

DICHIARAZIONE SOSTITUTIVA DI CERTIFICAZIONE

(art. 46 DPR 28/12/2000 n. 445)

Il sottoscritto NICOLA BARBUTI, nato a BARI (BA) il 25/03/1967, residente a BARI (BA) in via NAPOLI n. 104, C.F. BRBNCL67C25A552K, ricercatore a tempo indeterminato presso il Dipartimento di Ricerca e Innovazione Umanistica (DIRIUM) dell'Università degli Studi di Bari Aldo Moro, consapevole delle responsabilità penali previste per le ipotesi di falsità in atti e dichiarazioni mendaci così come stabilito negli artt. 75 e 76 del DPR 28/12/2000 n. 445,

con riferimento al contributo dal titolo *A Pilot of Smart Digital Library Used-Centered: The Project SMARTER*, del quale questa Dichiarazione costituisce allegato integrante e asseverativo ai sensi della normativa in vigore,

DICHIARA:

- che è responsabile scientifico della ricerca della quale nel contributo si presentano stato dell'arte e risultati;
- che è autore principale del contributo e ne ha redatto e scritto integralmente il testo originale in tutte le sue parti;
- che i co-autori hanno collaborato alla redazione definitiva dei seguenti paragrafi:
 - Prof. Stefano Ferilli: parr. 3.1; 3.2.2; 3.2.5
 - Dott. Tommaso Caldarola: parr. 3.2.3; 3.2.4

Bari, 29/05/2023

In fede

A Pilot of Smart Digital Library Used-Centered: The Project SMARTER

Nicola Barbuti¹, Stefano Ferilli¹, Tommaso Caldarola²

¹ University of Bari Aldo Moro, Piazza Umberto I, n. 1, Bari, Index, Italy

² D.A.BI.MUS. Ltd., Piazza Umberto I, n. 1, Bari, Index, Italy

Abstract

The paper presents the results of the national PoC project SMARTER, aimed at prototyping a smart DL for the management, interaction, and preservation of digitized and born-digital resources, related to ancient printed and manuscript artefacts. The research focused on the development of an innovative metadata schema structured by integrating the languages of the semantic web with conceptual ontologies, and on the experimentation of the applicability of the ICRPad intelligent recognition system (Pat. UIBM n. 0001407881) to large collections of digital objects, with the aim of making them interoperable with each other and, at the same time, usable also through direct interaction with the contents of the metadata. For the interaction with the digital collections, a set of innovative methods and technologies for management and display on the web has been designed, which allow users to interact with the digitized content through advanced tools.

Keywords

Smart Digital Library, SMARTER DL, Graph DB, ICRPad

1. State-of-the-art

The digitization of cultural heritage (CH) and the creation of digital libraries (DL) are fields of renewed and growing interest, especially related to galleries, libraries, archives, and museums (GLAM). The recent PNRR also provides for substantial financing on these fields, aiming to regenerate the relationships and interactions between the citizens and the CH by pouring online digitized collections.

The newborn Central Institute for the Digitization of Cultural Heritage – Digital Library has recently started to rethink the entire ecosystem of the digitization of heritage, starting from the redefinition of the processes of creation and publishing digital objects¹.

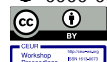
Although there is a twenty-year tradition of scientific studies on this topic [1] [2], even the recent scientific literature is focused on standardized models and processes, especially relating to data management and user interaction [3] [4]. The CH digitization methodology and practices used are still heterogeneous and dissimilar [5] [6]. The indexing of digital objects with metadata is only focused on the description of the original artifacts represented in the layouts of the data [7]². The users enjoy the digital collection by a passive consultation of pre-packaged data, with scarce interaction almost limited to the download of the PDF files of the digital objects.

So, today several DLs float on the network which completely improper are considered Digital Cultural Heritage (DCH) only because the digital objects reproduce cultural artifacts in their layouts. Instead, as poor attention is paid to information relating to the digital objects, which are original creations and require specific descriptive criteria, these DLs basically contains *digital twins* of the

IRCDL 2022: 18th Italian Research Conference on Digital Libraries, February 24–25, 2022, Padova, Italy

✉ nicola.barbuti@uniba.it (N. Barbuti)

ORCID 0000-0003-0817-4235 (N. Barbuti)



© 2022 Copyright for this paper by its authors.

Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

¹ <https://digitallibrary.cultura.gov.it/notizie/le-nuove-digital-libraries-siano-imperfette-ma-stimolanti-e-coerenti/> (last consultation: 7 October 2021).

