Unlocking Historical Knowledge: A Semantic Web Approach to Medieval Notarial Document Analysis

García-Menéndez, Ángel[1,*,†], Labra-Gayo, José Emilio[2,†] and Gayo-Avello, Daniel[3,†]

[1]University of Oviedo (Department of Computer Science), West Departmental Building (Office 1.B.22), Polytechnic School of Engineering (Gijón), Asturias, Spain

[2]University of Oviedo (Department of Computer Science), Science Building (Office 206), Llamaquique Campus (Oviedo), Asturias, Spain

[3]University of Oviedo (Department of Computer Science), Science Building (Office 57), Llamaquique Campus (Oviedo), Asturias, Spain

Abstract

The vast amount of historical knowledge embedded in medieval notarial documents remains largely inaccessible without modern computational methods. This paper presents a new methodology for annotating and extracting information from these documents using semantic web technologies, such as RDF, Shape Expressions (ShEx), and SPARQL. By replacing traditional XML-based workflows with a flexible RDF-based ontology, we aim to provide historians with a more powerful and efficient tool for data extraction and analysis. This approach not only facilitates the replication of previous research but also enables the formulation of new research questions, offering insights that were previously unattainable. Additionally, the extracted information will populate a knowledge graph, allowing historians to explore legal and economic structures of the past with greater precision and interconnectedness. Our proposal bridges the gap between historians with limited technical expertise and computer scientists, fostering a collaborative approach to the study of historical documents and the broader field of digital humanities.

Keywords

Semantic web, Digital humanities, Knowledge representation

1. Introduction

Computer science has played a crucial role in the development of various fields of human knowledge. History is no exception, with the computer playing a crucial role[1], first as a mere tool to facilitate traditional process, then putting its calculation power to work in data analysis and, more recently, applying the new technologies in what has come to be known as digital humanities.

In this context, the DocuLab research group[1] from the University of Oviedo has been applying computer science to the study of historical documents by means of different XML-based technologies[2] to annotate and extract information, with a focus on medieval notarial documents.

This research group is currently collaborating with the WESO research group [2] from the same institution to improve their technological stack[3] and find new ways to improve their research capabilities and knowledge management[4]. The main contribution of this position paper is the proposal of a new methodology, based on semantic technologies, that will allow extracting and analyzing the knowledge currently trapped in historical documents.

Our methodology advances the state of the art in analyzing medieval notarial documents by addressing key limitations in current approaches. It leverages Shape Expressions (ShEx)[5], a schema language to validate RDF, SPARQL for improved cross-referenced querying, and introduces a reusable ontology

EKAW 2024: EKAW 2024 Workshops, Tutorials, Posters and Demos, 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), November 26-28, 2024, Amsterdam, The Netherlands

*Corresponding author.

garciamenangel@uniovi.es (G. Ángel); labra@uniovi.es (L. J. Emilio); dani@uniovi.es (G. Daniel)

https://flecktarn121.gitlab.io/ (G. Ángel); https://labra.weso.es/ (L. J. Emilio); https://danigayo.prof (G. Daniel)

0009-0006-3283-8125 (G. Ángel); 0000-0001-8907-5348 (L. J. Emilio); 0000-0002-4705-6891 (G. Daniel)

[1]https://doculab.grupos.uniovi.es

[2]https://www.weso.es/

specifically designed for medieval notarial documents. These innovations enable historians to explore interconnected legal, social, and economic data more comprehensively, enhancing both the depth and breadth of historical analysis. The approach is adaptable for different document types and extendable to broader domains within historical document analysis, filling a gap in existing semantic web ontologies.

2. Related work

The application of semantic technology to create corpuses of historical texts, making their information, a part of our collective cultural heritage, accessible and manageable is not unheard of[6] [7][8][9].

