a world, a periodic reconsideration of ethics is judicious; it remains ever a timely topic with violations resulting in vast scandals and increasing public distrust (most recently the bout of data breaches, such as Uber's - Shaban, 2017).

References

Alvarez, R. M. (2016b). Introduction. In R. M. Alvarez (Ed.), *Computational Social Science: Discovery and Prediction* (pp. 1-24): Cambridge University Press.

Conte, R., Gilbert, N., Bonelli, G., Cioffi-Revilla, C., Deffuant, G., Kertesz, J., Helbing, D. (2012). Manifesto of computational social science. *European Physical Journal-Special Topics, 214*(1), 325-346. doi:10.1140/epjst/e2012-01697-8

Giglietto, F., & Rossi, L. (2012). Ethics and Interdisciplinarity in Computational Social Science. *Methodological Innovations Online, 7*(1), 25-36. doi:10.4256/mio.2012.003

Manovich, L. (2011). Trending: The Promises and the Challenges of Big Social Data. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (1 ed., Vol. 1, pp. 460-475): University of Minnesota Press.

Reid, A. (2011). Graduate Education and the Ethics of the Digital Humanities. In M. K. Gold (Ed.), *Debates in the Digital Humanities* (1 ed., Vol. 1, pp. 350-367): University of Minnesota Press.

Shaban, H. (2017). *Uber is sued over massive data breach after paying hackers to keep quiet*. *The Washington Post*. Retrieved 28 November 2017, from https://www.washingtonpost.com/news/the-switch/wp/2017/11/24/uber-is-sued-over-massive-data-breach-after-paying-hackers-to-keep-quiet/

Watts, D. J. (2013). Computational Social Science: Exciting Progress and Future Directions. *The Bridge: Linking Engineering and Society, 43*(4), 5-10.

Bridging the Gap: Digital Humanities and the Arabic-Islamic Corpus

Dafne Erica van Kuppevelt
d.vankuppevelt@esciencecenter.nl
Netherlands eScience Center, The Netherlands

E.G. Patrick Bos
p.bos@esciencecenter.nl
Netherlands eScience Center, The Netherlands

A. Melle Lyklema
a.m.lyklema@uu.nl
Utrecht University, The Netherlands

Umar Ryad
amr.ryad@kuleuven.be
University of Leuven, Belgium

Christian R. Lange
C.R.Lange@uu.nl
Utrecht University, The Netherlands

Janneke van der Zwaan
j.vanderzwaan@esciencecenter.nl
Netherlands eScience Center, The Netherlands

Introduction

Despite some pioneering efforts in recent times, the *longue durée* analysis of conceptual history in the Islamic world remains largely unexplored. Researchers of Islamic intellectual history still tend to study a certain canon of texts, made available by previous Western researchers of the Islamic world largely based on considerations of the relevance of these texts for Western theories, concepts and ideas. Indigenous conceptual developments and innovations are therefore insufficiently understood, particularly as concerns the transition from premodern to modern thought in Islam.

What, then, are the silenced continuities, transformations and major fault lines in Arabic-Islamic discourses? The Islamic tradition offers a vast textual corpus for exploring this question from a *longue durée* perspective, but its very breadth poses substantial problems for the individual scholar seeking to survey the literature by traditional methods. In the last decade, vast collections of digitized classical Arabic texts have become available online (Muhanna 2016, pp. 11-64). This marks the "beginning of what could become a methodological revolution in the fields of Arabic and Islamic Studies", as noted by Peralta and Verkinderen in the very first edited volume on Digital Humanities and the Arabic-Islamic corpus (Muhanna 2016, pp. 199).

This paper presents ongoing research to use state-of-the art Digital Humanities approaches and technologies to make pioneering forays into the vast corpus of digitized Arabic. This is done along the lines of three case studies, each of which examines a separate genre of Arabic and Islamic literary history.

Case studies

(1) Islamic law: This case study analyzes the corpus of digitally available (Sunni) legal works (furu' al-fiqh) from premodern to modern times (ca. 150 digitized works with ca. 75 million words, extracted from the OpenITI corpus[1] to investigate *longue durée* shifts in concepts and idioms employed in Muslim juridical discourse. The scholarly questions pursued relate to the history of the senses and of sense perception in the Islamic world, and of the human body more broadly speaking. Digital humanities methods applied to this corpus will include topic modelling (around the five senses) and computer-supported statistical analysis in historical perspective, that is, by comparing legal teachings throughout the fourteen centuries of Islamic law.

(2) Modern Islamic proselytizing literature: This case study analyses a largely neglected corpus of Arabic texts written between the 19th and 21st centuries (approx. 500 titles) on Islamic missionary activities (da'wa). The focus of the analysis will be to identify continuities and changes regarding the key concept of da'wa and the discursive idioms used to express them, and identify, graph and visualize the transnational networks involved with the discourses on da'wa.

(3) Arabic poetry: This case study will investigate the digital corpus of Arabic poetry (estimated 2,5 billion words, extracted from the OpenITI corpus). Poetry is an especially apt corpus to study the history of the senses and of sense perception in the Islamic world. What senses were favored by Arabic poets over the course of centuries? What kind of semantic fields are constructed in Arabic poetry around, for example, the sense of vision, and how does this contrast with, for example, legal constructions of vision?

Method

Most of the research projects in Digital Humanities have focused on Western Europe and the Americas, leaving a gap between state-of-the-art Digital Humanities tools and the Arabic text corpus. Many current initiatives in Arabic Digital Humanities seek to teach programming languages to humanities scholars. We pursue a different strategy to move Arabic Digital Humanities forward, by developing a freely accessible, user friendly interface to Digital Humanities technology, based on existing software.

The development of the technology is at an early stage, and we aim to present a first version of an Arabic-specific Digital Humanities toolkit at the conference.

1 Romanov, M, OpenITI. http://alraqmiyyat.github.io/OpenITI/

The toolkit integrates existing tools for stemming and morphological analysis in Arabic, as such as the Khoja stemmer (Khoja, Garside and Knowles, 2001), Tashaphyne stemmer[2] and the AlKhalil morphological analyzer (Boudchiche *et al.*, 2017). We will use the SAFAR software (Jaafar and Bouzoubaa, 2015) to compare these libraries and integrate the most relevant tools in a pipeline for humanities research. The resulting tagged datasets will be made available in an existing search engine, such as BlackLab[3]. All software developed for this paper is published open source[4].

We will present the development of the Arabic-specific Digital Humanities toolkit, including challenges that emerge from developing text mining tools specific for Arabic, with proposed solutions. It will also present early findings from the three case studies.

References

Boudchiche, M., Mazroui, A., Ould Abdallahi Ould Bebah, M., Lakhouaja, A. and Boudlal, A. (2017) 'AlKhalil Morpho Sys 2: A robust Arabic morpho-syntactic analyzer', *Journal of King Saud University - Computer and Information Sciences*. Elsevier, 29(2), pp. 141–146. doi: 10.1016/J.JKSUCI.2016.05.002.

Jaafar, Y. and Bouzoubaa, K. (2015) 'Arabic Natural Language Processing from Software Engineering to Complex Pipeline', in *2015 First International Conference on Arabic Computational Linguistics (ACLing)*. IEEE, pp. 29–36. doi: 10.1109/ACLing.2015.11.

Khoja, S., Garside, R. and Knowles, G. (2001) 'An Arabic tagset for the morphosyntactic tagging of Arabic', in *Corpus Linguistics*. Lancaster University. Available at: http://eprints.lancs.ac.uk/11985/ (Accessed: 24 April 2018).

Muhanna, E. (ed.) (2016) *The Digital Humanities and Islamic & Middle East Studies*. Berlin, Boston: De Gruyter. doi: 10.1515/9783110376517.

Off-line sStrategies for On-line Publications: Preparing the Shelley-Godwin Archive for Off-line Use

Raffaele Viglianti
rviglian@umd.edu
Maryland Institute for Technology in the Humanities, University of Maryland, United States of America

Digital scholarly editions and archives are typically published on the web, which makes it possible to create interactive and reading experiences with the potential of reaching worldwide audiences. When text is encoded with care, for example by adopting the Text Encoding Initiative standard, it becomes possible for the same encoded content to be delivered in other formats and media, such as e-book and PDF for print. Web-based interactive digital editions, however, are the most efficient in utilizing the interactive and interconnected features of the web for presenting both text and the editorial scholarship that produced it. Ongoing scholarship around minimal computing and minimal editions has pointed out some important, yet addressable, flaws of many TEI digital editions. Bloatedness of infrastructure, for example, particularly when paired with rapid technical obsolescence and changes in funding, can hamper long-term preservation efforts; weighty resources may not be easily accessible from slower connections; and online-only access to a digital edition can be an obstacle to the world-wide access potential highlighted earlier.

The Shelley-Godwin Archive (S-GA) has taken steps to reduce its infrastructure footprint by generating a static site: in its production form, with the exception of its search index, S-GA is a collection of TEI, HTML, CSS, and JavaScript that can be hosted on any server without needing to set-up any server-side component (see Fig. 1).

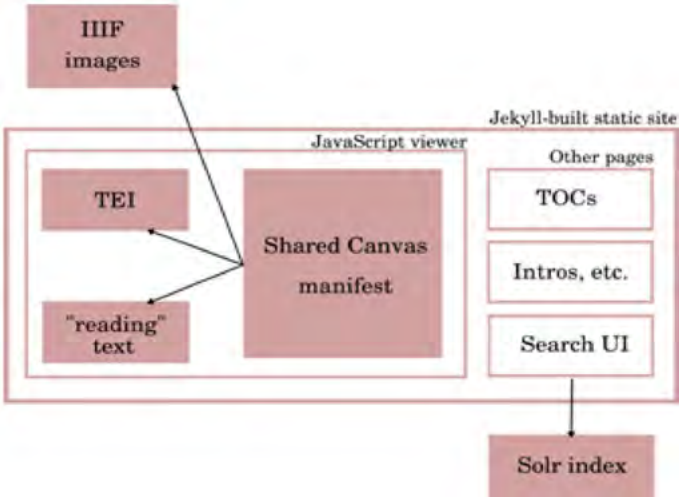


Fig. 1 The architecture of the S-GA website, built using Jekyll and static assets. Images are served primarily by the Oxford Bodleian Library using the International Image Interoperability Framework API.

This approach also makes it possible to bundle resources together for off-line use. This poster will show three potential approaches to creating off-line resources for an on-line publication: a one-document HTML bundle, a compressed archive of resources and an Electron desktop application. Unlike a PDF or e-book version, these downloadable resources will preserve the functionality of S-GA's website (with the exception, for now, of full text search), thus making the archive more usable in a poten-

2 T. Zerrouki, Tashaphyne, Arabic light stemmer, https://pypi.python.org/pypi/Tashaphyne/0.2
3 http://inl.github.io/BlackLab/
4 https://github.com/arabic-digital-humanities

tially greater number of cases, including increasing access for users with slow or no internet connections.

Academy of Finland Research Programme "Digital Humanities" (DIGIHUM)

Risto Pekka Vilkko
risto.vilkko@aka.fi
Academy of Finland, Finland

Digital Humanities (DIGIHUM) is a four-year research programme funded by the Academy of Finland. Its aim is to address novel methods and techniques in which digital technology and state-of-the-art computational science are used for collecting, managing and analysing data in humanities and social sciences research as well as for modelling humanities and social science phenomena.

Finland has a strong tradition in digital humanities. By bringing together the existing best knowledge and skills in digital humanities, Finland aims to put itself in a strong position to become a world leader in this rapidly evolving field. The programme is grounded in the needs of basic research, but technological advances in this area also have great potential for practical applications that warrant research.

The development of research in this area is based on broad collaboration involving not only researchers in the field but also technology experts, representatives of memory organisations (libraries, archives) and database administrators and developers. One aspect of the programme is to examine digitalisation as a cultural and social phenomenon.

The programme has three thematic areas:

1. Research into digital interaction and digital services
2. Employing open, multiform and/or real-time data in research
3. Data-based analysis and modelling of humanities and social sciences phenomena.

The programme produces new and more comprehensive knowledge and understanding about the themes under investigation. It fosters dialogue and exchange between a wide variety of scientific fields and disciplines, for example, by integrating methodologies and networking at national and international level. The programme encourages interdisciplinary or multidisciplinary projects that combine two or more fields of scientific research employing different methodologies and approaches. The aim is to promote:

1. collaboration among producers, processors and users of humanities and social sciences data
2. the development of research methods
3. ethical examination of the research field
4. the usability and awareness of datasets.

The poster will interactively introduce the programme's themes and objectives as well as the six research consortia that form the core of the programme:

- *Profiling Premodern Authors* (Prof. Marjo Kaartinen et al., University of Turku)
- *Interfacing Structured and Unstructured Data in Sociolinguistic Research on Language Change* (Prof. Terttu Nevalainen et al., University of Helsinki)
- *Citizen Mindscapes – Detecting Social, Emotional and National Dynamics in Social Media* (Prof. Jussi Pakkasvirta et al., University of Helsinki)
- *Computational History and the Transformation of Public Discourse in Finland, 1640–1910* (Prof. Hannu Salmi et al., University of Turku)
- *Digital Face* (Prof. Janne Seppänen et al., University of Tampere)
- *Digital Language Typology: Mining from the Surface to the Core* (Prof. Martti Vainio et al., University of Helsinki).

The poster will also introduce the four additional projects related to the Trans-Atlantic (T-AP) Platform *Digging into Data Challenge*:

- Digging into Manuscript Data (Prof. Eero Hyvönen, University of Helsinki)
- Analyzing Child Language Experiences Around the World (Prof. Okko Räsänen, Aalto University)
- Oceanic Exchanges: Tracing Global Information Networks in Historical Newspaper Repositories, 1840–1914 (Prof. Hannu Salmi, University of Turku)
- Digging into High Frequency Data: Present and Future Risks and Opportunities (Prof. Peter Sarlin, Hanken School of Economics).

The Academy of Finland is a government agency within the administrative branch of the Finnish Ministry of Education, Science and Culture. Its mission is to fund high-quality scientific research, provide expertise in science and science policy, and strengthen the position of science and research.

Modeling the Genealogy of Imagetexts: Studying Images and Texts in Conjunction using Computational Methods

Melvin Wevers
melvin.wevers@huygens.knaw.nl
DH Group, KNAW Humanities Cluster, The Netherlands

Thomas Smits
t.smits@let.ru.nl
Institute for Historical, Literary and Cultural Studies, Radboud University, The Netherlands

Leonardo Impett
leonardo.impett@epfl.ch
Digital Humanities Institute, EPFL, Switzerland

In his influential article "There are no Visual Media", theorist of visual culture W.J.T Mitchell argues that "all media are mixed media" (Mitchell, 2005). In earlier work, Mitchell already noted that composite works—media formats that consist of both image and text—cannot be adequately studied by comparing the meaning of these two forms of representation separately (Mitchell, 1994, p. 89). The subject matter of these "imagetexts", is, rather, the "whole ensemble of relations between media" (Mitchell, 1994, p. 89). In other words, the meaning of one of the components of an imagetext, be it either the image or the text, can only be understood in relation to the other. This paper combines methods from text mining, computer vision, and information theory to increase our understanding of this relationship throughout several historical datasets.

Several scholars have observed that Digital Humanities research mainly focuses on (large-scale) textual analysis (Champion, 2017; Meeks, 2013). Erik Champion, for instance, notes that the influential definition of Digital Humanities by the University of Oxford is entirely "text based and desk based" (Champion, 2017, p. 25). While he rightly claims that research in the Digital Humanities is centered on text, in recent years an increasing number of researchers have started studying visual material, in which has been called "visual big data" (Ordelman et al., 2014; Smith, 2013). Scholars increasingly rely on computational methods to analyze these large digitized visual datasets in innovative ways. Important examples are the work of Seguin (Seguin et al., 2017) on visual pattern discovery in large databases of paintings, Impett and Moretti's (Impett and Moretti, 2017) large-scale analysis of body postures in Aby Warburg's Atlas Mnemosyne, and Wevers' (Wevers and Lonij, 2017) and Smits's (Smits, 2017; Smits and Faber, 2018) analysis of visual trends in advertisements and images in newspapers. These projects were all presented at DH2017, some during the well-attended pre-conference workshop of the Special Interest Group AudioVisual Material in Digital Humanities (AVinDH).

The recent upsurge of large-scale analysis of visual material shifts the focus in Digital Humanities research away from texts. However, this has also led researchers to approach text and images as disjointed entities. Computational techniques can analyze similarity and change in both textual and visual discourse. Our project applies techniques from both textual and visual computational analysis to a dataset of advertisements for cars extracted from the widely-read Dutch newspaper *De Volkskrant* between 1945 and 1995, which we extracted from the large collection of digitized newspapers maintained by the National Library of the Netherlands. By juxtaposing change points in text and visual material, we show that the meaning of imagetexts can be studied by looking at the relation between the two forms of representation. Put differently, how does change and continuity in the visual correspond to changes in the textual and vice versa?

Using Kleinberg's burst algorithm, we detected bursty words in the textual content of advertisements (Kleinberg, 2002). These bursts indicate possible change points in advertising discourse that call for closer examination of the advertisements and can be cross-examined with possible changes in the visual content. Also, topic modeling (LDA) was used to detect clusters of advertisements based on textual context. These clusters were compared to cluster based on visual aspects.

Trends, similarities, and points of inflection in the image sets will be traced using a subspace learned by training a Generative Adversarial Network (GAN; see Goodfellow et al. 2016), which has been shown to generate semantically-meaningful vector subspaces. GANs work best with regular sets of images - our visual analysis process is thus twofold. First, we use a pretrained Mobilenet CNN (Howard et al. 2017) to detect objects (cars, trucks, people, etc), and then train individual GANs to explore the visual-semantic space of each object through time.

Whereas a traditional CNN can only encode from image to vector, a GAN can also decode from any vector to generate artificial images; trends or clusters hypothesized in a vectorial subspace can therefore be subjected to a 'close reading' of the corresponding artificial images. This generative hermeneutic avoids the 'black box' nature of traditional neural network image analysis.

The ability to detect how changes and continuity between text and images correlate increases our understanding of the function of imagetexts in modern culture. It also helps us understand whether the relationship between the two forms of representation became more entangled over time, or whether this entanglement is specific to particular products or specific periods.

References

Champion, E.M. (2017), "Digital humanities is text heavy, visualization light, and simulation poor", *Digital Scholarship in the Humanities*, Vol. 32 No. 1 sup, pp. i25–i32.

Howard, A.., et al. (2017) "Mobilenets: Efficient convolutional neural networks for mobile vision applications." arXiv preprint arXiv:1704.04861.

Impett, L. and Moretti, F. (2017), "Totentanz", *New Left Review*, No. 107, pp. 68–97.

Kleinberg, J. (2002), "Bursty and Hierarchical Structure in Streams", P*roceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, Edmonton, Canada (2002), pp. 91–101.

Meeks, E. (2013), "Is Digital Humanities Too Text-Heavy?", *Digital Humanities Specialist*.
Mitchell, W., 1994. Picture Theory. Essays on Verbal and Visual Representation, University of Chicago Press, Chicago.
Mitchell, W.J.T. (2005), "There Are No Visual Media", *Journal of Visual Culture*, Vol. 4 No. 2, pp.257–266.
Ordelman, R., Kleppe, M., Kemman, M. and De Jong, F. (2014), "Sound and (moving images) in focus – How to integrate audiovisual material in Digital Humanities research", ADHO 2014.
Seguin, B., di Leonardo, I. and Kaplan, F. (2017), "Tracking Transmission of Details in Paintings", ADHO 2017.
Smith, J.R. (2013), "Riding the Multimedia Big Data Wave", *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, NY, USA, pp. 1–2.
Smits, T. (2017), "Illustrations to Photographs: using computer vision to analyse news pictures in Dutch newspapers, 1860-1940", ADHO 2017.
Smits, T., Faber, W.J. (2018), "*CHRONIC* (Classified Historical Newspaper Images)", *KB Lab*, 21 March, http://lab.kb.nl/dataset/chronic-classified-historical-newspaper-images.
Wevers, M. and Lonij, J. (2017), "Siamese", *KB Lab*, 15 October, http://lab.kb.nl/tool/siamese.

