Extracting Entities and Events from Archives Textual Metadata*

Davide Varagnolo[1], Cássio Rodrigues[1], Ana Martins[1], Dora Melo[2,3][0000−0003−3744−2980], and Irene Pimenta Rodrigues[1,3][0000−0003−2370−3019]

[1] Department of Informatics, University of Évora, Portugal
[2] Coimbra Business School—ISCAC, Polytechnic Institute of Coimbra, Portugal
[3] NOVA Laboratory for Computer Science and Informatics, NOVA LINCS , Portugal
d.varagnolo@studenti.unipi.it, cassiorodrigues@outlook.com, anacatarinasmartins@hotmail.com, dmelo@iscac.pt, ipr@uevora.pt

Abstract. A method for extracting events and entities from ISAD(G) metadata archives elements, that contain text descriptions, is presented. The method consists of applying a set of processing rules according to text fields classification, with the objective of populating the CIDOC-CRM ontology with additional information about events and entities, like persons, locations, dates, relations, baptisms, births, etc. An illustrative example of a 'baptism' classification, extraction, and representation is also presented.

Keywords: Text Classification · Natural Language Processing · Semantic Web · Archives Linked Data Semantic Representation.

1 Introduction

EPISA (Entity and Property Inference for Semantic Archives) is a research project involving the Portuguese National Archives - Torre do Tombo, archival experts, and Information and Computer Science researchers The project aims to design a prototype, as an open-source knowledge platform, aiming to represent archival information on a linked data model. One of the project's major tasks is the semantic migration, i.e, the process to extract and represent the relevant entities and their properties from the existing records in the actual DigitArq, [13]. The DigitArq platform is the Portuguese National archive system that uses well-established description standards, namely the ISAD(G) (General International Standard Archival Description) [3] and ISAAR(CPF) (International Standard Archival Authority Record for Corporate Bodies, Persons and Families) [14] with a hierarchical structure adapted to the nature of archival assets.

To accomplish the migration process an automatic semantic migration prototype, based on Knowledge Discovery, from Digital Archive metadata to populate an ontology in CIDOC-CRM was developed. CIDOC-CRM (Conceptual

* This work is financed by National Funds through FCT - Foundation for Science and Technology I.P., within the scope of the EPISA project - DSAIPA/DS/0023/2018.

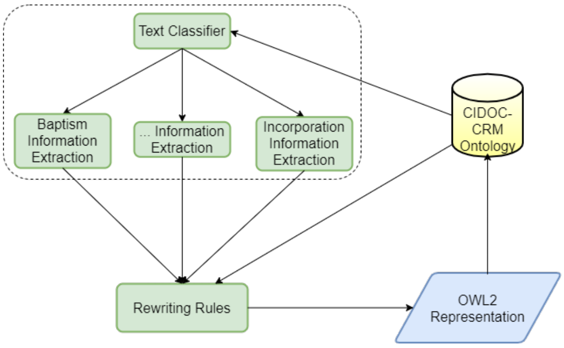
Text Classifier
Baptism Information Extraction
... Information Extraction
Incorporation Information Extraction
CIDOC-CRM Ontology
Rewriting Rules
OWL2 Representation

Fig. 1. Architecture of the Events and Entities Ontology Population.

Reference Model) standard, an ontology developed for museums by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM) [10, 2], was used to build the data model and description vocabularies [4, 11].

The resulting dataset is an OWL knowledge base representation of the existing information in DigitArq as it is, where each DigitArq representation of metadata archives units has a scheme complying to ISAD(G) [3] and ISAAR [14] recommendations. The DigitArq information is organized according to a set of fields and their values. Among this set of fields, there are some that present atomic values, such as the ”Reference code”, the ”Title”, or the ”Recipient”, that do not require further interpretation, and the migration process was done by applying a predefined set of rules establishing the mapping between ISAD(G) elements and CIDOC-CRM classes and properties. There are other fields, such as 'Scope and content' and Archival and Custodial History that are characterized by having additional information describing its unit, and it is in text format. These texts, usually, have a structure that can be recognized, by using Natural Language Processing (NLP) tools, and giving as output a feature value list that will be the input for the additional ontology Population.

The information extraction from text is not intended to extract all the information, but only parts of information considered important, such as baptisms, births, inventories due to death, incorporation of documents between archives, institutions, persons, and places involved in those events.

Methodologies to extracted general information from text into ontologies are presented in several works, such as [8, 1, 9, 12, 7, 5, 6]. In particular, OntoPrima is a NLP-based Ontology Population system that extracts instances of concepts and relations from text to populate an ontology using NLP techniques.

The goal of this paper is to present the process of applying a set of processing rules according to text fields classification, with the objective of populating the ontology with additional information about events and entities, like persons, locations, dates, relations, baptisms, births, etc.