² See the Italian standard MAG and international standards Dublin Core, METS, MODS, xDams.

heritage [7], which cannot be defined as cultural entities. This carelessness heavily affects the quality of digital objects, whose metadata always lack information on their life cycle, and their preservation over time, that is still an unsolved issue of the digitization [8] [9].

In this scenario, some recent DL projects arise as interesting attempts to evolve towards models for the enhancement of digital cultural objects, as the resources are linked by LODs³ based on the RDF model⁴. But once again, the descriptions are focused only on the artifacts represented in the layouts, with scarce information on the digital objects.

2. Motivation

The project SMARTER DL (SDL) we outline in this paper started in 2019 by these ongoing scenarios. The goal of the project was to study a pilot of interactive smart DL model for the management of, and advanced interaction with digital resources related to the GLAM heritage, evolving the architecture of a DL designed and tested in 2014 [10].

The research addresses a double perspective of innovation:

- prototyping an SDL model for managing digital resources related to cultural heritage with an innovative metadata schema, studying triples based on conceptual abstractions related among themselves, with the goal to overcome the standards, and to generate a model of description of digital objects that can be dynamically implemented over time;
- testing the tools of a patented system for the intelligent recognition of ancient handwritten and printed texts represented in the digital objects, ICRPad [11] [12], to extract and automatically index the hypertext, improving the users' interaction with the objects by querying and retrieving information from the layout.

The metadata schema has been designed developing formal ontologies that extend the classes, relations and attributes typically used in the cultural sphere. The schema expands the context information of the original artifacts both through conceptual relationships that connect it to other resources, and by collecting and making available information on *provenance* representative of the data life cycle [13]. The goal has been to outline a metadata structure that dynamically add cultural content to digital artifacts, in order to evolve current databases into *knowledge bases*, in which the descriptions of the original artifacts are expanded by the conceptual relationships, and at the same time they continue to be related to the physical and contextual descriptions of the data.

The use of Artificial Intelligence (AI) solutions was also a design-driving choice. Several kinds of AI techniques were involved: semantic technologies to describe and manage the library items, user modeling and profiling to tailor the behavior of the library to the single users, semantic information retrieval to overcome limitations of lexical approaches, machine learning and data mining to extract knowledge from documents and users, to group documents and users, etc.

The dynamic recording of information in the metadata schema aligns to the perspective of making the SMARTER DL available in the long term, giving digital resources the function of records consistent with the *FAIR* principles [14] [15] which, over time, can assume the value of the new DCH.

3. The SMARTER DL (SDL)

3.1. Metadata Schema

The innovative metadata management approach includes a technological platform providing both facilities for data storage and manipulation and an ontological layer opening new possibilities of automated reasoning for best serving the needs and purposes of its different kinds of users (managers, librarians, end users, etc.). This platform is named GraphBRAIN [16].

The technological platform relies on a Graph DataBase. For GraphBRAIN we adopted Neo4j [17]. Compared to traditional Relational DBs, more oriented toward batch processing of structured information, Graph DBs boost performance of instance-oriented data processing, allowing to navigate through data items efficiently and effectively by exploiting different kinds of binary associations

³ See the recent Catalogo Generale dei Beni Culturali of the ICCD <https://catalogo.beniculturali.it/> (last consultation: 7 October 2021).

⁴ https://it.wikipedia.org/wiki/Resource_Description_Framework (last consultation: 7 October 2021).

between them. Neo4j is based on the so-called Labeled Property Graph (LPG) model, where nodes and arcs in a graph may be labeled (usually the label represents the type of the instances) and associated to sets of attribute-value pairs. This is a very powerful data representation model, successfully adopted by big players in the industry. Differently from traditional DBs, Neo4j does not work based on a scheme for the data to be stored. Any label and any set of attributes can be associated to each node or arc, and nodes or arcs of the same type may involve completely different attributes.

Since schemes are extremely important to determine a meaningful structure in DBs, and to help data designers and managers properly organize their operations, in our platforms we provided for allowing schemes in the form of ontologies to be superimposed on the Graph DB, so that only information that is compliant with the scheme can be added. An additional facility of GraphBRAIN is allowing several schemes to be superimposed on the same graph, to express different domain-dependent perspectives on the same data.

While much research is available on ontologies in Computer Science, especially in the Knowledge Representation and Reasoning branch of Artificial Intelligence, the current standard formalism for representing ontologies have some idiosyncrasies with respect to the underlying LPG model. In fact, the standard formalism consists of triples <Subject,Relation,Object> (usually expressed in RDF format) where the components are atomic. This is partly incompatible with the LPG model (e.g., properties cannot be attached to relations, unless particular workarounds are adopted).