History for Everyone/Historia para todos: Ancient History Encyclopedia

James Blake Wiener
james.wiener@ancient.eu
Ancient History Encyclopedia (AHE), United Kingdom

Gimena del Rio Riande
gdelrio@conicet.gov.ar
IIBICRIT, CONICET, Argentina

The most important publication to emerge from 18th-century Europe was arguably the *Encyclopedia* by Denis Diderot (1713-1784 CE). The *Encyclopedia* sought to "change the general way of thinking," challenging all forms of bigotry, repression, fanaticism, and misinformation (Fowler, 2011). Through his commission of articles on a variety of topics, Diderot endeavored to summarize and disseminate the world's information in order to help human society progress to new heights of accomplishment while also mitigating the sufferings of civilization. Helped by the mathematician Jean Le Rond d'Alembert (1717-1783) as well as Montesquieu (1689-1755 CE) and Voltaire (1694-1778), Diderot's *Encyclopedia* was very much a collaborative project, which reflected the "party of humanity" in a new age of international and informational exchange (Micale and Dietle, 2000).

Ancient History Encyclopedia (http://www.ancient.eu) was founded in 2009, in the spirit of the Enlightenment, with the mission to improve education through the creation of the most complete, freely accessible, and reliable online, historical resource in the world. As scholars and supporters of the digital humanities, the contributors at Ancient History Encyclopedia felt a responsibility to construct a site in which users not only found what they were looking for, but also one which stressed the importance of global cultural heritage and world history. Our knowledge and interpretation of history shapes how we define ourselves as nations and as cultures, and it influences how we see other cultures as well. Whether through its interactive map of the ancient world, online videos, or its carefully curated articles and definitions, Ancient History Encyclopedia digitally imparts knowledge in new and creative ways.

Before Ancient History Encyclopedia's inception, most of what was available online pertaining to ancient history was scattered across various websites, illegible due to poor presentations, targeted exclusively at academic audiences and hidden behind paywalls, or tainted by a distinct nationalistic agendas. While Wikipedia undeniably advanced and pushed the aims of the Open Access movement, it sometimes remains riddled with inaccuracies and occasional bias. Omnipresent too is the lack of available content in major world languages like Spanish, Russian, Mandarin Chinese, Arabic, Portuguese, and Hindi. Other sites, like La guía de historia (https://www.laguia2000.com/), do not afford proper attributions to sources and lack curated multimedia libraries of pictures, videos, and other interactive learning tools. Over the last two decades, open access publishing has become increasingly widespread with the help of the Internet. The Open Access movement helps researchers, students, and educators access the latest research and data without restrictions. It is a movement defined by high standards, the exchange of information, the development and synchronization of models, and the promotion of innovation in technology and research methodology.

Through a shared commitment to Open Access Education, Ancient History Encyclopedia and its partners create interactive tools that facilitate historical and media literacy, build models of data exchange, and serve a broader community rather than solely those in academia. In this sense, Ancient History Encyclopedia is acting in unison with the principles of Open Access, Open Education, and Open Research. These are positive developments, but it is not nearly enough: historians, researchers, publishers, museums, and other institutional bodies must move beyond the paradigm of simply making it free or available only in English. Ancient History Encyclopedia is an international project with contributors from Germany, the United States, Hungary, India, Argentina, the United Kingdom, and Australia. Through Ancient History Encyclopedia's collaborations with other digital humanities pro-

jects and organizations, including the Pelagios Commons, Europeana's Eagle Project on ancient Roman epigraphy, Humanidades Digitales del Centro Argentino de Información Científica y Tecnológica (HD CAICYT-CONICET), and Laboratorio de Innovacíón en Humanidades Digitales at Madrid's Universidad Nacional de Educación a Distancia (LINHD-UNED), Ancient History Encyclopedia has aided in making important academic research and datasets available and digestible to Anglophone audiences.

In this poster, Ancient History Encyclopedia and Humanidades Digitales CAICYT-CONICET (http://www.caicyt-conicet.gov.ar/micrositios/hd/) review Ancient History Encyclopedia's encyclopedic model and successes, while also sharing plans for future projects that will include the translation and publication texts at CAICYT and the joint use of map data from Pelagios Commons (http://commons.pelagios.org/

References

Fowler, J. (2011). *New essays on Diderot*. Cambridge: Cambridge University Press.

Micale, M. S., and Dietle, R. L. (2000). E*nlightenment, Passion, Modernity. Historical Essays in European Thought and Culture*. Redwood City: Stanford University Press.

Princeton Prosody Archive: Rebuilding the Collection and User Interface

Meredith Martin
mm4@princeton.edu
Princeton Prosody Archive, Princeton University Center for Digital Humanities, United States of America

Meagan Wilson
mrwilson@princeton.edu
Princeton Prosody Archive, Princeton University Center for Digital Humanities, United States of America

Mary Naydan
mnaydan@princeton.edu
Princeton Prosody Archive, Princeton University Center for Digital Humanities, United States of America

The PPA collects and displays historical documents prior to 1923, bringing to light little-known texts about the study of language, the study of poetry, and where and how these intersect and diverge. By gathering these documents into one place, the PPA tracks the development of English poetry as a subject of study and shows how this development bridges a variety of discourses, most prominently the rise of linguistic nationalism and linguistic imperialism, but also the advent of stadial history and historiography, the rise of phonetic science and the beginnings of historical linguistics, and a variety of related pedagogical movements that evolve from rhetoric through to elocution and the study of "speech." The PPA is the only large-scale corpus focused specifically on the study of poetry in the English language. Materials in the archive include grammar handbooks, poetic treatises, versification manuals, elocution guides, histories of literature, editorial introductions, phonetic tracts, and journal articles pertaining to the measure and pronunciation of poetry. By viewing prosody broadly and collecting these materials into one archive, scholars can finally see how the histories of English poetics and linguistics are intertwined, and how the story of English poetic development, alongside the development of historical linguistics, increasingly borrowed, co-opted, imitated, erased, or "civilized" poetic forms from other languages.

Critical attention to these poetic histories and debates are the foundation of Historical Poetics. In addition to scholars of Historical Poetics, the PPA's audience is teachers of poetry, scholars of poetry, linguists, practicing poets, historians of language, historians of pedagogy, scholars of sound studies, scholars of rhetoric, and lexicographers—all of whom can use the PPA to discover the emergence of a disciplinary term, trace its evolution, or determine its ties to national or political debates. Finally, computer scientists and digital humanists are eager to run textual analytic algorithms on a curated data set that might reveal previously unknown or unexpected results such as the most frequently reprinted poetic example or the most frequently repeated (perhaps without attribution) definition of a particular term.

"Rebuilding the Collection and User Interface," the PPA's poster and interactive demonstration for DH2018, showcases the immense data-refinement and metadata-cleaning performed by the PPA since its DH2014 poster session. After launching our new website in May 2018, we are well-positioned to discuss the strengths and struggles of curating and designing an interactive website that relies on HathiTrust Digital Library content. In this way, the PPA sees itself as a project similar to Early American Cookbooks, recently published as a HathiTrust case study in *Code4Lib*. "Legacy MARC data for early books held in special collections presents particular challenges," Gioia Stevens writes; "Cleaning and standardizing this legacy data is an essential step in analyzing special collections metadata as a dataset rather than as individual records" (Stevens, 2017). This has proven especially germane to the PPA. From 2015 to 2017, the PPA refined its core collection by eliminating 3,729 duplicate works through a complex and painstaking metadata cleaning process. These duplications were the result of our initial file transfer from HathiTrust and the replicas were skewing users' search results. The PPA offers a case study in the challenges posed by working with unstandardized metadata. In addition to addressing the benefits and drawbacks of our collaboration with HathiTrust, our poster session aims to highlight how our new interface

guides users toward the database's implicit and explicit arguments, highlights unusual content, and provides pathways for discovery.

References

Stevens, Gioia. (2017). "New Metadata Recipes for Old Cookbooks: Creating and Analyzing a Digital Collection Using the HathiTrust Research Center Portal." *Code4Lib* **37**, http://journal.code4lib.org/articles/12548 (accessed 1 May 2018).

ELEXIS: Yet Another Research Infrastructure. Or Why We Need An Special Infrastructure for E-Lexicography In The Digital Humanities

Tanja Wissik
tanja.wissik@oeaw.ac.at
Austrian Academy of Sciences, Austria

Ksenia Zaytseva
ksenia.zaytseva@oeaw.ac.at
Austrian Academy of Sciences, Austria

Thierry Declerck
declerck@dfki.de
Austrian Academy of Sciences, Austria

In this presentation, we will discuss the recently started European project ELEXIS – European Lexicographic Infrastructure and its potential in the context of digital humanities.

The use of the computer in modern lexicography is intertwined with the history of the digital humanities (c.f. Schreibmann et al. 2004) and the lexical data have grown to be is indispensable in more and more DH projects, especially with the rise of the Semantic Web and Linked Open Data (c.f. Oldman et al. 2016).

However, current lexicographic resources, both modern and historical, have different levels of structuring and are not equally suitable for the application in other fields, such as Natural Language Processing, and thus not directly usable in DH projects for Semantic Web applications and methods.

Therefore, ELEXIS will develop strategies, tools and standards for extracting, structuring and linking lexicographic resources to unlock their full potential for Linked Open Data and the Semantic Web, as well as in the context of digital humanities.

The ELEXIS project is carried out by a consortium of partners from various fields (e.g. lexicography, computational linguistics, natural language processing, digital humanities, and artificial intelligence). The consortium consists of the following scientific institutions, language institutes, standardisation bodies, and publishing houses: "Jožef Stefan" Institute (Slovenia), Lexical Computing CZ s.r.o. (Czech Republic), Instituut voor de Nederlandse Taal (Netherlands), La Sapienza University of Rome (Italy), National University of Ireland, Galway (Ireland), Austrian Academy of Sciences (Austria), Belgrade Center for Digital Humanities (Serbia), Hungarian Academy of Sciences, Research Institute for Linguistics (Hungary), Institute for Bulgarian Language »Prof Lyubomir Andreychin« (Bulgaria), Universidade Nova de Lisboa (Portugal), K Dictionaries (Israel), Istituto di Linguistica Computazionale "A. Zampolli" (Italy), The Society for Danish Language and Literature (Denmark), University of Copenhagen, Centre for Language Technology (Denmark), Trier University, Center for Digital Humanities (Germany), Institute of the Estonian Language (Estonia), Real Academia Española (Spain).

The ELEXIS project aims to integrate, extend and harmonise national and regional efforts in the field of lexicography, both modern and historical, with the goal of creating a sustainable infrastructure which will enable efficient access to high quality lexical data in the digital age, and bridge the gap between more advanced and lesser-resourced scholarly communities working on lexicographic resources.

ELEXIS intends to take an innovative approach of production and development of lexico-semantic resources by creating intelligent applications for crucial tasks such as linking lexical resources, word sense disambiguation and cross-lingual mapping on the basis of applied methods and techniques in the fields of NLP and Artificial Intelligence fields.

The ELEXIS infrastructure will help researchers create, access, share, link, analyse, and interpret heterogeneous lexicographic data across national borders, paving the way for ambitious, trans-national, data-driven advancements in the field, while significantly reducing the duplication of efforts across disciplinary boundaries. In order to ensure the sustainability of the technical infrastructure after the end of the project, the created infrastructure will be integrated into the already existing infrastructures CLARIN and DARIAH, since most of the partners are members of CLARIN and DARIAH national consortia.

Besides the technical infrastructure, ELEXIS will establish a network for knowledge exchange and will develop and implement free online training courses for lexicography. Furthermore, ELEXIS will give researchers and research teams trans-national access to research facilities and lexicographical resources which are not fully accessible online or where professional on the spot expertise is needed in order to ensure and optimise mutual knowledge exchange. The trans-national access will have impact especially for under-resourced languages and will

all in all strengthen the infrastructure and collaborative network provided by ELEXIS.

Even though the infrastructure is at the moment planned as a European infrastructure, there are thoughts to expand it beyond Europe in order to cater for the needs of DH researchers around the globe.

References

Schreibman, S., Siemens, R. and Unsworth, J. (eds.) (2004). *A Companion to Digital Humanities. Oxford: Blackwell*. http://www.digitalhumanities.org/companion/

Oldman, D., Doerr, M. and Gradmann, S. (2016). Zen and the Art of Linked Data: New Strategies for a Semantic Web of Humanist Knowledge. In Schreibman S. et al. (eds.) (2016). *A New Companion to Digital Humanities*, 2nd Edition. Oxford: Wiley-Blackwell.

"Moon:" A Spatial Analysis of the Gumar Corpus of Gulf Arabic Internet Fiction

David Joseph Wrisley
djw12@nyu.edu
New York University Abu Dhabi, United Arab Emirates

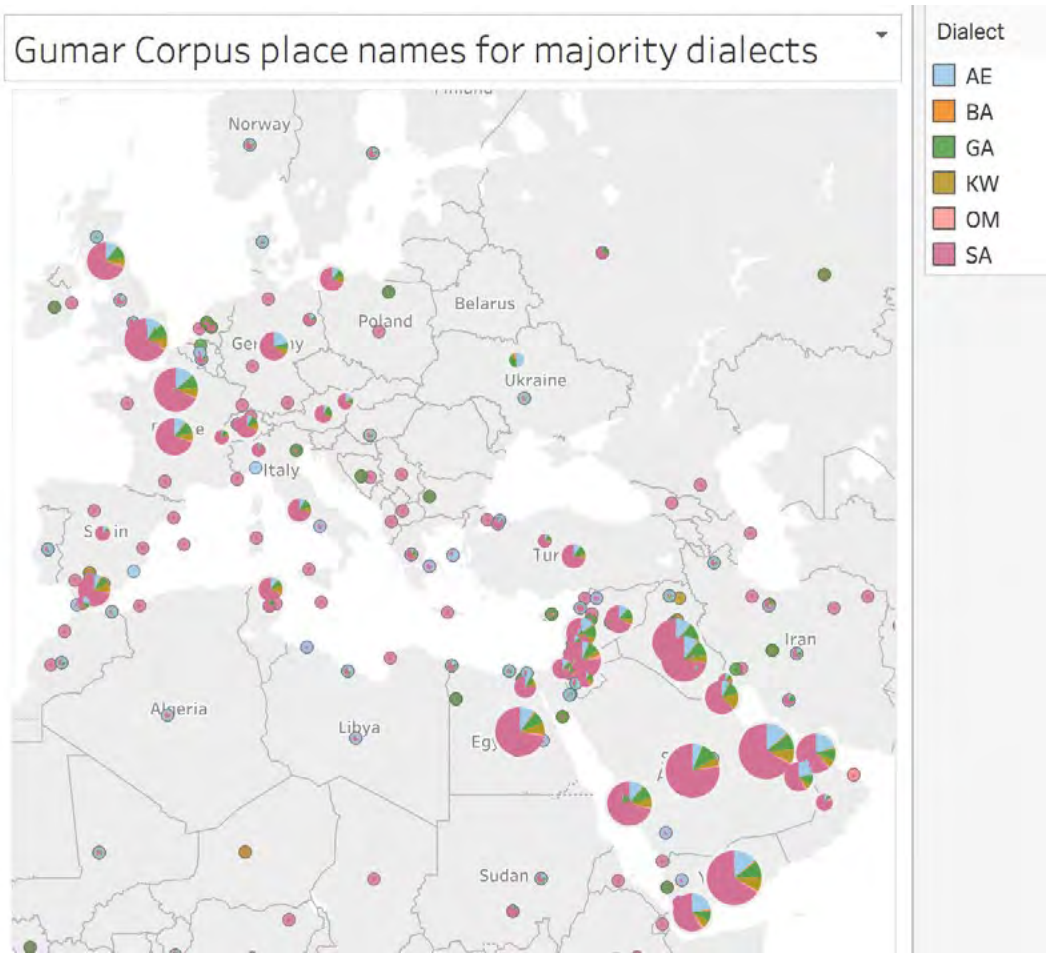
Hind Saddiki
hind.saddiki@nyu.edu
Mohammadia School of Engineering,
Mohammed V University in Rabat, Morocco; Computational Approaches to Modeling Language Lab,
New York University Abu Dhabi, United Arab Emirates

The Gumar Corpus (https://camel.abudhabi.nyu.edu/gumar/) consists of 110 million words from 1,200+ Internet forum novels written in a conversational style about romantic topics. Whereas the corpus was originally harvested and annotated for use within the context of dialectal Arabic (DA) natural language processing, the material is also of cultural and sociological significance concerning popular culture in the Gulf Arab region. The corpus' name comes from the Gulf Arabic word for moon {gumɛr}, a popular Arabic term of endearment. Whereas the genre is all but unknown outside of the Arab world, the Arabic blogosphere and social media are full of discussions about these "net novels," the authors of which are purportedly young women. In addition to Modern Standard Arabic (MSA), we have five dialect varieties mapping to roughly 12 national sub-varieties of dialectal Arabic--usually only one tag is assigned to each internet novel. Our poster is a very first attempt to tap into the cultural richness of the corpus using methods adapted to the Arabic language, in particular from the angle of spatial analysis of corpora.

The internet novels sometimes identify their country of origin in a short prologue, but there are additional clues as to their provenance including the fact that they are all written in DA, which is not necessarily the native dialect of the author. Much progress has been made in information extraction and NLP for Arabic in the last decade, but in dialectal forms much work remains to be done to catch up to Western languages. Even though we would not expect significant variance in toponyms in DA, initial attempts at extracting place names directly from the Arabic corpus posed a methodological challenge, particularly for disambiguation. Practical workarounds, often translingual and through English, are sometimes adopted in such cases with Arabic, as in the case of BetaCode that uses English script to deal with the vocalization, or partial vocalization of texts (Romanov, 2015).

With the goal of extracting place name entities from Gumar, our pilot study carried out morphological analysis and disambiguation on the texts using MADAMIRA (Pasha et al., 2014), a tool that currently functions for both for MSA and Egyptian Arabic. The configuration for Egyptian has been shown to outperform the MSA setting when compared to a manually annotated sample of 4K words from the Gumar corpus (Khalifa et al., 2015). Since the MADAMIRA morphological annotation provides both the lemma of a word and the English translation of the lemma, we build an English approximation of the novels and run them through Stanford Named Entity Recognizer to detect locations (Finkel et al., 2005).

Using Stanford NER, 19000+ occurrences of some 400+ distinct locations were extracted from the aggregate of the novels. Having English versions of the place names made the geocoding a relatively straightforward process. Geovisualization shows that the highest frequency of place names are found in the Arabian Gulf, Iraq, *bilad as-Sham* and Al-Andalus (southern Spain), as well as in England, France and Germany. Given that about sixty percent of the novels are identified as the dialect of Saudi Arabia, the high frequency of mentions of the Kingdom seems predictable. On the other hand, the places are not specific locales, as in the case of the city-level geographies of Palestine and Iraq. Other more detailed analysis about such specificity of place needs to be carried out through subsequent close reading of the corpus.



