Document Processing and Knowledge Management for Digital Libraries

Stefano Ferilli[1,*,†], Eleonora Bernasconi[1,†] and Domenico Redavid[1,†]

[1]University of Bari, Bari, Italy

Abstract

This contribution reports the findings of research carried out by our research unit within the CHANGES project, ongoing under the NRRP program funded by the NextGenerationEU initiative of the EU. Specifically, the unit worked on the "Digital Libraries, Archives and Philology" sub-project in 3 directions: digitization of sources, transcription and analysis of texts, and creation of a digital environment and digital library. These findings concern the development and use of a new framework for knowledge bases representation and exploitation, of new tools for handwriting recognition, and an advanced interface for knowledge search, browsing and consultation.

Keywords

Document Processing, Knowledge Graphs, Character Recognition, Cultural Heritage

1. Introduction

The Italian Ministry of University and Research funded several research projects, each in a different area, under the National Recovery and Resilience Plan (NRRP) program funded by the NextGenerationEU initiative of the EU. Each project tackles a different area, and is organized in several sub-projects ('spokes'), all coordinated by a 'hub'. The project connected to Cultural Heritage is "CHANGES" (Cultural Heritage Active innovation for Next-GEn Sustainable society)[1], one of whose spokes (Spoke 3) deals with "Digital Libraries, Archives and Philology". A unit from the University of Bari, including both computer scientists and philologists, participated to the activities of Spoke 3. In particular, the Computer Science team, with experience in Artificial Intelligence solutions and applications to Document Processing, Digital Libraries and Archives, was involved in 3 WorkPackages of Spoke 3[2]:

WP1 Digitizing archival, bibliographic, textual and illustrated sources

WP2 Towards an automated transcription and analysis of handwritten texts and book forms

WP3 Creating a digital philology environment and digital libraries of authorized texts

Now that the project is approaching its completion, this paper showcases the findings of the team in these directions. While most of the research results were published in isolation and at specialized venues, this is also an opportunity to provide for the first time an integrated, overall description of the work carried out and of the connections among the specific endeavors. Technical details and/or experimental results for each of them can be found in the cited publications, or in their references.

In the following, one section will be devoted to each WorkPackage, before proposing some conclusions and outlining future work perspectives.

IRCDL 2026: 22nd Conference on Information and Research Science Connecting to Digital and Library Science, February 19-20, 2026, Modena, Italy

*Corresponding author.

†These authors contributed equally.

stefano.ferilli@uniba.it (S. Ferilli); eleonora.bernasconi@uniba.it (E. Bernasconi); domenico.redavid1@uniba.it (D. Redavid)

0000-0003-1118-0601 (S. Ferilli); 0000-0003-3142-3084 (E. Bernasconi); 0000-0003-2196-7598 (D. Redavid)

[1]https://pnrr.sns.it/progetto/changes/

[2]Albeit not formally involved in them, our unit also contributed to the other two WorkPackages of the Spoke, concerning linguistics, by proposing a knowledge graph-based approach to linguistic analysis, and specifically to diachronic analysis of semantic change. The findings of this part will not be described in this paper.

2. Digitizing archival, bibliographic, textual and illustrated sources

For WP1, the main contribution was not on digitization itself, but on the advanced representation and exploitation of the digitized materials, so as to better satisfy the needs of scholars, researchers, professionals and enthusiasts in the *Humanities*. The aim was to overcome the limitations of both traditional static approaches to description and cataloguing of bibliographical/archival resources, and those of the most widespread graph-based approach (the Semantic Web), mostly related to fragmentation and complex readability of the representation, to limited support to privacy/data protection (often required by the owners of cultural heritage) and personalization, and to limited reasoning capabilities [1].

