

Knowledge Discovery from ISAD, Digital Archive Data, into ArchOnto, a CIOC-CRM based Linked Model

Dora Melo¹ ^a, Irene Pimenta Rodrigues² ^b and Inês Koch³ ^c

¹*Coimbra Business School - ISCAC, Polytechnic Institute of Coimbra, Portugal*

²*Department of Informatics, University of Évora, Portugal*

³*INESC-TEC, Faculty of Engineering of the University of Porto, Porto, Portugal*

Keywords: Natural Language Processing, Knowledge Representation and Reasoning, Archives Ontology, Semantic Migration, ISAD(G), CIOC-CRM, Linked Data.

Abstract: This paper presents an automatic semantic migration prototype based on Knowledge Discovery from Digital Archive Data for ontology population in the domain of Archives metadata, ISAD(G). Natural Language Processing (NLP) techniques are used for language processing and Semantic Web techniques for querying and updating the Ontology ArchOnto, a CIOC-CRM (Conceptual Reference Model) extension. This work is done in the context of project EPISA (Entity and Property Inference for Semantic Archives) where the Portuguese National Archives, Torre do Tombo (ANTT) is one of the partners. The data model and description vocabularies we adopted are built upon the CIOC-CRM standard, an ontology, developed for museums by the International Committee for Documentation (CIOC) of the International Council of Museums (ICOM). A detailed example of a baptism document metadata migration is presented to highlight the challenges on the natural language interpretation and the ontology representation.

1 INTRODUCTION

This work is done in the context of the EPISA project (Entity and Property Inference for Semantic Archives), a research project involving the Portuguese National Archives, Torre do Tombo (ANTT), the archival experts from ANTT, and Information and Computer Science researchers. EPISA intends to design a prototype, an open-source knowledge platform, to represent archival information on a linked data model. One of the project 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 (Ramatelho and Ferreira, 2004), the archive national system that uses well-established description standards, namely the ISAD(G) (General International Standard Archival Description) (Scifleet, 2001; International Council on Archives, 2000) and ISAAR(CPF) (Vitali, 2004) with a hierarchical structure adapted to the nature of archival assets.

The data model and description vocabularies we adopted are built upon the CIOC-CRM (Conceptual Reference Model) standard, an ontology, developed for museums by the International Committee for Documentation (CIOC) of the International Council of Museums (ICOM) (Meghini and Doerr, 2018; ICOM/CIOC-CRM Special Interest Group, 2019). The ArchOnto ontology¹ (Koch et al., 2019; Koch et al., 2020) extends the CIOC-CRM ontology to represent archives.

Our goal is to build a prototype for the automatic migration of the information in existing DigitArq records to the ArchOnto representation.

In some of DigitArq records that represent objects metadata and ArchOnto properties and relations, it is possible to identify homogeneity of the information among them, such as the reference code, title, dates, dimension and support, language, and physical location (de Almeida and Runa, 2018; Koch et al., 2020). The semantic mapping of ISAD(G), Archival Metadata, to the CIOC-CRM Ontology can be straightforward in some cases (Bountouri and Gergatsoulis,

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

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

^c  <https://orcid.org/0000-0002-9363-9713>

¹ArchOnto OWL available at https://github.com/feup-infolab/archontology/tree/master/ArchOnto_2020.

2011; Oldman, 2014). However, in some records the information in DigitArq is a text field that must be interpreted in order to extract the entities, events, locals, dates, relations and proprieties to populate the ArchOnto ontology. In those text fields, our approach uses natural language processing tools for Portuguese language to mark terms that are interpreted to insert new instances of concepts and relations into the existing ontology.

The use of natural language processing tools to automatically populate an ontology with the information extracted from text is a current investigation topic with many proposals (di Buono et al., 2014; Makki et al., 2008; Makki, 2017; Maynard et al., 2008).

This paper describes our approach to automatically populate the ArchOnto with the ISAD(G) format of archives metadata.