While the corpus is a rare opportunity both to work with contemporary popular culture Arabic in the textual digital humanities and to experiment with named entity recognition methods for non-Western contexts, caution must be exercised in our interpretations since the methods which work well for western languages are much more tentative in the (regional) Arabic case. For example, some cross checking was done against the Arabic texts in the corpus and revealed errors where DA colloquialisms {kif} ("what?") and {bliz} ("please"), generated some high frequency false locations "Kiev" and "Belize." As our research evolves, we intend to benchmark other Arabic-only tools for entity recognition to test their stability and performance on the set of materials in question (Gahbiche-Braham et al. 2013; Shaalan 2014). Time permitting, we would like to begin to do some correlations between topic and geography, what has been recently labelled a "geospatial semantics" (Gavin/Gidal, 2017) but for the transnational, multiregional context of Arabic. Our hope is to use the Gumar corpus to take on more in-depth analysis of a Gulf Arabic geopoetics of romance.

References

Finkel, J. R., Grenager, T. and Manning, C. (2005). Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling. In *Proceedings of the 43nd Annual Meeting of the Association for Computational Linguistics* (ACL 2005), pp. 363-370.

Gahbiche-Braham, S., Bonneau-Maynard, H., and Yvon, F. (2013). Traitement automatique des entités nommées en arabe: détection et traduction. *Traitement Automatique des Langues*, 54(2): 101-32.

Gavin, M., Gidal, E. (2017). Scotland's Poetics of Space: An Experiment in Geospatial Semantics, *Cultural Analytics*. http://culturalanalytics.org/2017/11/scotlands-poetics-of-space-an-experiment-in-geospatial-semantics/ (accessed 30 April 2018).

Khalifa, S. et al. (2016). A Large Scale Corpus of Gulf Arabic, In *Proceedings of the Language Resources and Evaluation Conference* (LREC), Portorož, Slovenia, pp. 4282-89.

Pasha, A. et al. (2014). MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In *Proceedings of the Language Resources and Evaluation Conference* (LREC), Reykjavík, Iceland, pp. 1094-1101.

Romanov, M. (2015). BetaCode for Arabic, *Al-Raqmiyyat*, https://maximromanov.github.io/2015/02-07.html (accessed 30 April 2018).

Shaalan, K. (2014) A Survey of Arabic Named Entity Recognition and Classification, *Computational Linguistics*. 40(2): 469-510.

A New Methodology for Error Detection and Data Completion in a Large Historical Catalogue Based on an Event Ontology and Network Analysis

Gila Prebor
gila.prebor@biu.ac.il
Bar Ilan University, Israel

Maayan Zhitomirsky-Geffet
maayan.geffet@gmail.com
Bar Ilan University, Israel

Olha Buchel
obuchel@gmail.com
Faculty of Information and Media Studies, University of Western Ontario, Canada

Dan Bouhnik
dan.bouhnik@gmail.com
Bar Ilan University, Israel; Jerusalem College of Technology, Israel

Introduction

The catalogue of Historical Hebrew Manuscripts, curated by the National Library of Israel, represents the largest collection in the world of over 130,000 Hebrew manuscripts that survived through the last millennium and are currently spread off in a variety of institutions all over the globe. The catalogue was created by many different classifiers during the long period of some 70 years. As a result, many of the fields are inconsistent and unorganized (Zhitomirsky-Geffet and Prebor, 2016). Moreover, a deeper examination of the data reveals missing and incorrect information (e.g. manuscripts with unknown date and place of writing). This missing and incorrect information poses a great pitfall for researchers who need reliable data to base their research on (Hric et al., 2016).

In this paper we present a novel approach for completion and correction of historical data from a large manuscript catalogue based on an event-based ontology and network communities' analysis. To resolve data inconsistencies in the catalogue, in the previous study we proposed an event-based ontology model (Zhitomirsky-Geffet and Prebor, 2016). The ontology model is shown in Figure 1.

Figure 1: Ontology model of the manuscript data.

Approach

- The proposed methodology comprises the following stages:
- Extraction of ontological entities from the catalogue data and ontology construction;
- Building networks of ontological entities based on direct and indirect ontological relationships between these entities, e.g. a network of censors who participated in the common Manuscript Censorship Events, or a bipartite network of manuscripts and people related to them through some events;
- Automatic community identification in the constructed networks (Blondel et al., 2008);
- Outlier detection among the related events in the network or in the closest community, i.e. if the manuscript creation event's date is later than its censorship event's date;
- Semi-automatic inference of missing data based on the ontological relationships and communities in the network, e.g. inferring a censor/author's missing time and place of living from the corresponding data of his peers in the community.

Results

Here we present preliminary results of the proposed approach applied on the case of Censorship Events of Hebrew manuscripts in medieval Italy. In the context of the Counter-Reformation, during the 16th-18th centuries, the Catholic Church closely supervised written and printed literature. The Church appointed censors (most of them apostates and experts in the Hebrew language) to censor and approve the Hebrew books.

The diagram in Figure 2 emphasizes the most influential censors and demonstrates the strengths of collaborations.

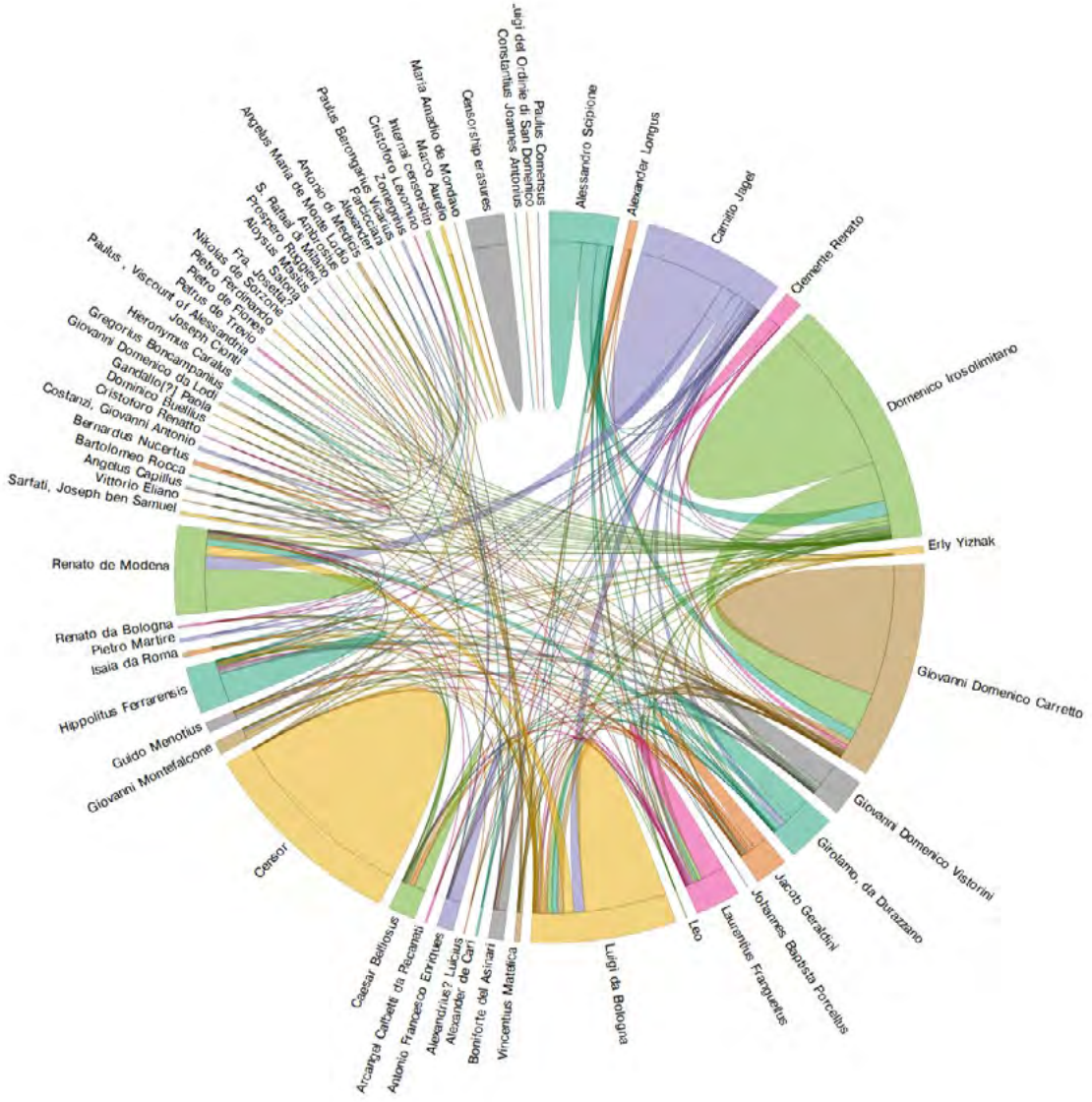


Figure 2: A chord diagram of censors related by common manuscripts.

The total number of relationships in the network depicted in Figure 3 is 2,037, and the number of nodes (representing Italian censors) is 62. For 37% of the censors we can observe the discrepancies between explicitly mentioned in the catalogue and automatically inferred periods of activity, the inference was based on the dates of individual censorship events in the ontology. In 5% of the cases the dates of their activity in the catalogue were incorrect (e.g. in cases where the activity period of 50 and more years). In addition, given Censorship Events' related entities, such as, a censor name and date and manuscripts censored by him and their related dates and locations we could infer its location and as a result, we obtained places of censorship for 53.9% out of all the Censorship Events in the corpus. Eight of the inferred places did not appear in the original list of 12 places that have been recorded in the catalogue.

Grey links in Figure 3 show that the most influential censors were censoring manuscripts one after another. Such collaborations can probably be regarded as waves of censorship. Two timelines under the map display which locations were inferred and which ones were specified explicitly in the catalogue. Comparing these two timelines allows researchers to identify mistakes about time and gives suggestions for unknown locations.

To conclude, our preliminary findings show that ontology-based network analysis and detection of communities provide an effective tool for correction and completion of missing or incorrect historical data.

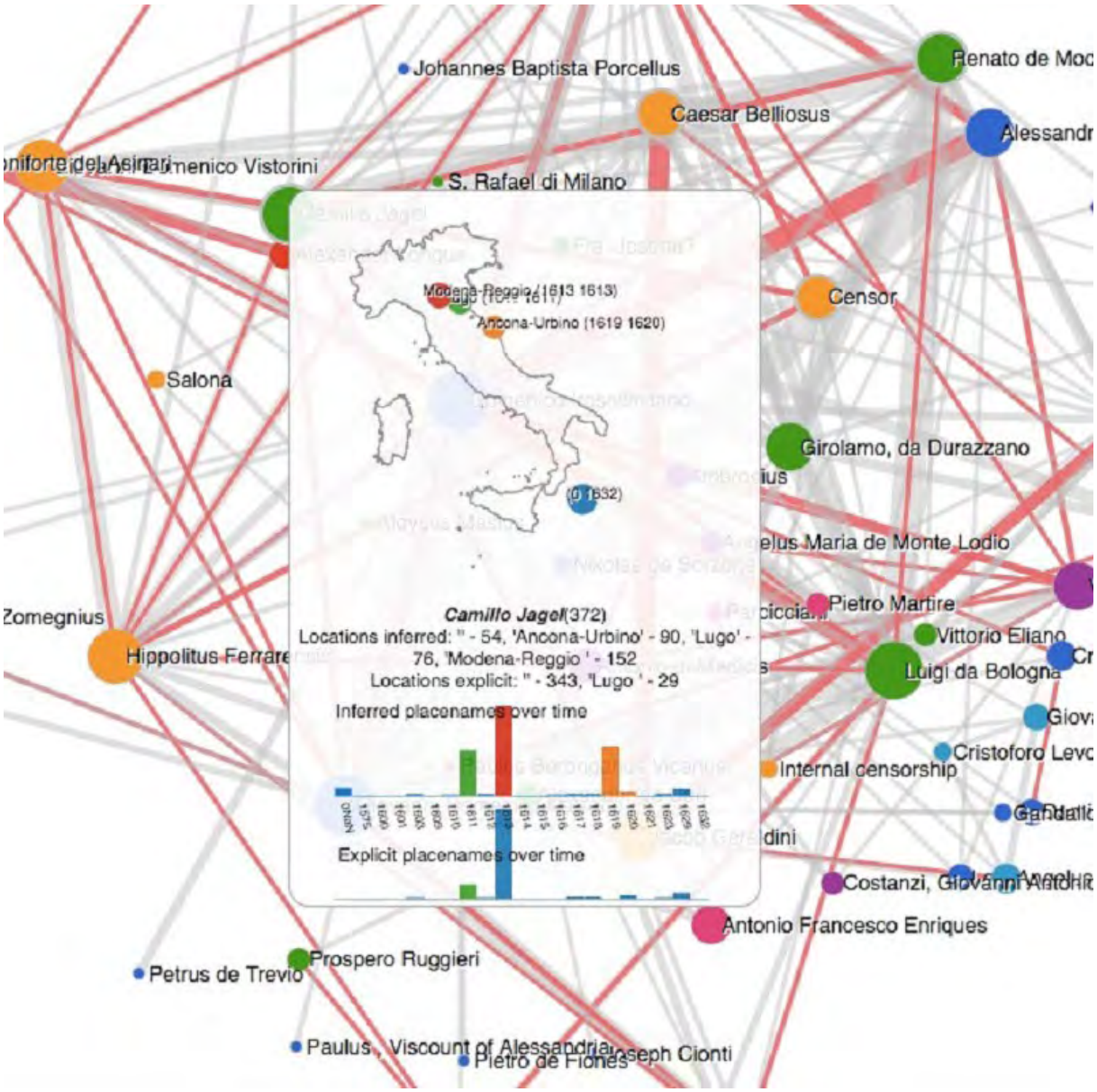


Figure 3: A net*work of censors in Italy with two relationship types – censors who worked on the same manuscript at the same time and censors who worked on the same manuscripts at different time periods (represented by red and grey links, correspondingly). Line thickness represents the number of joint manuscripts for a pair of censors. The censors were divided into seven communities by the Louvain algorithm (Blondel et al., 2008) (represented by different colours of nodes). Clicking on nodes shows time maps of censors.*

References

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, & E. Lefebvre. (2008). Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008(10): P10008, Oct. 2008.

Hric, D., Peixoto, T. P., & Fortunato, S. (2016). Network structure, metadata, and the prediction of missing nodes and annotations. *Physical Review X*, *6*(3), 031038.

Zhitomirsky-Geffet M. & Prebor G. (2016). Towards an ontopedia for historical Hebrew manuscripts. *Frontiers in Digital Humanities, section of Digital Paleography and Book History*, *3*, 3. http://dx.doi.org/10.3389/fdigh.2016.00003.

Preconference Workshops

Jumpstarting Digital Humanities Projects

Amanda French
amandafrench@gwu.edu
George Washington University, United States of America

Anne Chao
annechao@rice.edu
Rice University, United States of America

Marco Robinson
mtrobinson@pvamu.edu
Prairie View A&M University, United States of America

Brian Riedel
riedelbs@rice.edu
Rice University, United States of America

Brief Summary

"Jumpstarting Digital Humanities Projects" is a half-day pre-conference workshop on various aspects of beginning a digital humanities project: scoping and planning a sizable project; determining when to use institutional infrastructure and when to go beyond the institution; winning cooperation from institutional authorities and collaborators; collecting and digitizing materials; and designing for iterative development and efficient feedback loops. Our sessions will focus on the common type of digital humanities project that consists of a assembling a database of source material and generating interactive interpretations such as maps and visualizations from that database. Five scholars from different disciplines and institutions, each a participant in the Mellon-funded Resilient Networks for Inclusive Digital Humanities initiative, will give short tutorials, and workshop attendees will spend an hour on exercises in which they can begin planning a digital humanities project with help from the instructors.

Description of Content

"Jumpstarting Digital Humanities Projects" is a half-day pre-conference workshop on various aspects of beginning a digital humanities project: scoping and planning a sizable project; determining when to use institutional infrastructure and when to go beyond the institution; winning cooperation from institutional authorities and collaborators; collecting and digitizing materials; hiring students and technologists; and designing for iterative development and efficient feedback loops. Our sessions will focus on the common type of digital humanities project that consists of a assembling a database of source material and generating interactive interpretations such as maps and visualizations from that database. Five scholars from different disciplines and institutions, each a participant in the Mellon-funded Resilient Networks for Inclusive Digital Humanities initiative, will give presentations apiece of 30-45 minutes, and workshop attendees will spend an hour on exercises in which they themselves can begin planning their own digital humanities project with individualized help from the instructors. We will end the day with a brief group discussion on how humanities scholars at institutions without digital humanities centers can best form networks and advocate for infrastructure at their own institutions to support digital scholarship.

Scoping and Planning

Workshop leaders will discuss the collaborative and creative processes by which they determine what is achievable in a given project, and how they found the most optimal paths towards achieving their goals. These presentations will not be didactic but exploratory, the "leaders" having at this stage, on average, only begun to execute their workflows. This will provide an ideal space for attendees at various stages in their projects to feel invited to ask questions and contribute to strategies for determining what can be achieved within the specific constraints of budget, time, skills, and archival resources.

Institutional and Extra-Institutional Infrastructure

One of the major decisions projects have to make in their beginning stages is where to host content. Digital humanities projects of the type we are discussing in this workshop require a website, yet many if not most institutions do not provide server space for humanities scholars. Increasingly, libraries will host and manage digital humanities projects, but not all libraries provide this service, and those that have provided it in the past often find that as software and systems age, the cost in labor of maintaining digital humanities projects is a disincentive to provide such services for future projects. Commercial hosts such as GoDaddy and HostGator are one option, and an increasingly well-known option is Reclaim Hosting, founded by instructional technologists by and for educators, but many humanities faculty members are either not aware of these options or do not know how to choose between them. Workshop leaders will discuss their own choices and the relative advantages and disadvantages of each, balancing speed, efficiency, cost, support, sustainability, and longevity.

Feedback loops & iterative design

Collaborative humanities projects depend on the gathering of diverse skills in the pursuit of complex goals. While it is difficult in institutional settings to achieve appropriate parity, this sort of cross-department and cross-strata project work can form alternative modes of collecti-

ve intellectual labor that takes seriously the input of all stakeholders. The appropriate site for this integration of viewpoints in the context of project work is what we call "design." By negotiating over what a thing does and how, a team comes to understand better what it is they are doing in the first place. A project often looks different at the end than it did in the earliest planning stages, and this aspect of the discussion will invite participants to think more creatively about the possibilities of interdisciplinary and inter-departmental collaboration.

Achieving and Maintaining Buy-in

The differences in institutional situations between the different groups represented by collaborating members in an interdisciplinary project necessarily create communicative friction and potential divergences in goals and perceptions. While this on some level represents differences in commitments, the perceived shared goal of any project is what brings collaborators to the table in the first place, and a flexible orientated-ness is what maintains buy-in. Workshop leaders will lead open-ended discussions about experiences in this process.

Collecting and Digitizing Materials

Many digital projects in the humanities begin with non-digital materials, such as the images and documents in the county archives of Waller County, Texas. Projects that include oral histories such as the Houston Asian American Archive now usually capture recordings in born-digital formats, but comprehensive archives of this nature may also need to convert analog audio and video materials from earlier eras. Libraries and archives have a great deal of knowledge about digitization and metadata standards and conversion and migration technologies that can be of use to humanities scholars, so partnering with library and archives professionals early on can be of great benefit. Workshop leaders in this section will discuss their practices with digitizing and collecting materials, especially in partnership with librarians.