Adopting a graph-based approach to knowledge storage, handling and fruition, and a mix of network analysis, data mining and automated reasoning solutions from research in AI, the team proposed GraphBRAIN [2], an innovative general-purpose framework that covers all stages and tasks in the lifecycle of a knowledge base, including knowledge acquisition, organization, and (personalized) fruition[3]. It pursues the same objectives as the Semantic Web [3], but using a more structured knowledge data model (Labeled Property Graphs, or LPGs) and state-of-the-art DBMS technology, to ensure efficient data handling and allow a wider range of AI tools to be applied. By using ontologies as data schemes, it associates a clear semantics to the graph nodes, arcs and attributes, and enables formal reasoning. Several ontologies, representing different perspectives on the knowledge stored on one graph, can be used and combined, allowing the system to connect knowledge across domains. The ontologies are expressed in a proprietary format, specifically tailored for LPGs. All the functions of the framework are exposed as services of an API that will be released for free use by third-party applications.

Standard Create/Read/Update/Delete functions are available to populate and consult the knowledge base, both focusing on single items (entity or relationship instances, or their attributes) or by posing complex queries. The knowledge base can also be fed in batch by automatic knowledge extraction from documents and other kinds of resources. Both ontologies and instances may be imported from, or exported to, the standard Semantic Web format, in order to support their interoperability and reuse as linked open data (LOD) [4]. Instances may also have attachments, making GraphBRAIN a digital library, whose content is organized according to formal ontologies, fostering interoperability with other systems. Users may add, show, or delete attachments. In a collaborative spirit, users may add comments on, or approve/disapprove, each piece of information in the knowledge base, including the ontological items. This feedback is used to assign a trust value to the users.

Several algorithms for graph mining, network analysis, information extraction, automated reasoning, natural language processing, etc. can be run on the available knowledge. Some of these algorithms are reused from the literature; others are original. Currently included functions are:

- assess relevance of nodes and arcs in the graph, and extract the most relevant ones;
- extract a portion of the graph that is relevant to some specified starting nodes;
- extract frequent patterns and associated sub-graphs;
- predict possible links between nodes;
- recommend relevant knowledge items;
- translate into natural language the information content of a portion of the graph.

If available, a user profile can be used to personalize the results of all these functions. Also, a data protection strategy is included to ensure that privacy and intellectual property are enforced. These features are typically not available in standard SW/LOD settings.

Within CHANGES, GraphBRAIN was extended, improved and refined, and applied to the representation and manipulation of Cultural Heritage knowledge. For this, a holistic approach was proposed, in which knowledge from several fields was connected, and standard metadata were expanded with information about content, context, physical features or intangible aspects, and usage of the cultural items. Its top-level ontology, defining very general and highly reusable concepts and relationships (e.g., Person, Place; Person.wasIn.Place; etc.), was extended by defining new ontologies concerning *Libraries*,

[3] A demo of the system can be found at `http://digitalmind.di.uniba.it:8088/GraphBRAIN/`

Table 1
Statistics on the elements in the ontologies involved in CHIPS&BITS

Domain	Ontology	Classes+Subclasses	Attributes	Relationships	Attributes
Top-level	general	13+128	107	33 (157)	43
Libraries, Archives & Museums	ifla	12+60	128	23 (77)	22
	lam	7+45	88	32 (143)	26
Education	lom	12+43	88	14 (22)	11
	oerschema	6+14	15	11 (31)	2
	alocom	5+35	6	5 (8)	0
	intelleo	33+102	156	33 (101)	6
	education	4+16	6	17 (35)	26
Tourism & Traditions	food	8+21	13	13 (34)	3
	tourism	4+118	29	7 (104)	12
Core	computing	7+101	66	22 (81)	33

Archives, *Education*, *Linguistics*, *Vintage Computing*, and *Book History*. While the latter was aimed at collaboration with the humanities teams involved in the project, the others were combined to support an application to the domain of Vintage Computing aimed at supporting the needs of the many kinds of stakeholders, including those related to the Web Economy [5]. These ontologies were developed, validated and evaluated by experts involved in the project, and the knowledge graph populated by collaborative work, resulting (to date) in 340582 nodes, 496844 arcs, and 837426 attributes.