In section 2, our approach to the automatic migration process from DigitArq records to ArchOnto concepts, relations, and proprieties is presented. Section 3 describes in detail our approach to represent the extracted information from DigitArq text fields in the ArchOnto concepts and proprieties, the examples refer mostly to the representation in CIDOC-CMR concepts and proprieties. Finally, in section 4, we draw conclusions, further work and a future evaluation.

2 AUTOMATIC DOCUMENT METADATA MIGRATION

The automatic migration of DigitArq records into ArchOnto is based on simple translation rules for the fields where there is a mapping between ISAD(G) and CIDOC-CMR or its extension in ArchOnto, this case will be detailed in the next section.

When there is no mapping between concepts, we need to recognize what is written in text fields to obtain concepts and its proprieties to populate the ArchOnto.

Figure 1 shows the architecture of our approach to populate the ontology with the information of the DigitArq HTML records including the information extracted from the text fields.

In the following subsections the main steps of our approach are explained.

2.1 Text Analysis

DigitArq database contains a huge and diverse amount of records, currently over a million. Although the database is structured, using a well-established standard description, namely the ISAD(G) and ISAAR(CPF) with a hierarchical structure adapted to

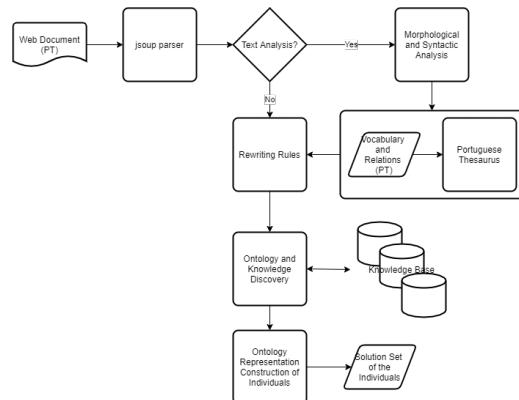


Figure 1: Automatic Migration Architecture.

the nature of archival assets, it is a hard task to analyse and establish the information in text fields that needs to be extracted and decide how to represented it in the ontology.

Along with the development of the DigitArq database, a web-based search engine was developed to allow local and remote users to find and browse the Archive's collections. The result is a well-structured and normalised web page for each record showing the whole information needed to be considered in the migration process.

Each record's web page has a standardized scheme following the ISAD(G) definitions, with the information organized according to a set of known fields and their values. Among this set of fields, there are some that present atomic values, such as "Reference code", "Title", or "Recipient", and others, that do not need further interpretation and the migration process is directly performed by applying a predefined set of rules. Even for fields whose values are not atomic, the migration process is done automatically in the same way as the atomic values. However, from non-atomic values, additional information can be extracted after an adequate analysis and interpretation. For this purpose, it is applied the Morphological and Syntactic Analysis to the non-atomic values. The non-atomic values have a normalized structure, facilitating the analysis and interpretation process.

For instance, consider a baptism record and the non-atomic field "Scope and content". The information of this field is structured in a normalised way to present the names of the parents, the grandparents, and the godparents, as well as the birth date of the person who is baptized, the recipient of the record.

The extraction of the web page content from specific fields is made by using the jsoup library², a Java HTML Parser that provides a very reliable, user-friendly, and easy configuration and parameter adjust-

²<http://jsoup.org>

ments capabilities, for connecting to URLs and extracting and manipulating data.

The use of jsoup library to extract information from web pages to be analyzed and interpreted is not new and can be found in (Fragkou et al., 2016; Cavalcanti et al., 2017), the first one presents a solution for querying Greek governmental site, and the second one presents a solution to extract semi-structured information from web pages in the context of the innovation environments of the state of São Paulo, Brazil.