Early in this century we can find examples of the semantic web being used to make the information in historical documents accessible and searchable. Such examples can include the Europeana[10] and Orlando[11] projects, that use RDF to represent the information in their datasets, and the LODE framework[12], which allows to process manuscripts, annotate them using RDF and provide tools to explore and enrich the content. More recently, we find efforts to help preserve and analyze the tragic events of the holocaust[13].

There also exists several proposals to create ontologies to represent cultural heritage information, which may include CIDOC-CRM[14], PROV-O[15] or FRBR[16], each of them centered on different aspects of the information.

The Pelagios Network is also worth mentioning[17], as it groups together several partners that work on the research and use of historical data. Among the activities of the network, we can find the semantic annotation of texts to explore and link information, in a wide range of topics, ranging from World War II testimonies to medieval sea charts[3].

Another recent example of semantic technologies applied to the study of history is the Warsampo project[18]. Information from Finnish World War II documents is extracted and semantically analyzed in order to populate a knowledge graph[19]. This provides a useful tool that allows researchers to study wartime history as a succession of events on which diverse actors participate. This research group has also developed other similar projects, regarding legislation [20], parliamentary activity[21] and archaeology[22].

3. Methodology

In figure 1 we can see the current workflow when analyzing the documents. They are transcribed and marked manually, using a TEI inspired[23] XML format, and then are validated using an XML Schema[24]. XPATH queries[25] are used to extract the information from the XML files and answer research questions.

This semi-manual technique allows historians to perform information extraction and analysis that would be almost impossible to do traditionally. Nonetheless, this approach shows some limitations. The TEI specification, although suitable for a wide range of texts, lacks some characteristics desirable in paleography and diplomatics. Furthermore, the necessity of manual transcription by domain experts limits the potential corpus size.

Therefore, we propose a new methodology, based on well established and standardized semantic technologies, that can be summarized by figure 2.

The TEI-based markup will be replaced by a more flexible and powerful RDF alternative[26] using a custom ontology designed using the Neon Project methodology[27].

The transition to RDF will allow, not only to enrich the existing texts and identify key pieces of information, but also to create a web of links between the data, a knowledge graph (KG). Therefore, we will have a tool to represent the different pieces of knowledge and their connections, offering a more precise depiction of reality and its complexity. Plus, the use of RDF will ease the publishing of that knowledge as Linked Open Data, to be used and combined with other available information by the research community.

