Knowledge extraction, management and long-term preservation of non-Latin cultural heritages - Digital Maktaba project presentation

Riccardo Martoglia[1], Sonia Bergamaschi[1], Federico Ruozzi[1], Matteo Vanzini[1,*], Luca Sala[1] and Riccardo Amerigo Vigliermo[1]

[1]*University of Modena and Reggio Emilia*

Abstract

The services provided by today's cutting-edge digital library systems may benefit from new technologies that can improve cataloguing efficiency and cultural heritages preservation and accessibility. Below, we introduce the recently started Digital Maktaba (DM) project, which suggests a new model for the knowledge extraction and semi-automatic cataloguing task in the context of digital libraries that contain documents in non-Latin scripts (e.g. Arabic). Since DM involves a large amount of unorganized data from several sources, particular emphasis will be placed on topics such as big data integration, big data analysis and long-term preservation. This project aims to create an innovative workflow for the automatic extraction of information and metadata and for a semi-automated cataloguing process by exploiting Machine Learning, Natural Language Processing, Artificial Intelligence and data management techniques to provide a system that is capable of speeding up, enhancing and supporting the librarian's work. We also report on some promising results that we obtained through a preliminary proof of concept experimentation. (Short paper, discussion paper)

Keywords

Cultural heritages, Non-Latin alphabets, Knowledge extraction, Machine Learning, Natural Language Processing, Big data management, Long-term preservation, Big data integration, Named Entity Recognition,

1. Introduction

Multiculturalism's linguistic and social effects cannot be ignored in any field today. Texts in non-Latin alphabets were previously found only in a few specialist libraries; nowadays every library must adapt to the new needs of heterogeneous users, but they are often unable to do so because of data management difficulties. Hence, the urgency of a global sharing of multicultural heritage: an activity made easier by technology that can create semi-automatic solutions to enhance the readability of documents, comprehend their content, preserve it through time and enable advanced digital use with sophisticated consultation and search functions.

19th IRCDL (The Conference on Information and Research science Connecting to Digital and Library science), February 23–24, 2023, Bari, Italy

*Corresponding author.

riccardo.martoglia@unimore.it (R. Martoglia); sonia.bergamaschi@unimore.it (S. Bergamaschi); federico.ruozzi@unimore.it (F. Ruozzi); matteo.vanzini@unimore.it (M. Vanzini); luca.sala@unimore.it (L. Sala); r.a.vigliermo@unimore.it (R. A. Vigliermo)

0000-0003-4643-6128 (R. Martoglia); 0000-0001-8087-6587 (S. Bergamaschi); 0000-0003-2729-5016 (F. Ruozzi); 0000-0003-0471-1101 (M. Vanzini); 0000-0002-4833-8882 (L. Sala); 0000-0001-9914-3295 (R. A. Vigliermo)

The challenges of devising automated information extraction solutions when dealing with non-Latin materials w.r.t Latin ones are essentially of two main orders: from one hand the graphic-linguistic and on the other hand the advancement of the state of the art for the Arabic script OCR. From the graphic-linguistic point of view the Arabic script is always cursive, reads from right to left, homographic (the majority of its graphemes are distinguished merely by one, two, or three diacritical dots above or below them), consonantic (*abǧad*, vowels are exclusively represented by diacritical signs). Moreover the shape of graphemes varies depending on the context and some of those *don't bind* leftward, resulting in words with two or more components joined without any ligature. From the State of the Art perspective, the studies on Latin script gained more attention with respect to Arabic, due mainly to the aforementioned graphic challenges. In fact, only in the last two decades more attention has been posed on the question even though Arabic OCR systems still perform poorly compared to Latin script OCRs.[1]

