Knowledge Extraction and Cross-Language Data Integration in Digital Libraries

Luca Sala[1]

[1]*First Year PhD Sudent, ICT Doctorate at DBGroup, University of Modena and Reggio Emilia, Modena, Italy*

Abstract

Digital Humanities (DH) is an interdisciplinary field that has grown rapidly in recent years, requiring the creation of an efficient and uniform platform capable of managing various types of data in several languages. This paper presents the research objectives and methodologies of my PhD project: the creation of a novel framework for Knowledge Extraction and Multilingual Data Integration in the context of digital libraries in non-Latin languages, in particular Arabic, Persian and Azerbaijani. The research began with the Digital Maktaba (DM) project and continued within the PNRR ITSERR infrastructure, in which the DBGroup[1] participates. The project aims to develop a two-component framework consisting of a Knowledge Extraction Subsystem and a Data Integration Subsystem. The case study is based on the DM project, which seeks to create a flexible and efficient digital library for preserving and analyzing multicultural heritage documents by exploiting the available and ad-hoc created datasets, Explainable Machine Learning , Natural Language Processing (NLP) technologies and Data Integration approaches. Key challenges and future developments in Knowledge Extraction and Data Integration are examined, which involve leveraging the MOMIS system for Data Integration tasks and adopting a microservices-based architecture for the effective implementation of the system. The goal is to provide a versatile platform for organizing and integrating various data sources and languages, thereby fostering a more inclusive and accessible global perspective on cultural and historical artefacts that encourage collaboration in building an expanding knowledge base.

Keywords

Data Integration, Cross-Language Record Linkage, Knowledge Extraction, Long-term Preservation

1. Introduction

Digital Humanities (DH) is an interdisciplinary field that aims to integrate humanities research with digital technologies. In recent years, DH has grown in importance and led to new methods of understanding and analyzing cultural and historical artifacts, enabling scholars and researchers to access, interpret, and share knowledge across multiple disciplines. This requires the collection and integration of data from various sources written in different languages, including non-Latin ones. To meet this need, Cross-Language Data Integration has emerged as a crucial process for fusing data from multiple sources and providing a unified virtual view, allowing academic users to access and analyze vast amounts of information from multiple sources, regardless of the language in which they were originally published. This is particularly relevant for digital

[1] https://dbgroup.unimore.it/site/home.html

SEBD 2023: 31st Symposium on Advanced Database System, July 02–05, 2023, Galzignano Terme, Padua, Italy

luca.sala@unimore.it (L. Sala)

0000-0002-4833-8882 (L. Sala)

libraries, which could be considered as massive information warehouses that require efficient organisation and access solutions to meet different user needs. Starting from the fact that data integration is a non-trivial task [1, 2, 3], the difficulty of dealing with different languages adds another layer of complexity to the whole project.

In response to this, my research will focus on creating a framework for merging knowledge from increasingly different sources and languages. The project also features a case study [4, 5, 6, 7], that investigates the development of a digital library where ML techniques are involved in order to develop an automatic cataloguing system for texts written in non-Latin languages, particularly in Arabic scripts (Arabic, Persian and Azerbaijani). In this context, some remarks on the languages under consideration must be made both from the point of view of the challenges posed by the Arabic writing system and with regard to the state of the art of certain techniques such as OCR and language resources. As for the first point, there are some difficulties intrinsic to the Arabic alphabet itself (cursive writing, presence of diacritical dots and marks, homography, etc.). While for the state of the art, e.g., of OCR systems for the Arabic alphabet, there is a significant backwardness compared to languages with Latin or other alphabets (e.g., Chinese, Japanese). The ultimate purpose, however, is not confined to library integration. Furthermore, it seeks to simplify the integration of various data types as well as overcome linguistic barriers in order to collect more information for future generations, ensuring also Long-term preservation.

In this paper, I first provide a brief overview of related works in Section 2. Then, the Digital Maktaba (DM) case study is introduced in Section 3, describing its context and objectives. In Section 4, I outline the key challenges that need to be addressed in order to achieve efficient Knowledge Extraction and Data Integration in DM, and I discuss the expected developments that will enable to overcome these obstacles. In the concluding section, final observations on the anticipated benefits and advantages of this project will be illustrated.

2. Related Work

Digital Humanities (DH) has seen a surge in interest, with various research projects focusing on Knowledge Extraction, Data Integration [8], and multilingual data management. For instance, [9] explored NLP techniques for processing non-Latin languages, while [10] addressed the challenges of Cross-Language Data Integration in digital libraries. This research builds upon these works, proposing a novel framework that combines state-of-the-art techniques for efficient Knowledge Extraction and integration of multilingual data.