2 Events and Entities extraction overview

The extraction of information from documents text elements depends on the type of the information, such as baptisms, births, inventories due to death, incorporation of documents in archives, institutions, persons, dates and places involved in those events. The ontology representation of this information and the corresponding mapping rules are defined manually and are presented in the next subsection.

Each type of information and extraction process is defined using GATE (https://gate.ac.uk/), and to decide the type of information in a text linked to a document, an automatic text classifier is used.

Figure 1 presents the architecture of the proposed system that has 3 phases. The first one runs a classifier over the national archives information OWL representation. In the second phase, for each classified text linked to a document (E_{31} Document), the text and the corresponding document reference code are sent to an information extraction process. Finally, the information extraction process extracts a set of relations, that using the OWL rewriting rules (mapping rules), will represent the information extracted in CIDOC-CRM linked to the document that had the information extracted.

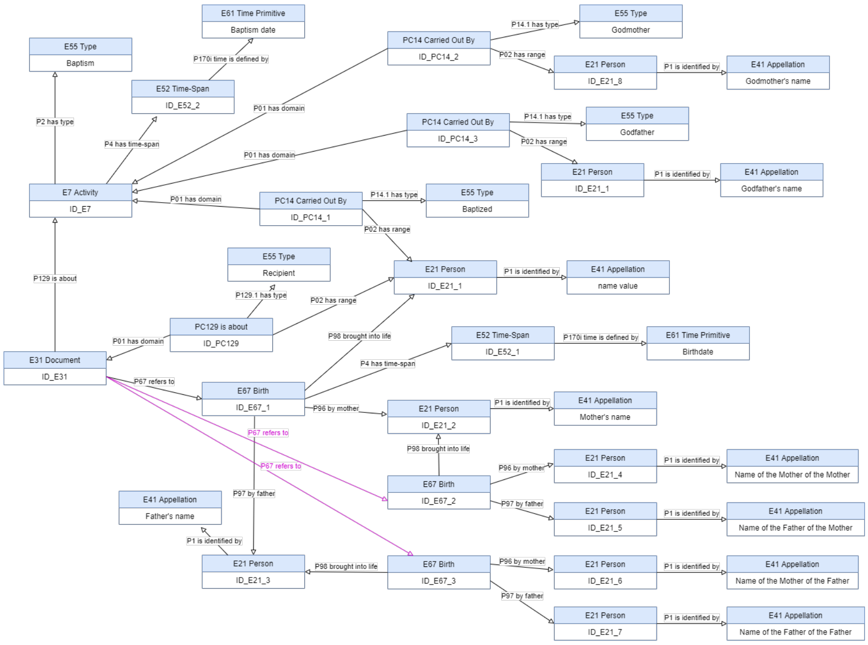

Fig. 2. CIDOC-CRM Representation of a Baptism Activity.

2.1 Representation of events and entities extracted

Each classified text has a set of mapping rules assigned to it, which allows to represent the information extracted in CIDOC-CRM. For better understanding, consider in particular the 'baptism' type of information.

The representation of a baptism in CIDOC-CRM is presented in Figure 2. A 'baptism' is represented as a CIDOC-CRM activity (E_7 activity) linked to the document representation by property 'P_{129} is about'. The 'E_7 activity' has type 'baptism' and can have: a date, the person baptized that is the document receipt, and the persons and their roles in the baptism context like the Godmother and the Godfather. In a baptism description the birth is also described, the birth is represented by the CIDOC-CRM class birth (E_{67} Birth) linked to the document by the property 'P_{67} refers to'. The birth class has properties to represent the parents, 'P_{96} by mother' and 'P_{97} by father'. To represent the grandparents, new birth events are used to represent the birth of the parents. All these births are linked to the document with the text description by property 'P_{67} refers to'.

This representation is automatically generated from the information extracted from the text and the document reference code.

2.2 Text metadata automatic classification

To build an automatic classifier to determine if a natural language text contains some information, such as the description of a baptism, a marriage, a passport request, some material transference between archives, or an enumeration of entities, a sample of Fonds with its hierarchical dependent documents was chosen and semi-manually tagged. The classified texts were then used to build the datasets for the automatic classifiers.

Corpus for the Classifications A set of 3,800 Portuguese National Archives documents was selected from four district archives, and for each document the information was represented as showed in Table 1 when adequate.

The text values of the ISAD(G) elements, such as "Scope and content" or "Archival and Custodial History", from the sample is used to define the automatic text classifier that will be applied to decide if a text, defined as a string and linked to a document in the Archives OWL2 representation, should be used to extract events and entities to populate the ontology.