This is the challenging scenario of ITSERR[1], a NextGenerationEU-funded project through the National Recovery and Resilience Plan (PNRR) started in November 2022 with the goal of enhancing the European Research Infrastructure RESILIENCE [2] in response to the demands of the scientific community in Religious Studies, supporting the current national infrastructures and elevating it to a more mature state in terms of technology integration and capacity to increase innovation, quality and variety of the knowledge produced by the field of Religious Studies. The premise of ITSERR is that the humanities would provide extremely diversified datasets whose complexity will challenge technological experts and ICT researchers. The development of digital tools will concern both instruments for text editions through a historical-critical approach as well as tools designed to support each phase of the research in the field of Religious Studies. Digital Maktaba (in Arabic "maktaba", "library": the "place where books are located", henceforth DM), which started as a Proof of Concept in collaboration with the Foundation for Religious Studies (FSCIRE), a national infrastructure located in Bologna and Palermo and the innovative startup mim.FSCIRE [3, 4], will be a research project carried out within ITSERR. It is part of the RESILIENCE framework and is intended to offer a helpful and innovative solution to libraries specialised in religious studies that need to manage multilingual and multi-alphabetic cultural heritage documents. In particular, its goal and expected contributions are the following:

- DM aims to develop intelligent extraction and data management processes, to help managing libraries and archives and to create virtuous cataloguing models that can handle non-Latin alphabets documents. The final goal is to deliver an intelligent (and semi-automated) system able to extract high-quality information from documents in different languages, including rich metadata content, thus supporting the manual work usually required for the cataloguing procedure;
- DM will have a rather unique and exclusive case study, offered by the Giorgio La Pira Library (FSCIRE) in Palermo, specialized on History and Doctrines of Islam, including specific knowledge, non-latin alphabets and multilingual variations in a comprehensive digital corpus of more than 200000 documents;

[1]The project involved the University of Modena and Reggio Emilia, CNR, University of Palermo, University of Turin and University of Naples "L'Orientale"

- Even if originally conceived for religious sciences, DM will embrace the difficulties presented by such alphabets (in Arabic, Persian and Azerbaijani languages) with regard to data extraction, huge data management, cataloguing and librarianship, ultimately aspiring to become a reusable tool that helps librarians and researchers to manage and study documents in a variety of contexts.

DM activity program involves interdisciplinary skills and experience of varied professionals. This synergy will be essential to effectively address the challenges of a technologically advanced, multicultural community like that of the European Union, placing it in the frame of the conservation and enhancement of its cultural heritage.

The paper is structured in the following way: in the next section, we will provide a review of the most relevant studies in the field, followed by a detailed description of DM in the ITSERR project framework (Section 3). Section 4 reports on some initial and promising results obtained in a preliminary proof of concept experimentation. In the conclusive section, final remarks on auspicable benefits and advantages will be drawn.

2. Related Works

Although the fields of Arabic script Natural Language Processing (NLP), Information Retrieval (IR), and Optical Character Recognition (OCR) have advanced significantly over the past few decades, there have not been many efforts to take advantage of these breakthroughs in order to create cutting-edge digital libraries. We are aware of very few noteworthy projects within the languages taken into consideration, compared to other automation challenges (e.g. automatic extraction of metadata, semi-automated cataloguing with Machine Learning approaches). In 2009 the Alexandria library created the Arabic Digital Library as a part of the Digital Assets Repository (DAR) project, with text extraction tools for Arabic language characters implemented [5]. Another similar project is Arabic Collections Online, a multi-institutional project mainly aiming to digitize, preserve and provide free open access to a wide variety of Arabic language books on various subjects [6]. From a more strict digitization standpoint it is worth to mention other few important projects on Arabic and Persian manuscripts that involve handwritten text recognition, such as The British Library projects on arabic [7][8] and persian manuscripts[9][10]. More recently, similar projects concerning the digitization and the building of Arabic and Persian text corpora have been developed, such as the Open Islamicate Text Initiative (OpenITI) [11], which is a multi-institutional effort to construct the first machine-actionable scholarly corpus of premodern Islamicate texts collected from open-access online libraries such as Shamela [12] and Shiaonline library [13]. From a character recognition standpoint, the OpenITI project expolits the Kraken OCR useful both for handwritten and printed text recognition [14, 15]. Two further intriguing initiatives from OpenITI are KITAB [16] and the Persian Digital Library (PDL) [17]. The first one is focused on discovering the relationship in the Arabic rich textual tradition with interesting Machine Learning (ML) solutions such as stylometric analysis [18] and subgenre classification [19]. The latter is focused on building a scholarly-verified corpus and an Optical Character Recognition (OCR) system for handwritten Persian texts. PDL has already created an open-access corpus of Persian poems collected from the Ganjoor site [20] and integrated with a lemmatizer [21] and a digital Persian dictionary [22].