3. Digital Maktaba (DM): Case Study

In this paper, the DM project, which is identified as Work Package 5 (WP5) within the ITSERR project, is used as a case study and a starting point to demonstrate the challenges and potential solutions for digital libraries managing multilingual data. DM focuses on developing a novel workflow for automatically classifying documents in non-Latin languages such as Arabic, Persian and Azerbaijani, attempting to provide an effective system for information and metadata extraction. This process is a key part in the building of a digital library that will include as further steps also the implementation of Explainable Machine Learning, Natural Language Processing (NLP), and Data Integration technologies.

Given that, the primary goal of DM was the development of smart extraction and data management processes for non-Latin alphabet documents, as well as the development of a semi-automated system for high-quality information extraction. The large collection of digital books made internally available by the "Giorgio La Pira" library in Palermo (over 200,000 documents), which is a hub of FSCIRE foundation, dedicated to history and doctrines of Islam was taken into account as the project test case. Despite being initially focused on religious sciences, the DM project intends to become a reusable framework for organising and analysing documents in a variety of contexts, thereby aiding librarians and scholars who work with multicultural heritage documents. Recommendations based on user feedback, as well as previous input data and metadata, will aid librarians in their data entry tasks through automatic cataloguing suggestions. Within the Knowledge Extraction Subsystem (Figure 1), to obtain optimal output from the images, various OCR systems will be explored and evaluated, while Supervised ML models will facilitate automatic category identification and ensure systematic data organization and classification. Incremental ML algorithms will enable the system to "learn" from previous actions, thereby becoming increasingly automated and efficient. The aim is to amplify librarians' work by positioning them at the heart of the system, capitalizing on their expertise and abilities in accordance with the "AI in the loop, human in charge" paradigm [11]. Both conventional and Deep Learning techniques will be considered, with implementations on parallel architectures for expedited execution. Furthermore, to provide transparency and explain the ML suggestions, particular attention will be given to Interpretable Machine Learning (IML) and Explainable Machine Learning algorithms, studying both off-the-shelves solutions and developing new methods where necessary. These have gained prominence in fields like healthcare [12] but have rarely been applied to cultural heritage contexts. The framework will promote the sharing and preservation of multicultural patrimony through integration with other libraries and expanding the range of information available to librarians and users. The approach that will be employed is hybrid, since some of the data will be retrieved at query time, while others will be materialized and stored in advance. A detailed discussion of this concept can be found in Section 4.3. This collaborative method will promote knowledge accumulation over time, resulting in an expanding and comprehensive knowledge base for digital humanities study.

Another noteworthy aspect of the project is the modular design of its software architecture, which allows for flexibility and adaptability when integrating with different systems or including new features, making it easier to upgrade and modify the framework based on individual needs. The microservices design of the entire framework will also support scalability, allowing the system to accommodate larger datasets, and library collections.

4. Goals, Challenges and Future Developments

The PhD project's goal is to develop a two-component framework:

1. *Knowledge Extraction Subsystem*
   As depicted in Figure 1, the initial subsystem processes unstructured documents and generates output that is both structured and semantically enriched.

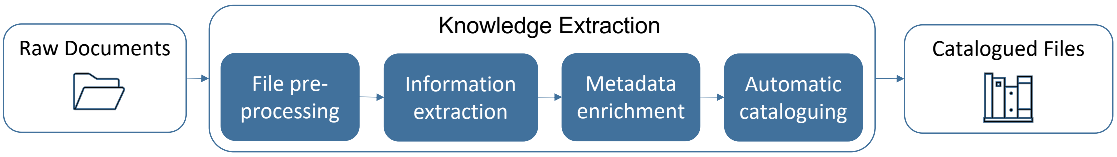

Figure 1: Knowledge Extraction Subsystem

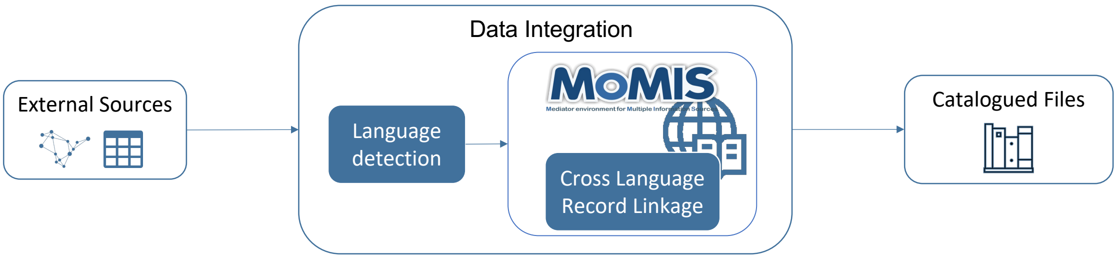

Figure 2: Data Integration Subsystem

2. *Data Integration Subsystem*
   The second subsystem, presented in Figure 2, seeks to integrate different data sources, typically in different languages, within the context of digital libraries.