Description of Audience

Humanities scholars in the early planning stages of large projects that require a broad array of technical and scholarly competencies. While Digital Humanities is of course a conference for advanced practitioners, we hope in this session both to entice "analog" humanities scholars to commingle with more experienced digital humanities scholars and to encourage experienced digital humanities scholars to think about how best to foster the spread of their methods.

Technical Requirements

This workshop requires a digital projector with audio capabilities, preferably one that can be used with instructor laptops: it requires no special software or hardware. We will expect attendees to bring laptops, and we hope that the workshop room will have sufficient power outlets for attendees.

Length, Format, and Budget

"Jumpstarting Digital Humanities Projects" will be a one-day workshop on the following schedule:

9am-12:30pm: Presentations of 20 to 30 minutes by course instructors
12:30pm-1:30pm: Lunch
1:30pm-3:30pm: Guided exercises in digital humanities project planning
3:30pm-4:15pm: Reflections on the day and discussion of institutional support needs for digital humanities projects

The Resilient Networks for Inclusive Digital Humanities project can fund the registration and travel of instructors. We would prefer a cost of no more than $25 USD for participants, especially since this workshop is meant to appeal chiefly to relative beginners in digital humanities.

Workshop Leaders

Anne Chao
Title: Manager, Houston Asian American Archive
Email: annechao@rice.edu
Phone: 713-202-5599
Address: 3970 Inverness Dr., Houston, TX 77019

Anne Chao is manager of the Houston Asian American Archive at Rice University. She oversees Rice student interns to conduct interviews with Asian Americans in Houston and the greater metropolitan area. Since 2010, HAAA has accumulated over 160 oral history interviews spanning diverse ethnicities from East, to Southeast, and South Asian-Americans. The collection of primary source materials details the contribution of Asian Americans in the building of greater Houston since the Jim Crow era, and provides new insight into the history of the region. Working with the archivist at the Fondren Library, HAAA uses the Omeka platform and includes GIS mapping to plot the life trajectories of the interviewees. The interviews are fully transcribed and time-stamped, synchronized, indexed with key words through the use of the Oral History Metadata Synchronizer (OHMS).

Amanda French
Title: Director, Resilient Networks for Inclusive Digital Humanities

Email: amandafrench@gwu.edu
Phone: 720-530-7515
Address: GWU Libraries, 2130 H Street NW, Washington, DC 20052

Amanda French's particular expertise consists of making humanities content (both cultural content and scholarly interpretation of that content) openly available online, as well as introducing scholars to the various methods of and issues with making humanities content openly available online. She held the CLIR Postdoctoral Research Fellowship at NCSU Libraries from 2004-2006. From 2010-2014, she was first Coordinator and later Principal Investigator for the Mellon-funded initiative THATCamp (The Humanities and Technology Camp), an international unconference that has seen more than 300 events to date attended by more than 7000 people. She often speaks and sometimes writes about open access, the scholarly publication landscape, Omeka, Scalar, Hypothes.is, THATCamp, the Digital Public Library of America, Wikipedia, grant-writing, and alternative careers for humanities PhDs. Her most recent digital research project is a catalog with accompanying exhibits of the personal library of the American poet Edna St. Vincent Millay, available at http://steepletoplibrary.org.

Brian Riedel
Title: Professor in the Practice of Humanities; Associate Director, Center for the Study of Women, Gender, and Sexuality – Rice University
Email: riedelbs@rice.edu
Phone: 713-348-2162
Address: CSWGS, MS-38 | 6100 Main St | Houston, TX | 77005-1892

Brian Riedel received his Ph.D. in Anthropology from Rice University. His research and teaching focus on engaged research and lesbian, gay, bisexual, transgender, and queer social movements, particularly in Greece and the United States. Two of his current projects use GIS to examine the historical connections of place and sexuality. One project examines the histories of the Montrose neighborhood of Houston, Texas, and the uses to which they are put. A core component of that project is a GIS visualization of Houston's LGBT-centered businesses from 1945 to 2015. The other project, conducted in collaboration with the African American Library at the Gregory School (part of Houston Public Library) and Rice Century Scholar Cameron Wallace, documents Houston's formal red-light district known as the "reservation," which operated from 1908 to 1917. Although freed slaves had settled on that land since Emancipation, the city claimed the area held "only a few Negro huts." The project uses GIS and StoryMaps to meld primary resources like census, city directory, and tax record data.

Marco Robinson
Title: Assistant Professor of History, Prairie View A & M University, Prairie View, Texas
Email: mtrobinson@pvamu.edu
Phone: 936-261-3219
Address: Division of Social Work, Behavioral, and Political Sciences, Prairie View A&M University, P.O. Box 519; MS 2203, Prairie View, TX 77446-2203

Marco Robinson is an Assistant Professor of History at Prairie View A & M University, Prairie View, Texas. Marco's research is centered around capturing the social, political, economic, and cultural histories of communities in the American South through collecting, preserving, and analyzing archival and oral history data. As it relates to digital humanities, Dr. Robinson uses this data to tell digital stories, for mapping using GIS and the digitization of historical artifacts. His most recent publication and project are "Telling the Stories of Forgotten Communities: Oral History, Public Memory, and Black Communities in the American South" (Collections: A Journal for Museum and Archives Professionals, Volume 13, Number 2, (Spring 2017): 171- 184.) and Using Interactive Maps and Apps to Preserve Local History: Digitizing the Black Experience in Waller County, Texas.

New Scholars Seminar

Geoffrey Rockwell
geoffrey.rockwell@ualberta.ca
University of Alberta, Canada

Rachel Hendery
r.hendery@westernsydney.edu.au
Western Sydney University, Australia

Juan Steyn
juan.steyn@nwu.ac.za
South African Centre for Digital Language Resources, South Africa

Elise Bohan
elise.bohan@gmail.com
Edith Cowan University, Australia

The New Scholars Symposium has been running since DH2015. It brings together graduate students and recent graduates in a one day "unconference" where they can develop their own research agenda and prepare for the conference. The NSS also includes an opportunity to meet with digital humanities leaders and a mentoring opportunity for the new scholars.

In the last three years centerNet has supported the NSS along with CHCI. The CHCI funding has come to an end, which is why we are applying as a workshop. The Kule Institute for Advanced Study at the University of Al-

berta (CHCI member) has and will provided support for organizing this seminar on behalf of centerNet. Rachel Hendery (Associate Professor of Digital Humanities, Western Sydney University) and Geoffrey Rockwell (Director, Kule Institute for Advanced Study, University of Alberta, Canada) have acted as conveners of the New Scholars Seminar. We propose to build on our experience with this format but add new workshop leaders including Juan Steyne from North-West University, South Africa. CenterNet will also be more directly involved with running of the symposium this year through the assistance of the CenterNet secretary, Elise Bohan.

Target Audience

For the purposes of the Seminar a "new scholar" is defined as someone who is either a graduate student or someone who has received their PhD within the last 5 years (or longer if a case is made for career interruption). Postdoctoral fellows and people in alternative academic positions are welcome to apply.

Participation is by reviewed application and participation is limited to a maximum of 20 people. Typically we support 10 from outside the target continent and 10 from inside, many of whom are students at the hosting university.

Deadline and application process

Applications have usually been due in April. We intend this to be the case this year too, if we the workshop is accepted in time for this to be feasible. Otherwise we will select the earliest feasible deadline. Applications include i) a Statement of Research that outlines their research interests in digital humanities; ii) a letter of support from a centerNet centre/institute director if applicable; and iii) a short two-page CV. Applications are sent to the Kule Institute for Advanced Study <kias@ualberta.ca> at the University of Alberta, a centerNet member. The applications will be reviewed by the following committee:

Geoffrey Rockwell (University of Alberta, Canada)
Rachel Hendery (Western Sydney University, Australia)
Juan Steyn (Northwest University, South Africa)
Elise Bohan (Macquarie University, Australia)
Adam Dombovari (University of Alberta, Canada)

Brief Outline: Intended length and format

The programme for the seminar is developed by the participants once accepted and coordinated by the workshop leaders. The idea is to empower new scholars to develop their own research directions and collaborations. This has previously been very successful, developing a program with a diversity of themes that could not have been anticipated by the workshop leaders. There are therefore typically two phases:

Before the Seminar there is an online gathering component using the University of Alberta eClass (Moodle) platform. Participants share their Statements and discuss what they are interested in discussing together. Clusters of research interests emerge which form the intellectual backbone of the Seminar. We encourage leadership to emerge from within the group so that the actual structure of the on-site days will be primarily organized by the participants.

The on-site portion of the Seminar then takes place in the days before the DH conference. Ideally we would have a day and a half for this, but it could be reduced to one day if necessary. The program that we find works includes three components:

Short presentations by participants of their research and interests followed by a social event the evening before the unconference. This helps break the ice and introduce everyone.

The unconference where we spend an initial hour identifying the key issues/sessions that participants want to organize followed by breakout sessions. The sessions are participant-designed and facilitated. When we reconvene, reporters from the sessions report back to the whole group. This can be structured to fit the time available by increasing or decreasing the number of sessions and running more or fewer of them parallel to each other.

Topics for these small sessions on the unconference day in previous years have included:

- DH pedagogy
- Amplifying diverse voices in DH
- Working with archival materials
- Working with databases
- Quantitative vs qualitative data
- Artificial Intelligence
- Crowdsourcing
- Web scraping
- Creating Twitterbots

One of the sessions from 2016 produced a Manifesto on Student-Driven Research that has since been further developed by the participants and submitted to the *Debates in the Digital Humanities* new series on 'Institutions, Infrastructures at the Interstices'.

c. Mentoring during the DH conference around careers or opportunities in the digital humanities. This last year (2017) we organized mentoring with senior scholars in the field of digital humanities. Before the Seminar participants identified the sort of mentoring they would like and we (Rockwell and Hendery) then contacted people we knew would be at the DH conference and asked them if they were willing to meet for coffee or lunch with a new scholar. The participant and mentor then arranged to meet at their convenience. This was a new feature of the NSS this last year and those that took advantage of it reported that they appreciated the

opportunity. In previous years it has taken the form of e.g. a panel discussion about careers with senior DHers, or small group discussion time with such people. This year we propose to connect the New Scholars with leaders from the centerNet community, both through a networking event sponsored by centerNet, and through one-to-one mentoring opportunities.

Budget

The NSS has secured support from centerNet and SADiLaR. CenterNet will provide catering for breaks and lunch. CenterNet will also provide support for the mentoring component and invite the NSS participants to a networking event with centerNet leaders. SADiLaR has provided assistance with organisation via Juan Steyn and will further provide full support for one participant from South Africa.

In previous years thanks to CHCI funding we have been able to offer participants a significant funding package to assist them to attend. Many of our students and ECRs have said in evaluations they would not have been able to get to the ADHO conferences without this. As we are unable to offer that this year, we would like to find other ways to lessen the burden on these participants. We do not charge any registration fee for this workshop. We also hope that the conference organizers and/or ADHO might provide discounts on registration for our New Scholars. We will also work with participants from outside North America to find travel and conference support for them from other sources where possible.

Special requirements for technical support

We would need space for parallel break-out sessions – usually a total of three spaces is sufficient. A single room can work if it is large enough that small groups can sit in separate corners and hold discussions without disturbing each other too much. Apart from this we only need a projector and a whiteboard.

Getting to Grips with Semantic and Geo-annotation using Recogito 2

Leif Isaksen
l.isaksen@exeter.ac.uk
University of Exeter, United Kingdom

Gimena del Río Riande
gdelrio.riande@gmail.com
CONICET, Argentina

Romina De León
rdeleon@conicet.gov.ar
CONICET, Argentina

Nidia Hernández
nidiahernandez@conicet.gov.ar
CONICET, Argentina

This workshop introduces *Recogito 2*, a tool developed by Pelagios Commons that enables annotation of geographic place references in text, images and data through a user-friendly online platform. Perhaps the most notable feature of Recogito 2 is the ability to produce semantic data without the need to work with formal languages directly, while at the same time allowing the user to export the annotations produced as valid RDF, XML and GeoJSON formats.

The availability of born digital data as well as digitised collections, is changing the way we study and understand the humanities. This amount of information has even greater potential for research when semantic links can be established, and relationships between entities highlighted. The work of Pelagios Commons has shown that connecting historical data according to their common reference to places (expressed via URIs stored in gazetteers) is a particularly powerful approach: information about material culture, archaeological excavations, ancient texts and related scholarship can be connected and cross referenced through the geodata.

Producing semantic annotations usually requires a certain amount of knowledge of digital technologies such as RDF, ontologies and/or text encoding. These techniques can sometimes act as a barrier for users that are not already familiar with Semantic Web theory. The Recogito annotation tool aims to facilitate the creation and publication of Linked Open Data by dramatically reducing some commonly encountered obstacles. First developed in 2014, the community-oriented philosophy behind Pelagios Commons has made users an active agent in shaping its functionality and interface. A dedicated forum on the Pelagios Commons website gathers feedback and suggestions. Recogito code is Open Access and available through GitHub where discussions of Recogito's more technical aspects are held. After a year of intensive redevelopment from the ground up, Recogito 2 was launched in December 2016 and now has almost 1,500 registered users. Introductory documentation is available in English, Spanish, German and Italian with the interface itself being translated into multiple languages in February 2018.

Recogito now supports both additional image standards (such as IIIF) and text standards (TEI export). This allows researchers to use the annotation tool as either a starting or intermediate point for their workflow in the production of semantic annotations that can be then built upon with other technologies. While the initial release already enabled collaboration among users, Recogito 2 features a more refined series of options to manage degrees of collaboration, from private annotations that can only be accessed by their creator, to collaborative and public ones that anyone can see and download. These options offer the opportunity to collaborate, but leaves users free to choose the degree of openness that best suits their materials at different stages of research.

Originally conceived for data related to the ancient world, Recogito 2 has become a valuable tool for annotating many other kinds of historical and modern sources, especially (but not confined to) those containing geographical information. Recogito 2 facilitates the annotation of any named entity. Where applicable, they can be resolved against a number of aligned digital gazetteers, including the ancient world (Pleiades) and modern (Geonames). Although the annotation of geographical information is its most principal focus, Recogito 2 also allows "people" and "event" references to be annotated (currently without semantic resolution), and the opportunity to add tags and comments to disambiguate and refine later searches. Two different colour-coding options makes it easy to identify the different kind of annotations (places, people or events) or different status of the geographic annotations.

This workshop walks participants through all stages of using Recogito 2 to annotate different types of source documents: from uploading a file to the online platform, through annotation, to the download of the annotations in the available data formats. More specifically, the workshop will show practical examples of:

Annotation of sources in text format

Attendees will learn how to benefit from Recogito's automatic recognition of named entities, and how to refine it manually. They will create annotations ex novo, and check or modify those identified by Recogito. The geo-annotations produced on the text can then be plotted on a digital map, through a user-friendly visualisation mode. The relevance of each place is displayed on the map proportionally to the number of annotations that the place has received. Places are linked, via a pop-up window, to all their annotations in the same document, and users are able to browse each annotation in a short, essential context, or to see them in the full text.

Annotation of images and tables

After beginning with text files, attendees will work on the semantic annotations of images. Maps are especially well suited to geo-annotation but Recogito 2 can also be used for the annotation of other types of image, such as photographs or even textual sources in the form of digitised manuscripts. Users will upload images to the Recogito platform and be able to select, transcribe, annotate and, georesolve toponyms within the image. Workshop attendees will also see how Recogito can import and annotate or align tabular (CSV) data such as that derived from spreadsheets, databases or gazetteers.

Exporting data from Recogito

Finally, participants will learn how to export data from Recogito in a variety of formats suitable for visualizing and analysis in other tools, such as spreadsheets, databases or GIS.

To maximise the benefit of the workshop, participants are invited to bring their own data and documents to annotate. Recogito currently has greatest support for ancient and modern sources (including most languages). Materials from other periods can also be annotated but the level and quality of georesolution may vary. The workshop will provide sample texts, imagery and data for attendees without their own datasets. The workshop will show examples of annotations of different kind of sources, and discuss their specific challenges. Throughout the workshop there will be opportunities for participants to discuss how Recogito 2 might be used to support their own research.

Visualising and contextualizing geographical information within documents can be an important step in reaching a deeper understanding of their content, potentially highlighting phenomena that would have been otherwise difficult to identify. It is also an effective tool for engaging students when encountering historical texts and collections. The design of Recogito 2 is intended to make the production of semantic annotations easy and intuitive, opening the door of the Linked Open Data ecosystem to a wide range of users, including without prior experience of semantic technologies.

Semi-automated Alignment of Text Versions with iteal

Stefan Jänicke
stjaenicke@informatik.uni-leipzig.de
Leipzig University, Germany

David Joseph Wrisley
djw12@nyu.edu
New York University Abu Dhabi, United Arab Emirates

Overview

Our half-day tutorial proposed for DH2018 concerns the semi-automatic alignment of different witnesses in complex textual traditions, with demonstration of specific use cases, a discussion of the relevance of the implemented system to particular textual problems relevant to the participants as well as a hands on discovery of the system. Alignment is a relatively simple task for modern languages with orthographic stability and relatively similar texts, but when there is a degree of instability of textual transmission as in oral literatures, popular music or poetry, or other complex texts with partial repetition the task becomes more difficult. Whereas methods of hand aligning and visualizing texts exists in TEI, we focus on the possibility of computational alignment for the purpose of exploratory textual visualization. Scholars who are

interested in visualizing scaled forms of reading will be interested in this tutorial.

Our visual analytics environment *iteal* supports the computational alignment of textual similarities and is not English-specific. It was originally implemented using orally inflected medieval French poetic texts (with test cases of the fabliaux and epic) and so is known to work on texts in Latin alphabets with inconsistent orthography.

This half-day tutorial aims at introducing *iteal* to the DH community for which the questions of multi-text problems, spelling variance and debates about distant forms of reading are currently quite salient. Many language processing and visualization tools do not work well with languages beyond English. Our environment is known to work with languages beyond English will be of interest those interested in expanding innovative techniques in the textual humanities across the North/South divide. Participants of the tutorial will be led in a step-by-step, hands-on approach through the full cycle of an *iteal*-based text alignment workflow, and they will finally have the opportunity of testing the tool with their own data. Although proven to be effectively useful for text variants of medieval poetry, we will not focus only on this type of text as *iteal* can be used to determine alignments among texts of a different kind in any language and in multiple genres. Currently, *iteal* works with plain text in utf8.

iteal consists of two major modules:

First, it automatically determines line-to-line alignments pairwise between all given text editions based on user-configurable parameters including:

- **Edit distance:** Variant spellings are taken into account by this function. We define two words as spelling variants if they have the same first letter, and if the string similarity of the remaining substrings is higher than a user-configurable threshold.
- **Coverage:** In order to ensure that a specific proportion of words of both lines are aligned, the user can configure a minimum coverage value of the line.
- **N-grams:** The user can configure the minimum required n-gram size n that is the largest number of subsequent word matches of both lines.
- **Broken n-grams:** Quite often, the only difference between two lines is a single word in the middle of a line that is either inserted, synonymous, or a transposed stopword. Large n-grams, from this perspective are not achieved. Thus, we allow the user for considering broken n-grams, which is the total number of word matches among both lines.