3. Towards an automated transcription and analysis of handwritten texts and book forms

Populating a knowledge base for philological purposes may require to transcribe historical documents, which has been the main focus of OCR and transcription tools so far [6, 7, 8]. From another, less investigated, perspective the very characters and symbols in the document might be the object of study. Hence, the need to characterize, recognize, compare or study various kinds of symbols in digitized documents, especially handwritten text. The current state-of-the-art for OCR, focused on transcription of standardized characters, is often unable to deal with the variability and peculiarities of such sources, and not suitable to support the activities and specific needs of researchers. This motivated the development, for WP2, of two tools for symbol recognition. Both rely on established document image processing algorithms and techniques [9], and both place a strong emphasis on Explainable AI [10] aspects, in order to support the user in interpreting the outcomes and understanding which parameters were more or less relevant in the recognition process and decision.

The former [11] consisted of an on-line platform to analyze and segment the layout of ancient manuscripts, extract character symbols from them, automatically identify groups of symbols standing for the same character, and use Convolutional Neural Networks to learn models for such characters. Tools for setting up and running experiments, and for evaluating performance, are also provided. Finally, the learned models can be used for OCR on other documents of the same kind. The platform is specifically designed to allow its users to carry out batch processing on the documents.

A different perspective was used for the other tool [12]. It was an innovative solution for symbol recognition, specifically aimed at supporting the needs of scholars and researchers, rather than those of standard transcription. It leverages the specific shape features of the symbols, both external (contour) and internal (pixels layout). It is based on prototypes, so that even one example of a symbol is enough to have a model, which also improves efficiency. It is very flexible, and able to accommodate several kinds of variations and distorsions in the shapes, but at the same time extremely precise in recognizing perfect matches in the shapes, or in highlighting the causes of the differences. It is interpretable, meaning that the users can understand why a symbol was (or was not) recognized. It can provide an assessment of the similarity between a symbol to be recognized and a prototype, and return a ranking of possible

