NotaryPedia: Knowledge Graph representation and Visualization of Cultural Heritage Texts

Department of Artificial Intelligence
University of Malta, Malta

Charlene Ellul, Joel Azzopardi, Charlie Abela

The Notarial Archives in Valletta is frequented by different users varying from notaries, historical researchers, students, artists and sometimes also civilians that are interested in their family history, genealogy or possessions. Curating over 20,000 historical manuscripts is not an easy feat and provides a constant challenge.

NotaryPedia is an exciting project aimed at automatically mining data from historical Notarial Acts and uses a scalable approach to store this data. Extracted entities and relations will provide better insight into our past and the multi- disciplinary enrichment of the data will aid researchers to complete the jigsaw puzzle of our ancestors. Thus apart from preserving the information contained within notarial acts, NotaryPedia aims to bring the past back to life by linking people, places and keywords together.

Starting from the oldest and only transcribed registers in the collection[1], dating back to the 15th century, machine learning (ML) techniques are used to automatically extract dates, people, places and keywords. These registers are mainly written in Latin but one can also find some medieval Sicilian and Maltese words where lack of a better word could not be found in Latin. Document annotation is done using the indices of the publications and ML models were trained on this data. The results were outstanding with regards to the automatic recognition of people's names, places, popular keywords and key phrases. Keywords and key phrases such as faldellas, olej, asinj show the subject of the document and provide insights about what the Maltese citizens were trading. The indiction methodology of representing dates is common in these documents and a rule based approach was used to extract such dates.

Furthermore, deed classification was deemed important especially when the scribe omitted the type of deed or when only partial text of the deed is available or is legible. Deeds are classified as Apoca, Debitum, Dos.

The extracted information is stored using a scalable approach where no predefined schema is required. This allows for further enrichment of the data from the documents themselves, from other archives related to different disciplines (like paleography, conservation, Maltese linguistics), from Open Data available on the Web and ultimately also from the findings of historians themselves.

Visually, the extracted entities can be represented as bubbles while relations between these entities can be represented as links between them. This structure is better known as a Knowledge Graph and is shown in Figure 1.

The relations between the entities represent genealogical relationships such as mother of, son of, widow of, the origin of people such as Jacobo-lives in-Hal Manin and limits of geographical areas such as Hal Leu-limits of-Hal Manin. Further enrichment techniques are being investigated to find other relations between buyers, sellers and traded items or services.

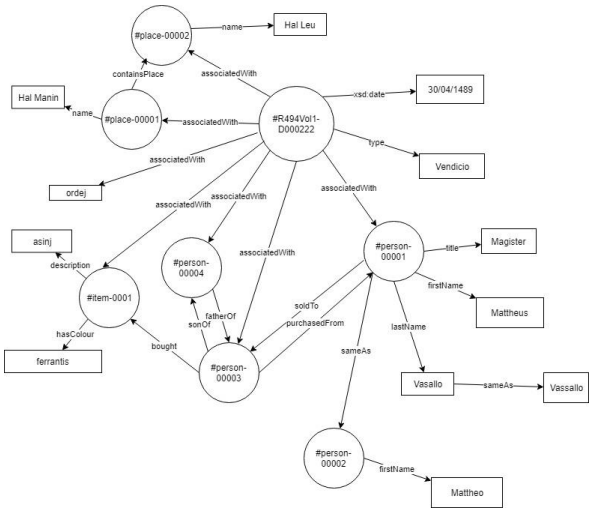

Fig. 1. A subgraph of the knowledge graph representation with relations between entities

A user-friendly web interface will soon be available to visualize the data and enable users to navigate through the Knowledge Graph. Furthermore, crowd sourcing in a controlled manner is going to be gradually introduced whereby users will also be able to directly contribute to the growth of this space.

Follow us on our social media:
https://www.facebook.com/notarialarchives/
https://www.instagram.com/notarialarchivesfoundation/

*This work is partially funded by project E-18LO28-01 as part of the collaboration between the Notarial Archives in Valletta and the University of Malta.

[1]Compiled by Dr Stanley Fiorini in the series Documentary Sources of Maltese History