Second, for the purpose of analyzing the determined alignment we provide interactive visualizations for different text hierarchy levels (examples for all three views can be found in Figures 1, 2 and 3, and a teaser outlining a brief workflow with *iteal* can be found at https://vimeo.com/230829975):

- **Distant Reading:** In order to get a rough overview of alignment patterns throughout the observed text versions, we draw a miniature representation for each version in the form of a vertical bar reflecting its number of verse lines in contrast to the other shown versions. For us, this is the most distant form of reading, where the text itself is not visualized, but rather abstract depictions of textual similarity point to patterns worth discovering.
- **Meso Reading:** Since multiple texts are displayed in synoptic views, the visualization is able to convey more complex patterns of textual relationship. We call this a meso reading that might be said to connect multiple close readings all the while transmitting information that lies beyond the scope of a close reading. Here, we use the intuitivity of stream graphs to connect aligned verse lines among different versions. For a more detailed inspection of an individual alignment, clicking on a stream opens a popup window for line-level close reading.
- **Close Reading:** Next to plain text, the close reading view provides word level alignments for the corresponding verse lines in the form of two Variant Graph visualizations. Within the close reading view, individual alignments can be confirmed with user input, so that it gets persistently stored in the backend.

Target audience: Anyone studying variance in the textual digital humanities and its visualization would be interested in our tutorial. It will be offered in English, but can accommodate data in a variety of languages. Potential participants in the tutorial are encouraged to be in touch with the presenters in advance of DH2018 to provide some sample data that can used to provide a mashup. Required for this step is a version of at least two documents sharing some text in common, of at least 20 lines.

Tutorial Schedule

Part I (1 hour + break time)

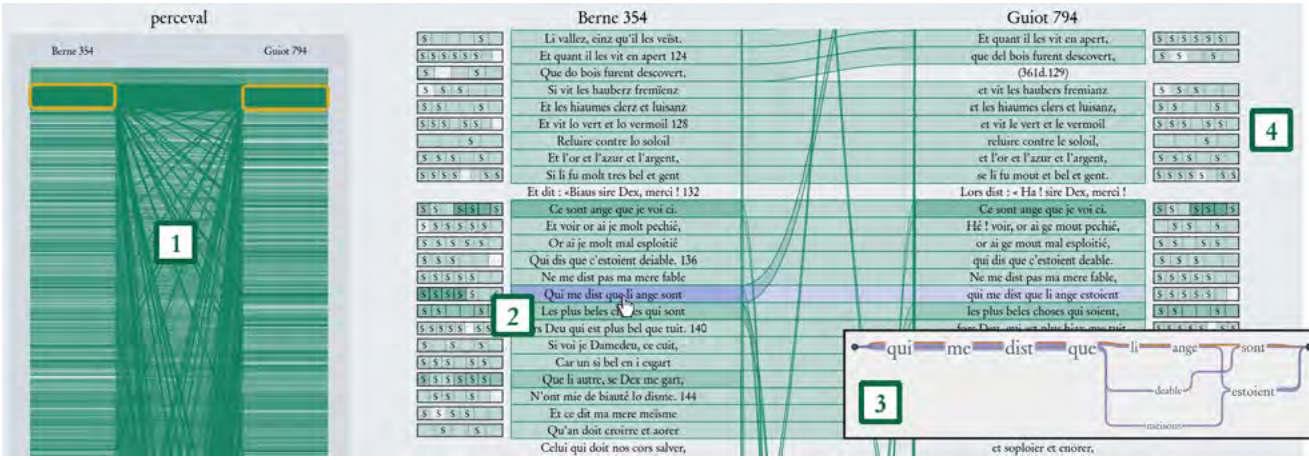
– *iteal* introduction: purpose, functionality, configuration, visualization (Stefan Jänicke)
– Medieval French poetry as an *iteal* use case (David J. Wrisley)
– Further use cases, future work, questions (Stefan Jänicke & David J. Wrisley)

Break

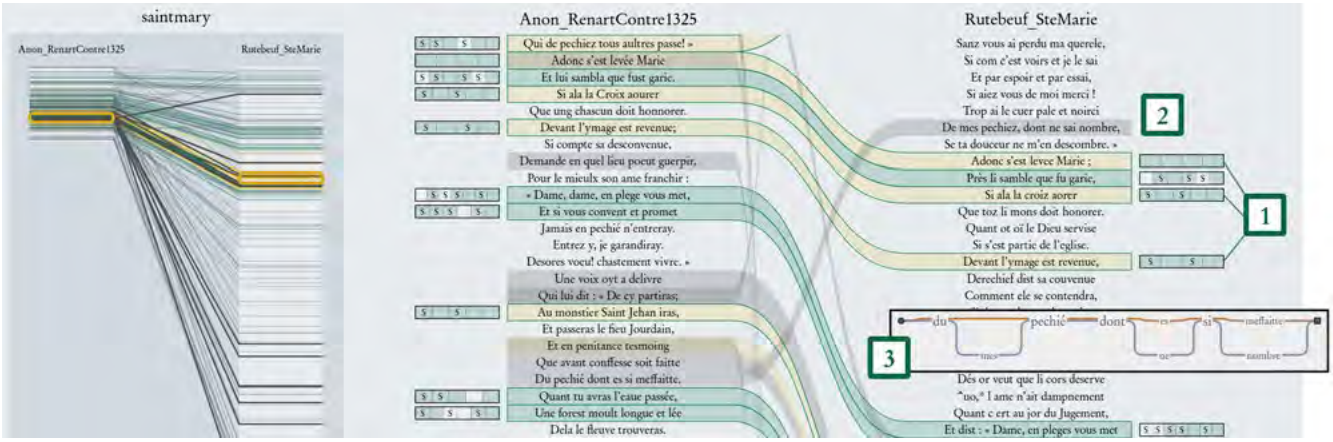
Part II (2 hours - break time)

– Step-by-step hands-on session with texts brought by tutorial participants
– wrap up, feedback and steps forward

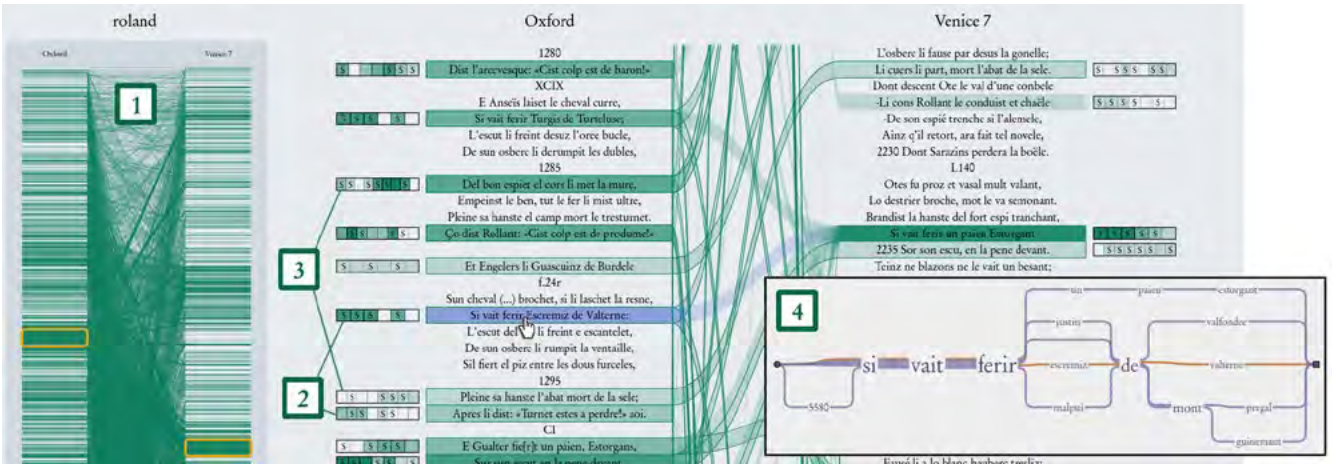
Sample images from iteal



Aligning two editions of Perceval with iteal



Aligning two editions of La vie de saint Marie l'Egyptienne with iteal



Aligning two editions of the Chanson de Roland with iteal

Biographies of the Tutorial Organizers

Stefan Jänicke (stjaenicke@informatik.uni-leipzig.de): Dr. Stefan Jänicke is a post-doctoral researcher at the Image and Signal Processing Group at Leipzig University, Germany, where he leads a text visualization group focusing on applications in the digital humanities. Over the last years, he has gained experience in developing information visualization and visual analytics techniques within a number of digital humanities projects. His PhD thesis investigates the utility of visualization techniques to support the comparative analysis of digital humanities data, and his current research relates to information visualization with a focus on applications for text- and geovisualization in digital humanities. *Homepage:* http://stjaenicke.vizcovery.de

David Joseph Wrisley (djw12@nyu.edu): Dr. David Joseph Wrisley is Associate Professor of Digital Humanities at New York University Abu Dhabi. His research interests include the creation of open, inclusive corpora in medieval studies, corpus-based geovisualization as well as visual exploration of variance in poetic traditions. Furthermore, he is interested in the challenges in humanities data stemming from both multilingual environments and social data creation. *Homepage:* http://djwrisley.com

References

S. Jänicke, A. Geßner, G. Franzini, M. Terras, S. Mahony and G. Scheuermann (2015). TRAViz: A Visualization for Variant Graphs. In: *Digital Scholarship in the Humanities* 30, suppl 1, pp i83–i99.

S. Jänicke, G. Franzini, M. F. Cheema and G. Scheuermann (2015). On Close and Distant Reading in Digital Humanities: A Survey and Future Challenges. In: Eurographics Conference on *Visualization (EuroVis) - STARs*. The Eurographics Association.

S. Jänicke and D. J. Wrisley (2017). Visualizing Mouvance: Towards a Visual Analysis of Variant Medieval Text Traditions. In: *Digital Scholarship in the Humanities* 32, suppl 2, pp ii106–ii123.

S. Jänicke, A. Geßner, M. Büchler and G. Scheuermann (2014). Visualizations for Text Re-use. In: Proceedings of the *5th International Conference on Information Visualization Theory and Applications (VISIGRAPP 2014)*, pp 59–70.

S. Jänicke and D. J. Wrisley (2017). Interactive Visual Alignment of Medieval Text Versions. In: *IEEE Conference on Visual Analytics Science and Technology, IEEE VAST 2017*.

S. Jänicke, A. Geßner, M. Büchler and G. Scheuermann (2014). 5 Design Rules for Visualizing Text Variant Graphs. In: *Conference Abstracts of the Digital Humanities 2014*.

S. Jänicke and D. J. Wrisley (2016). Visualizing Mouvance: Towards an Alignment of Medieval Vernacular Text Traditions. In: *Conference Abstracts of the Digital Humanities 2016*.

Innovations in Digital Humanities Pedagogy: Local, National, and International Training

Diane Katherine Jakacki
diane.jakacki@bucknell.edu
Bucknell University, United States of America

Raymond George Siemens
siemens@uvic.ca
University of Victoria, Canada

Katherine Mary Faull
faull@bucknell.edu
Bucknell University, United States of America

Angelica Huizar
ahuizar@odu.edu
Old Dominion University, United States of America

Esteban Romero-Frías
erf@ugr.es
University of Granada, Spain

Brian Croxall
brian.croxall@byu.edu
Brigham Young University, United States of America

Tanja Wissik
tanja.wissik@oeaw.ac.at
Austrian Academy of Sciences, Austria

Walter Scholger
walter.scholger@uni-graz-at
University of Graz, Austria

Erik Simpson
simpsone@grinnell.edu
Grinnell College, United States of America

Elisabeth Burr
elisabeth.burr@uni-leipzig.de
Universität Leipzig, Germany

Context: as the digital humanities take firm root in the humanities curriculum, institutions around the world are now committing significant resources toward developing DH and integrating it in standalone courses, graduate degrees and undergraduate majors and minors within and across departments. With this commitment comes the realization that such formal implementation of DH and its siblings (e.g. digital social sciences, digital media, etc.) at a degree-granting level requires articulation of core requirements and competencies, identification and hiring of faculty who are capable of teaching DH in a variety of learning environments (coding, systems, application of methods), evaluating a broad spectrum of student work,

and beyond. It also changes the foundational principles of the work of those in our network, as training increasingly involves learning how to teach competencies at the same time as we ourselves develop and maintain them in light of fast-paced advances.

2018 Focus, and Call for Proposals: at the 2017 mini-conference, attendees reached consensus about forming an ADHO Special Interest Group (SIG) dedicated to DH Pedagogy in all its forms. In support of this, for our 2018 mini-conference and meeting, we continue in inviting proposals for lightning talks on all topics relating to digital pedagogy and training -- and especially this year for those that will lead us to substantial discussion about how a SIG could support instructors, students, practitioners, and administrators. Mini-conference talks will take place in the morning, and the afternoon member meeting will be dedicated to work on a collaborative draft of the SIG proposal. In particular, we welcome proposals with a focus on:

- Ways in which individual universities, colleges, and other educational institutions are extending DH in the classroom.
- Implementing DH pedagogical frameworks locally and working across institutions and training institutes to develop and collaborate on materials that can inform ways in which DH offerings and programs are formalized.
- Assessment techniques in DH curriculum. What types of assessment should occur in digital humanities courses? And, significantly, how might these assessment practices challenge existing university or community-based outcomes? We particularly desire talks that include involvement of students who have been assessed.
- DH training in an international context-how do we articulate/coordinate/collaborate across international boundaries? What can we learn from our differences?
- Developing a multilingual lexicon for teaching DH.
- Discussion of pedagogical materials, pre-circulated for critique and consideration. We are particularly interested in the submission of specific syllabi, tutorials, exercises, learning outcomes, assessment and rubrics that attendees might complete during the workgroup portion of the mini-conference.
- Any topics that might further inform our discussion about DH training.

Machine Reading Part II: Advanced Topics in Word Vectors

Eun Seo Jo
eunseo@stanford.edu
Stanford University, History Department, United States of America

Javier de la Rosa Pérez
versae@stanford.edu
Stanford University, Center for Interdisciplinary Digital Research, United States of America

Scott Bailey
scottbailey@stanford.edu
Stanford University, Center for Interdisciplinary Digital Research, United States of America

Fernando Sancho
fsancho@us.es
Dept. of Computational Sciences and Artificial Intelligence, University of Seville, Spain

Description

This half day workshop is an introduction to word vectors and text vectorization broadly. We will focus on building intuition of how word vectors work, incorporating visualization methods, using pre-trained vectors, and exploring applications of word embeddings. We will teach you both the high-level concepts and the practical usages of these widely used analytical tools for text analysis in digital humanities (DH). It is a hands-on workshop with practical activities for the participants starting with a review of word vectors by way of visualization, an overview of downloadable word vectors, and examining the potential pitfalls of using word vectors in humanistic analysis and the methods for mitigating these issues. Given the general applicability of machine learning models in real life, addressing issues concerning biased models, datasets, and algorithms, is of vital importance for correct interpretation of their applications.

We will provide a Python Jupyter Notebook and an accompanying text corpus that we will work through as a group. By the end of the workshop, the participants will have working knowledge of how and where to download or train word embeddings and the caveats of using them.

Relevance to the DH Community

Since the apparition of analytical approaches to distant reading and macro-analysis, popularized by Moretti and Jockers, and the possibility of access to huge amounts of textual data and long-term studies such as Culturomics, new tools were needed to tackle the increasing complexity of large corpora. Borrowing from advances in machine learning and computational linguistics, digital humanists have experimented with various methods of text quantification for interpreting macro contours of culture and language. In particular, word vectors have gained recognition for their versatility in DH studies. Scholars have used word vectors in a variety of tasks such as measuring similarity in word meaning (Caliskan et al., 2017), authorship attribution (Kocher et al., 2017), or dialogism in novels (Muzny et al., 2017).

This workshop is both a theoretical and practical introduction to humanist applications of these methods.

Those interested in large scale text-analysis of any corpora will learn the basics of transforming textual data into numerical form.

Instructors

Eun Seo Jo researches the language of American foreign relations in historical contexts and applications of NLP and ML in history. She is a PhD candidate in history at Stanford University where she is also a member of the Literary Lab and a Digital Humanities Fellow. She has presented at various DH conferences and is a DH methodology consultant at Stanford.

Scott Bailey is a Research Developer in the Center for Interdisciplinary Digital Research in the Stanford University Libraries. He collaborates and consults on research projects across the humanities and social sciences, and teaches workshops on tools and methods in digital scholarship, such as natural language processing. His research ranges from vulnerability in the context of theological anthropology to computational approaches to systematic and historical theological works, such as Karl Barth's *Church Dogmatics*.

Javier de la Rosa is a Research Engineer at the Center for Interdisciplinary Digital Research, a unit at the Stanford University Libraries focused on digital scholarship. He is an active member of the DH scholarly community at Stanford and regularly participates in conferences, professional organizations, and teaches workshops and tutorials to faculty and graduate students. He holds a Post-doctorate research fellowship and a PhD in Hispanic Studies at Western University, Ontario, where he also served as Tech Lead for the CulturePlex Lab. He completed both his MSc. in Artificial Intelligence and BSc. in Computer Engineering at University of Seville, Spain. His work and interests span from cultural network analysis and computer vision, to text mining and authorship attribution in the Spanish Golden Age of literature.

Fernando Sancho is an Associate Professor at the Dept. of Computational Sciences and Artificial Intelligence at the University of Seville, and holds a PhD by the same university. He has worked in topics ranging from complex systems, and data analysis to cultural objects studies. He has regularly collaborated with the CulturePlex Lab at the University of Western Ontario, and the Complex Systems Modeling Group at University of Central Ecuador.

Target Audience and Prereqs

Post-docs, faculty, and advanced graduate students with Python prerequisites. Although the main concepts will be overviewed, knowledge of basic word embeddings and word2vec specifically would be desirable. In order to participate fully in all activities, participants must have working knowledge of basic programming concepts, the Python language, data structures, and the Numpy library.

- Technical Support: Microphones and Projector
- Proposed Length: Half-day (4 hours; 4 sessions)
- Medium: Notebook (Jupyter)
- Libraries: Numpy, Pandas, Textacy, SpaCy, Gensim, scikit-learn, matplotlib

Workshop Outline

The workshop is split into four 50 min sessions with 10 minutes breaks in-between. We teach several methods in each unit with increasing difficulty. The schedule is broken down below:

Understanding Word Vectors with Visualization

This unit will give a brief introduction of word vectors and word embeddings. Concepts needed to understand the internal mechanics of how they work will also be explained, with the help of plots and visualizations that are commonly used when working with them.

- 0:00 - 0:20 From word counts to ML-derived Word Vectors (SVD, PMI, etc.)
- 0:20 - 0:35 Clustering, Vector Math, Vector Space Theory (Euclidean Distance, etc.)
- 0:35 - 0:50 [Activity 1] Visualizations (Clustering, PCA, t-SNE) [We provide vectors]

Word Vectors via Word2Vec

This unit will focus on Word2Vec as an example of neural net-based approaches of vector encodings, starting with a conceptual overview of the algorithm itself and end with an activity to train participants' own vectors.

- 0:00 - 0:15 Conceptual explanation of Word2Vec
- 0:15 - 0:30 Word2Vec Visualization and Vectorial Features and Math
- 0:30 - 0:50 [Activity 2] Word2Vec Construction [using Gensim] and Visualization(from part 1) [We provide corpus]
-

Extended Vector Algorithms and Pre-trained Models

This unit will explore the various flavors of word embeddings specifically tailored to sentences, word meaning, paragraph, or entire documents. We will give an overview of pre-trained embeddings including where they can be found and how to use them.

- 0:00 - 0:20 Overview of other 2Vecs & other vector engineering: Paragraph2Vec, Sense2Vec, Doc2Vec, etc.
- 0:20 - 0:35 Pre-trained word embeddings (where to find them, which are good, configurations, trained corpus, etc.)
- 0:35 - 0:50 [Activity 3] Choose, download, and use a pre-trained model

Role of Bias in Word Embeddings

In this unit, we will explore an application and caveat of using word embeddings -- cultural bias. Presenting methods and results from recent articles, we will show how word embeddings can carry historical bias of the corpora trained on and lead an activity that shows these human-biases on vectors and how they can be mitigated.