[3]Available here: https://pelagios.org/case-studies/

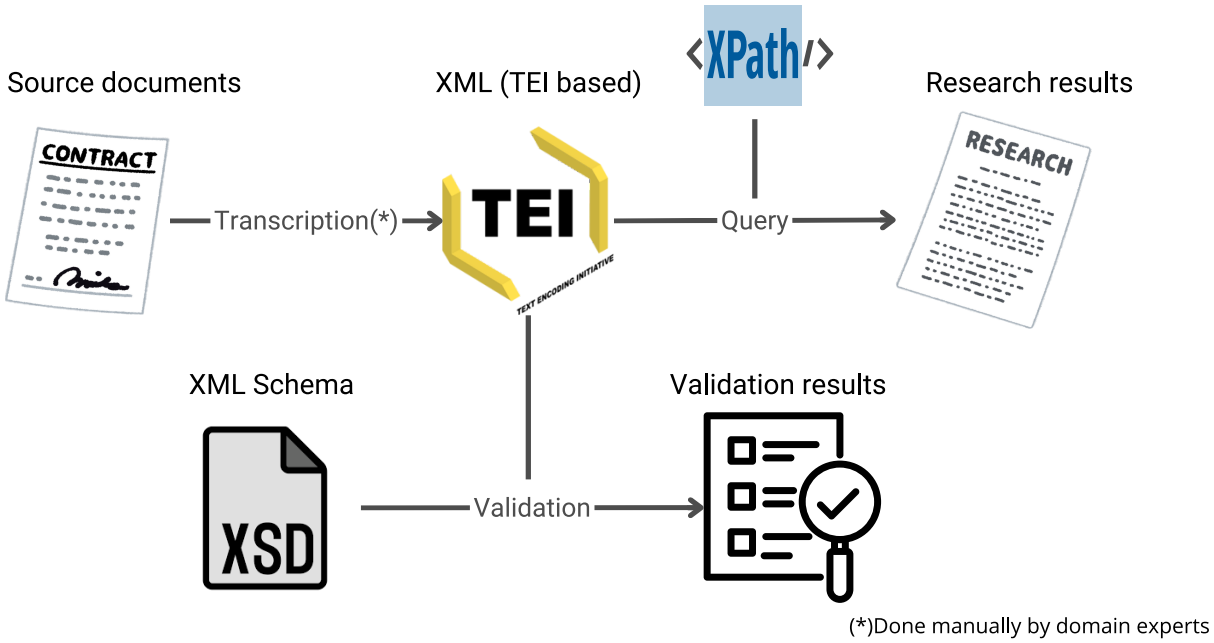

Figure 1: Original workflow using XML-based technologies

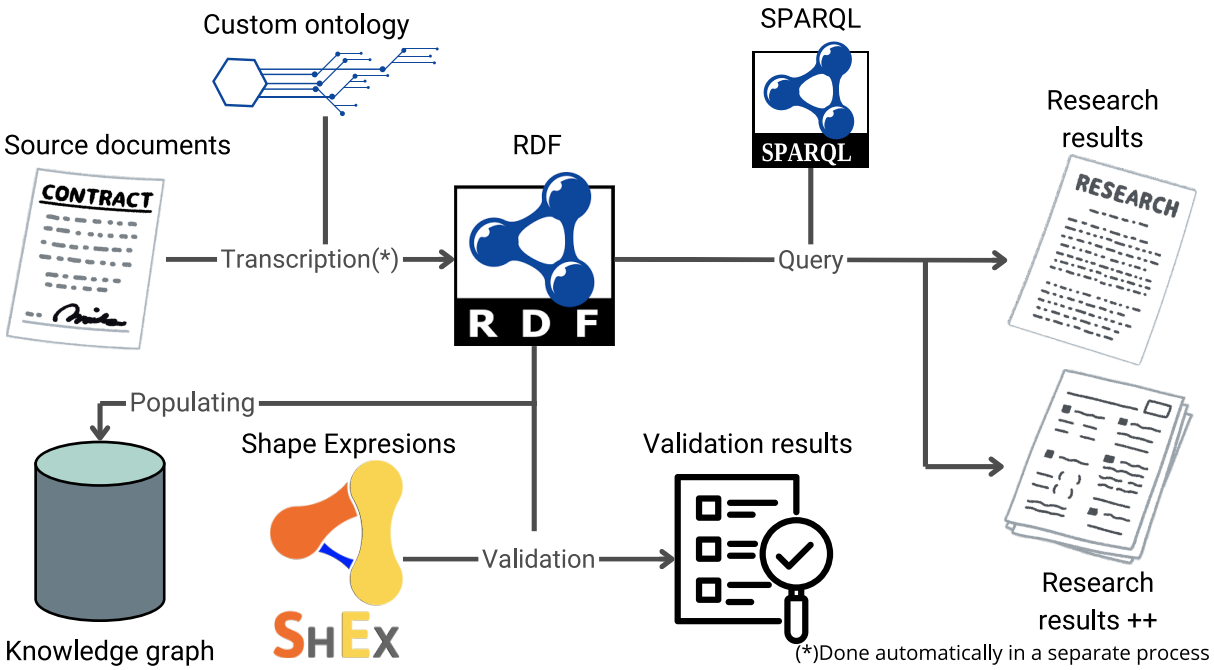

Figure 2: Proposed workflow using semantic technologies

The necessity of a custom ontology is justified by the lack of a suitable alternative for the domain. There are several generalist representations for historical documents (like the TEI specification itself) but none provides the level of detail and expressivity needed for the domain.

