

Evaluating and Exploring Text Fields Information Extraction into CIDOC-CRM

Davide Varagnolo¹, Guilherme Antas³, Mariana Ramos³, Sara Amaral³, Dora Melo^{2,3}^a
and Irene Pimenta Rodrigues^{1,3}^b

¹*Department of Informatics, University of Évora, Portugal*

²*Polytechnic of Coimbra, Coimbra Business School—ISCAC, Coimbra, Portugal*

³*NOVA Laboratory for Computer Science and Informatics, NOVA LINCS, Portugal*

Keywords: Natural Language Processing, Knowledge Representation, Knowledge Discovery, Semantic Web, Archives Linked Data Semantic Representation.

Abstract: This paper presents a method for extracting information from ISAD(G) elements, that contain semi-structured text descriptions. Natural language processing is done using Gate environment and defining the set of Jape rules necessary to process the text and extract the intended information. The evaluation of the information extraction processes is done in a sample of 800 records for each type of information, and a dataset that is manually built for each type of information considered, such as baptisms, passport requisitions testaments, etc. The implementation of several automatic information extraction processes enables the population of the CIDOC-CRM knowledge base with new linked events and entities automatically. The exploration of the information, migrated from DigitArq and extracted from text descriptions represented in CIDOC-CRM, is done through SPARQL queries enabling new visualisations of the archival records and the retrieval of information collected in different records from different archives.


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, (Ramalho and Ferreira, 2004). The DigitArq platform is the Portuguese National archive system that uses well-established description standards, namely the ISAD(G) (General International Standard Archival Description) (International Council on Archives, 2011) and ISAAR(CPF) (International Standard Archival Authority Record for Corporate Bodies, Persons and Families) (Vitali, 2004) with a hierarchical structure adapted to the na-

ture 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 (Melo et al., 2022b). CIDOC-CRM (Conceptual Reference Model) standard, an ontology developed for museums by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM) (Meghini and Doerr, 2018; ICOM/CIDOC, 2020), was used to build the data model and description vocabularies (Koch et al., 2019; Melo et al., 2020; Melo et al., 2022b). The semantic migration process is supported by a set of description mapping rules that models the information in CIDOC-CRM representation.

The resulting dataset is an OWL knowledge base representation of the existing information in DigitArq, where each DigitArq representation of metadata archives units has a scheme complying to ISAD(G) (International Council on Archives, 2011) and ISAAR (Vitali, 2004) recommendations. The DigitArq information is organized according to a set of fields and their values. Among this set of fields,

^a <https://orcid.org/0000-0003-3744-2980>

^b <https://orcid.org/0000-0003-2370-3019>

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 process of extracting information from text is not intended to extract all the information, but only parts that are considered to be important, such as passports, baptisms, and births records, inventories due to death, incorporation of documents between archives, institutions, persons, and places involved in those activities.

Methodologies to extract general information from text into ontologies are presented in several works, such as (Makki, 2017; di Buono et al., 2014; Maynard et al., 2008; Petasis et al., 2011; Lubani et al., 2019; Kordjamshidi and Moens, 2015; Leshcheva and Begler, 2020). In particular, Onto-Prima 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 an evaluation of the automatic extraction of information about events and entities, like persons, locations, dates, relations, baptisms, births, etc; and the exploration of this information to obtain a new organization of the archives information.

The remainder of this paper is organized as follow. Section 2 provides an overview about the events and entities that are important to be extracted from text ISAD(G) elements. The proposed approach for extracting events and entities from semi-structured text is presented in Section 3. The evaluation of the extraction information process is detailed in Section 4. In addition, the exploration of the CIDOC-CRM knowledge base, through SPARQL queries, and the advantages of the new information representation are also presented. Finally, in Section 5, conclusions and future work are drawn.

2 EXTRACTION OVERVIEW

The events and entities to be extracted from the documents text ISAD(G) elements depends on the type

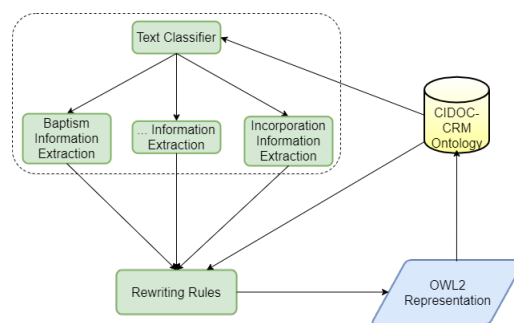


Figure 1: Architecture of the Events and Entities Ontology Population.

of the information provided by those documents. It is possible to identify various types, such as passports and baptisms records, births, inventories due to death, incorporation of documents in archives, institutions, persons, dates, and places involved in those activities. The CIDOC-CRM ontology representation of this information and the corresponding mapping rules are defined manually and are presented in Section 3.2.