Most of the above-mentioned projects aim to fully digitize a (relatively) small library of books, frequently involving a significant amount of manual labor or only focusing on a tiny subset of the languages taken into account by DM. Additionally, DM will enable the extraction of a rich array of metadata rather than just text content, fully supporting non-latin alphabet metadata.

3. Digital Maktaba project description

The project is composed by two macro-phases:

a. Definition of data, metadata and knowledge extraction techniques.

The first macro-phase aims to investigate, define, implement and test the techniques required to obtain text and metadata from the documents, which means to extract the significant knowledge that will be used in the second macro phase to build a tool for supervised cataloguing.

First of all, an *analysis of the operating scenario and materials* task will be conducted from the IT perspectives on one side and the historical-linguistic one on the other. This preliminary step will include a state-of-the-art analysis to identify useful techniques and tools, on which several tests will be executed to evaluate their strength and limitations. In particular, we are interested in analyzing:

- Available OCR technologies for the languages considered by the project to extract the text contained in the document;
- Linguistic tools (multi-lingual resources such as dictionaries, thesauri, corpora, etc.) to enrich the obtained metadata from both syntactic and semantic standpoints;
- Additional text mining techniques to enhance and refine the knowledge extraction phase.

The subsequent *Development of algorithms for automatic text recognition, metadata and knowledge extraction* task pursues the goal of defining the following innovative techniques to be applied to the extensive digital library heritage provided by the Giorgio La Pira library:

- *Text acquisition/OCR techniques* for assisting/automating text extraction, exploiting object fusion techniques to combine the best OCR tools available for each language and produce a high-quality output while leveraging the unique benefits of each engine;
- Further *knowledge extraction techniques* in order to collect several useful metadata which is seldom offered by available state-of-the-art tools (and that will be key for powering the intelligent cataloguing assistance features of the final tool):
  - *Syntactic metadata* that include information on text regions, detected language(s) and character(s), text size and location on page, and self-assessed extraction quality using an ad-hoc score;
  - *Linguistic metadata* that incorporate references to external linguistic sources that offer helpful data, such as word definitions for additional (semantic) processing;
  - *Cataloguing metadata* gathered through intelligent approaches to automatically identify the different cataloguing fields present on the frontispiece of a document (e.g., title, authors, category, etc.).

Finally, validation of the established recognition and extraction procedures will be performed using broader corpora from the literature and from partner institutions, as well as samples of the materials from other WPs in the ITSERR project.

b. Building a complete tool for supervised cataloguing.

The second macro-phase involves the project and implementation of a supervised cataloguing tool, offering a complete solution for knowledge extraction (exploiting the extraction techniques defined in the previous macro-phase), data management, storage and access. To this end, the *Data Management, Interactive Search and Supervised Cataloging* sub-task will include:

- *Database and data management* design. The database will store the extracted data and metadata and accommodate the library's actual demands but also be capable of managing additional data acquisitions that have different characteristics from those present in the La Pira library. Special emphasis will be placed on Big Data Management for internal (other ITSERR activities) and external (other institutions) use, Data Integration, long-term preservation [23] (to ensure that data will be available and accessible also in the future) as well as Data Exchange aspects in order to enable interoperability with catalogued data from other libraries, but also to improve the accessibility and future re-usability of the managed data. Moreover, Entity Recognition (ER) [24] aspects will be deepened to disambiguate newly acquired information;
- Definition of *advanced searching techniques* (including approximate and full-text search) to effectively supply the needed information and improve the search speed and scope compared to standard cataloguing tools;
- Definition of *intelligent and AI-based techniques* to enhance and semi-automate the cataloguing process. Suggestions based on user feedback and previously entered data (and metadata) will be used to assist the librarian in the data entry task. Supervised ML models will enable automatic category recognition and provide systematization and classification of data in accordance with the topographical design of the La Pira library. Incremental ML algorithms will allow the tool to "learn" from past actions, making it more automated and efficient over time. The aim of the team is to enhance the work of librarians by placing them at the center of the system, taking advantage of their contribution and skills, following the paradigm "AI in the loop, human in charge" [25]. Both traditional and deep learning methods will be taken into account, deploying them on parallel architectures for faster execution. Moreover, to go beyond the black box nature of ML suggestions and explain them, a special focus will be given to Interpretable Machine Learning (IML) algorithms, which are becoming more and more important in contexts such as healthcare [26] but which have seldom been applied to cultural heritage;
- Design of a *web user interface* for cataloguing new documents and searching the archive.