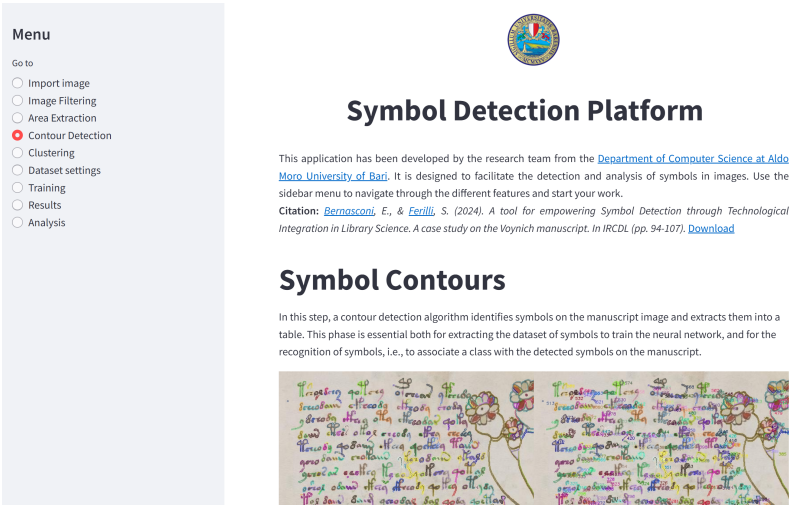

Figure 1: Interface of the batch symbol recognition and handling tool

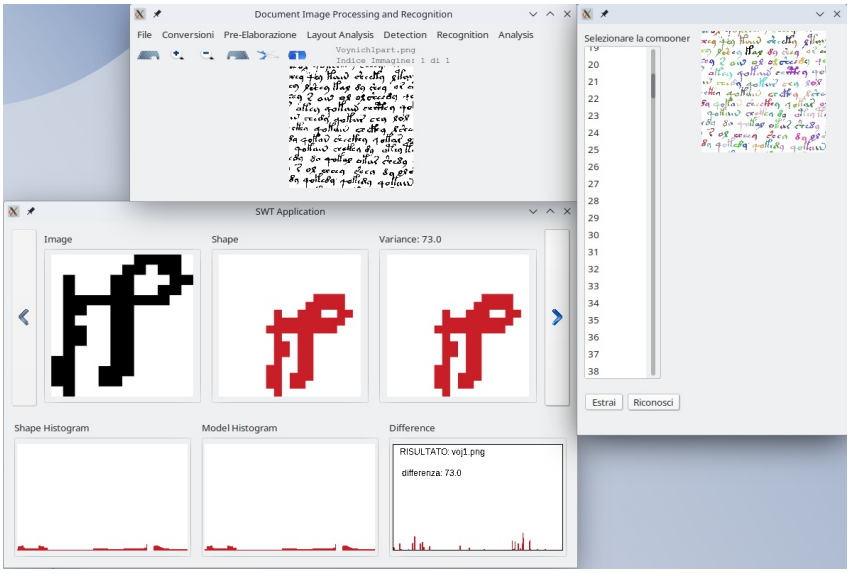

Figure 2: Interfaces of the prototype-based symbol recognition and handling tool

recognition options, each associated with its similarity value.

For both solutions, suitable Graphical User Interfaces were developed, featuring a wide range of image manipulation and enhancement functions that provide the user with full control on the document pre-processing phases, in addition to the specific controls for managing the model learning and application processes. In all cases, the user can interactively execute the various steps of the pipeline, and fine-tune the parameters in order to get the best results.

Figure 1 shows a screenshot of the Web-based interface for the batch symbol recognition tool. The menu on the left allows the user to select the various processing steps: image import, filtering, area extraction, contour detection, clustering, dataset setting, training, results and analysis. By selecting a step with the corresponding radio button, the panel on the right displays the associated information, control widgets and outcomes. In the screenshot, the contour detection step is selected, and the panel reports the document image in which each contour is highlighted using a different color.

Figure 2 shows a screenshot of the interfaces for the stand-alone, prototype-based symbol recognition tool. On the top of the screen, the binarized version of the document after pre-processing is shown. From this representation, the symbols are extracted and displayed in a different window (on the right of the screenshot), each in a different color, so that the user may act on each of them separately. After

selecting one and asking the system to recognize it, the recognition and manipulation window on the bottom-left is shown, where the user can see in the top row (from left to right) the extracted symbol, a graphical representation of its parameters, and the prototype responsible for recognition (normalized to fit the extracted symbol); on the bottom row, the histograms of the parameters of the symbol and the prototype are displayed, and the recognition result along with the information useful to interpret it (name of the recognized prototype, similarity value, and histogram of the differences between the parameters). Arrow buttons on the left and right of the window allow the user to scroll a ranked list of candidate prototype, by decreasing similarity.

As a testbed, both tools were applied to the Voynich manuscript [13], a book found in the XX century, but dating back to the XV century, reporting some images and handwritten text from an unknown script. This proved that they can be applied even to non-standard characters and when only few samples are available for learning. The latter tool was also applied to early printed books and to ancient writing on papyrus rolls. The extracted sets of symbols, along with their relationships to alphabets and to documents in which they were found, are stored in the GraphBRAIN knowlege graph.

4. Creating a digital philology environment and digital libraries of authorized texts

Our main aim in WP3 was overcoming the lack of user-centered environments that could give the user full control on the various steps of processing, also providing insights and explanations on the automated behavior, especially the AI part. Several modules were designed and implemented for this.