The type of the information, conveyed by a text linked to a document, is determined by using an automatic text classifier. After the classification, a proper extraction process is applied. Some of these written texts follow a semi-structured scheme, allowing the definition of a methodology for extracting information based on this scheme. The GATE¹ (General Architecture for Text Engineering) environment is used in this context, see Section 3 for more details.

Figure 1 presents the architecture of the proposed approach 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, which, using the OWL rewriting rules (mapping rules), will represent in the CIDOC-CRM the extracted information linked to the document associated with the target text of the extraction.

3 EXTRACTION PROCESS

The information extraction process from semi-structured text is defined using GATE, a Java suite of tools to perform natural language processing tasks over corpus. The tasks are managed by applications that include several language processing resources. The main application is ANNIE (A Nearly-New Information Extraction System), which is a set

¹<https://gate.ac.uk/>

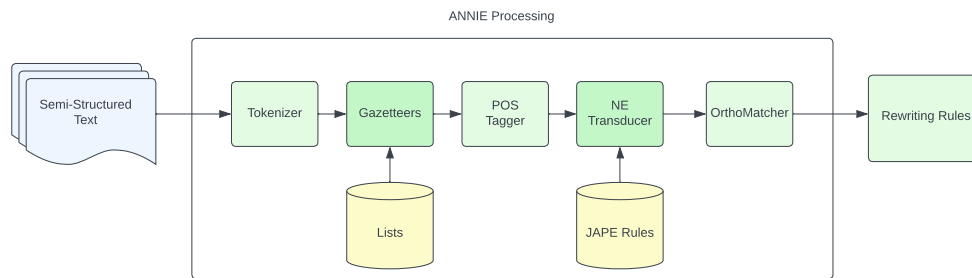


Figure 2: Architecture of the Extraction Process from Semi-structured Text.

of modules comprising a tokenizer, a gazetteer, a sentence splitter, a part-of-speech tagger, a named entities transducer, and a coreference tagger. Figure 2 presents the architecture of the information extraction process from semi-structured text. 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, defined as a geographical dictionary or a directory used in conjunction with a map or atlas, are interpreted, in this work, as entity dictionaries used in the Named Entity Recognition task. In ANNIE application, these entity dictionaries help to tag different words in the texts.

The Named Entity Transducer is a technique of Named Entity Extraction via Finite State Transducer, managed by rules. In ANNIE application, it is possible to create and modify rules, using Jape² (Java Annotation Pattern Engine). 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 the 'Scope and content' of 'José Augusto Teixeira' record³, classified as a 'passport'. It is possible to identify a predefined structure scheme in the written text. For instance:

- The identification of the parents of the applicant


```
{Filho de}: {father} e {mother}
```

- The occupation of the applicant

```
{Profissão}: {occupation}
```

This type of structured text is suitable for defining rules that aid in the extraction of entities. The pattern, as presented in this kind of text, is sufficient to carry out the extraction process only by applying the defined Jape rules, and there is no need to apply named entity recognition techniques.

After the Gazetteer processing, the first-degree relative and the places 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 names of the parents, birthplaces, parishes, and counties. 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.

As an example, the following Jape rules allows to extract the father, the mother, and the occupation of the applicant, from text classified as passport:

```
Rule: Parents Priority: 100
{{Token.string == "[Ff]ilho"}}
{{Token.string == "de"}} {{Token.string == ":"}}:start
{{Person.kind == fullName}:father
{{Token.string == "e"}} {{Token.string == "e"}}
{{Token.string == "de"}}:and
{{Person.kind == fullName}:mother
-->
:father.Recipient = {field= "Father", recipient = "1" },
:mother.Recipient = {field= "Mother", recipient = "1" }

Rule: Profession Priority: 100
{{Token.string == "[Pp]rofissão"}}
{{Token.string == ":"}}:start
(((Token.category == NN)|
{{Token.category == NNP}}):occupation
-->
:occupation.Recipient = {field= "Occupation", recipient = "1" }
```

Finally, the output of 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 with the properties that link those individuals. The rewriting rules represent the information extracted in CIDOC-CRM. These new individuals are also linked to the document where they were mentioned.

