AI-Innovative digitization and management processes for Digital Libraries and Archives cultural heritage: toward an inclusive and sustainable phygital ecosystem

Stefano Ferilli [1] Annastella Carrino [2] Giuliano De Felice [2] Paolo Fioretti [2] Pietro Silanos [2] and Eleonora Bernasconi [1]

[1] *Università di Bari – DIB, Via E. Orabona 4, Bari, 70125, Italia*
[2] *Università di Bari – DIRIUM, Piazza Umberto I 1, Bari, 70121, Italia*

Abstract

In its most recent interpretation, the concept of 'cultural heritage' goes beyond any physical, chronological or identity classification, taking the form of a network of relationships based on the dialectic between heritage and possible communities. An archaeological site, an archive, a library, a collection or a museum cannot be fully understood or studied without relating them to the context in which they were produced and to the consequences they generated. Without a human context of reference, no cultural heritage can be managed, protected, safeguarded and enhanced. This paper describes a project, funded by the Italian National Recovery and Resilience Plan (NRRP), aimed at developing a truly inclusive and sustainable phygital ecosystem, which can combine not only digital with cultural heritage, but also the tangible and intangible components of the latter. AI plays a key role in the project.

Keywords

Digital Libraries & Archives, Written Cultural Heritage, Digitization, Management

1. Introduction

The digital transition of both research objects and methodologies seems to respond to the specific needs and practices of several fields. Textual Scholarship together with synchronic and diachronic studies on language and multilingualism are privileged sectors of application of ICT. The outcomes - e.g., digital corpora and editions, descriptive metadata, and logic representations of their semantic relationships - are increasingly creating knowledge and knowledge networks. This paper describes the contribution, in these directions, of a multidisciplinary group of researchers working in Spoke 3: "Digital Library, Archives and Philology" of project CHANGES (Cultural Heritage Active Innovation for Next-Gen Sustainable Society). CHANGES won the competition issued by the Italian Ministry of University and Research for the Creation of Enlarged Partnership 5 "Humanistic culture and cultural heritage as laboratories of innovation and creativity" extended to Universities, Research Centres, Enterprises and aimed at funding basic research projects under the National Recovery and Resilience Plan (NRRP), funded by the EU under the NextGenerationEU programme. Its general objectives are:

1. the creation of a multi-technological transdisciplinary hub of international reference for training, research, technology transfer, in order to (a) enhance the attractiveness of the CH assets of Italy; (b) implement an excellent public-private model for collaboration and stable partnerships; and (c) offer a pole of attraction for companies, public and private Institutions.

[1] *2nd Italian Workshop on Artificial Intelligence for Cultural Heritage (IAI4CH2023, https://ai4ch.di.unito.it/), co-located with the 22nd International Conference of the Italian Association for Artificial Intelligence (AIxIA 2023). 6-9 November 2023, Roma, Italy.*

stefano.ferilli@uniba.it (S. Ferilli); annastella.carrino@uniba.it (A. Carrino); giuliano.defelice@uniba.it (G. De Felice); paolo.fioretti@uniba.it (P. Fioretti); pietro.silanos@uniba.it (A. 5); eleonora.bernasconi@uniba.it (E. Bernasconi)
0000-0003-1118-0601 (S. Ferilli); 0000-0002-1852-5783 (P. Silanos); 0000-0001-8416-9564 (G. De Felice) 0000-0002-4929-8886 (P. Fioretti); 0000-0001-6249-1886 (P. Silanos); 0000-0003-3142-3084 (E. Bernasconi).

2. the implementation of a plan of action and structural interventions that over a decade leads to a progressive increase, in the three mentioned areas of strategic interest, of people, laboratories, collaborations and investments focused on research and education in the thematic areas of interest for CH.
3. Relaunch of the economy and territorial development of the CH sector in centre-southern Italy after the pandemic crisis.

