Excavating the Data Pit: the Portuguese Parish Memories (1758) as a Gold Standard

Ivo Santos[1 [0000−0001−5152−6027]], Fernanda Olival[1 [0000−0003−4762−3451]], and Ofélia Sequeira[1 [0000−0003−2376−1817]]

CIDEHUS - University of Évora, Portugal cidehus@uevora.pt
This work is funded by national funds through the Foundation for Science and Technology, under the project UIDB/00057/2020
http://www.cidehus.uevora.pt/
{ifs,mfo,osequeira}@uevora.pt

Abstract. The common approach to research in History and Archaeology tends to the continuous development of new databases, completely independent of each other with the consequence of data fragmentation, atomisation of knowledge, and ultimately the creation of *data silos*. This happens because of academic tradition, but also because these disciplines work with fragmented information to understand historical data, the contexts, which enables the creation of multiple narratives and interpretations. However, for these disciplines, the *context* is a key aspect that always should be preserved.
The *Memórias Paroquiais* (Parish Memories) correspond to a survey, organized in 3 major parts (land, mountain and river) and are an essential source for obtaining a radiography of Portugal in 1758-1761. We believe that this primary source could reach a new exponent if worked from a different approach: semantically annotated, processed and modeled.
We propose that the Portuguese Parish Memories, due to their intrinsic characteristics, should constitute a Knowledge Base (KB) to connect with other historical sources and research outputs. Ultimately, the Parish Memories could be a Gold Standard for the Natural Language Processing with impact on the research on other historical sources of Early Modern History Portugal, regardless of the knowledge domain.

Keywords: Parish Memories · Knowledge Base · Gold Standard · Natural Language Processing · Open Linked Data.

1 Introduction

The *Memórias Paroquiais* (Parish Memories) are an essential source for obtaining a radiography of Portugal in 1758-1761. They correspond to a survey, organized

in 3 major parts (the locality itself, the mountain and the river), which was printed and sent to those responsible for the dioceses of the country, by notice (*aviso*) of the Kingdom's Secretary of State of 18 January 1758. Following the tradition of other earlier examples, the inquire was limited to the Kingdom and not even the Atlantic Islands were comprised.

The survey included a total of 60 questions: 27 questions about the locality, 13 about the mountain and 20 other about the river. Not surprisingly, the parish priests only responded to what suited their territory. The questions were not limited to historical aspects; they inquired about administrative and jurisdictional (ecclesiastical and secular) issues, demographic data, major 'fruits of the locality', fairs, 1755 earthquake impact, existence of seaport or ramparts. This is about the land or the parish itself. Data about the mountains is fundamental for a study of the natural landscape and the use of resources (they asked about size of the mountains, rivers, special water source, medicinal herbs, mines, lagoons, villages, monasteries and churches). On the river part, the detail was also great: size and dexterity of flow, navigability, direction of the current, fish and fishery related activities, bridges, mills, and cultivation of the margins, etc. And each of the three parts closed with an open request: 'And all that is worthy of memory' and was not included in the survey. It was an invitation to describe what was specific and relevant about each place.

It was up to each bishop or prelate, and in case the See became vacant to the Cathedral chapter, to distribute the inquiries. Many of them turned to the vicars forane and the vicars-general to make the task, sending the survey to the parish priests. It was the latter, or their representatives, who responded by giving information about their parish. Therefore, it is data presented by someone who generally knew well the target territory and mastered writing skillfully.

Some answers denounce that some priests heard knowledgeable persons and that the task deserved some care. It was the case of the parish of Nossa Senhora da Graça de Monforte, in the South of Portugal (Alentejo region): "A printed paper was handed to us from Your Most Illustrious Reverend [the Cathedral Chapter of Elvas]. It arrived trough the Reverend Vicar Forane of this Village, Doctor Francisco Mendes Madeira. So we should answer each question individually, what we knew and what we asked for information. And because it was the time of Lent and the fun of history was embarrassing for us in our occupation, as well as the investigations on antiques and, due to the lack of libraries and old papers, we rely on people of certain trust who have served us willingly." [1, p. 1175]. Some priests listened to the oldest people in the locality [4, p. 393] and others did not consult anyone. There were parish priests who indicated a vast bibliography on the parish [3, p. 287-88]; others, such as the parish priest of one of the parishes of Lisbon (Campo Grande), lamented that his church had no archives with "documents or other old papers from which much news could be drawn" [5, p. 318].

The parish priests had 3 months to respond and many did so quickly, although there were still responses dating from 1759, 1760 and even 1761, espe-

cially in Lisbon, at the time struggling with the impact of the 1755 earthquake [5, p. 29-30].