To bridge the gap between RDF and LPG, we defined a specific formalism for GraphBRAIN schemes [18], that can fully exploit the power and flexibility of LPGs while still allowing a relevant and useful mapping to RDF and its ontological declinations such as OWL.

By applying the ontological schemes to the data in the graph DB, we obtain a so-called Knowledge Graph, that allows advanced exploitation of the information by expanding the possibilities of standard graphDB Manipulation Languages (Cypher for Neo4j) with actual high-level logical reasoning that can infer information not explicitly expressed in the database. In particular, we envision the joint use and cooperation of various kinds of inference strategies: ontological, deductive, abductive, argumentative, etc.

Given these premises, for the SDL we started the design of a GraphBRAIN schema that goes beyond the standard bibliographic metadata currently in use, expanding them in several directions. The most important ones are allowing to store the entire lifecycle of the cultural objects in the library, and several kinds of contextual information. The former is important to fully capture and exploit the peculiarities of digital cultural heritage, that is to be considered a cultural object by itself, not just a representation of some physical item. The latter is important to allow an exploitation of the information that is not limited to single records independent of each other, but provides several kinds of interconnections between different items, direct (e.g., having the same contributor, or period, or publisher) or indirect (e.g., citing persons who lived in the same place in the same period). The former would allow to preserve the history and unity of an object along the years and through different interpretations and exploitations, which is yet more difficult for digital items that do not have a physical identification. The latter would support researchers, scholars, managers, or simple users in their activities, suggesting relevant items to consult or even proposing non-obvious research directions worth investigation.

Especially relevant to this ‘holistic’ and contextual perspective is the possibility of combining, in GraphBRAIN, different schemes. So, for instance, using together the GLAM, Tourism and Food schemes might support the specific consultation and exploitation of the library’s content for touristic purposes, joining the cultural aspects and the more experiential ones connected to the folklore and traditions of the place being visited. Also, depending on the different schemes that are combined, different connections can be found between objects in and around the library collection.

The defining ontology provides for a set of concepts and relationships that go beyond what has been proposed so far, both in the description of analog artifacts and in the LOD context. One part is aimed at describing the life cycle of digital objects, including the activities and actors who are part of it in various capacities. Other elements expand the range of context information used today. Elements are provided to describe and manage not only GLAM assets, but also users and their characteristics, to allow AI technologies to adapt the behavior of the system to the specific needs and purposes of each user.

The schema is undergoing further development and expansion, and to date contains information relating to the GLAM, tourism and history of computer science domains [18]. Although still under construction, the ontology includes 61 classes and 161 reports. The overall consistency is 336483 class

instances, described by a total of 1875571 attribute values, and 496564 relationship instances, described by a total of 41301 attribute values. Figure 1 shows a section of the graph connecting entity instances via relationship instances. Different colors of the nodes represent different classes.