As an illustration, consider the baptism record³ selected from the DigitArt database, which describes the event baptism of the person called "Ana". Using a set of jsoup functions, it is possible to extract information like the title, the fields and their values, presented in the web page. For instance, the function *title()* allows to extract the title of the record, the function *getElementsByClass()* allows to extract the information per fields, and the function *text()* allows to extract the values of each field. jsoup also provides functions to connect and parse directly the web page source, such as *connect()* followed by *get()*, and *parse()*, respectively. The following fragment of Java code illustrates how this is performed.

```
Document document = Jsoup.connect("http://pesquisa.adporto.arquivos.pt
/details?id=1374655").get();
print(document.title());
Elements fields = document.getElementsByClass("Field");
for(Element f: fields) print(f.text());
```

The function *text()* allows extracting from a field a String containing the field's name followed by its value, which facilitates the task of applying rules to analyse, interpret and migrate the information expressed by these fields. The Strings generated for the "Ana"'s baptism record are

```
Description level Item
Reference code PT/ADRT/PRQ/PPRT01/001/0004/00005
Title type Formal
Date range 1812-02-12 to 1812-02-12
Dimension and support 120x210mm ; papel
Recipient Ana
Scope and content Pais: Manuel de Oliveira e de Rufina Maria Avos
maternos: Manuel da Fonseca e Rosa da Silva Avós paternos: José
de Oliveira e Jacinta de Oliveira Padrinhos: Manuel Martins
Ramos e Maria Francisca Data de nascimento: 10 de Fevereiro
de 1812
Physical location E/20/6/3 - 9.4 - fl. 3 v.,ass.5
Original numbering B1
Language of the material Por (português)
Creation date 5/22/2012 12:00:00 AM
Last modification 3/19/2013 10:55:46 AM
```

The information extracted is adequately analysed, where each fields' name and their values are identified, the adequate ontology representation is established, and the respective ontology individuals are then generated. For this purpose, the non-atomic values go through the Morphological and Syntactic Analysis process to obtain a proper representation that allows and facilitates the ontology representation of

³<https://pesquisa.adporto.arquivos.pt/details?id=1374655>

the additional information that can be extracted from these fields. This process is explained in the next section.

2.2 Morphological and Syntactic Analysis

The Morphological and Syntactic Analysis step consists of the text interpretation by means of lexical and syntactic rules for the Portuguese language, which allows identifying terms and concepts that can be related to the ontology concepts and properties.

The complete grammatical analysis is needed when the information to be interpreted is not directly related with the homogeneous representation of the records structure. It is also needed to translate the terms identified from Portuguese to the English language, once the DigitArq records are expressed in Portuguese and the ArchOnto ontology is represented in the English language.

Consider again the baptism record mentioned before, which describes the event baptism of the person called "Ana". Also, consider the sentence "Pais"⁴: Manuel de Oliveira e de Rufina Maria" extracted from the field "Scope and content".

Using the Portuguese VISL parser⁵ (Bick, 2014), the result of the grammatical analyses of the sentence is

```
pais [pai] <*> <Hfam> N M P @NPHR
:
Manuel de Oliveira [Manuel=de=Oliveira] <hum> <*> PROP M S @NPHR
e [e] KC @CO
de [de] PRP @ADV1
Rufina Maria [Rufina=Maria] <hum> <*> PROP F S @P
```

By applying the parser, the sentence was divided into three parts, the first one "Pais" ("Parents") reflects the relation assigned to the other two parts identified as names, the names of both parents. It also gives, for instance, information about the multiplicity or the gender of the terms, which are useful for future interpretation of the corresponding concepts and proprieties of the ontology.

2.3 Ontology and Knowledge Discovery

The Ontology and Knowledge Discovery step consists in finding the concepts and properties represented in ArchOnto that are related to the terms found at morphological and syntactic analysis level. To broaden the search spectrum, it is also considered the synonyms of the terms found. The ontology concepts and properties found will be then filtered based

⁴Translation of "Pais" is "Parents".

⁵<https://visl.sdu.dk/visl/pt/parsing/automatic/parse.php>

on the interpretation of the sentence and the existing relations between their terms or parts. This process is made using Apache Jena⁶, an open source Semantic Web framework for Java, that provides a programming environment for RDF (Resource Description Framework), RDFS (a general-purpose language for representing simple RDF vocabularies on the Web) and OWL (Web Ontology Language). In particular, it provides a set of functionalities to extract data from and write to RDF graphs, intended as an abstract model, which can be queried through SPARQL (a semantic RDF query language).