²<https://gate.ac.uk/wiki/jape-repository/>

³<http://digitalq.adbpc.arquivos.pt/details?id=1298186>

3.1 Passports Requisitions

Consider the DigitArq record of the passport requisition of 'José Augusto Teixeira'. The ISAD(G) 'Scope and content' element from the passport record contains information about the applicant, such as parents names, age, occupation, literacy, marital status, and also the travel destination. The written information contained in the 'Scope and content' has an intentional scheme in mind, with the aim of facilitating the recording of information by archivists, as well as by those who wish to consult it.

The information that can be extracted from the ISAD(G) 'Scope and content' element of passports records:

Recipient - Name of the person that requested the passport.

Father Name - Name of the father of the recipient.

Mother Name - Name of the mother of the recipient.

Place of Birth - Local, Parish and County Names in Portugal.

Age - Age of the recipient

Literacy - Normally the annotation is: can write or cannot write

Marital Status - The recipient marital state: married, single, widow or divorced

Occupation - The recipient occupation, such as farmer, domestic, mason, tailor, student, etc.

Place of Destination - A country or a country and a town where the recipient is going to.

Passport Emission Date - The ISAD(G) 'date range' element of the requisition document.

3.2 CIDOC-CRM Representation

Each text classification type 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 'passport' type of information and the example previously presented in Section 3.1.

The representation of a 'passport' record in CIDOC-CRM is presented in Figure 3. A 'passport' requisition 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 'passport' and can have: a date, the applicant person that is the document recipient, his parents, the destiny place, and other information in the context of a DigitArq passport record, see Section 3.1 for more details. In a passport record description the birth is also

implicitly described, when persons are referred and the relative relations between them. 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'. 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 using the information extracted from the 'Scope and content' text, and from the document reference code and title.

4 EVALUATION

The evaluation of the different extraction processes, such as baptisms or passports descriptions, is done individually for each process topic considered. Therefore, for each process classification type, a set of records was randomly selected from DigitArq and with information accordingly classified with the process topic. A dataset, a CSV file, is manually built with the information of each set of records. 800 records are used for each topic. Another dataset is build containing, for each set of records chosen, the information represented in CIDOC-CRM and retrieved through SPARQL queries. The retrieved information is saved in a CSV file with the same structure of the dataset.

Dataset. A dataset for a specific topic contains the information that the researcher considers to be adequate to extract from the text. The dataset file is organized by a set of columns with the values of each record in each row. Each topic is independently and individually worked by 3 researchers. When the researchers finish a topic, a final dataset for the topic is obtained by combining the information of the 3 datasets produced by them. The information presented in the ISAD(G) 'Scope and content' element may depend on how the Archives are used (Portugal has one Archive for each District and there are other institutions with their own archives, each one of them following their own orientations on how to display and update the digital information). For instance, in some archives the 'literacy' is always displayed in the passport records, but there are others that do not include this information.

Evaluation Methodology and Results. The evaluation uses as metrics the precision and the recall. In each CIDOC-CRM dataset row, the non-empty value