To develop a reproducible and reusable tool that enables a straightforward cataloguing workflow while overcoming linguistic and geographic challenges, all the above-mentioned strategies will be tested and integrated within the final *Integration of the proposed solutions* sub-task. The tool will be designed to work in many libraries with several languages and cataloguing requirements. This last task also deepens the hypothetical integration of the solutions developed by DM with

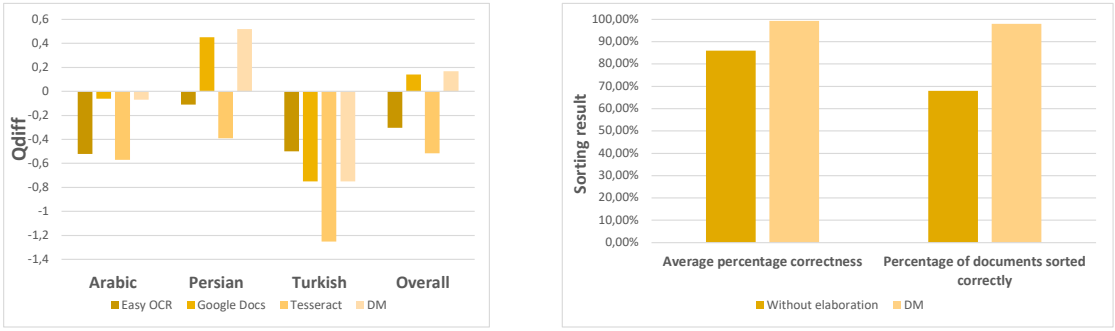

Figure 1: Proof of concept test results: performance of the text extraction techniques with respect to open source OCR tools (left part) and of the text-boxes sorting and merging algorithms (syntactic metadata extraction, right part).

other ITSERR developed tools closely related to infrastructural services. Finally, a use case workshop will be produced in order to demonstrate the capabilities of the tool.

4. Digital Maktaba: Proof of Concept preliminary results

In this section, we will summarize some of the preliminary results we have obtained from the tests performed in the past months on a partial proof of concept of the DM tool focused on text and metadata extraction. The proof of concept allowed us to verify the feasibility and potential of the project before the official start of the activities. Our tests were carried out on a sample of 100 documents from the project library, selected to be representative of the complete collection (both in terms of variety and linguistic contents). With respect to the results presented in [3, 4], the tests have been updated with the latest findings (latest versions of the proof of concept prototype and external tools).

Figure 1 (left) presents the obtained results as to the text extraction techniques performance with respect to freely available OCR tools. The considered OCR tools are the ones that have been selected in our preliminary analyses as the ones best supporting the involved languages: they are EasyOCR[2], GoogleDocs[3] and Tesseract[4]. "DM" represents the proof of concept solution that exploits multiple OCR engines and automatically selects the results that are found to be more promising (also exploiting self-assessed quality scores relying on external linguistic resources such as the Open Multilingual WordNet thesauri[5]). The evaluation was performed through an ad-hoc metric we defined to take into consideration both the output quality (*oq* in a range [0,2]), based on the actual correspondence with ground truth defined by experts, and the input quality (*iq*, range [0,2]), based on a manually assessed quality of the scan/image: *qdiff* (range [-2,2]) simply expresses the relationship between output quality w.r.t. input quality, where 0 indicates in line, a positive value equals higher than input quality, a negative value equals lower than