As an example of this, consider the following phrase:

> Who ever he were, be he cursed by thy Lord and with Judas condemned to the pits of hell

There is no ontology that allows us to express that this is a penal clause, of spiritual type, that can be located in any part of the document (notaries often forgot clauses and added them at the end of the contract). And similar examples of really specific scenarios can be found all around these documents.

This is why we are working on building a custom ontology. The choice of NeOn as a methodology is not accidental, as it provides a scenario based framework, with common ontology design patters. It contains phases with processes and activities for the creation, maintenance and expansion of ontologies, plus methods for reusing and incorporating elements from already existing ontologies (such as those mentioned in 2).

The validation will be performed using Shape Expressions. Even though ontologies tell how the information should be represented, via the structure of classes and properties, they lack the depth necessary to express every schema feature in a KG. SheEx solve this problem, as they can serve as a tool for both validating and modeling the KG.

SPARQL queries[28] will serve as the tool for information extraction, providing a query language capable of linking different data types with a higher expressivity level than XPATH. It is also a powerful tool to deal with both geo-spacial[29] and temporal uncertainty, very common in historical contexts.

Even though the transition to these technologies is considerable leap forward, there is a remarkable downside: usability. The learning curve for XML, XSD and XPATH was not so stiff, and the technical barrier could be overcome by many historians, specially those who are digital natives, provided some training. However, RDF, ShEx and SPARQL are not so accessible, specially when taking into account knowledge graphs. It will be therefore necessary to create user-friendly tools for history researchers to be able to use the proposed.

In order to evaluate the proposed methodology, the research already performed by DocuLab is replicated, using the new tools and techniques. We are aiming at, not only obtaining the same results, but also at improving them and being able to answer new research questions, impossible to answer previously.

4. Discussion and future work

Currently, much remains to be accomplished. We are defining the model to start working in both the ontology and the RDF representation.

Once that is done, two main tasks would remain. On the one hand, the definition of the competency questions necessary to evaluate the methodology, using research previously conducted by Doculab as a starting point. Then, the definition of new and interesting questions to be answered, alongside our historian colleges.

On the other hand, we would work towards creating a knowledge graph with the extracted information. We would take the Warsampo project as a reference, aiming at providing a useful tool for other researchers to access usable information.

5. Conclusion

As previously stated, historical knowledge is in many cases trapped inside paper, with great potential to be unlocked. We are aiming at unlocking that knowledge, focusing on notarial documents, which can provide useful insights about the legal and economic landscape of the past. In addition, the creation of a knowledge graph will provide a powerful tool for the study of this section of history, allowing for a better access to information and facilitating its analysis.

Plus, with this project we intend to provide resources and tools for two types of people: the historians that lack the technical knowledge to create these kinds of tools by themselves and the computer scientists that might discover a new field of interest.

6. Acknowledgements

We would like to thank Miguel Calleja-Puerta and the DocuLab research group for their help as domain experts and for providing the corpus to work with.

References

[1] M. Peponakis, S. Kapidakis, M. Doerr, E. Tountasaki, From calculations to reasoning: History, trends and the potential of computational ethnography and computational social anthropology, Social Science Computer Review 42 (2024) 84–102. URL: https://doi.org/10.1177/08944393231167692. doi:10.1177/08944393231167692. arXiv:https://doi.org/10.1177/08944393231167692.

[2] M. Calleja-Puerta, G. Fernández Ortiz, El ordenador como herramienta para la investigación diplomática: evolución y perspectivas, Documenta &; Instrumenta - Documenta et Instrumenta 21 (2023) 13–35. URL: https://revistas.ucm.es/index.php/DOCU/article/view/88103. doi:10.5209/docu.88103.

