Converting Asturian Notaries Public deeds to Linked Data using TEI and ShExML

Herminio García-González[1], Elena Albarrán-Fernández[2], Jose Emilio Labra-Gayo[1], and Miguel Calleja-Puerta[2]

[1] Deparment of Computer Science, University of Oviedo, Oviedo, Asturias, Spain
garciaherminio@uniovi.es, labra@uniovi.es
[2] Deparment of History, University of Oviedo, Oviedo, Asturias, Spain
albarranelena@uniovi.es, mcalleja@uniovi.es

Abstract. Comprehension of past events and its reconstruction is one of the tasks performed by historians. With the introduction of computer-aided methods the way in which historians perform their work has been transformed. One of these inclusions is the Semantic Web which can act as an alternative for publication, conciliation, standardisation and integration. Asturian notaries public contracts are a valuable material to understand the society of this epoch, specially in the Middle Ages where a renovation process in the institution was taking place. Therefore, in this work we explore the transformation of TEI-based XML transcriptions of notarial contracts to RDF by means of an heterogeneous data mapping tool which can improve the mechanism to publish Linked Data from existing transcriptions.

Keywords: Notaries public · TEI · XML · Linked Data · RDF

1 Introduction

Elucidating past events and bringing them to our days is one of the tasks performed by historians which search evidences to reconstruct historical discourses. As in many fields, the introduction of computer aided methodologies has opened a new dimension in historical research in which has been coined as Digital Humanities. One field that has had a good reception is the Semantic Web whose technologies have been envisaged as a mean of publication, conciliation, standardisation and integration in the Humanities field [9].

A particular main problem, in Digital Humanities, is the generation of Linked Data from historical material. Although many procedures were proposed, most of them involve creating *ad-hoc* solutions that could only handle with a certain model, and modifications would need much effort or a software expert. Another problem is how to deal with heterogeneous models without creating one solution per schema. In the Semantic Web community new tools that try to deal with data transformation and, also with heterogeneity—of formats and data models—have appeared [3]. Moreover, they try to offer tools that can be used by domain experts without the constant implication of a software expert. Therefore, this kind of

tools can bring a new dimension on Linked Data generation from historical sources.

Among other historical resources, notarial contracts are particularly valuable for the study of western mediterranean societies in the Middle Ages [1]. This kind of documents and the notarial institution itself represented the conjunction of romanist legal traits—inherited from Ancient times—in a profoundly religious culture that was already showing evident signs of transformation. In Asturias, a small northern region of the Castilian Crown, the overall renovation process of the notarial institution in the mid XIII century transformed its own writing tradition.

In this paper we describe our work on transforming transcriptions of medieval Asturian notarial contracts encoded with Text Encoding Inititive (TEI), how this can be achieved with heterogeneous data mapping tools and what are the challenges that this methodology poses.

2 Related Work

In the last years several works have explored the idea of transforming XML-based historical artefacts to Linked Data. This is the case of [11] which explores a conversion from XML/TEI to RDF/XML on historical documents, [7] which discusses the use of XTriples[3] to model, link and visualise XML corpora as Linked Data and [4] which presents a transformation of TEI-XML annotated Latin medieval texts to RDF by means of XSLT.

Regarding the notarial deeds, the most similar work to the one presented in this paper is presented in [5] where the authors produce a dataset of Notarial Archives in Valleta extracting entities, keyphrases and relations from notarial deeds. However, the extraction of entities using artificial intelligence techniques instead of basing the creation on transcriptions suppose a difference between both works.

Although these works explore different ways of transforming data to RDF, none of them tackle the use of heterogeneous data mapping tools, which are capable to integrate—in the same script—data in various forms and formats. This proposal could lead to a faster transformation process due to the centralisation in one tool, the higher flexibility against model changes and addition of other sources of information—with heterogeneous formats like: CSV, HTML, JSON, etc.—and the improvement on learning time from the domain experts.

3 Historical background

In the XIII century, during the reign of king Alfonso X, a renewed doctrine influenced the elaboration of a legal frame fitted to the times and the particularities of the Castilian Crown.

[3] https://xtriples.lod.academy/index.html

This frame included a new legal corpus and the transformation of judicial and documentary practices. In this context, the traditional scribes—most of them coming from clerical institutions—were replaced by public notaries.

For Asturias, an ancient kingdom located in the north of Castile, the new policy established by king Alfonso X meant several changes, as it was the king's intention to modernise not only the administration of his realm, but to transform rural areas into a more dynamic urban ones.

Public notaries assumed their predecessors' role with a whole new meaning: first, written culture no longer belonged exclusively to the Church; second, they offered their services to everyone, no matter their economical solvency or social background; third, their profession was defined by the law; thus, they recorded every single legal action and contract in the daily life of the Asturian society.