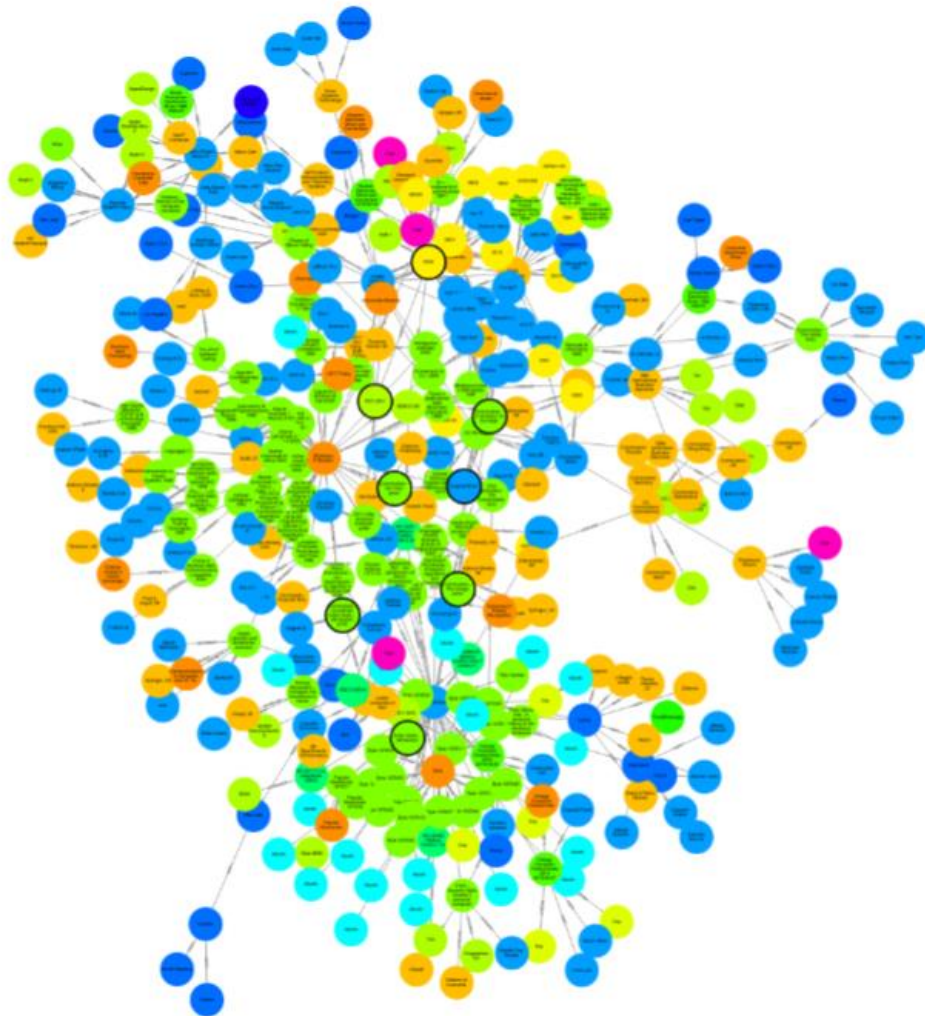


Figure 1. Section of the graph

3.2. The SMARTER DL (SDL) prototyping

The definition of the new metadata scheme made it possible to focus the design of the SDL pilot in a user-centered perspective, providing a set of methodologies and technologies for managing and displaying resources on the web aimed at encouraging advanced user interaction with the collections.

From a functional point of view, the project focused on two main lines of activity:

1. analysis, choice, optical acquisition, metadata solutions both of the semantic web and of conceptual ontologies, conservation of digital contents relating to documentary cultural heritage;
2. enhancement of the contents available through display on the web with innovative interactive consultation and search functions, usable with normal browsers and prepared for use through special apps on mobile technologies.

The DL prototyped in the 2014 project, whose architecture was developed using the open source DLMS dSpace, represented the basis of piloting, outlined according to the following steps:

1. analysis and choice of digital objects in relation to different types of ancient textual content represented in the layout (manuscripts and printed);

2. testing the integrability of the functionality of the ICRPad application in the DL and interoperability with metadata;
3. analysis of AI solutions for the conceptual description of digital resources and the interactive and dynamic use of contents;
4. design of the prototype;
5. analysis and evaluation of the prototyping effectiveness.

The architecture and functionality have been enriched as shown in the following paragraphs.

3.2.1. The SDL Architecture

In the design of the SDL the three-level architecture of the previous model have been preserved, adding some further subsystems.

The *Application level* includes the access tools both to the back-end area and to the front-end. The back-end allows the upload and modification of digital content and associated metadata, the management of digital services and of users who access them. The front-end allows the visualization, rendering and use of digital objects and services through a web interface that can combine all the required digital formats, also offering interaction with multimedia and multi-channel resources thanks to collaborative tagging. Web-responsive tools can enable adaptive viewing, rendering and interaction via mobile devices. By monitoring tools, it is possible to manage reports on the interaction with the system by users, in order to better define their needs and request for the inclusion of new content or the activation of digital services. Interoperability tools manage the interfaces that allow the exchange of metadata through the OAIS, OAI-PMH, OAI-ORE and Z39.50 protocols. Furthermore, it has been further implemented to manage digital objects indexed with Open Data (OD) and Linked Open Data (LOD), and with the new schema above described.

The *Management Logic Level* includes modules for implementing system functionality and basic tools for configuration and logging. The level can provide the *information retrieval* by both simple and advanced query; the management of *collaborative tagging* to improve user interaction with digital objects; the *access management* through authentication and advanced user profiling tools; advanced user interaction applying ICRPad plugins for intelligent recognition and text extraction from the layout of objects; the management of indexing and description of digital objects with the schema we are implementing; the possibility of georeferencing digitized resources through interoperability with the main online open source platforms (e.g., Open Street Map), and of searching for contents based on spatial queries.