[3] H. García González, E. Albarrán Fernández, J. E. Labra Gayo, M. Calleja Puerta, et al., Converting asturian notaries public deeds to linked data using tei and shexml, in: CEUR Workshop Proceedings, 2020.

[4] J. Álvarez Fidalgo, Applying Semantic Technologies and Knowledge Graphs to improve the representation of historical information, Theses, University of Oviedo, 2023. Ongoing.

[5] E. Prud'hommeaux, J. E. Labra Gayo, H. Solbrig, Shape expressions: an rdf validation and transformation language, in: Proceedings of the 10th International Conference on Semantic Systems, SEM '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 32–40. URL: https://doi.org/10.1145/2660517.2660523. doi:10.1145/2660517.2660523.

[6] I. Nishanbaev, E. Champion, D. A. McMeekin, A survey of geospatial semantic web for cultural heritage, Heritage 2 (2019) 1471–1498. URL: https://www.mdpi.com/2571-9408/2/2/93. doi:10.3390/heritage2020093.

[7] B. O'Neill, L. Stapleton, Digital cultural heritage standards: from silo to semantic web, AI & SOCIETY 37 (2022) 891–903. URL: https://doi.org/10.1007/s00146-021-01371-1. doi:10.1007/s00146-021-01371-1.

[8] A. Meroño-Peñuela, A. Ashkpour, M. Van Erp, K. Mandemakers, L. Breure, A. Scharnhorst, S. Schlobach, F. Van Harmelen, Semantic technologies for historical research: A survey, Semantic Web 6 (2015) 539–564.

[9] E. Hyvönen, Using the semantic web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery, Semantic Web 11 (2020) 187–193.

[10] V. R. Benjamins, J. Contreras, M. Blázquez, J. M. Dodero, A. Garcia, E. Navas, F. Hernandez, C. Wert, Cultural heritage and the semantic web, in: C. J. Bussler, J. Davies, D. Fensel, R. Studer (Eds.), The Semantic Web: Research and Applications, Springer Berlin Heidelberg, Berlin, Heidelberg, 2004, pp. 433–444.

[11] J. Simpson, S. Brown, From xml to rdf in the orlando project, in: 2013 International Conference on Culture and Computing, 2013, pp. 194–195. doi:10.1109/CultureComputing.2013.61.

[12] T. Sztyler, J. Huber, J. Noessner, J. Murdock, C. Allen, M. Niepert, Lode: Linking digital humanities content to the web of data, in: IEEE/ACM Joint Conference on Digital Libraries, 2014, pp. 423–424. doi:10.1109/JCDL.2014.6970206.

[13] H. García-González, M. Bryant, The Holocaust Archival Material Knowledge Graph, in: T. R. Payne, V. Presutti, G. Qi, M. Poveda-Villalón, G. Stoilos, L. Hollink, Z. Kaoudi, G. Cheng, J. Li (Eds.), The Semantic Web – ISWC 2023, Springer Nature Switzerland, Cham, 2023, pp. 362–379. doi:10.1007/978-3-031-47243-5_20.

[14] M. Doerr, The cidoc conceptual reference module: An ontological approach to semantic interoperability of metadata, AI Magazine 24 (2003) 75. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/1720. doi:10.1609/aimag.v24i3.1720.

[15] T. Lebo, S. Sahoo, D. McGuinness, K. Belhajjame, J. Cheney, D. Corsar, D. Garijo, S. Soiland-Reyes, S. Zednik, J. Zhao, Prov-o: The prov ontology, W3C recommendation 30 (2013).

[16] B. Tillett, What is frbr? a conceptual model for the bibliographic universe, The Australian Library Journal 54 (2005) 24–30. URL: https://doi.org/10.1080/00049670.2005.10721710. doi:10.1080/00049670.2005.10721710. arXiv:https://doi.org/10.1080/00049670.2005.10721710.