As no notarial registers from this early period remain today, we are working with documents issued by these notaries. Most of them were preserved by Asturian monasteries and cathedral, as they frequently used notaries' services to record their economical activities. Nowadays, a great amount of these documents are still guarded by an ecclesiastical institution—the monastery of San Pelayo of Oviedo[4] is one of the richest private archives in the region—while many are also preserved at the Archivo Histórico Nacional[5]—the biggest public-state archive in Spain.

4 Methodology

Asturian notarial contracts from XIII and XIV centuries are held by several private and public-state archives which, in some cases, can hinder their access. Furthermore, in many cases digitised versions are not available. But even with digitised versions, it is still to be proven that accuracy of promising state-of-the-art Optical Character Recognition (OCR) techniques [6] can be transposed to regional variations (i.e., differences in typographies can pose a problem to these OCR techniques). Therefore, the work of an editor is essential to transcribe the writings of this era.

As a first step manuscripts are transcribed to TEI-XML using vocabulary features plus some additions which cover diplomatic elements [1]. This first phase holds the digitised content plus some structure information about the manuscript itself and meta-data[6]. However, entities such as places and names are neither represented unambiguously nor linked with other existing entities. Therefore, this step corresponds with the creation phase of historical information life cycle as proposed by [9].

This version of manuscripts transcription can be queried and published but it has the problem of entities reconciliation and integration with other datasets. As a way to solve these lacks, the translation of these TEI-XML transcriptions

[4] http://sanpelayomonasterio.org/

[5] http://www.culturaydeporte.gob.es/cultura/areas/archivos/mc/archivos/ahn/portada.html

[6] An example manuscript transcribed to XML can be seen on: https://github.com/albarranelena/AsturianNotaries/blob/master/AAA_7.xml

to Linked Data is explored. In this work we have decided to use ShExML as it offers a simple syntax and, as being developed by two of the authors, it can be tuned if it is necessary.

5 Transformation process

The transformation process begins with the creation of the transformation script in ShExML syntax[7].

To create the data model we have taken the schema.org vocabulary to define the general attributes. This vocabulary, in its pending branch[8], offers new types that are suitable for generics attributes of works like the one presented in this paper. Therefore, the archive component type is used to model the content and meta-data of each TEI-XML transcription.

Some of these attributes require to have another type in the object part. For these cases a shape link is made which is a mechanism to define a new shape with a new form that will be linked to the upper one. This is the case of the `schema:locationCreated` which is a `schema:Place` and has a name and a link to a Linked Data Cloud[9] entity. Here, we have linked the `schema:sameAs` attribute with their Wikidata counterpart entity. This process is made using the ShExML matchers feature[10] which allows to replace a string for another string of our choice. For instance, the town of Avilés can be linked with its Wikidata[11] entry. Therefore, linking shapes we are able to create links between generated triples and model schema.org types inside ShExML. With the iterator nesting we are able to cover the tree structure and, also, multiple children from one parent which must be considered as a one triple generation per child.

Once this script is generated we can use the ShExML engine[12] to convert an arbitrary number of files following this encoding model to their RDF counterpart. To check the conversion presented in this paper we also offer an online demo[13] where we can upload the generated script and select the "Convert to RDF" option to generate the RDF output[14].

6 Limitations and challenges

Although this conversion can cover a lot of what is described in the TEI-XML transcription there are some limitations—which are also in line with some limitations encountered in TEI vocabulary and related formal ontologies derivatives

[7] Script available on:
http://herminiogg.github.io/whiseIII-paper-2020/notariesShort.shexml
[8] http://pending.schema.org
[9] https://lod-cloud.net/
[10] http://shexml.herminiogarcia.com/spec/#matcher
[11] https://www.wikidata.org/wiki/Q14649
[12] https://github.com/herminiogg/shexml
[13] http://shexml.herminiogarcia.com/editor
[14] Full output available on:
http://herminiogg.github.io/whiseIII-paper-2020/notariesShort.ttl

[2]. The first problem was shown in the Office shape where the people belonging to an office cannot be represented using schema.org. The most likely relation is the schema:employee; however, the relations in medieval times cannot be understood as being an employee of an organisation but as a guild. It is also a problem that there are not defined relations between the different roles inside a notarial office and there is not a procedure to create these roles.

This problem increases when a diplomatic study is raised. In this case, modelling aspects such as the legal action described in the contract, the tradition of the act or the number and role of the participants cannot be achieved with the current vocabulary. Although this limitation do not restrict the conversion to Linked Data[15] and, it can also be queried through SPARQL queries, it is true that it can limit future inferences and, moreover, it could limit integration with other graphs which is the final goal of Linked Data.

Other ontologies like FRBR [10], NIE-INE[16], RiC[8] and ROAR[17] can define similar concepts to schema.org with more specificity or flexibility. However, they tend to focus in general concepts and meta-data but not on the domain specific content. To the best of our knowledge, there is no domain specific ontology nor vocabulary which defines this topic, and the closest one is the CEI [12] vocabulary which still do not define all the concepts present in our corpora. Therefore, in order to increase transformation inference capability and standardisation, it arises that a new ontology definition for this topic should be tackled.