Back to the sentence and the terms identified, the ArchOnto ontology is questioned about the terms "parent", and its synonyms "father", and "mother". For each term to be searched, a SPARQL query is constructed and applied to the ontology. For instance, related to the "parent" the query is:

```
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#>

SELECT ?subject
WHERE {
    ?subject rdfs:label ?label .
    FILTER REGEX(?label,"parent", "i") .
}
```

The output are the CIDOC-CMR properties P152 is parent of, with domain E21 Person and range E21 Person, and its inverse P152 has parent.

Based on morphological and syntactic analysis, as well as the property found, it is inferred that the names "Manuel de Oliveira" and "Rufina Maria" are related to the person who was baptized, and the relation "is parent of" makes them parents, more precisely father and mother, see below for more details, of "Ana", the person whose the baptism record refers to.

The same search is made for the terms "father" and "mother", and the resulting properties are, respectively, P97 from father (its inverse P97i was father for) and P96 by mother.

Both properties have as domain the class E67 Birth and as range E21 Person. As such, the existence of a relationship between the three persons is based on the existence of a birth event, the birth of "Ana". Therefore "Manuel de Oliveira" (male and singular) and "Rufina Maria" (female and singular) are respectively "Ana"'s father and mother.

Regarding the birth event and the way to relate it to the birth of a particular person, the CIDOC-CRM provides the object property P98 brought into life, with domain E67 Birth and range E21 Person.

Finally, all the identified names are related to the concept of being a person and querying the ontology

for classes or properties that are related to the person's concept, the only result obtained was the class E21 Person.

2.4 Individuals' Construction

The Construction of Individuals is the step of creating the instances to populate the ontology and define the relation between them.

After the identification of all the concepts and properties related to the text interpretation, it is possible to instantiate the individuals of each class and assign the properties between them.

Regarding to the example, at least the following statements are generated:

```
E21 Person("Ana")
E21 Person("Manuel de Oliveira")
E21 Person("Rufina Maria")
P152 is parent of(E21 Person("Manuel de Oliveira"),
E21 Person("Ana"))
P152 is parent of(E21 Person("Rufina Maria"), E21 Person("Ana"))
P98 brought into life(E67 Birth("Ana's Birth"), E21 Person("Ana"))
P97 from father(E67 Birth("Ana's Birth"),
E21 Person("Manuel de Oliveira"))
P97 by mother(E67 Birth("Ana's Birth"),E21 Person("Rufina Maria"))
```

3 ARCHIVAL INFORMATION DATA MINING THROUGH ArchOnto

Even though DigitArq has a great diversity of records, as mentioned before, it is possible to identify the homogeneity information among them, such as the reference code, title, dates, dimension and support, language, and physical location.

Taking into account the ISAD(G) concepts, the reference code and the physical location can be considered identifiers of the document, which are represented in the ArchOnto by the class named E42 Identifier. The other concepts have the following representation: dimension by E54 Dimension, support by E57 Material, titles by E35 Title, dates by E52 Time-Span, and language by E56 Language.

The description of events characterise a huge volume of documents placed in the archive. These events are mostly related to persons, places, and also time periods, which are subjects that are central in the archival description. ArchOnto, as an extension of CIDOC-CRM ontology, has a quite rich and complete representation regarding the concepts of event, people, place, and time, which allows to match and to automate the process of generating the individuals and populate the ontology.

⁶<https://jena.apache.org/index.html>

3.1 Person's Information

Consider again the "Ana"’s baptism record. As mentioned before, one of the information that is possible to extract from the baptism record, more precisely from the field named "Scope and content", is both names of Ana’s parents, which appears in the sentence "Pais: Manuel de Oliveira e de Rufina Maria", as seen before.

Consider the name of the father "Manuel de Oliveira", which is a full name, and refers to a person. ArchOnto makes available the class E21 Person to represent the concept person. The class E21 Person is derived from the root E1 CRM Entity, which means that every person is an entity of the ontology. Therefore, it is possible to infer that every person has an appellation, using the property P1 is identified by, with domain E1 CRM Entity and range E41 Appellation, which is the representation of a person’s name.