The *Storage Level* manages the organization of digital objects and their metadata, the user information and associated permissions, and the status of the approval flow when someone insert digital objects into the DL. The level provides for the management of authentication, access and user profiling through specific tools. User management tools will verify the authentication and authorize access to the user profiling him. The authorized user can also propose the insertion of digital objects in the system through a dedicated function in the user interface. The proposal will be analyzed by *content manager* tool, which will validate it and activate the process for publishing it in the collection. Authorized users can also have access to monitoring data. For the harvesting operation during the fruition of the contents, interoperability and metadata converter tools can be provided, which can also be used for the physical transfer of metadata in batch mode and for exposure as OD and LOD.

All content and user data are managed through the storage API modules.

In the following paragraph the description of the designed modules is outlined.

3.2.2. Fruition

Fruition is the module that can provide the Web interface for the distribution and the interactive use of multimedia digital collections and services.

In the SDL, this module has been improved for supporting the set of multimedia digital objects, also providing multi-channel management and data protection. The collections and services can be described with adequate information consistent with the standard and good practices for managing digital resources of CH. Moreover, the descriptive tags associated with each one object allow an easy

information retrieval not only by querying author, date, topic referring to original artifacts, but also by other elements relating to usable formats, dimensions, resolutions and, in perspective, textual content of the digital objects.

The module will support the management of the following multiple classes of digital objects, with related formats and functions:

1. text files, e.g., TXT, PDF, OOXML, ODF, Djvu formats;
2. 2D high resolution and uncompressed formats (TIFF, GIF, FITS, PNG, BMP); tools for the basic post-processing of images can be integrated, aiming to produce by each object other formats with different resolution applying dynamic resizing to the required resolution, and to generate JPEG lossy versions for exposure online; a bidirectional conversion of storage formats TIFF to FITS and reverse, and export to JPEG can also be provided; the ICRPad tools for graphic matching and intelligent recognition will allow the real-time interaction both with the layout of the digital object, and with hypertext extracted by the content textual content, which can be also indexed with metadata schema we are implementing;
3. most used audio formats (e.g., WAV, MP3, ABS, MPA); web interfaces for fruition via streaming will integrate open-source libraries to optimize the quality of audio reproduction even on mobile technology; tools are provided for podcasting audio contents;
4. several audio-video formats (e.g., MPEG3, MPEG4, H26x, Quick Time, AVI, streaming formats); the module is prepared for integrating open-source libraries that will allow video playback of supported formats also on mobile technology; on the back end, the integration of specific open-source libraries will allow to insert captions, subtitles and alternative content into the digital objects; these can be also modified by changing the compression and extracting video sequences.

Tools have been designed for grouping digital objects into albums and alternative contents.

In future perspective, this module will also manage formats for interactive 3D contents, such as, for example, virtual routes, or panoramic images, etc., by integrating libraries for the 3D rendering and the user interaction.

Another innovation designed is the implementation of specific tools to support the interaction of Fragile Subjects with the digital collections by mobile devices. Open-source libraries will be integrated that will allow text-to-speech.

3.2.3. Recognition Subsystem (SSR)

The Recognition Subsystem (SSR) is a software module of the SDL that allows you to perform Intelligent Character Recognition (ICR) functions on documents in electronic format for handwritten and printed text. The main features offered by the SSR module are listed below.

Languages

The system allows you to extract text from images or documents in mixed language, however it is possible to specify the specific language as input. For the printed version the recognition is enabled for more than one hundred languages while for the handwritten the functionality is provided only for the Latin languages.

Natural order of recognition

For several ancient and modern languages, e.g. Latin, it is possible, optionally, to specify an input parameter that allows you to perform the recognition respecting the nature of the documents from left to right ensuring the sorting by columns.

Range of pages

If you want to extract a portion of text from a single or a multi-page document, or a collection of documents, you can specify a range in the layout of one image for obtaining exactly the occurrences by the overall set You have searched.

Supported Document Types

The supported file formats are: JPEG, PNG, TIFF, BMP and PDF. For PDFs and TIFFs, documents containing at most 1000 pages are allowed and the size of each file must not exceed 50 MB with minimum and maximum dimensions of 50 x 50 pixels and 10000 x 10000 pixels respectively.

The main technical requirements of the SSR module are listed below.

Asynchronism

The POST operation with respect to the GET operation takes place in asynchronous mode. The invocation to the input API returns an identifier to be used for the reading API.