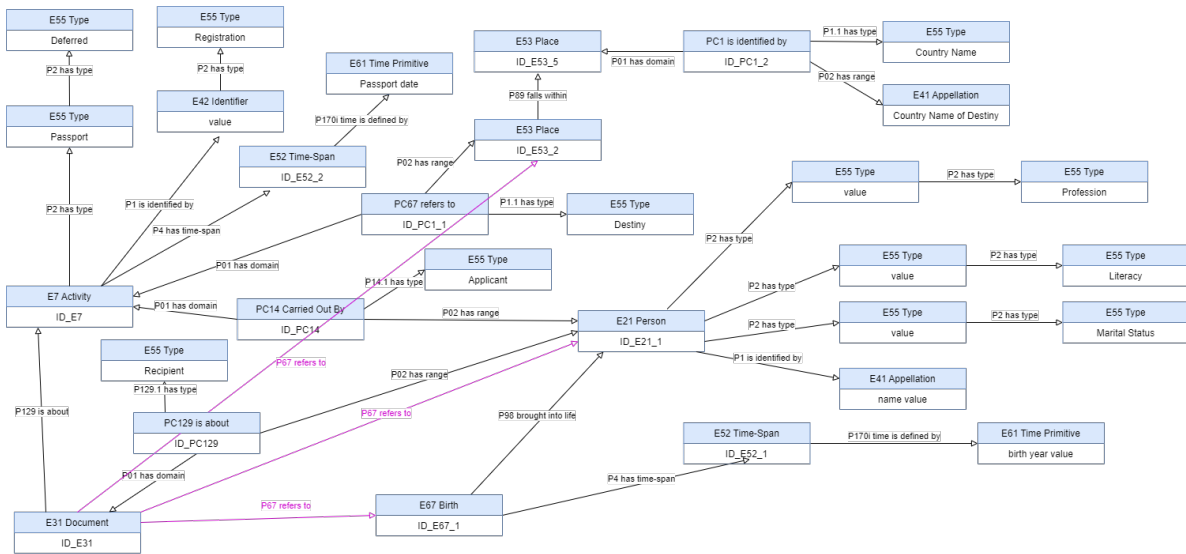


Figure 3: CIDOC-CRM Representation of a Passport Requisition Activity.

of a field (column) is correct if it matches the row and column value of the manually built dataset. In each CIDOC-CRM dataset row, the non-empty value of a field (column) is incorrect if it does not match the row and column value of the manually built dataset. Therefore, the Precision and Recall formulas are as follows.

$$\text{Precision} = \frac{\text{number of correct values}}{\text{number of correct values} + \text{number of incorrect values}} \quad (1)$$

$$\text{Recall} = \frac{\text{number of correct values}}{\text{number of non-empty values in the Manual dataset}} \quad (2)$$

The Precision measures the quality of the values extracted and the Recall measures the quantity of the information extracted.

CIDOC-CRM dataset annotation provides the information needed to refine information extraction processes, allowing developers to look for error patterns in natural language processing rules. New versions of the extraction processes can be evaluated in the manually built datasets to test their adequacy.

These information extraction specialised processes are still under development. New versions to improve precision and recall or new topics that need manually created datasets, new CIDOC-CRM representations, and new Jape rules are being built. However, preliminary evaluation of the process for baptisms and passports descriptions can achieve a precision of 98% and recall of 99%, when extracting dates, entities, birth events, baptism activities or passport requisition activities, with the correct roles assigned to the persons who participated in those events, as well as places.

4.1 Information Exploration

The information extracted from the semi-structured text fields of the Portuguese Archives and represented in CIDOC-CRM ontology enables new and different content searches and visualizations of the information taking into account the archives users needs (Henricke, 2013; Alma'aitah et al., 2020; Marlet et al., 2019; Francart,).

One example are the 'locals' extracted, i.e., places that are linked to activities or entities in some of these information extraction processes, such as the passports requisitions, marriages, divorces or testaments. In the archival records of these type of documents the ISAD(G) 'Scope and content' element allows the recognition of the birth place of the document recipient, and eventually the birth places of other persons mentioned in the document, such as the parents of the recipient. In particular and related to places:

- Passport** - describes the birth place of the passport receipt.
- Marriage** - describes the birth place of the groom and the bride as well as of their parents.
- Divorce** - describes the birth place of the ex-groom and the ex-bride as well as of their parents.
- Testament** - describes the residence place of who makes the testament.

After the extraction process applied over all these kind of documents and the updating of the CIDOC-CRM ontology population with the events and entities extracted, it is possible to obtain some historical information on the Portuguese administrative organization

for a year or a time period. For each processed document, the 'Local' is linked to the document representation, 'E31 Document', using the CIDOC-CRM property ' P_{67} refers to'. This property is inferred from the facts extracted and the set of Semantic Web Rule Language (SWRL)⁴ rules such as the following ones.

```
R1: erlangen-crm:P01_has_domain(?pcl29, ?doc)
    ^ erlangen-crm:P02_has_range(?pcl29, ?person)
    ^ erlangen-crm:P129.1_has_type(?pcl29, 'Recipient')
    -> erlangen-crm:P67_refers_to(?doc, ?person)

R2: erlangen-crm:P67_refers_to(?doc, ?person)
    ^ erlangen-crm:P96_brought_into_life(?person, ?birth)
    ^ erlangen-crm:P7_took_place_at(?birth, ?local)
    -> erlangen-crm:P67_refers_to(?doc, ?local)
    ^ erlangen-crm:P67_refers_to(?doc, ?birth)
```

Rule R1 infers that all the persons that are recipients from a document are referred by the document. Rule R2 infers that if a person referred by a document has a birth, then the birth and the place of birth are referred by the document.