[2]EasyOCR. https://github.com/JaidedAI/EasyOCR
[3]GoogleDocs. https://docs.google.com
[4]Tesseract. https://github.com/tesseract-ocr/tesseract
[5]Open Multilingual WordNet thesauri. http://compling.hss.ntu.edu.sg/omw/)

input quality results:

$$qdiff = oq - iq \tag{1}$$

The *qdiff* equation's core principle is to penalize the OCR results when input quality is higher and output quality is lower. The resulting values allow us not only to compare the systems' performance (the higher the better), but also to have an idea whether the quality of the output is better or worse than expectations. As we can see from Figure 1 (left), the quality of the freely available OCR engines is very much dependent on the language, and there is no engine that is superior to others for all the considered languages. Instead, the experiments revealed that the combined DM approach is able to give better results in terms of overall output quality. Moreover, DM is actually designed to extract the additional metadata described in Section 3, while available solutions often focus on bare text / text regions extraction.

Another test we present is about a specific problem that has been preliminarily analyzed, the sorting and merging of the text regions (boxes) extracted by OCR tools (part of the syntactic metadata extraction). The output offered by available tools is often not ordered correctly w.r.t. the meaning and reading rules (e.g., right to left) of the specific languages, moreover in some cases phrases are fragmented into different boxes. These problems can lead to low metadata quality and increased complexity of the subsequent knowledge extraction phases. A preliminary ad-hoc algorithm has been developed in order to solve the above-mentioned issues: it exploits the positions of the text boxes in order to perform horizontal grouping, merging, and renumbering of the boxes. Figure 1 (right part) shows the performance increase obtained by applying this approach. Performances are evaluated on two metrics w.r.t. a gold standard manually determined by experts: average percentage correctness — the percentage of boxes in each document having the right number, averaged on the whole document set, and percentage of correctly sorted documents — the percentage of documents without errors in the numbering of their boxes. As we can see, thanks to the current algorithm implementation, the average percentage correctness increases by 14%, while the percentage of correctly sorted documents increases by 30%, going from 68% to 98%.

5. Conclusion

As discussed in the previous sections, several advantages on a variety of interesting and innovative fronts are expected. First of all, from a broader standpoint, studies on cataloguing in contexts involving many languages should advance without relying exclusively on perplexing transliteration schemes. As the proof of concept already hinted, DM aims to overcome the current limitations in terms of text extraction over different non-latin languages, allowing the direct use of the documents' native language. Moreover, library services will be strengthened thanks to the design and implementation of intelligent features for user assistance as well as the exploitation of other available library catalogs. This will enable a faster cataloguing pipeline and, not less important, greater data consistency (also through time). Moreover, there will also be improvements in areas such as flexibility of data output/exchange, and explainability of the assistance techniques.

Overall, the project multidisciplinary and multicultural nature has the potential to significantly improve cultural heritage preservation and exploitation, thanks to a tool that is part of a

broader shared-knowledge framework (ITSERR), encompassing various languages, cultures, and religious realities.

Acknowledgement

This work is partially supported by the PNRR ITSERR project.

References

[1] W. Albattah, S. Albahli, Intelligent Arabic Handwriting Recognition Using Different Standalone and Hybrid CNN Architectures, Applied Sciences 12 (2022) 10155. URL: https://www.mdpi.com/2076-3417/12/19/10155. doi:10.3390/app121910155, number: 19 Publisher: Multidisciplinary Digital Publishing Institute.

[2] M. Büchler, S. Riegert, F. Alpi, F. Cadeddu, Towards Big Religious Data: RESILIENCE Research Infrastructure for Data on Religion in the Digital Age, in: Proceedings of the 2nd International Conference on Digital Tools & Uses Congress, DTUC '20, Association for Computing Machinery, New York, NY, USA, 2020. URL: https://doi.org/10.1145/3423603.3424007. doi:10.1145/3423603.3424007.

[3] S. Bergamaschi, R. Martoglia, F. Ruozzi, R. A. Vigliermo, S. De Nardis, L. Sala, M. Vanzini, Preserving and Conserving Culture: First Steps towards a Knowledge Extractor and Cataloguer for Multilingual and Multi-Alphabetic Heritages, in: Proceedings of the Conference on Information Technology for Social Good, GoodIT '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 301–304. URL: https://doi.org/10.1145/3462203.3475927. doi:10.1145/3462203.3475927.