Note that the parish, headquartered in the main church, was the smallest organizational division of the territory at that time. It was also "the most cohesive and homogeneous territorial and social unit of the life of the Portuguese populations of the past" [3, p. 14]. In this way, the Parish Memories collect data with great relevance, by the micro scale used and due to the fact that the survey cover the whole country, described from the same grid of questions.

The *Torre do Tombo* holds 44 volumes of these handwritten Memories, all available online. In fact, only 41 volumes correspond to the papers sent by the parish priests. To these were added an index volume and two summary data on approximately 500 parishes. These data may be of interest to mayors and local councillors, demographers, botanists, zoologists, architects, and, most of all, historians and archaeologists.

2 Workflow

2.1 Motivations

According to R. Kummer, "Historians do not want to find database records; they want to understand historical contexts" [9]. By other words, History and Archeology work fragmented information to understand historical data, the contexts, which enables the creation of multiple narratives and interpretations. Thus, databases should be prepared for characteristics such as subjectivity, plural views, fragmented data, and uncertainty.

After the advent of Digital Humanities, there is a progressive attempt to apply computer methods to History, however, in the Portuguese case, there is still some difficulty in finding pre-processed, standardized and open-sourced historical data available for analysis. Similarly, not all research projects seek to make their data available, which, in extreme cases, may render the conclusions impossible to reproduce. This gap makes it difficult to verify and discuss results, which is essential to the scientific process. In part, this difficulty has born from the same academic tradition that values the effort needed to find novel and unpublished documentation or to analyze sources already known in an original and innovative way.

Among other data sets, the transcripts made by the Portugal 1758 project are available in CIDEHUS Digital [7] with a textual search, faceting and the possibility of direct navigation for each question, in each parish. This search (based on Apache Solr), while useful for generic research in Parish Memories, does not programmatically reflect relationships between entities and events visible in the text, nor even characterizes relationships present in data or helps the user to understand the context of relevant information. On other hand, the linking of this data with other sources, such as the Portuguese Corography, by Padre António Carvalho da Costa (published between 1706-1712), was only tested briefly through toponymy, limiting more complex analyzes for some users, as they imply previous data processing, knowledge and time for the effort.

In the situation described, the approach chosen is limiting the study of the context, something primarily essential to heritage sciences. This limit is found in studies of History and Archaeology, but also in other research focused on other domains, since, in isolation, any study will tend to *ignore* existing relationships that do not fall within the scope of that same study. This means that the completely independent development of new databases consequence is more data fragmentation, the atomisation of knowledge, and ultimately the creation of *data silos* [10]. In the case of the Portuguese Parish Memories, the current practice is, generally, limited to extracting relevant information for the researcher own interests and, therefore, implies infinite iterations of construction of listings / databases (e.g. the presence / absence of industry or foodstuffs). Ultimately, we consider that research done in this way embodies the constant construction of *data silos* which, if differently done or related to each other, would give us a more complete view of Early Modern Portugal: the context.

2.2 Proposal

We believe that Parish Memories as a source of information could reach a new exponent if worked from a different perspective: semantically annotated, processed and modeled. With this approach, it will be able to become a Knowledge Base (KB) to connect with historical sources and research outputs. We therefore propose that:

- all transcripts and notes should be available online and according to Open Science principles;
- it should be possible for anyone interested in the topic to collaborate (crowdsourced);
- the collaboration is independent of the geographical area and previous background knowledge of the user (with different levels of access to control the quality of data processing);
- is essential to semantically annotate the Parish Memories and construct an ontology to represent this Knowledge Base;
- the annotations should be associated with a thesaurus;

To implement this proposal, a paradigm shift in Parish Memories research practices is required, respecting research interests and academic tradition, but with a normalizing workflow for its study. This way, we intend to provide a route to remove each piece of knowledge constructed from this historical source from it's present day *isolation*, maintaining its original context, both historical and scientific.

2.3 Methods

The proposed workflow is based on the platform INCEpTION. This platform, answers to most of our requirements, allows interactive and collaborative semantic annotation, and there may be different levels of access. These features alone

allow any user to collaborate and to review the process of annotation by another hypothetically more experienced user. Thus, it is possible, for example, that students from Digital Humanities, Palaeography, or even any interested citizen can contribute to the transformation of the Portuguese Parish Memories into a Knowledge Base in all its potentiality. Even in a traditional view, all this has numerous advantages, as a mean of mitigating errors on transcriptions or even valuing the base transcription with knowledge from users who know very well the territory under analysis.