These rules are applied to the representations obtained from specific kind of documents, such as passport requisitions, baptisms, and other events (E7 Activity), during the information extraction process from the ISAD(G) 'Scope and content' element.

Dates can be obtained from text descriptions or from the ISAD(G) 'Date Range' element, which means that the archival material was produced in the value of this field, usually a date range.

In the archives modelling in CIDOC-CRM (Melo et al., 2022b), every ' E_{31} Document' is linked to an object (' E_{22} Human Made Object') that may have a production event (' E_{12} Production') linked to it. The production event may have the date of production of the archival material. The production event links this date in CIDOC-CRM by the property ' P_{108i} was produced by' and the date is represented by an individual of ' E_{52} Time-span' that can be qualified by begin and end date.

As an example of the type of search that it is possible to achieve with this new representation of the information, consider the following cases.

Query 1: *What are the locals and their parishes located in the county 'Bragança',⁵ between 1900 and 1910?*

Figure 4 presents this question expressed as an SPARQL query diagram, based on the CIDOC-CRM representation of the events, such as passports requisitions (see Figures 3). This diagram can be straight forward translated into the SPARQL query presented as follows.

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
```

```
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX cidoc: <http://erlangen-crm.org/200717/>

SELECT ?local ?parish WHERE{
  ?doc cidoc:P70_documents ?hmObj .
  ?hmObj cidoc:P108i_was_produced_by ?prod .
  ?prod cidoc:P4_has_time-span ?t1 .
  ?t1 cidoc:P79_beginning_is_qualified_by ?ti .
  ?ti cidoc:P80_end_is_qualified_by ?tf .
  ?ti rdfs:label ?ti_datetime .
  ?tf rdfs:label ?tf_datetime .
  BIND(year(xsd:dateTime(?ti_datetime)) AS ?yeari) .
  BIND(year(xsd:dateTime(?tf_datetime)) AS ?yearf) .
  filter(?yeari >= 1900 && 1910 >= ?yearf) .

  ?doc cidoc:P67_refers_to ?loc .
  ?loc cidoc:P89_falls_within ?local0 .
  ?loc cidoc:P89_falls_within ?local1 .
  ?loc cidoc:P89_falls_within ?local2 .

  ?pc0 cidoc:P01_has_domain ?local0 .
  ?pc0 cidoc:P1.1_has_type ?type0 .
  filter(?type0 = "County Name") .
  ?pc0 cidoc:P02_has_range ?localApp0 .
  ?localApp0 rdfs:label ?local0 .
  filter(?local0="Bragança") .

  ?pcl cidoc:P01_has_domain ?local1 .
  ?pcl cidoc:P1.1_has_type ?type1 .
  filter(?type1 = "Place Name") .
  ?pcl cidoc:P02_has_range ?localApp1 .
  ?localApp1 rdfs:label ?local1 .