[17] L. Isaksen, R. Simon, E. T. Barker, P. de Soto Cañamares, Pelagios and the emerging graph

of ancient world data, in: Proceedings of the 2014 ACM Conference on Web Science, WebSci '14, Association for Computing Machinery, New York, NY, USA, 2014, p. 197–201. URL: https://doi.org/10.1145/2615569.2615693. doi:10.1145/2615569.2615693.

[18] M. Koho, E. Ikkala, P. Leskinen, M. Tamper, J. Tuominen, E. Hyvönen, Warsampo knowledge graph: Finland in the second world war as linked open data, Semantic Web – Interoperability, Usability, Applicability 12 (2021) 265–278. URL: https://doi.org/10.3233/SW-200392. doi:10.3233/SW-200392.

[19] M. Koho, E. Heino, P. Leskinen, E. Ikkala, M. Tamper, K. Apajalahti, J. Tuominen, E. Mäkelä, E. Hyvönen, Warsampo knowledge graph, 2020. URL: https://doi.org/10.5281/zenodo.3611322. doi:10.5281/zenodo.3611322.

[20] E. Hyvönen, M. Tamper, E. Ikkala, M. Koho, R. Leal, J. Kesäniemi, A. Oksanen, J. Tuominen, A. Hietanen, Lawsampo portal and data service for publishing and using legislation and case law as linked open data on the semantic web, in: AI4LEGAL-KGSUM 2022: Artificial Intelligence Technologies for Legal Documents and Knowledge Graph Summarization 2022, volume 3257, CEUR Workshop Proceedings, 2022, pp. 41–50. URL: http://ceur-ws.org/Vol-3257/paper5.pdf.

[21] E. Hyvönen, L. Sinikallio, P. Leskinen, M. L. Mela, J. Tuominen, K. Elo, S. Drobac, M. Koho, E. Ikkala, M. Tamper, R. Leal, J. Kesäniemi, Finnish Parliament on the Semantic Web: Using ParliamentSampo Data Service and Semantic Portal for Studying Political Culture and Language, in: Digital Parliamentary data in Action (DiPaDA 2022), Workshop at the 6th Digital Humanities in Nordic and Baltic Countries Conference, long paper, CEUR Workshop Proceedings, Vol. 3133, 2022, pp. 69–85. URL: http://ceur-ws.org/Vol-3133/paper05.pdf.

[22] H. Rantala, E. Ikkala, J. Tuominen, E. Hyvönen, V. Rohiola, E. Oksanen, M. Koho, Findsampo: A linked data based service for analyzing and disseminating archaeological finds, in: 6th Digital Humanities in Nordic and Baltic Countries Conference, poster paper, book of abstracts, 2022, pp. 118–119. URL: http://urn.kb.se/resolve?urn=urn%3Anbn%3Ase%3Auu%3Adiva-472170.

[23] TEI P5: Guidelines for Electronic Text Encoding and Interchange Version 4.3.0, Standard, The TEI Consortium, 2021.

[24] S. Gao, C. Sperberg-McQueen, H. Thompson, W3c xml schema definition language (xsd) 1.1 part 1: Structures, 2012.

[25] S. Gao, C. Sperberg-McQueen, H. Thompson, W3c xml path language (xpath) 3.1, 2017.

[26] D. Beckett, B. McBride, Rdf/xml syntax specification (revised), W3C recommendation 10 (2004).

[27] M. C. Suárez-Figueroa, A. Gómez-Pérez, M. Fernández-López, The NeOn Methodology for Ontology Engineering, Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 9–34. URL: https://doi.org/10.1007/978-3-642-24794-1_2. doi:10.1007/978-3-642-24794-1_2.

[28] S. Harris, Sparql 1. 1 query language, W3C Recommendation 21 (2013).

[29] C. Bernard, Immersing evolving geographic divisions in the semantic Web, Theses, Université Grenoble Alpes, 2019. URL: https://theses.hal.science/tel-02524361, issue: 2019GREAM048.