In this context, the objectives of spoke 3 include:

- Enhancement of the international quality standards for the Digital Imaging and Preservation of book-related sources (IFLA and FADGI), by promoting whole or selective Reflectance Transformation Imaging (RTI) on manuscripts, fragments, printed books, documents and image sources and fostering the adoption of international standards (such as IIIF).
- Setting up of open-source Web environment for the automated recognition (HTR and OCR) of the layouts and the characters of the textual sources; feeding the HTR engines for automated recognition of linguistic features and *loci critici* for the digital *recensio*.
- Creation of computational philology platforms in order to realize digital libraries and scholarly editions encoded in XML-TEI markup standard.

The most recent interpretation of Cultural Heritage (CH) goes beyond any physical (tangible vs. intangible), chronological (ancient vs. modern) or identity (ours vs. others) classification, taking the form of a network of relationships based on the dialectic between heritage and possible communities. Indeed, every artifact, landscape, tradition or memory, cannot be considered as CH as such, but only in relation to the links they are able to establish with society. In this perspective, our effort aims at developing a truly inclusive and sustainable phygital ecosystem, which can combine not only digital with CH, especially the written one, but also the tangible and intangible components of the latter.

In its hybrid, phygital indeed, dimension, this project aims at overcoming the logic of heritage as a passive and static object, instead promoting an approach that embraces that of participation. This solution allows to overcome the impasse between digital and heritage as expressed, e.g., in the Faro Convention, filling the gaps in terms of digital backwardness, but at the same time preserving the essential principles of sharing and community.

Today, users can establish a relationship with CH by connecting virtually to sites and containers, or even to each other through different platforms, such as websites, social media, mobile apps etc. These access points, or 'channels', are envisaged as an opportunity to convey content, communicate heritage, and tell stories to users that have acquired co-protagonist roles. In fact, terms customer-centrism or user-centricity indicate the centrality of the user in the process of knowledge transmission.

Digital technologies are a naturally unifying tool for all aspects of CH: they can store digitizations of tangible cultural assets and save records of intangible ones; moreover, they can store descriptive metadata of both, and manage all these aspects in an integrated and coordinated way. Under appropriate conditions the same digital records become, in turn, CH. So, the digitization of CH is combined with a new concept of usability of the works, which take on multiple lives, thanks to diversified communication, implementing the principles of solidarity and substantial equality.

The project, intrinsically multidisciplinary, is supported by scholars from various institutions (UniBA, Bocconi, UniCam, UniFG, Sapienza) including a wide range of disciplinary fields, so as to ensure coherence and in-depth study from multiple perspectives; it is also supported by several institutions and companies. Most prominent is the Superintendency of Archival and Library Assets (Soprintendenza dei beni archivistici e bibliotecari), but several industrial partners are also willing to contribute for an initial pilot application: Fondazione G. Di Vagno; Fondazione Gramsci di Puglia; IPSAIC, Puglia; Associazione per la storia della Puglia e del Mezzogiorno nell'integrazione europea; Alcide De Gasperi Centre on European integration history (European University Institute); Casa editrice Laterza; DABIMUS S.r.l., Fondazione Ampioraggio; Artificial Brain S.r.l.; Apulia Retrocomputing OdV; 3D Research srl.

The rest of this paper is organized as follows. After reporting in the next section related work, in Section 3 we provide motivations for our effort and discuss the challenges it involves. Then, Section 4 describes the objectives and expected impact of the project. Section 5 is devoted to introducing the project team and possible interdisciplinary connections, before concluding the paper in Section 6.

2. Related Works and New Perspectives

«A model that aims solely at storing and preserving cultural heritage, making it live only through a few very limited access points, is now insufficient: a cultural heritage lives or dies only to the extent that it is animated by a community of people» [6]. Jeffrey Schnapp, founder of Harvard University's metaLAB and leading global figure in Digital Humanities, thus identified one of the critical issues of existing CH. This matter identifies two levels of analysis, related to access to CH and to the narration of heritage to diverse audiences [17]. Within this framework, digitisation processes offer many opportunities.