  ?pc2 cidoc:P01_has_domain ?local2 .
  ?pc2 cidoc:P1.1_has_type ?type2 .
  filter(?type2 = "Parish Name") .
  ?pc2 cidoc:P02_has_range ?localApp2 .
  ?localApp2 rdfs:label ?parish }
```

This query evaluation lists every Local's name and its Parish's name that occur in the text fields or in other ISAD(G) elements, of the Portuguese National Archives records, corresponding to archival materials that were produced in the range from 1900 to 1910. Since this local's names and its parish's names are the ones used in the specified range, varying the range in this query, it is possible to obtain the nomenclature evolution in the administrative organization of Portugal regions, a result that can be useful for sociologists or historians. A dataset, containing an excerpt of the CIDOC-CRM representation of DigitArq records from Bragança District Archive, is available at (Melo et al., 2022a). The dataset also includes two SPARQL query examples to facilitate the exploration of the ontology.

At this moment, the ontology queries are performed in SPARQL, a difficult task even for experts. However, if graph query diagrams, like the one shown in Figure 4 are constructed, the task of translating it into a SPARQL query is facilitated.

⁴<https://www.w3.org/Submission/SWRL/>

⁵Bragança is a Portuguese town.

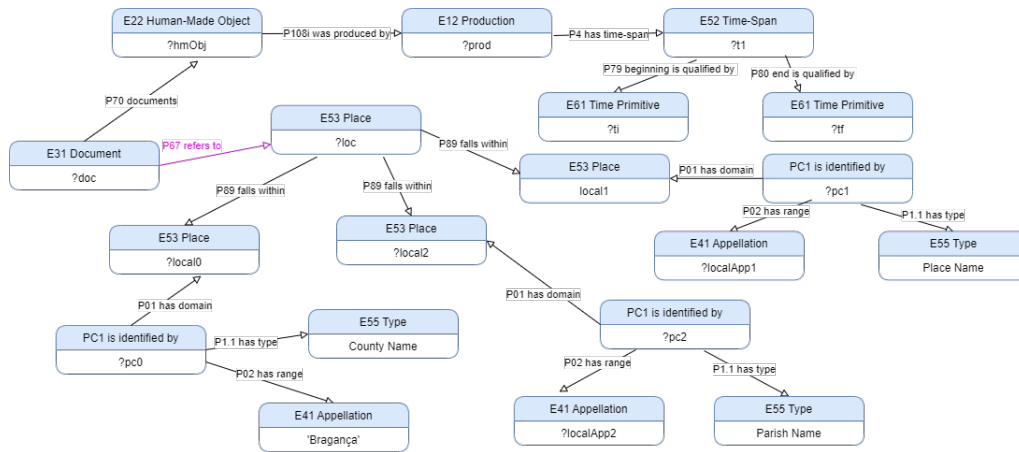


Figure 4: Graph Representation of the Query 1.

Query 2: *What is the number of children per couple, between 1800 and 1850?*

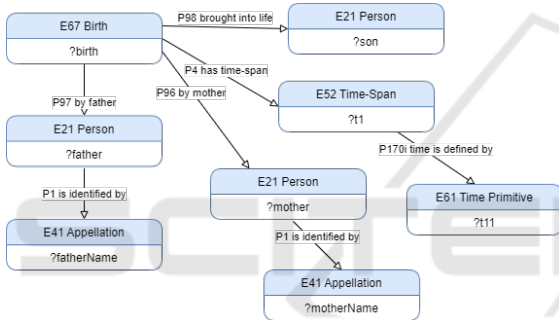


Figure 5: Graph Representation of the Query 2.

This query may be represented by the graph diagram presented in Figure 5, that can be translated into the following SPARQL query.