Classifiers performance Since the text categories present in the selected Fonds were unbalanced, from the 3,800 texts manually classified, a sample of 300 text for each category was selected. Then, Decisions Trees were used to build a classifier for each information type. The dataset for each Decision Tree was obtained by selecting 300 positive text occurrences and 300 negative occurrences selected from the other texts categories positive occurrences. The classifier was built dividing the dataset intp 2 subsets, one for training (70%) and another for testing (30%). The classifiers achieved 100% accuracy.

Table 1. Examples

Classification	ISAD(G) Element	Text
Incorporation	Archival and Custodial History	"Livros entrados no Arquivo por transferência do Arquivo da Universidade de Coimbra, onde se encontravam provisoriamente em 1976, e por incorporações do Cartório Notarial de Oliveira de Azeméis."
Incorporation	Archival and Custodial History	"Terá sido transferida da Direcção de Finanças de Beja e/ou da Tesouraria da Fazenda Pública para o ADBeja em 1988, ao abrigo do DL n.º 46350, de 22 de Maio de 1965."
Incorporation	Archival and Custodial History	"Proveniente da Direcção Escolar de Beja, em 2001."
Baptism	Scope and Content	"Pais: Manuel Martins Ramos e Ana Joaquina Martins Avos maternos: João Martins de Oliveira e Joana Francisca da Silva Avós paternos: Manuel Martins Ramos e Custódia Maria da Costa Padrinhos: António Martins de Oliveira e Josefina Maria de Jesus Data de nascimento: 9 de Novembro de 1811"

2.3 Events and entities extraction

The information extraction process is defined using GATE (General Architecture for Text Engineering), a Java suite of tools to perform natural language processing tasks over corpus. The tasks are managed by applications that include several language resources. The main application is ANNIE (A Nearly-New Information Extraction System), which is a set of modules comprising a tokenizer, a gazetteer, a sentence splitter, a part-of-speech tagger, a named entities transducer and a coreference tagger.

Despite the fact that GATE does not support (natively) the Portuguese language, the test results obtained in the information extraction tasks are very satisfactory and promising.

Each information extraction task is implemented by defining new rules for two modules:

The Gazetteer (originally, geographical dictionary or directory used in conjunction with a map or atlas) are entity dictionaries used in Name Entity Recognition task. In the Annie application, these entity dictionaries help to tag different words in the texts.

The Named Entity Transducer is a technique of Name Entity Extraction via Finite State Transducer, managed by rules. In Annie application, it is possible to create and modify rules, using Jape (Java Annotation Pattern Engine, https://gate.ac.uk/wiki/jape-repository/). A Jape grammar consists of a set of phases, each of which consists of a set of pattern/action rules. The

> phases run sequentially and constitute a cascade of finite state transducers over annotations. The left-hand-side (LHS) of the rules consist of an annotation pattern description. The right-hand-side (RHS) consists of annotation manipulation statements. Annotations matched on the LHS of a rule may be referred to on the RHS by means of labels that are attached to pattern elements.

Consider, as an illustrative example, the text presented in Table 1 and classified as a 'baptism'. This text has a predefined structure:

- All the relatives have the structure
  `{DegreeOfKinship}: {MaleRelative} e {FemaleRelative}`
- The date of birth has a similar structure too, but checking the possible cases it has two variants:
  `Data de nascimento: {DayInNumbers} de {NameOfMonth} de {YearInNumbers}`
  `Data de nascimento: {DayInNumbers}-{MonthInNumbers}-{YearInNumbers}`

This kind of structured text is adequate to define rules that helps to extract entities. To achieve this purpose, two lists of names were created in the Gazetteer to automatically tag the months and nouns designations of the degrees of kinship in Portuguese (months_pt.lst and relatives_pt.lst). Portuguese proper names are also extracted, but there is no need of applying named recognition techniques, because the pattern of this kind of texts allows the extraction process to be carry out directly with the Jape rules defined.

After the Gazetteer processing, all the degrees of kinship and the months that appears in the texts are tagged. Then and after the processing of the Sentence Splitter and the POS Tagger, the Jape rules are applied to extract the name of the relatives and the birthdates. Concerning the other text classification, specialized rules are defined to extract the information (in order to cover different patterns), that have a priority system to be triggered. The following Jape rule allows to extract the father in the baptism classification:

```
Rule: Pais Priority: 100
({Token.string ==~ "[Pp]ais"} {Token.string == ":"} ):intro
({Person.kind == fullName} ):pai
({Token.string == "e"}|({Token.string == "e"} {Token.string == "de"})
):and
({Person.kind == fullName, Token.string != "Avos", Token.string != "Data"}):mae
-->
:pai.Relative = {kind = "pai" }, :mae.Relative = {kind = "mae" }
```

Finally, the output of the each extraction task is a XML file with all the information extracted. This file is the input of the 'Rewriting Rules' module that will update the knowledge base with the new individuals, class instances, and properties that link them together and represent the information extracted in CIDOC-CRM. These new individuals are also linked to the document where they were mentioned.