Output format

The read operation returns information in JSON format decorated with the following information, in order to be able to apply, on the client side, the functional logic appropriately:

- lines: list of text lines for each page
- words: list of words by Line
- region: a BoundingBox for Line or Word that shows the type coordinates (X, Y) of the box that encloses the element within the page
- text: the text contained in the Line or Word
- score: the recognition goodness value between 0 and 1 for each element
- width: the value of the width of the single page
- height: the height value of the single page
- page: page number.

Storage

The document repository must be reachable via HTTPS by the service performing the acknowledgment. Even if it is read-only, it is advisable to protect the repository, for example as happens for the storage hosted by the main Cloud providers.

3.2.4. Customization

Even this module has been improved for better managing the access of users in the DL, and monitoring their interaction with innovative advanced services and digital collections. Each user will be identified by login and password.

Compared to the previous, this version implements advanced tools to profile user behavior during interaction with collections and digital resources, through the application of advanced AI techniques. From the data analysis, specific information about each user can be inferred, such as, for example, particular interests, interaction preferences, purposes, etc. By the AI tools, users and contents can be automatically grouped into clusters based on the information inferred, allowing to customize the services in relation to the different clusters and their levels of interaction, and to associate specific modules and functions to each one cluster.

To this goal, the use of *groups* is envisaged, consisting of explicit lists of users. By the groups, the different types of access to the system can be profiled, each user would be associated with a profile consistent with the role defined, information can be extracted, that will be used to aggregate groups into *clusters*, also in order to customize the services on different needs. Anyone who belongs to a group will get the privileges granted to it.

The DL will try to automatically infer the belonging of a group to a cluster, and to propose specific methods and tools for each cluster, based on the respective characteristics and needs.

Clusters can be identified based on a classification, e.g.:

1. Common user
2. Researcher
3. Paleographer
4. Historian
5. Archivist
6. Bibliologist

7. ...

A collection, as well as a document, can be associated with multiple groups and multiple clusters identified as relevant. It will also be possible clustering contents and users to automatically identify new bottom-up aggregations, based on similarity by description or by content of the instances.

If users interact with one or more objects of a collection by proposing customized implementations, a flow-check will be activated, that will allow one or more reviewers to make sure that the implementations are consistent with the other contents of the collection. The control will track all the changes that digital objects may undergo from the moment the user proposes the implementation, and it will prevent multiple users from simultaneously performing operations on the same object that could change its status.

The profiling and customizing data will be managed with metadata, with a view to recording and preserving information on the *provenance* and *life cycle* of the resources, favoring their *historicization*. Monitoring tools have been designed that will allow to store and analyze information on the interaction, in order to obtain statistics and reports to be used according to several aggregations, among them:

- number of users accesses;
- complete report of all actions performed;
- number of digital objects used;
- number of searches carried out;
- number of OAI requests;
- more frequent searches;
- log information;
- processing information.

The results of the statistical analysis can be collected in reports and used to identify the user needs. They may also be made available through the user interface in public mode or reserved for the administrators of the DL.

Tagging features have been designed to improve user interaction, so each one can assign one or more labels to different contents. The tools for inserting and displaying tags will be included in the responsive web interface.

3.2.5. Data management

The organization of data in the SDL has been prototyped based on a model that, in addition to allowing its population, will facilitate the indexing and description of digital objects through the metadata schema we are implementing. In the perspective to interlink the SDL with other digital libraries and collections already indexed, the metadata schema encloses elements consistent with the most important standards used, such as MAG, Dublin Core, METS, MODS, RDF.

Tools that support the interoperability of the schema with the OAI-PMH, OAI-ORE, OAIS and Z39.50 protocols have been designed, in order to have massive transfer of data to and from the SDL, both through back-office and through web query. In particular, the OAI-PMH protocol guarantees interoperability between the various providers to expose and collect metadata. This implementation will ensure the exchange of contents with the major national and international digital libraries (e.g., Biblioteca Apostolica Vaticana, Catalogo Generale dei Beni Culturali, EUROPEANA, Library of Congress, World Digital Library).

The extraction of hypertexts from digital objects by ICRPad tools will also allow for indexing and description of textual content, greatly expanding the potential for user interaction. Two modules have been designed to be plugged into the SDL: Document Repository Service module (DSR) and Recognition Service module (RS). The DSR module will expose the document repository to RS module through an URI (cloud or on premises). The RS module is composed by a set of API that can be invoked in order to perform recognition against documents stored into a repository.