The total blockade imposed to stem the COVID-19 contagion has reawakened citizens' new need for interaction with CH. Institutions have seen fit to meet it effectively by putting their digitisation projects online[2]. However, most of the exhibitions showed a digital production in line with the digitisation concept outlined in the 2019 Three-Year Plan, thus inevitably resulting from outdated technological approaches, focused exclusively on the visual aspect of digital resources and with minimal possibilities for interaction. Digitisation and digital creativity projects implemented in Italy focus exclusively on the visual components of digital objects. However, digitisation has always needed more attention to cultural objects' descriptive and informative components as indispensable elements to make digital resources intelligible and to historicise them as culturally valid.

A recent project that tried to go in this direction, conceived and developed by the Commission for the National Edition of Leonardo da Vinci's Manuscripts and Drawings in 2019[3], digitized the manuscripts that were part of Leonardo's library, making them accessible online in a single digital space, but also at offering a dynamic experience of their use [7].

Precisely concerning written cultural assets, new perspectives for digitisation are now opening up. The most frequently discussed one is based on the recognition and automatic transcription of writings (and thus of their contents): see, for example, the project *In Codice Ratio*[4] [8]. Others are also working on the extraction of knowledge from text, and on its formalization into so-called Knowledge Graphs [9,10], or investigating various tasks connected to document processing and management [11], knowledge representation for CH [12,13]. An innovative Intelligent Character Recognition software for text recognition, extraction, indexing and retrieval from digitized ancient handwritten or printed artifacts [14] is being developed at the University of Bari, which also invested in the recent creation of a master's degree course on "Digital Cultural Heritage: Museums, Archives, Libraries" that integrates the scientific tradition in archive, library and museum science with computational methodologies applied to innovation for CH.

A new perspective tries to go beyond a traditional acquisition, limited to specific components of written CH, but considers the manuscript book in its complexity. Text, writings, images and any decorations are inseparably linked to the physical aspects of their supports, form and function. A written Cultural Property can be understood and enjoyed only by considering its multidimensional materiality and even the container has its own inalienable historical value, different from all the others. The digitisation that can also capture their physicality, transforms each of these assets into true complex emotional activators that reveals the extraordinary richness and communicative power of cultures and societies that produced them.

Moreover, written CH is rapidly perishable due to the precarious media. It is, therefore, inevitable to devise and experiment with new protection and valorisation strategies, using innovative digital methodologies and techniques that integrate 2D optical documentary acquisition with the most advanced 3D and Augmented Reality technologies. The potential of such an innovative digitisation model of written CH can affect several levels: 1) study and knowledge; 2) valorisation and fruition, didactic and tourist-cultural; 3) protection and restoration. In this way, users of written CH would become visit-actors able to interact with the complexity of such Heritage, enjoying a User Experience that maximizes the historical-cultural impact and the communicative and cognitive richness.

The increasing application of disruptive technologies to CH has shown significant innovation potential. The Internet of Things, Artificial Intelligence (AI), Social Robotics, and Virtual,

[2] http://www.icom-italia.org/museums-and-covid-19-8-steps-to-support-community-resilience/

[3] https://bibliotecadileonardo.museogalileo.it/

[4] http://www.inf.uniroma3.it/db/icr/index.html

Augmented and Mixed Reality enable visitors to become protagonists in the experience of CH and collections and to foster engagement, which is at the core of both active citizenship and new cultural professions. Significant changes are produced by born-digital entities, which increasingly contribute to the constitution of tomorrow's CH and its services (computerized management of processes, dematerialised archives of public and private administrations, visual and performing art, etc.). The born-digital entities also must be coherently integrated within the reflection and practices of CH.

The use of digital to improve and democratize participation in CH is an open question. Even the Faro Convention [15,16] did not resolve it, leaving only a residual space for the digital, considered more as a danger than a resource. From this point of view, important reflections and experiments have been carried out especially in the field of archaeology, which has always been one of the CH sectors most inclined to move into innovative areas [1,2]. Our project sees in archaeology an important contribution, because of its specific focus on the physical dimension of heritage, and on imagination [3]. Some reflections developed by digital archaeology on the relationship between technologies, domain languages and creative approach are worth mentioning as a process of designing and building participatory environments [4, 5].