Another problem is how to identify and disambiguate persons' names which will require a mechanism to identify and disambiguate them from other people with the same name and surname. It is also problematic that these people are not registered in any other repository as may be, for example, the Kings of Spain (e.g.: Wikidata, DBpedia, etc.). This would involve the addition of an entity disambiguation mechanism in ShExML plus the creation of specific algorithms for this case. This kind of knowledge extraction from the text would imply in a simpler and faster process for the transcriber that can focus more on the transcription process and less in the identification and categorisation of entities.

7 Conclusions

In this work we have explored the possibility to apply heterogeneous data mapping tools to a TEI-based XML transcription of notarial contracts in order to convert them to RDF. This transformation was carried out using ShExML, which aims to offer a simple syntax to define these kinds of transformations, and using the schema.org vocabulary to assure the integration of these corpora with other existing or future resources. The process has shown that schema.org and other existing vocabularies are not able to synthesise what is needed for diplomatic

[15] Full ShExML script with diplomatic features: http://herminiogg.github.io/whiseIII-paper-2020/notariesFull.shexml
Full RDF result: http://herminiogg.github.io/whiseIII-paper-2020/notariesFull.ttl
[16] https://github.com/nie-ine/Ontologies
[17] https://leonvanwissen.nl/vocab/roar/docs/

studies. Therefore, we envisage the creation of a diplomatic ontology as an approach to cover this topic. Moreover, we highlight the need for an identification and disambiguation mechanism for person entities to favour further analyses.

Funding This work has been partially funded by the Principality of Asturias through the Severo Ochoa call (grants BP17-29 and BP16-51) and by the Ministry of Economy, Industry and Competitiveness under the calls of "Programa Estatal de I+D+i Orientada a los Retos de la Sociedad" (project TIN2017-88877-R) and "Proyectos de I+D de Generación de Conocimiento" (project PGC2018-093495-B-100).

References

1. Albarrán-Fernández, E.: A TEI-Based model to encode notarial charters (Asturias, 1260-1350 ca .) (Sep 2019). https://doi.org/10.5281/zenodo.3447525
2. Ciotti, F., Peroni, S., Tomasi, F., Vitali, F.: An OWL 2 formal ontology for the text encoding initiative. In: Digital Humanities 2016, DH 2016, Conference Abstracts, Jagiellonian University & Pedagogical University, Krakow, Poland, July 11-16, 2016. pp. 151–153. Alliance of Digital Humanities Organizations (ADHO) (2016)
3. De Meester, B., Heyvaert, P., Verborgh, R., Dimou, A.: Mapping languages analysis of comparative characteristics. In: First Knowledge Graph Building Workshop, part of ESWC2019. pp. 1–8 (2019)
4. Dhouib, M.T., Zucker, C.F., Zucker, A., Corby, O., Jacquemard, C., Draelants, I., Buard, P.Y.: Transformation et visualisation de données rdf à partir d'un corpus annoté de textes médiévaux latins. In: IHM'14, 26e conférence francophone sur l'Interaction Homme-Machine. Lille, France (Oct 2014)
5. Ellul, C., Azzopardi, J., Abela, C.: Notarypedia: A knowledge graph of historical notarial manuscripts. In: On the Move to Meaningful Internet Systems: OTM 2019 Conferences. pp. 626–645. Springer International Publishing, Cham (2019)
6. Firmani, D., Merialdo, P., Nieddu, E., Scardapane, S.: In codice ratio: OCR of handwritten latin documents using deep convolutional networks. In: Proceedings of the 11th International Workshop on Artificial Intelligence for Cultural Heritage co-located with the 16th International Conference of the Italian Association for Artificial Intelligence (AI*IA 2017), Bari, Italy, November 14, 2017. pp. 9–16 (2017)
7. Grüntgens, M., Schrade, T.: Data repositories in the humanities and the semantic web: modelling, linking, visualising. In: 1st Workshop on Humanities in the Semantic Web (WHiSe). pp. 53–64 (2016)
8. Llanes-Padrón, D., Pastor-Sanchez, J.: Records in contexts: the road of archives to semantic interoperability. Program **51**(4), 387–405 (2017)
9. Meroño-Peñuela, A., Ashkpour, A., van Erp, M., Mandemakers, K., Breure, L., Scharnhorst, A., Schlobach, S., van Harmelen, F.: Semantic technologies for historical research: A survey. Semantic Web **6**(6), 539–564 (2015)
10. Peroni, S., Shotton, D.: The spar ontologies. In: The Semantic Web – ISWC 2018. pp. 119–136. Springer International Publishing, Cham (2018)
11. Pollin, C., Vogeler, G.: Semantically enriched historical data. drawing on the example of the digital edition of the "urfehdebucher der stadt basel". In: Second Workshop on Humanities in the Semantic Web (WHiSe) (2017)
12. Vogeler, G.: Towards a Standard of Encoding Medieval Charters with XML. Literary and Linguistic Computing **20**(3), 269–280 (09 2005)