Hereafter, in different subsections, will be explained the challenges and future developments in Knowledge Extraction and Data Integration within and beyond the DM and ITSERR project by providing a comprehensive overview of the work that my PhD project will include.

4.1. Knowledge Extraction and FAIR Principles

The DM project, which is currently developing an early modular subsystem for extracting information from document images, as a first step in the design of a semi-automated digital library, as shown in Figure 1, must ensure compliance with FAIR principles to guarantee the success of the project.

Wilkinson et al. (2016) [13] established these principles, which highlight that data should be *Findable, Accessible, Interoperable, and Reusable.* Following this ensures that the system will be able to combine works from various collections while remaining Interoperable with future data sources. Moreover, by adhering to these standards, data will also be more accessible to both librarians and users, assuring the information's Long-term preservation. In addition to the present development goals for what concerns the Knowledge Extraction, there are plans to create microservices that can extract information from non-textual media sources such as images, videos, and audio files. Once the test case is completed, the development of these expanded capabilities will allow for the creation of an increasingly modular system capable of meeting the various needs of different projects while also enriching the digital library's knowledge base, making it more comprehensive and useful for specialised as well as unspecialised users.

4.2. Data Storage and Data Sources Integration

The acquisition of knowledge on how to properly store data gathered by the Knowledge Extraction Subsystem will be a subject of research, which will consist in considering a variety of models and techniques among Relational and/or NoSQL DBMS (DataBase Management Systems) technologies, e.g. RDF triple store and graph databases, depending on data characteristics. Furthermore, the integration of data from many sources will be critical in extending the digital library's comprehensive knowledge store. The QuranicThought library[1] is a potential source for data merging in the context of the ITSERR project. It is important to emphasise that addressing such external information needs will be crucial in database choice as well as in enriching the intellectual resources held within the digital library.

4.3. Leveraging MOMIS for Data Integration

Merging sources that lack a schema is a difficult task [14]. To overcome this issue the intention is to leverage MOMIS (Mediator envirOnment for Many Information Sources) [15], which is a system developed for semi-automated integration of data using thesaurus-based semantic annotation. The goal is to enhance its multilingual capabilities by examining language resources and thesauri. The DM test case will serve as an opportunity to repurpose these tools and demonstrate their effectiveness in such integration tasks. Cross-Language Record Linkage (CLRL) is a challenging task that involves identifying pairs of records referring to the same entity across multiple databases in different languages. To enhance the MOMIS system, it is essential to develop a microservice capable of accurately linking records across languages. Several studies [16, 17, 15], have proposed different mechanisms for addressing this challenge. Since MOMIS' Data Integration procedure is carried out at query-time, the system is extremely valuable as a baseline for the creation of the final framework. With the development of ad-hoc microservice, MOMIS can be exploited in two different ways:

1. The first approach is to employ MOMIS in its standard workflow, where information is retrieved at query time.
2. The second option is to partially materialize the information from one of the sources to improve its efficiency. This approach can also be used for data originating from sources that may not be available on the internet in the long run. In addition, machine learning techniques will be integrated into the query processing to determine the credibility of each source and provide accurate responses that are not merely a combination of different source instances [18]. Given the variety of sources involved in this project, addressing this issue is crucial to ensure Long-term preservation, especially for smaller entities.

As easily noticeable, each of the two alternatives has pros and cons. However, the second choice seems more appropriate given the heterogeneous nature of the sources involved. This is because digital archives may choose to provide temporary unrestricted access to their information, which would make materializing the data necessary. Furthermore, the integration of text documents during query time is also a significant topic, as its importance to the project is currently being discussed. Concluding, the adoption of a system like MOMIS combined

[1] https://www.quranicthought.com

with ongoing further research and development will enable the management of a variety of heterogeneous sources and provide a unified view of data to the users.

4.4. Microservices and Data Update Strategies

For its implementations, the presented endeavour will employ a microservices-based architecture, ensuring a flexible and modular architecture that can promptly adapt to changing requirements. Using microservices allows the system to be scaled, maintained, and updated independently of other components. Data sources within the project will be managed based on their nature, which means that they will be updated at regular intervals depending on the characteristics of each one. The updates can happen daily, weekly, or monthly. This scheduling provides an efficient data handling and keeps the system up to date with the most recent information available in a smart manner.