3. Motivations & Challenges

In the described context, there are still gaps and open problems, related to different perspectives:

- Research on CH has so far been conducted mostly separately from the various "vertical" perspectives: digitization, description, storage, and use. In order to add value, the different perspectives must be properly harmonized.
- Sectoriality also afflicts each of these perspectives "horizontally": various types of CH have been studied independently and have generated different, often incompatible solutions; this urges to treat in a uniform way all the existing digital assets.
- On the formal description side, in particular, the traditional standards for cataloging and description, or computerized metadata, of CH items are inadequate and insufficient to grasp and connect all aspects. A 'holistic' descriptive approach is needed, in order to support adaptive and personalized exploitation.
- The gaps on the descriptive side generate other gaps, including that of data analysis techniques that exploit the full power of 'holistic' descriptions; these could be used to discover types and communities of users, to build models of individual users and communities, and to extract from them understandable and useful information to meet the specific needs of each.
- Written CH has its own specificity: text, graphic shapes, physical aspects, form and function of the media, images and decorations are inseparable elements. The tight coupling between them requires a rethinking of the digitization of written CH to enhance and to preserve it.
- In 2020 the EU clarified the requirements that should characterize the new Digital Cultural Heritage (DCH): "There is a real need to establish a comprehensive picture of the studied assets, capturing and re-creating not only visual and structural information, but also stories and experiences (stored in language data), together with their cultural and socio-historical context, as well as their evolution over time".

Each of these points opens up opportunities for research and represents a motivation for this project, requiring a multidisciplinary research that takes into account all perspectives. It is necessary to study in depth and systematize the types of CH and the ecosystem that revolves around them; to define new descriptive schemes; to invent new types of computer-based representation, digitization, storage and exploitation, also for conservation and accessibility purposes, and allow new and more advanced and targeted types of use.

4. Objectives

From the IT point of view, the general objective of the project will be to produce integrated technologies, covering the entire life cycle of the generated digital objects, from digitization, to metadata, to storage in DBs, to syntactic and semantic retrieval, to knowledge discovery on the data

thus generated, to personalized support to users, to interfaces for the fruition of tangible and intangible CH. Specific objectives will be:

- Developing conceptual schemes and formalisms, which are instrumental to automatic processing by computers, so as to describe in a uniform and integrated way the different types of CH.
- Defining technological solutions and guidelines for the digitization of various types of CH and for the development of DBs that are able to preserve and correlate all the various types of digital content.
- Designing data analysis and management techniques based on AI, which are able to identify and extract relevant information to be provided to various types of users to improve and personalize fruition.
- Studying innovative digitization processes for written CH through which to experiment the integration of methodologies of optical 2D acquisition of documents with the most advanced 3D and augmented reality technologies.
- Studying the adoption of solutions for the storage of structured and unstructured data, heterogeneous in nature (text, images, 3D models, etc.), in relation to each other, also using techniques for the efficient retrieval of information.
- Applying the results of this research to pilot contexts.

While considering the whole landscape of AI approaches and solutions, the specific focus of the researchers at UniBA will be on symbolic representation and reasoning techniques, that are closer to the way in which humans represent and manipulate information, and thus can provide more understandable, interpretable, explainable and, as a consequence, trustworthy results to the users. Still, sub-symbolic approaches will be considered for tasks more related to intuition and perception, such as handwriting recognition or computation of similarity between cultural objects.

Major original and specific features of this project are that it looks specifically at cultural assets that are not immediately intuitive, nor the subject of special attention, preservation, and enhancement in the current landscape. These are archival and library assets, which are less central in policy and design solutions for enhancement, but hold enormous potential, and are just waiting to be harnessed and enhanced. Museums could also benefit from the development of these solutions, also in the perspective of more aware and experiential fruition. Research activities and scientific and cultural dissemination are, in fact, more and more interconnected with the virtual world in order to effectively transmit knowledge and reach a wider, more aware and participating public. Digital environments are attractive for the younger and older generations alike.