- 0:00 - 0:10 Algorithmic bias vs human bias
- 0:10 - 0:40 [Activity 4] Identifying bias in corpora (occupations, gender, ...) [GloVe] (Caliskan et al., 2017)
- 0:40 - 0:50 Towards unbiased embeddings; Examine "debiased" embeddings
- 0:50 - 0:60 Conclusion remarks and debate

References

Caliskan, A., Bryson, J.J., Narayanan, A., 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 183–186. https://doi.org/10.1126/science.aal4230

Kocher, M., Savoy, J., 2017. Distributed language representation for authorship attribution. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx046

Nanni, F., Dietz, L., Ponzetto, S.P., 2017. Toward a computational history of universities: Evaluating text mining methods for interdisciplinarity detection from PhD dissertation abstracts. *Digital Scholarship in the Humanities*. https://doi.org/10.1093/llc/fqx062

Interactions: Platforms for Working with Linked Data

Susan Brown
sbrown@uoguelph.ca
University of Guelph, Canada

Kim Martin
kimberleymartin@gmail.com
University of Guelph, Canada

Following on from a successful LOD workshop in Montreal that saw 30 plus people come together and discuss the potential for linked data in the humanities, we propose a workshop that focuses more specifically on interacting with Linked Data. There are many different platforms for working with linked data — for visualizing, creating, reconciling, cleaning, and analyzing it. Some of these tools have been developed from within the Digital Humanities community, and others have been developed beyond it but adapted to our purposes. We hope to create the opportunity for fruitful exchange by providing time for hands-on demonstration and discussion.

All participants will have the opportunity to submit the following in advance of the workshop:

1. Answering an online form that indicates the type of LOD tools or platforms, or features within these, that they wish to see discussed at the workshop.
2. Details on where their tool (if they have one) fits in, and a description of their work with LOD.
3. A description (1 page max) of their LOD platform with features that they wish to showcase during the workshop.

Outline/Schedule (Based on 6 hrs – Full Day)

Introductions	20 mins
Featured Tool Demos x4	30 mins
Coffee break + discussion	15 mins
Lightning Tool Demos x12	60 mins
Poster session and lunch	100 mins
(Participants wanting to do an afternoon workshop could shift to that at this point)	
Discussion of challenges, desiderata, next steps etc.	60 mins
Coffee break	15 mins
Time for breakout discussions (re: possible collaborations etc.)	60 mins

Workshop leaders:

Susan Brown (sbrown@uoguelph.ca) is a Canada Research Chair in Collaborative Digital Scholarship and Professor of English at the University of Guelph, and Visiting Professor at the University of Alberta. She researches Victorian literature, women's writing, and digital humanities. All of these interests inform *Orlando: Women's Writing in the British Isles from the Beginnings to the Present*, an ongoing experiment in digital literary history published by Cambridge UP since 2006 that she co-directs. She directs the Canadian Writing Research Collaboratory, an online repository and research environment for literary studies in and about. Her current research touches on a range of topics in the digital humanities including interface design and usability, visualization and data mining, semantic technologies, and humanist-centered tool development. She is increasingly engaged with inquiry into how linked open data can serve humanities research. She also works on the impact of new technologies in the literature of the Victorian period. Brown is President of the Canadian Society for Digital Humanities/Société canadienne des humanités numériques.

Kim Martin is the Ridley Post-doc in Digital Humanities and the Associate Director of THINC Lab at the University of Guelph. Her PhD thesis focussed on the serendipitous experiences of historians during their research

process. She is currently working on linked data projects for Canadian Writing Research Collaboratory (CWRC) and on the Mellon-funded Records of Early English Drama - London (REED-London) project. Kim's interests in serendipity and linked data tie neatly together with her work on HuViz — the humanities visualizer, a tool developed by the CWRC team for interacting with linked open datasets.

Target Audience/Expected number of participants: 40
Special requirements for technical support:
Projector and screen. Ideally also boards of some kind (black, white, paper) on which to write, though we can bring something if need be. Decent internet access.
We will need one room that is big enough for 40 people.
Proposed budget:
The organizers will provide funds for lunch. Estimated cost: $300-400.

Call for Participation: We will put out a call for participation based upon this proposal within two weeks of confirmation that the workshop will run.

Deadline for submissions: We will ask for submissions by April 30th. We will make the papers available through a Dynamic Table of Contexts edition that expands the one from last year to other prospective participants and mount the ranking poll by May 5th and run the poll until May 31st. Applicants will be informed if they will be presenting and/or participating in the workshop by June 1st.

Program Committee:

Susan Brown, Professor of English
University of Guelph

Sharon Farnel, Metadata Coordinator
University of Alberta Libraries

Lisa Goddard, Academic Systems Librarian
University of Victoria

Karl Grossner, Geographer, DH Researcher
University of Pittsburgh

Abigel Lemak, PhD Student, English
University of Guelph

Kim Martin, Postdoc in DH
University of Guelph

Deb Stacey, Professor, Computer Science
University of Guelph

Building International Bridges Through Digital Scholarship: The Trans-Atlantic Platform Digging Into Data Challenge Experience

Elizabeth Tran
etran@neh.gov
National Endowment for the Humanities, United States of America

Crystal Sissons
crystal.sissons@sshrc-crsh.gc.ca
Social Sciences and Humanities Research Council, Canada

Nicolas Parker
nicolas.parker@sshrc-crsh.gc.ca
Social Sciences and Humanities Research Council, Canada

Mika Oehling
mika.oehling@sshrc-crsh.gc.ca
Social Sciences and Humanities Research Council, Canada

This workshop will focus on how international partnership can benefit large-scale research projects in digital scholarship. During the workshop, participants will learn about the Digging into Data Challenge 4, an initiative of the Trans-Atlantic Platform (T-AP) for Social Sciences and Humanities, a network of public funders representing countries in Europe, North America, and South America. The Digging into Data Challenge invited international teams to undertake multidisciplinary projects that use techniques of large-scale data analysis and demonstrate how these can lead to new insights. The Digging into Data Challenge has had four rounds of funding, and offers an valuable opportunity to (1) see how the international dimension benefits the scholarship; (2) understand the challenges of working internationally on big data projects addressing questions in the humanities and social sciences; (3) understand how international funding initiatives might enable research in ways that domestic funding cannot.

This workshop is targeted at (1) individuals who are interested in "scaling up" their research efforts to include an international dimension and (2) funders who are interested in launching or joining international funding opportunities. The workshop will touch on various themes that impact digital researchers and international collaboration, including:

- legal considerations,
- the intellectual challenges for large scale research,
- big data skills,
- funding policies and processes, and
- the challenges to international research collaboration for researchers from both small and large countries.

The workshop is scheduled as a full-day event so as to allow ample time for conversation and networking.

In order to better incorporate and interests of workshop participants and foster dialog and discussion, participants may provide a brief one-page synopsis outlining their interest in international collaboration and what they hope to gain from the workshop. The synopsis should be sent to: odh@neh.gov.

Herramientas para los usuarios: colecciones y anotaciones digitales

Amelia Sanz
amsanz@filol.ucm.es
Complutense University, Spain

Alckmar Dos Santos
alckmar@gmail.com
Federal University of Santa Catarina, Brazil

Ana Fernández-Pampillón
apampi@ucm.es
Complutense University, Spain

Oscar García-Rama
ogarcia@supportfactory.net
Support Factory, Spain

Joaquin Gayoso
jgayoso@ucm.es
Complutense University, Spain

María Goicoechea
mgoico@filol.ucm.es
Complutense University, Spain

Dolores Romero
dromero@filol.ucm.es
Complutense University, Spain

José Luis Sierra
jlsierra@ucm.es
Complutense University, Spain

Desde el año 2010, el Grupo de investigación LEETHI (Literaturas Europeas del tExto al Hipermedia) desde la Facultad de Filología y el grupo ILSA (Implementation of Language-Driven Software and Applications) desde la Facultad de Informática de la Universidad Complutense de Madrid trabajan juntos afin de dar respuesta concreta a necesidades de la docencia y la investigación en Humanidades en la UCM. Han diseñado y probado sistemas y herramientas que permiten tanto la construcción de repertorios digitales para colecciones propias como el comentario detallado y a la edición creativa; han desarrollado repertorios como las bibliotecas Mnemosine con el grupo LOEP o Ciberia y herramientas de anotación como @Note que mereció uno de los 12 premios otorgados por Google en 2010 dentro de su *Google's Digital Humanities Award Program*. LEETHI e ILSA han colaborado en proyectos de carácter europeo como la COST Action INTEREDITION o el grupo de trabajo en DARIAH, Women Writers in History; han participado con "short papers" en las DH Conferences en Hambourg, Lausanne y Montréal.

Ambos grupos colaboran desde el año 2006 con el grupo NUPILL (Nucleo de Pesquisas con Informatica, Literatura i Lingüística) de la Universidad Federal de Santa Catarina (Brasil). El grupo tiene como vocación la exploración de las posibilidades que las tecnologías ofrecían al desarrollo de la investigación sobrela lectura y la escritura literaria en el medio electrónico. Así ha desarrollado la Biblioteca de Literaturas de Lingua Portuguesa, revistas como Texto Digital, ediciones como la de Machado de Asís. Estos grupos han organizado seminarios conjuntos en Florianópolis y en Madrid, y han participado en congresos en Brasil, Francia y España.

Support Factory es una empresa dedicada a la investigación y creación de software seguro y estable para entornos educativos a partir de un acompañamiento del usuario durante todo el proceso de configuración, prueba y utilización. Colabora con el Grupo LEETHI en el diseño de software para la edición electrónica amigable desde 2017.

Todos comparten una búsqueda de:

- estrategias de desarrollo de utilidades, sistemas y competencias de abajo arriba: desde las necesidades de los usuarios expertos y para los usuarios en la docencia y la investigación en Humanidades;
- cooperación horizontal que permita mutualizar y mancomunar tecnologías digitales en español y en portugués;
- desarrollos que generen masa crítica científica, pero también innovación social y beneficio comerical;
- respuestas a necesidades de investigación y de enseñanza localizadas: una solución tecnológica global no funciona necesariamente en todos los espacios, cuando sus modelos no corresponden con prácticas culturales y soberanía epistemológica de los lugares.

Los grupos están en condiciones de ser actantes en el campo de las Humanidades Digitales globales y de visibilizar sus prácticas, con el fin de contrastarlas, ponerlas al servicio de la comunidad investigadora que habla español y portugués y pluralizar así sus funcionalidades.

Propopen presentar las siguientes herramientas:

- CLAVY (http://clavy.fdi.ucm.es), desarrollada por ILSA, permite la administración de colecciones digitales heterogéneas, para:
  - agregar colecciones externas gracias a una potente arquitectura de importación;
  - integrar y unificar estas colecciones a través de una representación explícita de sus estructuras

mediante ontologías reconfigurables en un entorno amigable para expertos en Humanidades;
- transformar la arquitectura del plug-in a un nivel más profundo con la intervención de programadores;
- exportar las colecciones a plataformas externas gracias a una arquitectura de exportación para ello.

La plataforma también permite gestionar un sistema extensible para alimentar y refinar las colecciones, de forma que la arquitectura para integrar los plug-ins específicos de edición permita adaptar el flujo de autor a las necesidades específicas de cada campo. Así resulta posible integrar de manera sencilla, por ejemplo, la geolocalización de los objetos o la anotación de recursos.

La plataforma ha sido ya utilizada en Mnemosine y Ciberia. Está especialmente destinada a investigadores-profesores que quieren diseñar su propio repertorio y adaptarlo a sus propias necesidades según avanza su construcción.

- *@Note* (http://anote.fdi.ucm.es), desarrollada por ILSA/LEETHI, permite la anotación de textos digitalizados. Es el resultado del *Google's Digital Humanities Award Program* 2010. La herramienta permite a los profesores-investigadores crear actividades de anotación de textos en modo imagen apoyándose en ontologías con las que clasificar las anotaciones de forma colaborativa, tanto para definir las ontologías, como para realizar las anotaciones. La herramienta permite incluir notas multimedia e hilos de discusión asociados a cada anotación, así como la utilización de esas anotaciones para escribir ensayos, comentarios o realizar análisis. Está especialmente diseñada para el análisis de textos en la enseñanza universitaria y secundaria, al alcance de cualquier profesor con sus estudiantes.
- DLNOTES2 (http://www.dlnotese2.ufsc.br), creada por NUPILL, permite hacer anotaciones libres y semánticas en obras digitalizadas y en modo texto. Se proporcione a los alumnos la posibilidad de 1) hacer comentarios en cualquier extracto de un texto; 2) gestionar las anotaciones para análisis de la lectura (por parte del profesor o incluso del estudiante) o para elaboración de revisiones de la obra leída. En cuanto a las anotaciones semánticas („semántica" en el sentido que le atribuyen las Ciencias de la Computación), el propósito es asociar conceptos de la teoría literaria a cualquier secuencia de caracteres de la obra, lo que ayuda a los estudiantes a comprender esos conceptos y a desarrollar lecturas en perspectivas diferentes (y complementarias) de la lectura tradicional.
- AOIDOS(http://aoidos.ufsc.br), desarrollada por NUPILL, realiza escansiones automáticas en corpora textuales poéticos, señalando todos los fenómenos fonéticos necesarios a la realización rítmica de los versos, y posibilitando, además, que los lectores tengan acceso a todos los datos cuantificados y organizados de varias maneras. Su funcionamiento en portugués está ya demostrado y se mostrará una versión en español para este taller.
- CONTENT-AWAY: es una herramienta que permite a alumnos y profesores crear sus libros a medida, tanto físicos como digitales, mejorar el flujo de información entre profesores y alumnos, facilitar el acceso a materiales de estudio en differentes formatos y soportes. Partiremos del desarrollo de Enclave realizado para la Real Academia Española y mostraremos a los participantes como adaptar y comenzar a utilizar la herramienta para sus propias necesidades docentes.

El taller persigue dos objetivos:

- los usuarios podrán calibrar la posibilidad de elaborar su propio repertorio digital con sus propias categorías, su propio sistema de anotación de textos;
- intercambiaremos modelos y propuestas con otros investigadores y expertos de forma que sea posible compartir y enriquecer sistemas y herramientas.

Descripción de la audiencia:
- Por un lado, profesores-investigadores sobre literaturas en soporte electrónico que quieran conocer las funcionalidades de estas herramientas y adaptarlas a su propio uso; por otro, programadores que quieran conocer las claves de las herramientas para adaptarlas a su entorno.
- Podemos asumir hasta 60 participantes solo para presentaciones.

Necesidades para la celebración:
- Se requerirá una sala con ordenadores o, en su defecto, se requerirá a los participantes que aporten sus propios equipos. En cualquier caso, será necesaria una buena conexión a Internet.
- Todas las sesiones se desarrollarán en portugués y en español, según el equipo que presente cada herramienta.

References

Mittmana, A.; Samanta R. M. ; dos Santos A. L. Análise comparativa entre as escansões manual e automática dos versos de Gregório de Matos A comparative analysis between automatic and manual scansions of Gregório de Matos' verses. Texto Digital, Florianópolis, v. 1, n. 1, p. 157-179, jan./jun. 2017.

Gayoso-Cabada, J., Rodríguez-Cerezo D., Sierra, J.L. Browsing Digital Collections with Reconfigurable Faceted Thesauri. En J. Gołuchowski, M. Pańkowska, H. Linger, C. Barry, M. Lang & C. Schneider (Eds.), *Information Systems Development: Complexity in Information Systems Development (ISD2016 Selected and Extended Papers)*. Lecture Notes in Information Systems and Organisation Vol. 22, pp. 69-86.

Gayoso-Cabada, J., Rodríguez-Cerezo D., Sierra, J.L. Multilevel Browsing of Folksonomy-Based Digital

Collections. In Wojciech Cellary, Mohamed F. Mokbel, Jianmin Wang, Hua Wang, Rui Zhou, Yanchun Zhang (eds): *Web Information Systems Engineering - WISE 2016 - 17th International Conference, Shanghai, China, November 8-10, 2016, Proceedings, Part II. Lecture Notes in Computer Science 10042,* pg. 43-51. 2016.

Where is the Open in DH?

Wouter Schallier
wouter.schallier@un.org
UN/ECLAC, Chile

Gimena del Rio Riande
gdelrio@conicet.gov.ar
CONICET, Argentina

April M. Hathcock
april.hathcock@nyu.edu
New York University, United States of America

Daniel O'Donnell
daniel.odonnell@uleth.ca
University of Lethbridge, Canada

When it comes to promoting the importance of open scholarship, Latin America and the Caribbean stand out in a sense that the concept of "openness" is generally accepted all over the region. Several countries, such as Peru, Argentina, Brazil and Mexico, have shown real advances in terms of national laws that seek to make knowledge produced with public funds a common good, managed by the academic community. We can also highlight regional projects such as SciELO and redalyc.org that have played a unique role to make the production published in Ibero American and Latin American journals available free of charge. Open access is now established in Latin America and the Caribbean as the most extended communication model in the academic community, giving visibility and value to scientific production at a regional and global level.

Nevertheless, the question remains to what extent this wide acceptance of openness has influenced the work of digital humanists in Latin America and the Caribbean and beyond. Much of the most well-known digital humanities (DH) work in the world tends to focus on projects coming out of North America and Western Europe. And despite efforts by groups such as Global Outlook::Digital Humanities (GO::DH) and the Alliance of Digital Humanities Organisations more broadly, DH still remains a very English language centric interdiscipline (Fiormonte y del Rio Riande, 2017).

What would it take to bring DH into a more global openness, not only in terms of access but also in terms of methods, best practices and opportunities for collaboration? And what could this openness look like set against the backdrop of the long-standing and highly developed open access movement in Latin America and the Caribbean?

The workshop will analyse these challenges, as well as highlight initiatives and explore options to advance open in DH in Latin America and the Caribbean. It will begin by examining the aforementioned national laws and specific cases that illustrate the progress and challenges of open access as a movement in Latin America and the Caribbean, as well as in the global context and present a practical approach to deal with the „different open accesses in the world" (Curry, 2017; Babini, 2013). The workshop will then shift to focus on the ways these various infrastructures for open can be deployed to build a more globally open DH.

Furthermore, the workshop will highlight particular existing DH projects that have begun building openness, in access, methods, and collaboration. Instructors and facilitators will help attendees to explore examples from the Global North and South, such as the LEARN project (http://www.learn-rdm.eu/), CLACSO's activities (https://clacso.org.ar/), Red Argentina de Educación Abierta (AREA. http://a-rea.org/), Cientópolis (https://www.cientopolis.org/), Acta Académica (https://www.aacademica.org/), Humanities Commons (https://hcommons.org) , OpenCon (http://www.opencon2017.org/), FORCE11 (https://www.force11.org/), DARIAH (https://www.dariah.eu/), among others, to begin building a set of good practices, including examples of institutional policies and practical recommendations from Europe and Latin America and the Caribbean devoted specifically to DH projects. We will give examples of Open projects in DH, in this set of good practices, institutional policies and practical recommendations that will address project work, digital objects, Open Access publishing and research collaboration.

Finally, the workshop will place DH output modes, from collaborative web projects to traditional publications to research data, in the context of the larger open access movement, which is changing the face of academic research and society in a very profound way. This vision of open access is creating a global environment where researchers, innovators, and citizens can publish, find, use and reuse each other's data, tools, publications and other outputs for research, innovation and educational purposes.