The ICR tools can be plugged/unplugged into the system via simple configuration by using *plug-in* architecture. The model consists of two types of components, a *core system* and *plug-in modules*. Application logic is divided between independent plug-in modules and the basic core system, providing extensibility, flexibility, and isolation of application features and custom processing logic. Plug-ins can

be added or removed from the core anytime. One plug-in's addition or removal from the core does not affect the other plug-ins.

The digital resources will be searchable by the set of *information retrieval* techniques used for the targeted retrieval of information. The user who has a specific information need can query the system, that will perform a search in its internal indexes and will provide in response one or more digital items, which represent the entities containing the information. If the user will not be satisfied with the results inferred, because they are too scarce (silence) or huge, or contain incorrect and disturbing elements (noise), he can perform a new query or refine the previous one by inserting filters. The filter function will be based on Boolean operators.

If the digital resources lack the specific keyword entered by the user in the query, but the text has terms related to it, advanced indexing techniques based on semantics have been provided which greatly enhance the system's *information retrieval* functions.

For improving the interaction of users, the DL will support a *multiple layering* query interface:

- simple search: it will allow the search by entering keywords; the combined use of Boolean operators will allow to vary the extent of information retrieval in relation to the needs of the user;
- advanced search: it will allow to optimize the search by selecting some characteristics of interest; the function will be divided referring to the metadata, but the user can use them in a combined way to refine his searches, e.g. as it follows:
 - a. search for descriptive elements (author, title, and other descriptive elements included in the metadata standards that will be defined and used);
 - b. search by classes of digital objects (documents, images, audio, video, photographs, etc.);
 - c. search by metadata elements: if a common high-level schema is used using specific semantic web languages capable of associating a formal meaning to the metadata, the search will also allow the retrieval of information connected with the same descriptive metadata;
 - d. search by content: it can be carried out on a single digital object or on sets of simple or complex digital objects, both of the same class and of different classes; for this function the search engine will use the output of the ICRPad application;
 - e. document layout analysis aimed at the selective extraction of the content: this tool will allow the system to automatically select parts of the layout that identify the relevant information of the digital document, in order to limit the subsequent processing (in whole or in part) to these content, reducing processing times and increasing the quality of the result.

In order to ensure the easiest use of the SDL collections and services, it will be possible to aggregate the data in collections consistent with the areas of interest of each institution that uses the model. The collections can also be implemented on the basis of information detectable by user interaction.

Some areas identifiable for the designed SDL could include, among the other:

- Environment and Landscape;
- Archeology;
- Art;
- Creativity;
- Music;
- Cartography;
- Architecture;
- Literature;
- Show;
- Economy and society;
- Places of culture;
- Events;
- Food and wine;
- Handicraft;
- History and traditions;
- Sports.

3.2.6. Georeferencing

This module will allow to georeference all the contents stored in the SDL. In particular, for each stored content it will be possible:

1. to associate one or more geographical positions to the information dealt within the digital object;
2. to search for groups of objects using queries based on geographical proximity;
3. to record the georeferencing data in the metadata and store it.

A graphic component will be implemented, accessible via computer and mobile devices, which will also allow users to carry out the aforementioned activities. The graphic interface will be created using web frameworks, in order to facilitate access from a mobile device to the functions implemented. The generated positions will be displayed using open-source mapping services, such as Open Street Map.

Compared to the previous version, tools for *on-the-fly* georeferencing via mobile devices have been designed, in the perspective to allow users to georeference original artifacts related to digital objects displayed in the SDL by being in the physical place where they are located, by selecting a specific function on their device. In addition to greater precision, this function would improve the precise mapping of cultural objects located in positions not clearly identified through the conventional address system (eg, a stele in the open countryside).

3.2.7. License management

Each digital object, collection or service of the SDL will be associated with the type of release license, depending on whether the user is a person or a legal entity. During the data entry it will be possible to choose or customize an existing license or add a new license. Since the data entry and the activation of services pass through an approval flow, the mandatory association of one or more license agreements/forms to each content will be provided. Some content manager tools will check that the content sharing is fully or partially consistent with the relevant license agreement/form. The text of each license will be available on the web and mobile interfaces. During the interaction with any digital object or the fruition of a service, the licenses can be viewed through a link (or button) clearly identifiable in the layout of the digital object.

3.2.8. Repository

The preservation and persistence of data on relational databases will be provided by a storage layer (SL, e.g., Storage Resource Broker), having care of the integrity of data and metadata for saving and searching for stored contents.

The contents will be stored within the SL. The infrastructure of the SL will be used to create a distributed logical file system, where all the digital contents of the SDL will be stored. This solution can be useful as it allows to physically distribute and replicate content, and at the same time to keep backup copies of the contents without the need to purchase any backup service.