The 'holistic' perspective to the description of cultural heritage will open new perspectives to its fruition, making available to the user not only the items, but all the vast wealth of knowledge directly and indirectly related to them.

5. Expected Results & Impact

On the IT front, the project aims at attaining several important results:

- A 'holistic' ontology, to act as a metadata schema to describe the domain of tangible and intangible CH, including and correlating, in addition to the formal aspects traditionally considered in the community of practice, the aspects of content, structure, context, life cycle and actors; all these aspects are often neglected, but fundamental to be able to fully grasp CH in all facets of its form and substance. The ontology will allow the integration of advanced AI solutions, including Knowledge Representation and Reasoning, Machine Learning, Data Mining and support for personalized recommendation, advanced retrieval and browsing of knowledge, data analysis, etc..
- The definition of an AI-based architecture and guidelines for the construction of Knowledge Bases that accommodate digital descriptions of all 'holistic' aspects of CH.
- The definition of a paradigm that, based on the description and storage of CH, defines the steps and technologies required for the digitization of the originals and for the preservation and fruition of their digital surrogates. Given the complexity and variability of the application

domain, this outcome will also include a broad spectrum of AI solutions, in an innovative mixture.
- The application of all of the above to the digitization, representation, rendering and fruition of case studies of particular significance and interest to the CH community.

These results are expected to have a significant impact as an instrument of social welfare and economic growth, in various directions:

- Promote scientific culture starting from heritage and cultural activities in a perspective of "lifelong education".
- Start the process of digitization of CH and green transition to promote a cultural archive for future generations, scientific-technological progress (Apps, virtual and augmented reality), cooperation with the third sector for a more productive link with the academic area, fostering new trades, skills and market opportunities.
- Enrich the experience of users through the use of technological devices and products, thus encouraging interaction between the structure that educates and the public that learns, in a physical experience of visit enriched by multimedia content. In this way, larger and more heterogeneous portions of society will be sensitized to the CH of the territory and its potential as a tool for social cohesion and well-being, and as a driver of new inclusive participation procedures that also promote gender equality.

The project will provide many interaction opportunities with other research disciplines. E.g., for law, economics, and museums, to understand the extent of the norms on the reuse of public data, to identify a tradeoff between the needs of public institutions (museums, libraries, archives), digital professionals, and the public interest in cultural promotion and dissemination; to investigate the contractual policies of cultural institutions in order to find a balance between the freedoms to protect information, scientific research, expression of artistic thought and the right to cultural fruition and the interest of the public body in governing the use of digital versions of items according (also) to parameters of economy. This involves the theme of business archives, especially private publishing archives. Editorial archives are both a corporate asset (and therefore subject to the events of the business organization they belong to) and a strategic asset of collective CH (and therefore also a touristic asset). In this respect, the subject of investigation might be the historical aspects, the critical issues concerning intellectual property protection, the managerial, strategic and operational business aspects of the so-called culture-driven companies, and the ecological-merceological profiles.
The aim is to start a dialogue with the territorial bodies (associations, political and social organizations) by creating an ‘Observatory of Digital Technologies and Cultural Heritage’.
A positive impact is expected on sustainable cultural tourism development to be shared and planned with public administrations, supporting a new political governance that promotes circular and sustainable economy, trust in participatory democracy.

6. Conclusions

In its most recent interpretation, the concept of ‘cultural heritage’ goes beyond any physical, chronological or identity classification, taking the form of a network of relationships based on the dialectic between heritage and possible communities. This paper described a multidisciplinary research aimed at developing a truly inclusive and sustainable phygital ecosystem, which can also combine the tangible and intangible components of CH. The research will be carried out within the “Digital Libraries, Archives and Philology” branch of a larger CH project funded by the Italian National Recovery and Resilience Plan (NRRP). We provided the current landscape, motivations, objectives and expected results of the research. We believe this effort is one of the possible and practicable ways to valorise a heritage with a high intrinsic value, but which the logic of the market and of preservation and valorisation all too often leaves in the background, conceiving it as an area reserved for specialists and not deserving adequate attention and investment of resources and energy.