First, the administration interface for GraphBRAIN, in the form of a Web Application [2], was further developed and extended (some screenshots are shown in Figure 3). It reports statistics on the knowledge base (current content, quality, contributions and trustworthiness of the users), and allows the users to select the ontology to work with (screenshot on the top-left). It provides the users (especially the administrators) a tool to create, build and maintain the ontologies (screenshot on the top-right). It includes a form-based section that allows users to manually insert/update/remove instances and to query the knowledge base (screenshots in the middle row). Also, a dahsboard from which several graph algorithms and automated reasoning tasks can be run, and a panel that allows users to display a portion of the graph, browse it interactively, display detailed information about entity and relationship instances and, again, run network analysis and data mining algorithms (screenshots in the bottom row). So, both structured and serendipitous approaches to data are supported. Another version of the interface, that can work on standard Semantic Web repositories, albeit with a reduced set of functionalities, was developed [14] (a screenshot is shown in Figure 4).

Since GraphBRAIN is also a repository of documents, in addition to querying and browsing the knowledge, also adding functions to extract insights on the content of the stored documents, and to help interpreting them, may provide valuable support to scholars and researchers. To fulfil these needs, another product of research was a Web-based tool [15] to set up and run various kinds of analytics on the content of digital libraries, and to visualize the results of the analysis at a high level of abstraction. Leveraging external data and metadata repositories, and driven by ontologies, it uses state-of-the-art solutions in AI (including Natural Language Processing, Large Language Models, Topic Modeling) and Visualization (e.g., Trend Plots, Word Clouds and Intertopic Distance Maps) to carry out these functions on (possibly large) corpora of texts. AI plays a crucial role in effectively and efficiently automatizing these functions, that would be hard to carry out manually. Some examples are (see Figure 5 for screenshots of the tool interface associated with these functions):

- displaying the geographical location of the sources (a);
- showing the count of documents per year and associated trend, with several kinds of filtering parameters (b);
- plotting the relevance of topics along the years, and identify the associated trends (c);
- clustering the documents, displaying the clusters' topology and describing the clusters through their associated keywords and most frequent terms (d);