Among others, the semantic annotation capability provides "concept linking, fact linking, knowledge base population, semantic frame annotation" [8] and includes automatic learning algorithms to actively assist the annotation task. From corrected and learned annotations, INCEpTION can therefore suggest annotations in other processed or pre-processed texts. This approach, called machine-assisted interactive annotation (human-in-the-loop with Active Learning (AL)) [6], allows the concept, entity and fact linking task to be faster and more efficient [8]. At the same time, the tool creates the KB, on the fly, as annotations are created.

In addition to these features, it includes corpus search to facilitate annotation [2], is modular, and other algorithms are being developed to assess whether annotations entered are plausible [8].

We propose starting the annotation task by approaching related questions, for example the question "If the margins of the river are cultivated, and if there is a lot of fruit or wild trees" (River-10) and the question "If there are mills, olive oil presses, cloth beater, waterwheels, or some other device" (River-16). This makes it possible to restrict the work of annotations to specific domains of knowledge, in this case mainly economic and patrimonial and in which there are several specialists in CIDEHUS, who can contribute to a successful and smooth start of the task. On the other hand, it will allow to test the NLP detection of similar references in texts where the parish priests answered in free text and not structured by questions.

Taking advantage of the platform's capabilities, we also propose that the annotated transcripts should always include a thesaurus in order to somehow respect the original orthography they were written in, allow searches with current spelling and improve associated knowledge.

After an initial investment by researchers, it is important to engage the community into collaboration, especially the academic, but not only. Universities have a very relevant critical mass in other disciplinary areas that can and should be harnessed. They may not even like history, but they should be made aware that the realities they study have a past dimension, capable of generating new meanings or explanations.

3 Discussion

Context is essential. In the area of Humanities, but not only, it is a primarly essential aspect. As such, databases in History and Archaeology should reflect

all information present in the sources and their context. Technologies such as NLP or LOD are increasingly essential to overlap this issue.

The workflow proposal presented here is not innovative, but allows, for example, to characterize the predominance of fruits per parish and also to the micro-scale (locality, mountain and river); to link the information present in the Parish Memories to databases, to onomastic or prosopographic data, whether respecting the same period of time or not; or even to help to build a question-answer system related to the Early Modern Portuguese History.

The heterogeneity of Parish Memories may thus be its greatest richness as a KB. As such, the progressive implementation of semantic annotations from various knowledge domains will make it possible to provide a basis for linking to other coeval data, to other databases, and to the product of research, independently of the knowledge domain.

In short, we propose that the Portuguese Parish Memories, due to their intrinsic characteristics, should constitute a Gold Standard for the Natural Language Processing of other historical sources of Early Modern History Portugal, regardless of the knowledge domain, with the Portuguese Corography as being the next natural immediate step.

References

1. ANTT (ed.): Memórias Paroquiais, vol. 24, p. 1175
2. Boullosa, B., de Castilho, R.E., Laskari, N.K., Klie, J.C., Gurevych, I.: Integrating knowledge-supported search into the inception annotation platform. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 127–132 (2018)
3. Capela, J.V., Ferreira, A.d.C. (eds.): Braga triunfante ao tempo das Memórias Paroquiais de 1758. s.n., Braga (2002)
4. Capela, J.V., Matos, H. (eds.): As freguesias dos distritos de Aveiro e Coimbra nas Memórias Paroquiais de 1758 : memórias, história e património. Ed. José Viriato Capela, Braga (2011), http://repositorium.sdum.uminho.pt/handle/1822/19969, não contém as p. 129-699.
5. Capela, J.V., Matos, H., Castro, S. (eds.): As freguesias dos distritos de Lisboa e Setúbal nas 'Memórias Paroquiais de 1758': memórias, história e património. Casa Museu de Monção/Universidade do Minho, Braga (2016)
6. de Castilho, R.E., Ide, N., Kim, J.D., Klie, J.C., Suderman, K.: Towards cross-platform interoperability for machine-assisted text annotation. Genomics & Informatics **17**(2) (2019)
7. CIDEHUS: Portugal 1758 - CIDEHUS Digital (2019), http://www.cidehusdigital.uevora.pt/portugal1758, Last accessed on 2019-12-10
8. Klie, J.C., Bugert, M., Boullosa, B., de Castilho, R.E., Gurevych, I.: The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. pp. 5–9 (2018)
9. Kummer, R.: Archaeology and the semantic web—prospects and challenges. In: Frischer, B., Webb Crawford, J., Koller, D. (eds.) Computer Applications and Quantitative Methods in Archaeology (CAA). Proceedings of the 37th International Conference. pp. 178–190. Archaeopress (2010)

10. Migliorini, S., Grossi, P., Belussi, A.: An interoperable spatio-temporal model for archaeological data based on ISO standard 19100. Journal on Computing and Cultural Heritage **11**(1), 1–28 (12 2017). https://doi.org/10.1145/3057929, https://doi.org/10.1145%2F3057929