Acknowledgments

This research was partially supported by projects FAIR - Future AI Research (PE00000013), spoke 6 - Symbiotic AI, and CHANGES - Cultural Heritage Active innovation for Next-GEn Sustainable

society (PE00000020), Spoke 3 - Digital Libraries, Archives and Philology, under the NRRP MUR program funded by the NextGenerationEU.

References

[1] T. L. Evans, P. Daly (eds), *Digital Archaeology, Bridging method and theory*, London-New York 2006.
[2] E. C. Kansa, S. W. Kansa, E. Watrall (eds), *Archaeology 2.0: New Approaches to Communication and Collaboration*, Berkeley 2011.
[3] M. Shanks, *The archaeological imagination*, Walnut Creek 2012.
[4] G. De Felice, A. Fratta, Ordona XIII. Dalla città fantasma alla città virtuale, Bari 2022.
[5] E. Degl'Innocenti, D. Leone, M. Turchiano, G. Volpe (eds.), *Taras e i doni del mare*, Bari 2022.
[6] J. Schnapp, *Digital Humanities*, a cura di M. G. Mattei, Milano, Egea, 2015
[7] G. Cirnigliaro, *The Digital Reconstruction of Leonardo's Library: Revealing Formal Patterns in Early Modern Thought*, in *Special Issue on "Digital Humanities for Academic and Curatorial Practice"*, «Studies in Digital Heritage», 3/2 (2019), pp. 128-143
[8] D. Firmani, P. Merialdo, M. Maiorino, E. Nieddu, *Towards Knowledge Discovery from the Vatican Secret Archives. In Codice Ratio – Episode 1: Machine Transcription of the Manuscripts*, in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (KDD '18). ACM, New York, NY, USA, 263-272
[9] E. Bernasconi, M. Ceriani, M. Mecella. Exploring a Text Corpus via a Knowledge Graph. Proceedings of the XVIII Italian Research Conference on Digital Libraries, Central Europe (CEUR) Workshop Proceedings vol. 2816, 91-102, 2021
[10] E. Bernasconi, M. Ceriani, M. Mecella, A. Morvillo. Automatic Knowledge Extraction from a Digital Library and Collaborative Validation. In Linking Theory and Practice of Digital Libraries, Lecture Notes in Computer Science 13541, 480-484, Springer, 2022.
[11] S. Ferilli. Automatic Digital Document Processing and Management - Problems, Algorithms and Techniques. Advances in Pattern Recognition series, Springer, London, 2011
[12] S. Ferilli & D. Redavid. *An Ontology and Knowledge Graph Infrastructure for Digital Library Knowledge Representation*. In Digital Libraries: The Era of Big Data and Data Science, Communications in Computer and Information Science 1177, 47-61, Springer, 2020
[13] S. Ferilli. Holistic Graph-Based Representation and AI for Digital Library Management. In Linking Theory and Practice of Digital Libraries, Lecture Notes in Computer Science 13541, 485-489, Springer, 2022
[14] N. Barbuti, S. Ferilli & T. Caldarola. Un innovativo graphic matching system per la ricerca in database di manoscritti antichi. Umanistica Digitale, ISSN 2532-8816, 3:45-66, AIUCD, 2018
[15] R. Palmer (ed), Heritage and Beyond, Brussels 2009
[16] G. Volpe, Archeologia pubblica. Metodi, tecniche, esperienze, Roma 2020
[17] P. Silanos, *Una finestra aperta sul passato. Medioevo, musei, digitale e Public History*, in *Il medievista come Public Historian*, a cura di E. Salvatori, Roma, ISIME, 2022 (Nuovi Studi storici, 125), pp. 205-225