In addition to people interested specifically in the case of Latin America and the Caribbean, this course will be of comparative interest to people working in other regions in both the Global South and the Global North. We will encourage participants to engage reflectively with the material, bringing the own experiences to bear.

References

Arévalo, A. J. (2016). Análisis de los estudios sobre las ventajas del acceso abierto y la ventaja de cita. *Blog de la biblioteca de Traducción y Documentación de la Universidad de Salamanca.* https://universoabierto.org/2016/05/29/analisis-de-los-estudios-sobre-las-ventajas-del-acceso-abierto-y-la-venta-

ja-de-cita/ (last visit: 27 April 2018)
Alperin, J.P. (2015). The Public Impact of Latin American's Approach to Open Access. https://stacks.stanford.edu/file/druid:jr256tk1194/AlperinDissertationFinalPublicImpact-augmented.pdf (last visit: 27 April 2018)
Babini, D. (2013). Open access initiatives in the Global South affirm the lasting value of a shared scholarly communications system. *London School of Economics and Political Science Impact Blog* http://blogs.lse.ac.uk/impactofsocialsciences/2013/10/23/global-south-open-access-initiatives/ (last visit: 27 April 2018)
Curry, S. (2017). Why I don't share Elsevier's vision of the transition to open access http://occamstypewriter.org/scurry/2017/10/03/why-i-dont-share-elseviers-vision-of-the-transition-to-open-access/ (last visit: 27 April 2018)
Fernándes, P. and Vos, R. A. (2017). Open Science, Open Data, Open Source. https://pfern.github.io/OSOD-OS/gitbook/ (last visit: 27 April 2018)
Fiormonte, D. and del Rio Riande, G. (2017). Por unas Humanidades Digitales Globales. https://infolet.it/2017/10/09/humanidades-digitales-globales/ (last visit: 27 April 2018)
Packer, A. L. et al. (2018). Los criterios de Indexación de SciELO se alinean con la comunicación en la ciencia abierta. *SciELO en Perspectiva.* http://blog.scielo.org/es/2018/01/10/los-criterios-de-indexacion-de-scielo-se-alinean-con-la-comunicacion-en-la-ciencia-abierta/#.WocMbujwaM9 (last visit: 27 April 2018)
Suárez, A. V. and McGlynn, T. (2017). The fallacy of Open-Access Publication. *The Chronicle of Higher Education.* https://www.chronicle.com/article/the-fallacy-of-open-access/241786

Indexing Multilingual Content with the Oral History Metadata Synchronizer (OHMS)

Teague Schneiter
tschneiter@oscars.org
Academy of Motion Picture Arts & Sciences, United States of America

Brendan Coates
bcoates@oscars.org
Academy of Motion Picture Arts & Sciences, United States of America

Brief Description:

Are you in need of a way to provide access to oral histories not recorded in English? Do you have dreams of creating multilingual metadata for interviews recorded in English to non-English speaking communities/users? In late 2016 The University of Kentucky Nunn Center updated the OHMS application and viewer to have multilingual functionalities, creating the capability to synchronize both a transcript/translation, as well as to create a bilingual index, making all of these searchable and synchronized to the corresponding moment in the audio or video. In this mini-workshop OHMS power users Teague Schneiter and Brendan Coates will demonstrate the multilingual functionalities of OHMS. Through demonstration of a bilingual use case, instructors will walk attendees through each step of the indexing process to prepare a sample Spanish-English index. Instructors will also guide attendees to develop workflows to support multilingual indexing.

Workshop outline:

1. OHMS intro
    a. History
    b. Development
    c. Multilingual Functionality

OHMS basics - in conjunction with a worksheet sent prior to conference setting up an account
    a. Linking a video
    b. adding thesaurus/ ontology/ data dictionary
    c. End user functionality, switching languages, etc.

Basic indexing
    a. Instructors lead group in indexing a video
    b. in small groups or pairs, index a video

Multilingual indexing
    a. Instructors lead group in creating a multilingual index of a video

*Note - participants do not need to be bilingual, videos can be indexed as Index1 - Index2 instead of English - Non-English

Instructors

Teague Schneiter is an Audiovisual Archivist, Project Manager and Strategist who is currently working for the Academy of Motion Picture Arts and Sciences as Senior Manager (and founder) of the Oral History Projects department, instituting digital content initiatives around filmmaker oral histories. Her professional experience spans moving image preservation and access infrastructure, with the bulk of her experience in human rights and cultural heritage content. Her work at WITNESS solidified her interest in web knowledge management projects and in people organizing, and concretized a firm belief that communication technologies in the digital age should facilitate openness, innovation, participation among individuals and communities, and should further social change. In the past she has worked as a long-term consultant for indigenous media organization IsumaTV, focused mostly on outreach, strategic planning, knowledge-sharing and social media.

Brendan Coates is an Audiovisual Archivist and Preservationist, currently working at the Academy of Motion Pic-

ture Arts and Sciences, where he oversees the ingest, description, preservation, and dissemination of the Academy's Oral History holdings. Prior to this, he ran the UCSB Library's audiovisual digitization and preservation program, including its Cylinder Audio Archive, and its participation in the Library of Congress National Jukebox project and the Discography of American Historical Recordings (DAHR). His research interests are grounded in workflow automation and quality control, ensuring that video is digitized to appropriate standards and is playable and accessible long into the future.

Target Audience:

Our target audience is anybody working with video assets who would like to make subject/ language/ community specific, time-based metadata to describe them. We're anticipating about 20 people.

Technological support:

The workshop will require a computer with projection, WiFi and participants will need their own workstations or to bring their own laptops.

Sig Endorsed

Distant Viewing with Deep Learning: An Introduction to Analyzing Large Corpora of Images

Taylor Baillie Arnold
tarnold2@richmond.edu
University of Richmond, United States of America

Lauren Craig Tilton
ltilton@richmond.edu
University of Richmond, United States of America

Short Description

This tutorial provides a hands-on introduction to the use of deep learning techniques in the study of large image corpora. The TensorFlow and Keras libraries within the Python programming language are used to facilitate this analysis. No prior programming experience is required.

Image analysis tasks covered in the tutorial include object detection, facial recognition, image similarity, and image clustering. We will make three open-access image corpora (historic photographs, still frames from moving images, and scanned works of art) available in order to test these methods. Alternatively, participants may bring and use an image dataset of interest to them. At the conclusion of the tutorial, participants will have created an interactive website running locally on their machines. This website will provide tools for analyzing their selected dataset. Additional instructions for making the website publicly available will be provided.

Audience and Number of Participants

This tutorial is aimed at scholars who work with visual materials who want to integrate DH methods into their analysis of image corpora. Our tutorial is based off of lectures notes used in a non-major, undergraduate-level course at the University of Richmond. It is accessible to participants with little to no programming background. However, as the tutorial will focus on the methods behind image processing rather than low-level coding, it will also be interesting and useful for experienced programmers new to image processing.

Following the large number of participants at the AVinDH SIG sponsored Workshop in Montreal for DH20167 and our popular tutorial at DH2016 in Krakow, we expect the workshop participation to be equally popular with somewhere between 15 and 25 participants.

Presenter Information

Taylor Arnold is Assistant Professor of Statistics at the University of Richmond. A recipient of grants from the NEH and ACLS, Arnold's research focuses on computational statistics, text analysis, image processing, and applications within the humanities. His first book *Humanities Data in R* (Springer, 2015) explores four core analytical areas applicable to data analysis in the humanities: networks, text, geospatial data, and images. His second book, the forthcoming *A Computational Approach to Statistical Learning* (CRC Press 2018), explores connections between modern machine learning techniques with theories of statistical estimation. Numerous journal articles extrapolate on these ideas in the context of particular applications. Arnold has also released several open-source libraries in R, Python, Javascript and C. Visiting appointments have included Invited Professor at Université Paris Diderot and Senior Scientist at AT&T Labs.

Lauren Tilton is Assistant Professor of Digital Humanities in the Department of Rhetoric and Communications at the University of Richmond and a member of Richmond's Digital Scholarship Lab. Her current book project focuses on participatory media in the 1960s and 1970s. She is the Co-PI of the project *Participatory Media*, which interactively engages with and presents participatory community media from the 1960s and 1970s. She is also a director of *Photogrammar*, a web-based platform for organizing, searching and visualizing the 170,000 photographs from 1935 to 1945 created by the United States Farm Security Administration and Office of War Information (FSA-OWI). She is the co-author of *Humanities Data in R* (Springer, 2015). She is co-chair of the American Studies Association's Digital Humanities Caucus.

Detailed Outline

In this three hour tutorial we plan to spend the first 15 minutes getting all participants set up with the software and datasets required for the tutorial. The tutorial participants will we able to work on any reasonably recent version of Windows, macOS, or Linux. All of the software is free and open source. The remained of the workshop will consist of two 75-minutes sessions with a 15 minute break between them.

Each of the two 75-minute sessions will consist of working collectively through "labs" formatted as IPython notebooks. Participants will have the option of using one of three pre-compiled datasets during the workshop depending on their interests:

- historic photographs
- still frames from moving images
- scanned works of art

Alternatively, tutorial participants may alternatively work with their own collection of images.

The first session will focus on describing the potential difficulties of working with image data and explaining how deep learning can be used to address several of the-

se challenges. Working at a conceptual level we will work through the following tasks:

- how to structure a large collection of images as files on a computer
- how to load images into Python as multidimensional arrays
- the concepts behind applying neural networks to image data
- code for projecting images into the penultimate layer of the YOLOv4 neural network
- methods for visualizing the output projects from the neural networks

The second session will focus on how the features detect in the first session can be used to annotate higher level features and measure the similarity between images. Specifically:

- the application of image projections to image similarity metrics
- the application of image projections to object detection
- the application of image projections to face detection

In the final 30 minutes, we will discuss how these techniques ultimately can be used to address humanities questions. This will culminate in running Python code that will output the constructed annotations as an interactive website running locally on each user's computer. This will open up further possibilities for extending the methods of the tutorial without the need for an extensive programming background.

References

Arnold, T. and Tilton, C. (2015). *Humanities Data in R*. New York, NY: Springer.

Arnold, T., Kane, M., and Lewis, B. (2017). *A Computational Approach to Statistical Learning*. New York, NY: CRC Press.

The re-creation of Harry Potter: Tracing style and content across novels, movie scripts and fanfiction

Marco Büchler
mbuechler@etrap.eu
University of Göttingen, Germany

Greta Franzini
gfranzini@etrap.eu
University of Göttingen, Germany

Mike Kestemont
mike.kestemont@uantwerpen.be
University of Antwerp, Belgium

Enrique Manjavacas
enrique.manjavacas@uantwerpen.be
University of Antwerp, Belgium

The tutors

This one-day tutorial will be given by Marco Büchler, Greta Franzini, Mike Kestemont and Enrique Manjavacas.

Endorsement: This workshop is formally endorsed by the Special Interest Group on *Digital Literary Stylistics* (SIG-DLS).

Mike Kestemont (mike.kestemont@uantwerpen.be) is assistant research professor in the department of Literature at the University of Antwerp. He specializes in computational text analysis for the Digital Humanities, in particular stylometry and machine learning, topics on which he has given dozens of hands-on courses. Whereas his work has a strong focus on historical literature, his present research projects cover a wide range of topics in literary history, including classical, medieval, early modern and modernist texts. Mike currently takes a strong interest in representation learning via neural networks.

Marco Büchler (mbuechler@etrap.eu) is a computer scientist and leader of the *Electronic Text Reuse Acquisition Project* (eTRAP) research group at the University of Göttingen. Marco's research interests concern the processing of natural languages with a specialization in the detection of historical text reuse. Furthermore, he is interested in the mining process and the systematization of changes of text reuse. He has worked in this field for over eight years. Together with his eTRAP team, in the past three years he has organized ten text reuse tutorials.

Greta Franzini (gfranzini@etrap.eu) is a Classicist and member of the *Electronic Text Reuse Acquisition Project* (eTRAP) research group at the University of Göttingen. Greta's research interests concern the production of digital editions of texts as well as the combination of quantitative and qualitative methods to advance computational analyses and linguistic resources for Classical literature. Together with her team, Greta has already given eight text reuse tutorials.

Enrique Manjavacas (enrique.manjavacas@uantwerpen.be) is a PhD student at the University of Antwerp. He is associated with the Antwerp Centre for Digital Humanities and Literary Criticism. His current research focuses on sequential methods based on recurrent neural networks to develop semantically-infused models for Stylometry and text reuse detection. He is also interested in Natural Language Generation and has been involved in various projects around the concept of Synthetic Literature.

Description

Computer-assisted text analysis is a core research area in the Digital Humanities. It embraces a wide variety of applications (stylometry, text reuse detection, topic modelling, etc.) and can assist researchers in complex tasks, particularly when it comes to processing large amounts of text. This tutorial brings together two popular and complementary text analysis tasks, stylometry (the quantitative study of writing style) and text reuse detection. While stylometry typically focuses on stylistic similarities between texts (i.e. *how* texts are written), text reuse studies are geared towards the reuse of elements across works (i.e. *what* texts are written about). As such, both methodologies tie into the theoretical notion of *intertextuality* (Orr 2003), albeit in complementary ways.

Creativity and individuality are important phenomena at stake in both fields: are writers at liberty to escape their own 'stylome' - or unique stylistic fingerprint - and to which extent can they free themselves from the many predecessors to which they are intertextually indebted? (Harold Bloom (1973) famously spoke of the 'Anxiety of Influence' in this respect) This leads to interesting theoretical tensions: if authors are stylistically close to one another, does that imply that we can also expect a more elevated level of text reuse between them (and vice versa)? Or can authors frequently reuse textual elements while developing an independent stylistic profile? To which extent is it theoretically possible to oppose style and content?

In this workshop we offer a hands-on introduction to these topics using the case study of Rowling's Harry Potter novels. The vast body of academic scholarship of these writings attests to the relevance of this series, including the highly mediatized stylometric study by Patrick Juola (2013) unmasking Rowling as "Robert Galbraith", the pseudonym under which she temporarily managed to escape her own fame. Intertextuality is also a major concern of Rowling scholarship and scholars as Karin Westman (2007) have meticulously analyzed Rowling's nuanced indebtedness to British authors such as Jane Austen. Rowling herself has invited much intertextual offspring by now too, not in the least in the form of so-called fanfiction (Milli & Bamman 2016), the global phenomenon where (typically non-professional) writers read, reinterpret and expand literary universes (*fandoms*) originally created by acclaimed authors in their own writings (or *fanfics*).

The workshop's tutorial will focus on offering scholars the practical tools and skills to begin to tackle such complex issues. For text reuse detection, participants will learn how to operate TRACER, a language-independent suite of state-of-the-art Natural Language Processing (NLP) algorithms aimed at discovering text reuse in both historical and modern texts, helping users to identify different types of text reuse ranging from verbatim quotations to paraphrase. For the stylometric analyses and visualizations, participants will mainly use custom scripts that exploit the numerous possibilities of the popular Python library *scikit-learn* for Machine Learning. Stylometry with R (Eder et al. 2016), a software package for text analysis in *R*, is another tool that will be used in the introductory sessions.

Data

Participants will practise with data provided by the organizers to better familiarize themselves with the software. The texts under analysis will be the seven English language Harry Potter novels by J. K. Rowling (the so-called core canon of the fandom), a large corpus selection of Harry Potter fanfiction (harvested from *Archive of Our Own*) as well as the Harry Potter movie subtitles.

Objectives

The first objective of the tutorial is to introduce participants to two popular applications of text analysis that tie in closely with intertextuality studies, providing them with an understanding of some of the challenges, methods and strategies proper to this area of research. To this end, we use the illustrative Rowling case study to identify which proportion of the original novels and how much of their style the movies and fanfiction both retain. Additionally, the tutorial seeks to equip participants with the necessary knowledge to independently use the demonstrated software at home (and on their own corpora). Finally, it introduces visualization techniques to display results in an intuitive fashion, provoking new hermeneutic questions.

References

Bloom, H. (1973). *The Anxiety of Influence: A Theory of Poetry*. Oxford, New York: Oxford University Press.

Eder, M., Rybicki, J., Kestemont, M. (2016). Stylometry with R: A Package for Computational Text Analysis. *The R Journal*, 8: 107–121.

Juola, P. (2013). Rowling and "Galbraith": an authorial analysis. *Language Log*. http://languagelog.ldc.upenn.edu/nll/?p=5315 (accessed 2 May 2018).

Milli, S., Bamman, D. (2016). Beyond Canonical Texts: A Computational Analysis of Fanfiction. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2048–2053. https://doi.org/10.18653/v1/D16-1218.

Orr, M. (2003). *Intertextuality: Debates and Contexts*. Polity.

Westman, K.E. (2007). Perspective, Memory, and Moral Authority: The Legacy of Jane Austen in J. K. Rowling's Harry Potter. *Children's Literature*, 35: 145–165. https://doi.org/10.1353/chl.2007.0021.

Archiving Small Twitter Datasets for Text Analysis: A Workshop for Beginners

Ernesto Priego
efpriego@gmail.com
City, University of London, United Kingdom

Abstract

In this workshop for non - coders, participants will be guided through two tasks: the first task will guide participants in creating an application to tap into Twitter's API, in our case to get Twitter data. The second task will guide participants in the use of a Google spreadsheet to capture streaming (live) data from Twitter in order to archive it, download it and perform text analysis, data visualization and other studies. This workshop will include a brief introduction contextualizing social media data collection good practices including user data privacy issues.

Rationale

Twitter data can be very valuable for researchers of perhaps all disciplines, not just DH. Given the difficulties to properly collect and analyse Twitter data as viewable from most Twitter Web and mobile clients (as most people use Twitter) and the very limited short - span of search results, there is the danger of losing huge amounts of valuable historical material.

Tweets are like butterflies – one can only really look at them for long if one pins them down out of their natural environment. The reason why we have access to Twitter in any form is because of Twitter's API, which stands for Application Programming Interface. Free access to historic Twitter search results is limited to the last 7 days. This is due to several reasons, including the incredible amount of data that is requested from Twitter's API, and – this is an educated guess – not disconnected from the fact that Twitter's business model relies on its data being a commodity that can be resold for research. Twitter's data is stored and managed by Twitter's enterprise API platform.

For the researcher interested in researching Twitter data, this means that harvesting needs to be do ne not only through automated means but in real time. It also puts scholars without the required coding and data mining skills at a disadvantage. As a researcher, this basically means that there is no way to do proper research of Twitter data without understanding how it works at API level, and this means understanding the limitations and possibilities this imposes on researchers.

What's a n individual researcher without access to pay corporate access to do? The whole butterfly colony cannot be captured with the nets most of us have available. At small scale, however, and collecting in a timely fashion, it is still possible to capture interesting and more - or - less complete specimens using fairly simply, non - coding required methods. (The Library of Congress h s now 12 years' worth of text - only Tweets. However, as before, the Library of Congress Twitter collection will remain embargoed and there was no projected timetable for providing public access as of 26 December 2017).

Most researchers out there are likely not to benefit from access to huge Twitter data dumps. For researchers without much resources that are trying to do the talk whilst doing the walk, and conduct research *on* Twitter and *about* Twitter, this workshop and tutorial will guide participants into creating a Twitter application in order to tap into the Twitter API, followed

by the setting up of a Twitter Google Archiving Spreadsheet. Once a trial archive or dataset has been collected, we will attempt text analysis and basic visualisations using Excel and Voyant Tools. This workshop will include a brief introduction contextualizing social media data collection good practices including user data privacy and research ethics issues.