[4] S. Bergamaschi, S. De Nardis, R. Martoglia, F. Ruozzi, L. Sala, M. Vanzini, R. A. Vigliermo, Novel Perspectives for the Management of Multilingual and Multialphabetic Heritages through Automatic Knowledge Extraction: The DigitalMaktaba Approach, Sensors 22 (2022). URL: https://www.mdpi.com/1424-8220/22/11/3995. doi:10.3390/s22113995.

[5] Alexandria Library, DAR Project, http://dar.bibalex.org/webpages/aboutdar.jsf, Last accessed: November 25, 2022.

[6] V. Danielson, B. Russel, ACO - Arabic Collections Online, http://dlib.nyu.edu/aco/, 2017.

[7] The british library projects, https://www.bl.uk/collection-guides/arabic-manuscripts., Last accessed: November 28, 2022.

[8] QDL - Qatar Digital Library, https://www.qdl.qa/en/about, Last accessed: November 25, 2022.

[9] The british library projects, https://www.bl.uk/projects/digital-access-to-persian-manuscripts, Last accessed: November 28, 2022.

[10] Iran heritage, https://www.iranheritage.org/, Last accessed: November 28, 2022.

[11] M. T. Miller, M. G. Romanov, S. B. Savant, Digitizing the textual heritage of the premodern islamicate world: Principles and plans, International Journal of Middle East Studies 50 (2018) 103–109. doi:10.1017/S0020743817000964.

[12] al-Maktabah al Shamela, Shamela library, Last accessed: November 23, 2022. URL: https://shamela.ws/.

[13] Shiaonline, Shiaonline library, Last accessed: November 23, 2022. URL: http://shiaonlinelibrary.com.

[14] M. Romanov, M. Miller, S. Savant, B. Kiessling, Important new developments in arabographic optical character recognition (ocr), Al-'Usur al-Wusta 25 (2017). doi:10.7916/alusur.v25i1.6996.

[15] B. Kiessling, Kraken - a Universal Text Recognizer for the Humanities (2019). URL: https://doi.org/10.34894/Z9G2EX. doi:10.34894/Z9G2EX.

[16] A. K. U. International, Kitab project, Last accessed: November 24, 2022. URL: https://kitab-project.org/about/.

[17] U. o. M. Roshan Institute for Persian Studies, Persian digital library, Last accessed: November 23, 2022. URL: https://persdigumd.github.io/PDL/.

[18] A. K. U. International, Kitab project, stylometry, https://kitab-project.org/methods/stylometry, Last accessed: November 24, 2022.

[19] A. K. U. Interntional, Kitab project, subgenre classification, https://kitab-project.org/methods/sub-genre, Last accessed: November 24, 2022.

[20] Ganjoor.net, Ganjoor, Last accessed: November 23, 2022. URL: https://ganjoor.net/.

[21] Roshan-ai.ir, Hazm, baray-e pardazesh-e zaban-e farsi, Last accessed: November 23, 2022. URL: https://www.roshan-ai.ir/hazm/.

[22] F. J. Steingass, A Comprehensive Persian-English dictionary, including the Arabic words and phrases to be met with in Persian literature, Routledge & K.Paul, London, 1892. URL: https://dsal.uchicago.edu/dictionaries/steingass/.

[23] U. Borghoff, P. Rödig, J. Scheffczyk, L. Schmitz, Long-Term Preservation of Digital Documents: Principles and Practices, Springer Berlin Heidelberg, 2007. URL: https://books.google.it/books?id=sZpm0dBV5MwC.

[24] G. Simonini, L. Gagliardelli, S. Bergamaschi, H. Jagadish, Scaling entity resolution: A loosely schema-aware approach, Information Systems 83 (2019) 145–165. URL: https://www.sciencedirect.com/science/article/pii/S0306437918304083. doi:https://doi.org/10.1016/j.is.2019.03.006.

[25] Hai - stanford university, ai in the loop: Humans must remain in charge, https://hai.stanford.edu/news/ai-loop-humans-must-remain-charge, Last accessed: December 2, 2022.

[26] M. A. Ahmad, C. Eckert, A. Teredesai, Interpretable machine learning in healthcare, in: Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics, BCB '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 559–560. URL: https://doi.org/10.1145/3233547.3233667. doi:10.1145/3233547.3233667.