```
SELECT ?motherName ?fatherName
(count(?son) as ?numberSons)
WHERE {
  ?birth cidoc:P98_brought_into_life ?son .
  ?birth cidoc:P97_from_father ?father .
  ?father cidoc:P1_is_identified_by ?fatherName .
  ?birth cidoc:P96_by_mother ?mother .
  ?mother cidoc:P1_is_identified_by ?motherName .
  ?birth cidoc:P4_has_time-span ?t1 .
  ?t1 cidoc:P170i_time_is_defined_by ?t11 .
  ?t11 rdfs:label ?t1_datetime .
  BIND(year(?t1_datetime) AS ?year) .
  filter(?year >= 1800 && 1850 >= ?year)
} group by ?motherName ?fatherName
```

Birth events are extracted of many types of archival materials from the Portuguese Catholic Church Archives, referring to baptisms and marriages, or from the Portuguese Civil Administration Archives, on passport requisitions, testaments, marriages or divorces. The informa-

tion extracted from these type of documents is used in the above query, integrating the content of all Portuguese archives, including the catholic church ones that normally are incorporated in Civil archives.

5 CONCLUSIONS

A method for extracting information from ISAD(G) elements, that contain semi-structured 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 into CIDOC-CRM representation, the Jape rules necessary to process the text and extract the intended information, a sample of 800 records with the type of information, and the dataset manually built for each type of information considered. These steps allowed the definition and implementation of several automatic information extraction processes that can be both evaluated and populate the CIDOC-CRM knowledge base with new linked events and entities automatically.

With the information of the archival records represented in CIDOC-CRM, it is possible to retrieve information that was not easily available in the DigitArq, the relational database, that only has public access by a portal with simple and advanced search that does not include SQL queries. With the Portuguese archives information migrated into a CIDOC-CRM knowledge base, the information can be searchable using SPARQL or DLquery, allowing new visualisations of the archival records and the retrieval of information collected in different records from different archives. Two examples of these queries were presented that highlight the new types of retrieved in-

formation facilitated by the representation in CIDOC-CRM. The exploration of the CIDOC-CRM information is done in SPARQL or DLqueries, which can be difficult to use by researchers who are interested in exploring the archives content. In a near future, this project work will focus on a friendly interface capable of automatically generate the SPARQL queries to CIDOC-CRM from user natural language queries.

ACKNOWLEDGEMENTS

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.

REFERENCES

- Alma'aitah, W., Talib, A. Z., and Osman, M. A. (2020). Opportunities and challenges in enhancing access to metadata of cultural heritage collections: a survey. *Artificial Intelligence Review*, 53(5):3621–3646.
- di Buono, M. P., Monteleone, M., and Elia, A. (2014). How to populate ontologies. In Métais, E., Roche, M., and Teisseire, M., editors, *Natural Language Processing and Information Systems*, pages 55–58, Cham. Springer International Publishing.
- Francart, T. Sparnatural - Javascript SPARQL query builder. <https://sparnatural.eu/>. Accessed: 2022-09-7.
- Hennicke, S. (2013). Representation of archival user needs using cidoc crm. In *Conference Proceedings TPDFL: International Conference on Theory and Practice of Digital Libraries, Selected Workshops*, Valletta, Malta.
- ICOM/CIDOC (2020). *Definition of the CIDOC Conceptual Reference Model*. ICOM/CRM Special Interest Group, 7.0.1 edition.
- International Council on Archives (2011). *ISAD(G): general international standard archival description, Second Edition*. Springer Nature BV.
- Koch, I., Freitas, N., Ribeiro, C., Lopes, C. T., and da Silva, J. R. (2019). Knowledge graph implementation of archival descriptions through cidoc-crm. In Doucet, A., Isaac, A., Golub, K., Aalberg, T., and Jatowt, A., editors, *Digital Libraries for Open Knowledge*, pages 99–106, Cham. Springer International Publishing.
- Kordjamshidi, P. and Moens, M.-F. (2015). Global machine learning for spatial ontology population. *Journal of Web Semantics*, 30:3–21.
- Leshcheva, I. and Begler, A. (2020). A method of semi-automated ontology population from multiple semi-structured data sources. *Journal of Information Science*, 0(0).
- Lubani, M., Noah, S. A. M., and Mahmud, R. (2019). Ontology population: Approaches and design aspects. *Journal of Information Science*, 45(4):502–515.
- Makki, J. (2017). Ontoprima: A prototype for automating ontology population. *International Journal of Web/Semantic Technology (IJWesT)*, 8.
- Marlet, O., Francart, T., Markhoff, B., and Rodier, X. (2019). OpenArchaeo for Usable Semantic Interoperability. In *ODOCH 2019 @ CAiSE 2019*, Rome, Italy.
- Maynard, D., Li, Y., and Peters, W. (2008). Nlp techniques for term extraction and ontology population. In *Proceedings of the 2008 Conference on Ontology Learning and Population: Bridging the Gap between Text and Knowledge*, page 107–127, NLD. IOS Press.
- Meghini, C. and Doerr, M. (2018). A first-order logic expression of the cidoc conceptual reference model. *International Journal of Metadata, Semantics and Ontologies*, 13(2):131–149.
- Melo, D., Rodrigues, I. P., and Koch, I. (2020). 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*, pages 197–204. INSTICC, SciTePress.
- Melo, D., Rodrigues, I. P., and Varagnolo, D. (2022a). Portuguese Examples (Semantic Migration), Mendeley Data, V1. 10.17632/t2cx9stwf1.1.
- Melo, D., Rodrigues, I. P., and Varagnolo, D. (2022b). A strategy for archives metadata representation on cidoc-crm and knowledge discovery. *Semantic Web*, Pre-press(Pre-press):1–32.
- Petasis, G., Karkaletsis, V., Paliouras, G., Krithara, A., and Zavitsanos, E. (2011). *Knowledge-Driven Multimedia Information Extraction and Ontology Evolution: Bridging the Semantic Gap*. Springer Berlin Heidelberg, Berlin, Heidelberg.
- Ramalho, J. C. and Ferreira, J. C. (2004). Digitaraq: creating and managing a digital archive. In *Building Digital Bridges: Linking Cultures, Commerce and Science: 8th ICCCI/FIP International Conference on Electronic Publishing held in Brasília - ELPUB 2004, Brazil, June, 2004*.
- Vitali, S. (2004). 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.