Regarding to the person’s name, it is also possible to distinguish different identifications such as first name, surname, full name, nickname, etc.. These kind of identification are considered types of names, which are represented in the ontology by the class ARE7 Name Type (a subclass of the class E55 Type), and the property P2 has type, with domain E1 CRM Entity (super class of E41 Appellation) and range E55 Type, enables to represent the relation between the appellation of a person and all its related types of names.

Thereby, the person "Manuel de Oliveira", in addition to its full name, has also a first name "Manuel" and a surname "de Oliveira". Figure 2 shows the graph representation of the person "Manuel de Oliveira", using Protégé.

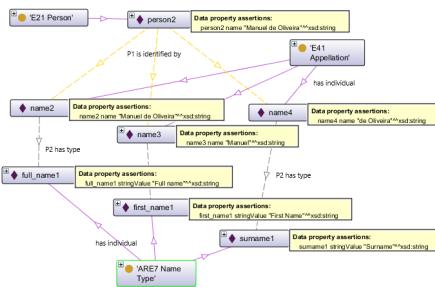


Figure 2: Person’s Instance "Manuel de Oliveira".

3.2 Temporal Information

The baptism record also provides information about the event’s date, as well as the birth date of the person who was baptized. Consider the birth date of

"Ana" set by the sentence "Data de nascimento⁷: 10 de Fevereiro de 1812⁸", information extracted from the same field as the parents’ names.

The ArchOnto model defines the temporal information, which refers to some time that is certain, as a time-span concept, with the class E52 Time-Span. As well as the class E21 Person also the class E52 Time-Span is derived from the super class E1 CRM Entity, which means that every time-span is an entity of the ontology. Therefore, the same way as a person, every time-span has an appellation, which can be set by using the property P1 is identified by mentioned before, where the appellation is the identification of the birth date.

The temporal information concept comprises a time interval. In this particular example, it comprises just one day, with a date-time representation. A time or temporal interval is defined by the class Interval (subclass of the class DataObject, which represents generically literal information, such as integer, string, date, etc.). The value of a temporal information is stated by using the object property hasValue, with domain E1 CRM Entity and range DataObject. Each temporal interval value has a type, which is certain, and is stated using the property P2 has type. Also each interval of time has a start and an end date values, which are represented with the data properties, respectively, startDateValue and endDateValue, both with domain Interval and range xsd:dateTime. The Figure 3 shows the definition of the ontology individual with birth date "10 de fevereiro de 1812".

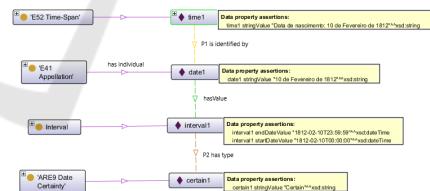


Figure 3: Time-Span’s Instance "Data de nascimento: 10 de fevereiro de 1812".

3.3 Well Known Events

The ArchOnto model describes and represents explicitly some known events, like birth or death, where the class E67 Birth (subclass of the class E5 Event, which is a E2 Temporal Entity) represents the concept birth. This fact facilitates the representation of

⁷The translation of "Data de nascimento" is "Birth date".

⁸"10 de Fevereiro de 1812" refers to the date "10 February 1812".

the known events with no further interpretation.

To establish that some birth event is related to the birth of a specific person, the ArchOnto provides the object property P98 brought into life, with domain E67 Birth and range E21 Person.

To define temporal relations between individuals, the ArchOnto ontology makes available the P4 has time-span property, with domain E2 Temporal Entity and range E52 Time-Span. Since the class E2 Temporal Entity is a super class of the class E5 Event, that in turn is a super class of E67 Birth, the P4 has time-span property allows to assign a date to a person's birth.

Back to the sentence "Pais: Manuel de Oliveira e de Rufina Maria" and as seen before in Subsection 2.3, to establish that the person "Manuel de Oliveira" is the father of "Ana", the ArchOnto provides the object property P97 from father, with domain E67 Birth and range E21 Person, and to establish that "Maria Rufina" is the mother of "Ana", the ArchOnto makes available the object property P96 by mother, also with domain E67 Birth and range E21 Person. As we have seen, the relations of being a father or a mother of someone else is established through the birth event, in this case through the birth of "Ana".