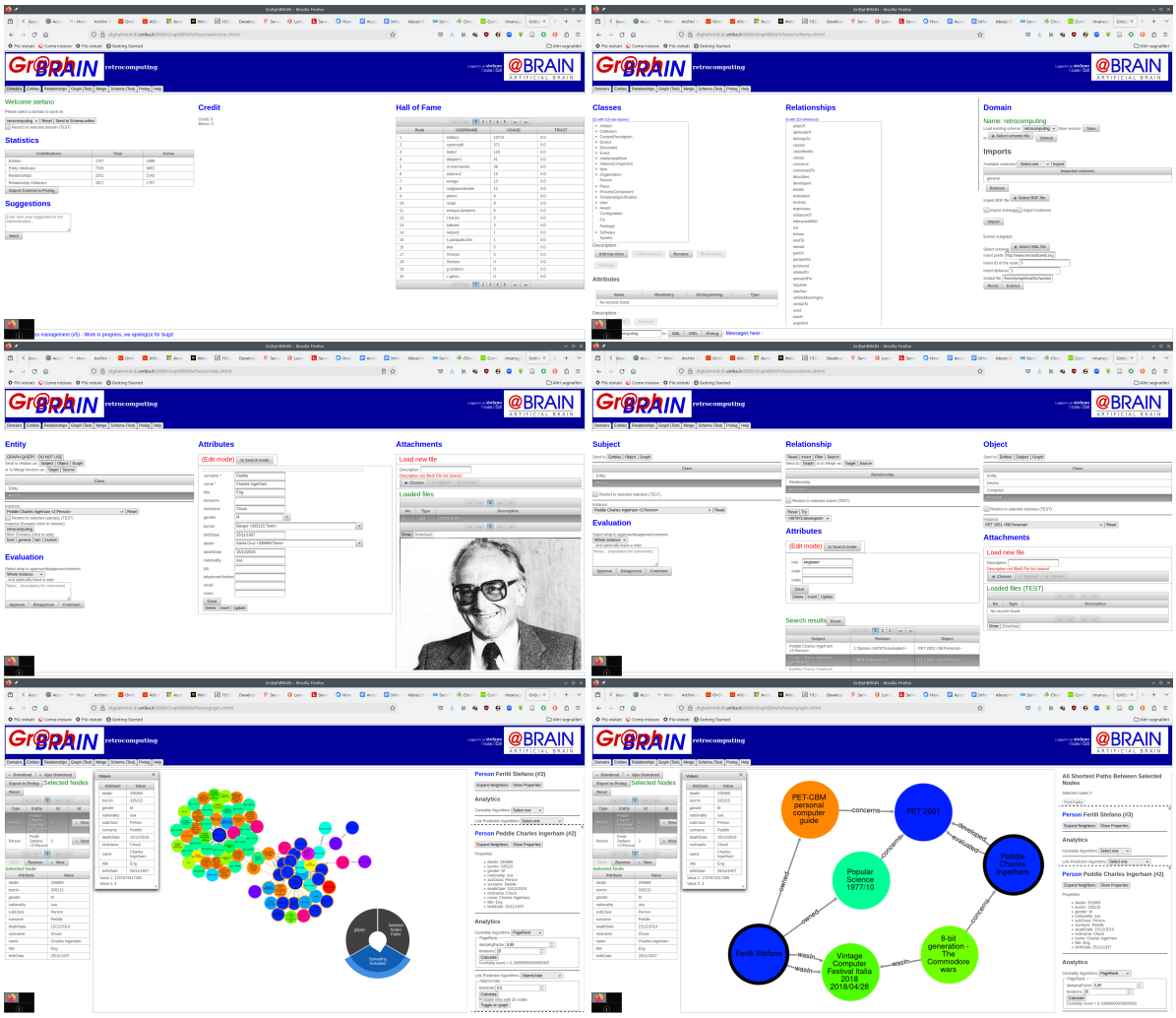

Figure 3: GraphBRAIN interface for browsing the knowledge graph and apply functions

- managing the collection of documents, providing at a glance keywords, topics and affiliations for each of them (e);
- describe and compare the library content along the years by means of word clouds (f).

The tool was successfully applied to analyze two decades of scholarly articles from IRCDL, an established conference on Digital Libraries, allowing us to understanding past and current academic research, and to predict future trajectories in this field [16].

5. Conclusions & Future Work

Application of Computer Science, and especially Artificial Intelligence, solutions to Cultural Heritage is still an under-explored field of research. Within the Italian NRRP project CHANGES, funded by the NextGenerationEU initiative of the EU, generally dealing with Cultural Heritage, one of the sub-projects ('spokes') specifically deals with "Digital Libraries, Archives and Philology". A research unit on AI from the University of Bari worked in this sub-project to develop innovative solutions that could overcome some shortcomings and limitations present in the state-of-the-art in this field. Specifically: a new framework for knowledge bases representation and exploitation was proposed and implemented; new tools for symbol recognition, more oriented towards research and philology than on transcription, were designed and built; and advanced interfaces for knowledge search, browsing, consultation and analysis were developed.

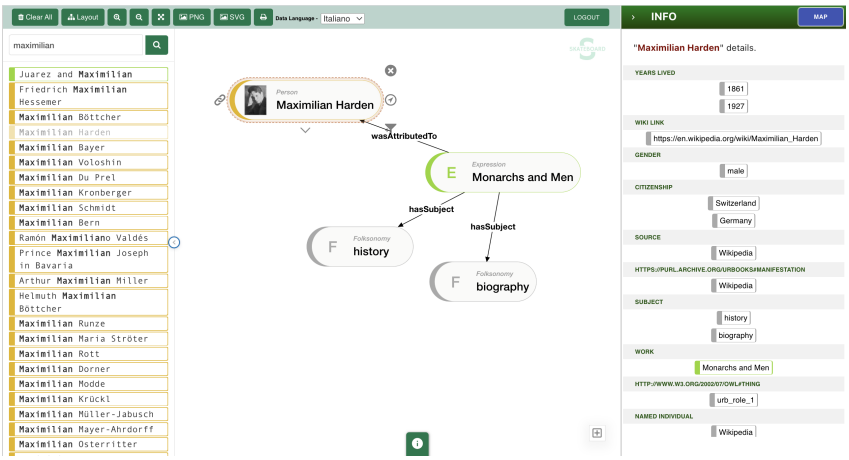

Figure 4: Interface of SKATEBOARD, the interface for querying and browsing Semantic Web repositories