5. Conclusions

This paper presented the research objectives and methodology of my studies which aim to establish a novel Knowledge Extraction and Multilingual Data Integration framework. As a case study, the DM project demonstrates the possibilities for building a flexible and efficient digital library that encompasses non-Latin languages such as Arabic, Persian, and Azerbaijani. Key challenges and future developments were outlined, including the improvement of Knowledge Extraction, optimising data storage and source integration by leveraging ad expand the MOMIS system for Data Integration. The project intends to overcome these difficulties while ensuring the Long-term preservation of multilingual and multicultural heritage by employing cutting-edge database technology and adopting a microservices-based strategy. This PhD project will hopefully contribute in a significant way by providing a versatile platform for organising and integrating various data sources and languages. This research aims to create a more inclusive and accessible global viewpoint on cultural and historical artefacts, while also fostering collaboration and the building of a constantly expanding knowledge base.

Acknowledgments

I would like to express my gratitude to my tutor Prof. Sonia Bergamaschi, my fellow colleagues within the DBGroup, and the project Maktaba team for their valuable contributions.

This work was supported by the PNRR project Italian Strengthening of Esfri RI Resilience (ITSERR) funded by the European Union – NextGenerationEU (CUP:B53C22001770006).

References

[1] X. L. Dong, D. Srivastava, Big data integration, Synthesis Lectures on Data Management 7 (2015) 1–198.

[2] L. Gagliardelli, G. Papadakis, G. Simonini, S. Bergamaschi, T. Palpanas, Generalized supervised meta-blocking, Proceedings of the VLDB Endowment 15 (2022) 1902–1910.

[3] V. Christophides, V. Efthymiou, T. Palpanas, G. Papadakis, K. Stefanidis, An overview of end-to-end entity resolution for big data, ACM Computing Surveys (CSUR) 53 (2020) 1–42.

[4] 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, 2021, pp. 301–304.

[5] R. Martoglia, L. Sala, M. Vanzini, R. Vigliermo, A tool for semiautomatic cataloguing of an islamic digital library: a use case from the digital maktaba project (short paper), Qurator (2021).

[6] 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).

[7] R. Martoglia, S. Bergamaschi, F. Ruozzi, M. Vanzini, L. Sala, R. A. Vigliermo, Knowledge extraction, management and long-term preservation of non-latin cultural heritages-digital maktaba project presentation (2023).

[8] S. Bergamaschi, D. Beneventano, F. Guerra, M. Orsini, Data integration, Handbook of conceptual modeling: theory, practice, and research challenges (2011) 441–476.

[9] K. Darwish, N. Habash, M. Abbas, H. Al-Khalifa, H. T. Al-Natsheh, H. Bouamor, K. Bouzoubaa, V. Cavalli-Sforza, S. R. El-Beltagy, W. El-Hajj, et al., A panoramic survey of natural language processing in the arab world, Communications of the ACM 64 (2021) 72–81.

[10] C. Yang, K. W. Li, Cross-lingual information retrieval: The challenge in multilingual libraries, in: Design and Usability of Digital Libraries: Case Studies in the Asia Pacific, IGI Global, 2005, pp. 153–170.

[11] Hai - stanford university, ai in the loop: Humans must remain in charge, https://hai.stanford.edu/news/ai-loop-humans-must-remain-charge, Accessed Drecember 2, 2022.

[12] 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.

[13] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.

[14] G. Simonini, L. Gagliardelli, S. Bergamaschi, H. Jagadish, Scaling entity resolution: A loosely schema-aware approach, Information Systems 83 (2019) 145–165.

[15] G. Simonini, L. Zecchini, S. Bergamaschi, F. Naumann, et al., Entity resolution on-demand, Proceedings of the VLDB Endowment 15 (2022) 1506–1518.

[16] Ö. Ö. Çakal, M. Mahdavi, Z. Abedjan, Clrl: Feature engineering for cross-language record linkage, in: Proceedings of Extending Database Technology, 2019, pp. 678–681.

[17] Y. Song, B. Batjargal, A. Maeda, Cross-language record linkage based on semantic matching of metadata, The Database Society of Japan English Journal 17 (2019).

[18] F. Benedetti, S. Bergamaschi, L. Po, et al., Online index extraction from linked open data sources, in: CEUR Workshop Proceedings, volume 1267, DEU, 2014, pp. 9–20.