Information Extraction Performance These information extraction specialized processes are still under development. However, preliminary evaluation of the process for baptisms descriptions can achieve a 98% precision and 99% recall, when extracting dates, entities, birth events and baptism activities with the correct roles assigned to the persons who participated in the events, as well as the places. Similar results on precision and recall were obtained for processes to extract the information from very well structured texts, but the experiments with texts classified as incorporation have a much lower precision and recall, due to the fact that those texts are not so well structured. These kind of texts require the use of syntactic information and semantic similarity to recognize the events and the roles in the events.

The experiments to evaluate the precision and recall of the extraction processes are done by choosing a set of documents, where an ISAD(G) element is classified as the process category and for each text a human decides if the information extracted is correct and if all the information was extracted. In a near future an improvement of this evaluation will be done by trying to automatize some parts of the human verification.

3 Conclusions and Future work

A method for extracting information from ISAD(G) elements, that contain text descriptions, was proposed in this paper. The method has 5 phases: the definition of the information to be represented, the definition of the rules to map the information in CIDOC-CRM, a classifier for selecting the texts that convey the information, a process for automatic extract the information, and its representation CIDOC-CMR using the mapping rules.

The experiments made in extracting events and entities from text documents, for some types of information, show that the proposed strategy is adequate even if some processes such as the extraction of incorporation information need to be improved. The classifiers evaluation shows that it is easy to build a classifier with a very good performance to classify the text in the text elements. The task of manually annotate the set of documents for the classifier is laborious and time consuming since it has to be done for at least a sample of 500 positive occurrences of texts for each information type.

The evaluation of the final results is done by evaluating the correctness of the representation of the events and entities in the final Ontology, which is also laborious and time consuming and requires specialized knowledge on the ontology representation. As future work, it is intended to improve the evaluation process.

References

1. di Buono, M.P., Monteleone, M., Elia, A.: How to populate ontologies. In: Métais, E., Roche, M., Teisseire, M. (eds.) Natural Language Processing and Information Systems. pp. 55–58. Springer International Publishing, Cham (2014)

2. ICOM/CIDOC: Definition of the CIDOC Conceptual Reference Model. ICOM/CRM Special Interest Group, 7.0.1 edn. (October 2020)
3. International Council on Archives: ISAD(G): general international standard archival description, Second Edition. Springer Nature BV (2011)
4. Koch, I., Freitas, N., Ribeiro, C., Lopes, C.T., da Silva, J.R.: Knowledge graph implementation of archival descriptions through cidoc-crm. In: Doucet, A., Isaac, A., Golub, K., Aalberg, T., Jatowt, A. (eds.) Digital Libraries for Open Knowledge. pp. 99–106. Springer International Publishing, Cham (2019)
5. Kordjamshidi, P., Moens, M.F.: Global machine learning for spatial ontology population. Journal of Web Semantics **30**, 3–21 (2015)
6. Leshcheva, I., Begler, A.: A method of semi-automated ontology population from multiple semi-structured data sources. Journal of Information Science **0**(0) (2020)
7. Lubani, M., Noah, S.A.M., Mahmud, R.: Ontology population: Approaches and design aspects. Journal of Information Science **45**(4), 502–515 (2019)
8. Makki, J.: Ontoprima: A prototype for automating ontology population. International Journal of Web/Semantic Technology (IJWesT) **8** (2017)
9. Maynard, D., Li, Y., Peters, W.: Nlp techniques for term extraction and ontology population. (2008)
10. Meghini, C., Doerr, M.: A first-order logic expression of the cidoc conceptual reference model. International Journal of Metadata, Semantics and Ontologies **13**(2), 131–149 (2018)
11. Melo, D., Rodrigues, I.P., Koch, I.: Knowledge discovery from isad, digital archive data, into archonto, a cidoc-crm based linked model. In: Proceedings of the 12th International Joint Conference on Knowledge Discovery, KEOD - Volume 2. pp. 197–204. INSTICC, SciTePress (2020)
12. Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., Zavitsanos, E.: Knowledge-Driven Multimedia Information Extraction and Ontology Evolution: Bridging the Semantic Gap. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
13. Ramalho, J.C., Ferreira, J.C.: Digitarq: creating and managing a digital archive. In: Building Digital Bridges: Linking Cultures, Commerce and Science: 8th ICCC/IFIP International Conference on Electronic Publishing held in Brasília - ELPUB 2004, Brazil, June, 2004. (2004)
14. Vitali, S.: Authority control of creators and the second edition of isaar (cpf), international standard archival authority record for corporate bodies, persons, and families. Cataloging & classification quarterly **38**(3-4), 185–199 (2004)