4. References

- [1] A. Salarelli, A. M. Tammaro, *La biblioteca digitale*, Milano: Editrice Bibliografica, 2006
- [2] A. M. Tammaro, User perceptions of digital libraries: a case study in Italy, *Performance Measurement and Metrics*, Vol. 9, 2 (2008), 130-137. doi: 10.1108/14678040810906835.
- [3] Xie, Iris, Matusiak, Krystyna, *Discover Digital Libraries. Theory and Practice*, 1st ed., Elsevier, 2016.
- [4] M. T. Biagetti, *Le biblioteche digitali. Tecnologie, funzionalità e modelli di sviluppo*, Milano: Franco Angeli, 2019.
- [5] J. Bloomberg, Digitization, digitalization, and digital transformation: confuse them at your peril, *Forbes*, 29th April 2018. URL:

<https://www.forbes.com/sites/jasonbloomberg/2018/04/29/digitization-digitalization-anddigital-transformation-confuse-them-at-your-peril>

- [6] N. Barbuti, M. De Bari, La digitalizzazione che non c'è, *Biblioteche Oggi Trends*, Vol. 7, n. 1 (2021), 71-80. doi: 10.3302/2421-3810-202101-071-1.
- [7] N. Barbuti, *La digitalizzazione documentale. Metodi, tecniche, buone prassi*, Milano: Editrice Bibliografica, 2022 (in press).
- [8] L. Duranti, E. Shaffer (eds.), The memory of the world in the digital age: digitization and preservation, in: *An International Conference on Permanent Access to Digital Documentary Heritage*, UNESCO Conference Proceedings, Vancouver, 26–28 September 2012. URL: http://ciscra.org/docs/UNESCO_MOW2012_Proceedings_FINAL_ENG_Compressed.pdf
- [9] L. Bailey, Digital Orphans: The Massive Cultural Black Hole on Our Horizon, *Techdirt*, 13th October 2015. URL: <https://www.techdirt.com/articles/20151009/17031332490/digitalorphans-massive-cultural-blackhole-our-horizon.shtml>
- [10] N. Barbuti, T. Caldarola, D. Re David and S. Ferilli, An Integrated Management System for Multimedia Digital Library, *Procedia Computer Science*, Vol. 38 (2014), 128-132. doi: 10.1016/j.procs.2014.10.021
- [11] N. Barbuti, T. Caldarola and S. Ferilli, A Graphic Matching Process for Searching and Retrieving Information in Digital Libraries of Manuscripts, in: G. Serra, C. Tasso (eds), *Digital Libraries and Multimedia Archives. Proceeding of the 14th IRCDL 2018*, Communication Computer and Science, Vol. 806 (2018), 139-150 doi: 10.1007/978-3-319-73165-0_14.
- [12] N. Barbuti, T. Caldarola, An Innovative Multifunction System for Text Recognition of Digital Resources Reproducing Ancient Handwritten and Hand-Printed Artifacts, *Proceedings of the 1st DTUC '18*, ACM (©2018). doi>10.1145/3240117.3240141.
- [13] F. Tomasi, La preservazione del contenuto degli oggetti digitali: formalizzare la provenance *Bibliothecae.it*, 6 (2017), 17–40. URL: <https://cris.unibo.it/retrieve/handle/11585/611249/303579/paper-2017.pdf>
- [14] N. Barbuti, Creating Digital Cultural Heritage with Open Data: From FAIR to FAIR5 Principles, in: M. Ceci, S. Ferilli, A. Poggi (eds.), *Digital Libraries: The Era of Big Data and Data Science. Proceedings of 16th IRCDL*, Communications in Computer and Information Science 1177 (2020), 1-9.
- [15] N. Barbuti, Ripensare i dati come risorse digitali: un processo difficile?, in: *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica*. Milano: Università Cattolica del Sacro Cuore (2020), pp. 19-23.
- [16] S. Ferilli & D. Redavid. The GraphBRAIN System for Knowledge Graph Management and Advanced Fruition, in: *Foundations of Intelligent Systems, Lecture Notes in Artificial Intelligence* 12117, 2020, 308-317.
- [17] I. Robinson, J. Webber, and E. Eifrem, *Graph Databases*, 2nd ed.; O'Reilly Media: Sebastopol, CA, USA, 2015.
- [18] S. Ferilli. Integration Strategy and Tool between Formal Ontology and Graph Database Technology, *Electronics*, 27 pp., MDPI, 2021.