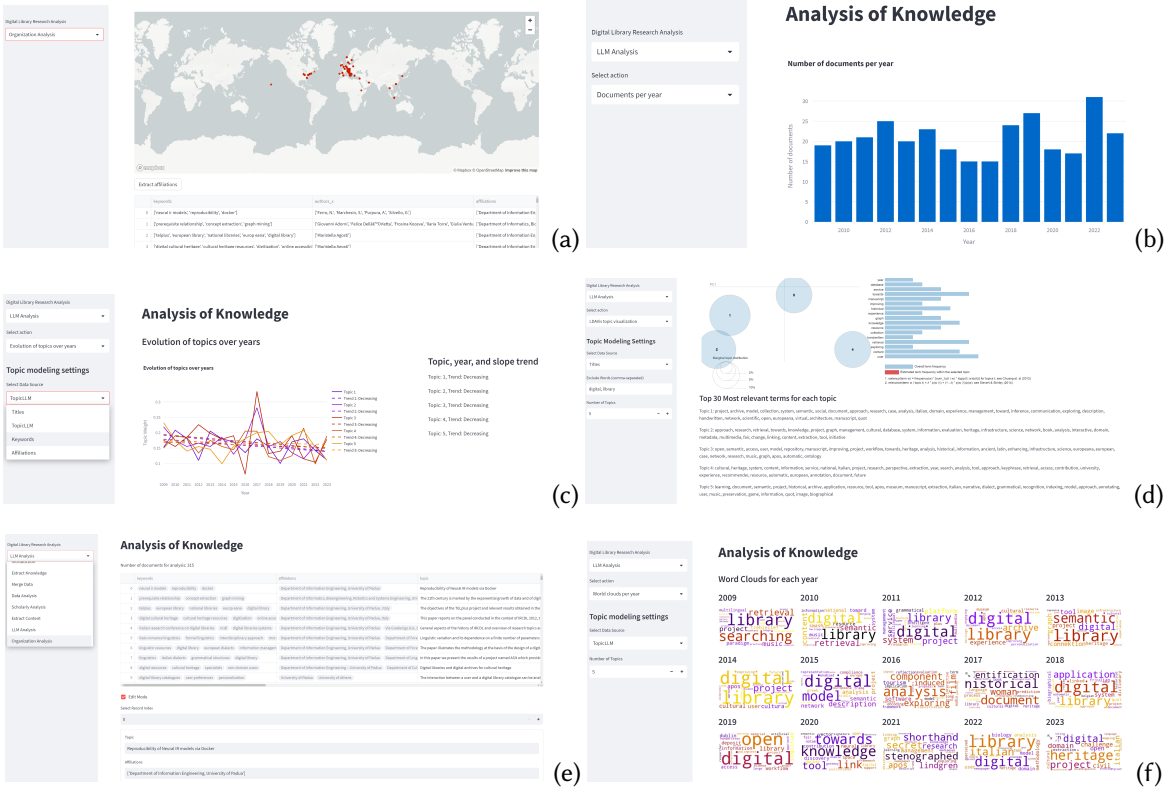

Figure 5: Functions in the digital library analysis tool

Future work after the project will carry on this research and integrate these tools in an overall platform that can provide advanced support to the users. Then, user studies will be carried out within the research and practice communities, in order to spot strengths and limitations on both usability and funcrionality aspects. Based on this, improvements and extensions of the tools will be designed and implemented.