References

Priego, E. 2018. #rfringe17: Top 230 Terms in Tweetage. https://epriego.blog/2017/08/05/rfringe17-top 230-terms-in-tweetage/Accessed 30 January 2018]

Priego, E., 2016. Bar Chart: Number of #DH2016 Tweets in Archive per Conference Day (Sunday 10 to Friday 15 July 2016 GMT). Available from: https://figshare.com/articles/Bar_Chart_Number_of_DH2016_Tweets_in_Archive_per_Conf erence_Day_Sunday_10_to_Friday_15_July_2016_GMT_/3490001/1 [Accessed 31 Jan 2018].

Priego, E. 2016. "Stronger In": Looking Into a Sample Archive of 1,005 StrongerIn Tweets. https://epriego.blog/2016/06/21/stronger-in-looking-into-a-sample-archive-of-1005- strongerin-tweets/ [Accessed 30 January 2018]

Priego, E. and Zarate, C., 2014. #MLA14 Twitter Archive, 9 - 12 January 2014. Available from: https://figshare.com/aticles_MLA14_Twitter_Archive_9_12_January_2014/924801/1 [Accessed 31 Jan 2018].

Bridging Justice Based Practices for Archives + Critical DH

T-Kay Sangwand
sangwand@gmail.com
UCLA, United States of America

Caitlin Christian-Lamb
caitlin.christianlamb@gmail.com
University of Maryland, United States of America

Purdom Lindblad
purdom@umd.edu
University of Maryland, United States of America

As scholars and practitioners in digital humanities, we create, analyze, trouble, and reference "the archive," though are often signaling vastly different (mis)understandings of archives, archivists, and archival practices. While both archivists and digital humanists engage critical questions around shared areas of practice (i.e. access, labor, privacy) these conversations often occur in parallel spheres with little recognition of the intellectual contributions in the distinct yet intersecting fields of archives and DH. This workshop aims to bridge the discourse occurring in critical archival studies and critical digital humanities by engaging participants in articulating justice based practices related to appraisal, access, description, pedagogy, privacy, provenance, and system design, as well as collectively contribute these suggested practices to expand existing resources on critical archives and DH (Caswell et al., 2017). At their best, archives and digital humanities center voices that have been obscured through negligence or violently silenced from mainstream narratives. In the face of increased criminalization of and violence towards people of color, immigrants, journalists, mounting militarization, consolidation of media outlets, the political, social, and material impacts of climate change, global capitalism, and white supremacy, we feel a renewed sense of urgency to surface, highlight, and empower narratives from marginalized groups as a tool for social justice and envision new critical archives and digital humanities realities while not recreating oppressive and exploitative power dynamics in the process. This workshop is inspired by Rasheedah Phillips call to articulate "oral futures" and "speaking into existence of what you want to happen" (Phillips, 2017) as well as Michelle Caswell's classroom exercise to "collectively strategize concrete steps to dismantle white supremacy" (Caswell, 2017). The workshop will address the following questions: What are the archival processes of appraisal, accession, description, and access that shape the materials that we can use/collect/analyze as digital scholars and practitioners? How do archivists exercise agency at these various points in an archives' life cycle? What power do researchers/users exercise in their use and (re)presentation of archives? How are communities represented in archives impacted by the use of their archives? What are our collective and individual responsibilities to issues of privacy, description, and access to the materials we collect, analyze, and publish? How can we interrogate archival and scholarly "best practices" and work towards ethical and just practices? How can investigating these overlaps better identify points of collaboration and promote better understandings of cultural heritage across a range of roles, disciplines, and publics?

References

Caswell, M. (2017). Teaching to Dismantle White Supremacy in Archives. *Library Quarterly: Information, Community, Policy*. 87 (3): 222-235. https://doi.org/10.1086/692299.

Caswell, M. et al. (2017). Critical Archival Studies: An Introduction. *Journal of Critical Library and Information Studies Special Issue: Critical Archival Studies*. 1 (2). https://doi.org/10.24242/jclis.v1i2.50.

Phillips, R. (2017). Time, Memory, and Justice in Marginalized Communities. Instagram post. April 23. https://www.instagram.com/p/BTODUEmBZpK/?taken-by=communityfutureslab

Academic Reviewers

Aalberg Trond
Abdul-Rahman Alfie
Adams Robyn Jade
Akbulut Muge
Akça Sümeyye
Albritton Benjamin Long
Alexander Marc
Allés Torrent Susanna
Alpert-Abrams Hannah
Alvarado Rafael
Alzetta Chiara
Anderson Deborah
Anderson Wendy
Anderson Clifford Blake
Andreev Vadim Sergeevich
Andrews Tara Lee
Antonijevic Smiljana
Appleford Simon James
Applegate Matt
Arbuckle Alyssa Emily
Armaselu Florentina
Arneil Stewart
Arora Shaifali
Arriaga Eduard
Arthur Paul
Auddy Purbasha
B Ferronato Priscilla
Babeu Alison L.
Bailey Christopher Scott
Baillot Anne
Baker James William
Bamman David
Bandmann Megyesi Beata
Bangert Daniel Fritz
Barbaresi Adrien
Bardiot Clarisse
Barnett Tully
Barth Florian
Barthonnat Céline
Batjargal Biligsaikhan
Bauer Jean Ann
Baumann Ryan Frederick
Beals M. H.
Beaudouin Valérie
Beelen Kaspar
Bégnis Hélène
Beierle Christoph
Bellandi Andrea
Bellia Angela
Bender Michael
Benedict Nora Christine
Bénel Aurélien
Berens Kathi Inman
Berra Aurélien
Beshero-Bondar Elisa
Bessette Lee
Bhattacharyya Sayan
Bizzoni Yuri
Blümm Mirjam
Bon Bruno
Bonds Elizabeth Leigh
Boot Peter
Borbinha José
Bordalejo Barbara
Borgna Alice
Bornet Philippe
Borovsky Zoe
Bosse Arno
Bouchard Matthew
Bourget Nicolle
Bourgne Gauvain
Boyd Jason Alexander
Boyles Christina
Bozia Eleni
Brando Carmen
Braunstein Laura
Brown Susan
Brumfield Ben
Brumfield Sara
Brussa virginia
Büchler Marco
Burghardt Manuel
Burr Elisabeth
Burrows Toby Nicolas
Cafiero Florian Raphaël
Calvo Tello José
Câmara Alexandra Gago
Campagnolo Alberto
Camps Jean-Baptiste
Cao Ling
Cardillo Elena
Carlton Patricia Lynn
Casarosa Vittore
Casenave Joana
Casties Robert
Caton Paul
Cavanaugh Erica Fallon
Cayless Hugh
Chammas Michel
Charles Valentine
Chartrand Louis
Château-Dutier Emmanuel
Chavez Villa Micaela
Chawla Swati
Cheesman Tom
Chen Shih-Pei
Chen Kuang-hua
Chen Jing
Chiaravalloti Maria Teresa
Childress Dawn
Chuang Tyng-Ruey
Chue Hong Neil

Ciotti Fabio
Ciula Arianna
Clavert Frédéric
Clement Tanya
Clivaz Claire
Cochrane Euan
Cohen Hart
Colavizza Giovanni
Conway Paul
Cooney Charles M.
Cordell Ryan
Cotarelo-Esteban Lucia
Courboulay Vincent
Cowan William
Cowan T.L.
Craig Hugh
Crawford Cole Daniel
Crompton Constance
Croxall Brian
Cummings James
Curado Malta Mariana
Dabbs Thomas Winn
Dadvar Maral
Daengeli Peter
Dahlstrom Mats
Dallachy Fraser James
Dalmau Michelle
Damerow Julia Luise
Davis Rebecca Frost
De la Cruz Fernandez Paula
De la Rosa Pérez Javier
De Roure David
De- Matteis Lorena Marta Amalia
Declerck Thierry
Degaetano-Ortlieb Stefania
Del Grosso Angelo Mario
Del Rio Riande Gimena
Delve Janet
Derven Caleb
Devaney Johanna
Di Bacco Giuliano
Di Cresce Rachel
Di Donato Francesca
Di Ludovico Alessandro
Dilley Paul
Dogruoz Seza
Dombrowski Quinn
Dorn Amelie
Duckett Victoria
Dunst Alexander
Dussault Jessica Valerie
Eccles Kathryn
Eckart Thomas
Eckert Kai
Eder Maciej
Edmond Jennifer C
Ehrmann Maud
Eichmann-Kalwara Nickoal
Eide Øyvind
Elli Tommaso
Endres Bill
Engel Maureen
Escandell-Montiel Daniel
Escobar Varela Miguel
Esteva Maria
Estill Laura
Falk Michael Gregory
Faull Katherine Mary
Fendt Kurt E
Fenlon Katrina Simone
Fernandez Riva Gustavo
Ferschke Oliver
Fields Paul J.
Finn Edward
Fischer Franz
Flanders Julia
Fokkens Antske
Forest Dominic
Forlini Stefania
Fornes Alicia
France Fenella Grace
Franzini Greta
Fredner Erik Christopher
French Amanda
Friedland Nancy E.
Froehlich Heather
Frontini Francesca
Gagarina Dinara
Gairola Rahul Krishna
Galina Russell Isabel
Galleron Ioana
Gallet-Blanchard Liliane
Gao Jin
Garcia-Fernandez Anne
Garfinkel Susan
Garnett Vicky
Gartner Georg
Gautier Laurent
Giannella Julia
Giannetti Francesca
Gil Alexander
Giovannetti Emiliano
Girard Paul
Giroux Amy Larner
Gius Evelyn
Gladstone Clovis
Glass Erin Rose
Gniady Tassie
Goddard Lisa
Gold Matthew K.
Gordea Sergiu
Gordon Tamar

Goto Makoto
Goudarouli Eirini
Grandjean Martin
Grant Katrina Caroline
Griffin Howard Kevin
Griggs Hannah C.
Grincheva Natalia
Grüntgens Max
Guido Daniele
Guiliano Jennifer Elizabeth
Gutiérrez De la Torre Silvia Eunice
Guzman Carina Emilia
Hackney S. E.
Hammond Adam
Han Myung-Ja K.
Heiden Serge
Hendery Rachel Marion
Hennicke Steffen
Henny-Krahmer Ulrike Edith Gerda
Henrich Andreas
Henry Geneva
Heppler Jason A.
Herrmann J. Berenike
Heuser Ryan James
Heuvel Charles van den
Heyer Gerhard
Hicks Benjamin Wesley
Hiebert Matthew
Higgins Devin
Hinrichs Uta
Hladík Radim
Ho Hou Ieong
Hodel Tobias
Hodošček Bor
Hoekstra Rik
Hoenen Armin
Holmes Martin
Homburg Timo
Hoover David L.
Horstmann Jan
Houston Natalie M
Hsiang Jieh
Hswe Patricia
Huculak John Matthew
Huijnen Pim
Huitric Solenn
Hulden Vilja
Hunter Jane
Hunter John
Hurtado Tarazona Alejandra
Hyman Christy
Idmhand Fatiha
Impett Leonardo Laurence
Isaksen Leif
Jacobs Hannah L.
Jakacki Diane Katherine
Jamison Anne
Janco Andrew
Jannidis Fotis
Jensen Thessa
Jett Jacob
Johnson Ian R.
Jones Michael Alastair
Jones Catherine Emma
Jones Madison Percy
Jordanous Anna Katerina
Juola Patrick
Kampkaspar Dario
Kane Julie
Karadkar Unmil
Kaufman Micki
Kawase Akihiro
Kelleher Margaret
Kemman Max
Kenderdine Sarah
Kermes Hannah
Kerr Sara Jane
Kessler Carsten
Khosmood Foaad
Kijas Anna Ewelina
Kim Minhyoung
Kim Evgeny Gamletovitch
King Lindsay
Kitamoto Asanobu
Kizhner Inna
Klein Lauren F.
Kleppe Martijn
Klinger Roman
Koho Mikko Kristian
Koolen Marijn
Körner Fabian
Koumpis Adamantios
Kretzschmar William
Kröger Bärbel
Kumar Ritesh
Kumari Ashanka
Kurlinkus Will
Lach Pamella R
Lahti Leo
Lana Maurizio
Lang Anouk
Lang Matthias
Laubrock Jochen
Lavagnino John
Lavrentiev Alexei
Leavy Susan
Leblay Christophe
Leem Deborah
Lester Connie Lee
Letricot Rosemonde
Levallois Clement
Licastro Amanda Marie

Lincoln Matthew
Lindblad Purdom
Lindquist Thea
Litta Eleonora
Liu Chao-Lin
Liu Jyi-Shane
Lopes Patricia
Lorang Elizabeth M
Losh Elizabeth
Madron Justin
Maeda Akira
Mäkelä Eetu
Makinen Martti
Malm Mats
Malta Joana
Manzanera Silva Norma Aida
Mapes Kristen
Marchetti Andrea
Martin Kim
Martinez-Canton Clara
Martins Bruno Emanuel
Maryl Maciej
Mas Joan
Mathiak Brigitte
Mattock Lindsay Kistler
Mauro Aaron Mathew
McDonald Robert
McGarry Shane Adam
McGrath Jim
Mehler Alexander
Melton Sarah
Mendoza Juan José
Meneses Luis
Menini Stefano
Menon Nirmala
Merritt Don
Meyer Eric T.
Meza Aurelio
Michlowitz Robert
Miller Ben
Milligan Ian
Mimno David
Miyagawa So
Monteiro Vieira Jose Miguel
Morán Ariel
Morgan Paige Courtney
Moritz Maria
Morlock Emmanuelle
Moro Jeffrey Tyler
Motilla José Antonio
Mpouli Suzanne
Murai Hajime
Murphy Orla
Murr Sandra
Murray-John Patrick David
Murrieta-Flores Patricia
Musgrave Simon
Mylonas Elli
Nagasaki Kiyonori
Nainwani Pinkey
Nanni Federico
Navarrete Trilce
Neovesky Anna
Nerbonne John
Neuber Frederike
Neuefeind Claes
Newton Greg T
Nieves Angel
Noordegraaf Julia
Nowak Krzysztof
Núnez Alexandra
Nurmikko-Fuller Terhi Maija
Nyhan Julianne
O'Connor Alexander
O'Donnell Daniel Paul
Ocampo Gutiérrez de Velasco Marat
Ochab Jeremi K.
Ohya Kazushi
Olsen Mark
Ore Espen S.
Orekhov Boris V.
Organisciak Peter
Orlowska Anna Paulina
Ortega Elika
Otis Jessica
Overbeck Maximilian
Padilla Thomas George
Page Kevin
Pagé-Perron Émilie
Pairet Laure
Palkó Gábor
Papadopoulos Konstantinos
Paquette-Bigras Ève
Paris Britt
Pawłowski Adam Tomasz
Peaker Alicia Rose
Peña Ernesto
Peña-Pimentel Miriam
Perez Isasi Santiago
Pernes Stefan
Peroni Silvio
Petersen Andrew
Pierazzo Elena
Pimenta Ricardo Medeiros
Piotrowski Michael
Poibeau Thierry
Polyck-O'Neill Julia Geneviève
Powell Daniel James
Preiser-Kapeller Johannes
Pretnar Ajda
Priani Ernesto
Priego Ernesto

Puren M.P.
Puschmann Cornelius
Radzikowska Milena
Ramos Adela María
Ray Murray Padmini
Rebora Simone
Reeve Jonathan Pearce
Rehberger Dean
Rehm Georg
Reiter Nils
Renault Arthur
Ribeiro Cláudia
Ricaurte Paola
Ricciardi Emiliano
Richards-Rissetto Heather
Riddell Allen Beye
Ridge Mia
Ridolfo Jim
Riondet Charles
Risam Roopika
Robertson Stephen Murray
Robey David
Robinson Peter
Robles-Gómez Antonio
Rochat Yannick
Rockwell Geoffrey
Rodighiero Dario
Rodríguez-Roche Sulema
Roe Glenn H
Roeder Torsten
Rogel Rosario
Rojas Castro Antonio
Romanello Matteo
Romary Laurent
Romero-López Dolores
Rosenblum Brian
Rosner Lisa
Rosselli Del Turco Roberto
Rotari Gabriela
Roueché Charlotte
Routsis Vasileios
Röwenstrunk Daniel
Rudman Joseph
Ruiz Fabo Pablo J
Rumyantsev Maxim
Rusinek Sinai
Rybicki Jan
Sahle Patrick
Saklofske Jon
Salvatori Enrica
Sanz Amelia
Saum-Pascual Alex
Sayers Jentery
Scharnhorst Andrea
Scheuermann Leif
Schich Maximilian
Schlarb Sven
Schlesinger Claus-Michael
Schl r Daniel
Schmidt Sara A.
Schmunk Stefan
Schöch Christof
Scholger Walter
Schommer Christoph
Schulz Sarah
Senier Siobhan
Senseney Megan Finn
Serantes Arantxa
Severo Marta
Sharpe CelesteTường Vy
Shaw Ryan Benjamin
Shep Sydney
Shepard David Lawrence
Shepherd Ammon
Sherratt Tim
Shibutani Ayako
Shimoda Masahiro
Shrout Anelise Hanson
Siders Anne R
Siemens Raymond George
Siemens Lynne
Silva Andrea
Sinclair Stéfan
Smithies James Dakin
Snyder Lisa M.
Song Yuting
Sostaric Petra
Spadini Elena
Spence Paul Joseph
Sperberg-McQueen Michael
Spiro Lisa
Sprugnoli Rachele
Stadler Peter
Stalnaker Rommie L
Stertzer Jennifer Elizabeth
Stewart Elizabeth Eleanor Rose
Steyn Zacharias Jacobus
Stokes Peter Anthony
Strötgen Jannik
Stutzmann Dominique
Subotic Ivan
Suire Cyrille
Sula Chris Alen
Swafford Joanna Elizabeth
Swanstrom Elizabeth Anne
Szabo Victoria
Takseva Tatjana
Tambassi Timothy
Tammaro Anna Maria
Tanasescu (MARGENTO) Chris
Teich Elke
Ter Braake Serge

Terras Melissa
Theibault John Christopher
Thomas Lindsay
Thompson Jeff
Thomson Christopher
Tilton Lauren
Tonelli Sara
Tonnellier Gaelle
Tonra Justin Emmet
Tournier Charlotte
Tracy Daniel G.
Travis Charles Bartlett
Tropea Rachel
Tsui Lik Hang
Tuffery Christophe
Tupman Charlotte
Turton Alexander Robert
Valverde Mateos Ana
van den Herik H. J.
van Eijnatten Joris
van Erp Marieke
Van Keer Ellen
Van Kranenburg Peter
Van Zundert Joris Job
Venecek John T.
Viana Vander
Viglianti Raffaele
Visconti Amanda
Vogeler Georg
Volkmann Armin
Volodin Andrei
von Waldenfels Ruprecht
Walkowiak Tomasz
Walkowski Niels-Oliver
Walsh John
Walsh Brandon
Walter Katherine L.
Warwick Claire
Webb Sharon
Weber Andreas
Weidman Robert William
Weidman Sean Gregory
Weigl David M.
Wernimont Jacqueline D
Wevers Melvin
Widner Michael Lee
Wieneke Lars
Wieringa Jeri
Wiesner Susan L.
Wilkens Matthew
Williams Patrick
Williams Helene C.
Wilms Lotte
Winder William
Wintergrün Dirk
Wisnicki Adrian S.
Wittern Christian
Wolff Mark
Worthey Glen
Wrisley David Joseph
Wulfman Clifford Edward
Würsch Marcel
Wuttke Ulrike
Yamada Taizo
Yang Bin
Yeates Stuart Andrew
Yin Xin
Youngman Paul
Zafrin Vika
Zeng Marcia Lei
Zhang Jinman
Zöllner-Weber Amélie
Zwarich Natasha

Digital Humanities 2018

dh2018.adho.org