Acknowledgments

This research was partially supported by project CHANGES “Cultural Heritage Active Innovation for Sustainable Society” (PE00000020), Spoke 3 “Digital Libraries, Archives and Philology”, funded by the Italian Ministry of University and Research NRRP initiatives under the NextGenerationEU program.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] S. Ferilli, Integration strategy and tool between formal ontology and graph database technology, Electronics 10 (2021). URL: https://www.mdpi.com/2079-9292/10/21/2616.

[2] S. Ferilli, E. Bernasconi, D. D. Pierro, D. Redavid, The graphbrain framework for knowledge graph management and its applications to cultural heritage, in: Proceedings of AISoLA 2023, volume 14129 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 144–161.

[3] T. Berners-Lee, J. Hendler, O. Lassila, The semantic web, Scientific American 284 (2001) 34–43.

[4] T. Heath, C. Bizer, Linked Data: Evolving the Web into a Global Data Space, Synthesis Lectures on the Semantic Web, Morgan & Claypool Publishers, 2011.

[5] S. Ferilli, E. Bernasconi, D. Redavid, G. D. Nunzio, Knowledge graphs for the web economy: The chips&bits project on cultural heritage, in: Proceedings of the XXI Conference on Information and Research science Connecting to Digital and Library science (IRCDL 2025), volume 3937 of *Central Europe (CEUR) Workshop Proceedings*, 2025, p. 13 pp. URL: https://ceur-ws.org/Vol-3937/paper5.pdf.

[6] N. Islam, Z. Islam, N. Noor, A survey on optical character recognition system, 2017. URL: https://arxiv.org/abs/1710.05703. arXiv:1710.05703.

[7] X.-F. Wang, Z.-H. He, K. Wang, Y.-F. Wang, L. Zou, Z.-Z. Wu, A survey of text detection and recognition algorithms based on deep learning technology, Neurocomputing 556 (2023) 126702.

[8] A. Adhwaith, I. Jossy, M. R. B. N. L. Koshy, K. Sreelekshmi, Handwritten text recognition: A survey of ocr techniques, International Journal of Advances in Engineering and Management (IJAEM) 6 (2024) 205–220.

[9] W. Burger, M. J. Burge, Digital Image Processing – An Algorithmic Introduction Using Java, 2nd ed., Springer, 2016.

[10] A. Holzinger, A. Saranti, C. Molnar, P. Biecek, W. Samek, Explainable AI Methods - A Brief Overview, Springer International Publishing, Cham, 2022, pp. 13–38.

[11] E. Bernasconi, S. Ferilli, Enhancing symbol recognition in library science via advanced technological solutions, Information 16 (2025) 25 pp. URL: https://www.mdpi.com/2078-2489/16/2/119.

[12] S. Ferilli, E. Bernasconi, D. Redavid, An interpretable technique for handwritten character recognition in historical documents, in: Post-proceedings of AISoLA 2025, Lecture Notes in Computer Science, Springer, 2026.

[13] J. H. Tiltman, The voynich manuscript: The most mysterious manuscript in the world, NSA Technical Journal XII (1967).

[14] E. Bernasconi, D. D. Pierro, D. Redavid, S. Ferilli, Skateboard: Semantic knowledge advanced tool for extraction, browsing, organization, annotation, retrieval, and discovery, Applied Sciences 13 (2023) 11782. URL: https://www.mdpi.com/2076-3417/13/21/11782.

[15] E. Bernasconi, S. Ferilli, Mining literary trends: A tool for digital library analysis, in: Linking Theory and Practice of Digital Libraries, volume 15177 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 342–359.

[16] E. Bernasconi, A. Mannocci, A. Tammaro, Exploring the Italian research landscape on Digital Library in the Conference IRCDL, in: Proceedings of the 20th Conference on Information and Research science Connecting to Digital and Library science (formerly the Italian Research Conference on Digital Libraries), volume 3643 of *CEUR Workshop Proceedings*, CEUR, 2024, pp. 230–245. doi:10.5281/zenodo.14923848.