The Figure 4 shows the birth event of the person "Ana", the relations established with her parents and the date of her birth.

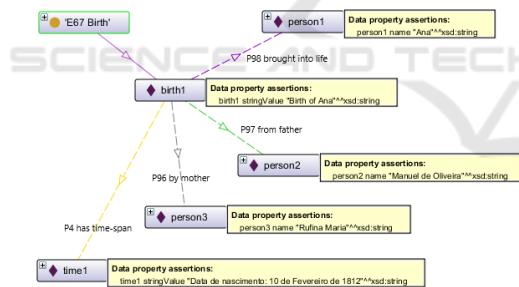


Figure 4: Birth of "Ana".

This kind of document provides information about someone's person birth from its baptism record, but it also provides information about the grandparents of the baptized person, which allows establishing more two birth events concerning to both parents of the person who was baptized. In particular, the "Ana"'s baptism record provides the sentences "Avós maternos": Manuel da Fonseca e Rosa da Silva" and "Avós paternos⁹: José de Oliveira e Jacinta de Oliveira", the first one identifies the grandparents from mother's side and

⁹The translation of "Avós maternos" is "Maternal grandparents".

¹⁰The translation of "Avós paternos" is "Paternal grandparents".

the second one identifies the grandparents from the father's side. However, with the information made available, it is not possible to know when both births occurred.

3.4 Sponsor's Relations

Information like grandparents and godparents is possible to extract from baptism records, but this kind of information is not directly represented by an ontology concept. However, the use of a ternary relation makes possible to define the nature of a person's participant on an event. The ternary relation is identified in the ArchOnto by the object property P14 carried out by, with domain E7 Activity and range E39 Actor, applied to the P14.1 in the role of property.

Consider again "Ana"'s baptism record and its field "Scope and content". The sentence "Padrinhos¹¹: Manuel Martins Ramos e Maria Francisca" provides the name of both godparents. For instance, consider the godfather, a person with the name "Manuel Martins Ramos". The relation established between the person "Ana" and the baptism event is that "Ana" is the person who was baptized. In the other case, the relation established between the person "Manuel Martins Ramos" and the baptism event is that this person is set has godfather of who was baptized. Since the baptism event is the same, it is possible to conclude and represent that the person "Manuel Martins Ramos" is the godfather of "Ana". Figure 5 shows the use of the ternary relation to define the role of a person as a godfather and the action carried out to turn that person into the godfather of someone else, i.e., the definition of "Manuel Martins Ramos" as the godfather of "Ana".

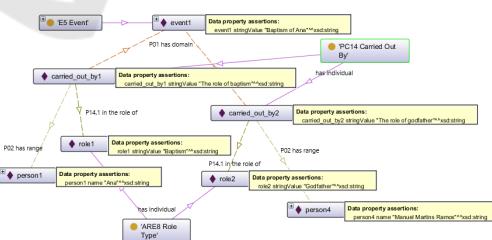


Figure 5: The role godfather.

In the same way, the person "Maria Francisca" is defined as the godmother of "Ana".

3.5 Events

Unlike birth, a baptism event has no concept representation available in ArchOnto. Therefore, the model

¹¹The translation of "Padrinhos" is "Godparents".

provides the class E5_Event to define all the information that are identified as events.

As already mentioned, a baptism record represents itself a description of an event, with a set of homogeneous information, such as the happening date (time-span instance), the identification of the persons that were present in the ceremony and are related to the person who was baptized. These persons are identified by their names and their relations to the baptized person. Each name identification generates a new person instance, in case it does not yet exist in the ontology population.

The relation between each instance and the baptism event is defined by using the proper object property. As showed before, in subsection 3.3, the object property P4 has time-span allows to assign a date to an event. The baptism event is the bond to define the godparents relations, as shown in the previous subsection 3.4. Finally, to state the relation that some event occurred in the presence of some person, the ArchOnto provides the object property P12 occurred in the presence of, with domain E5 Event and range E77 Persistent Item, which is a super class of the class E21 Person. The Figure 6 shows "Ana's baptism event.

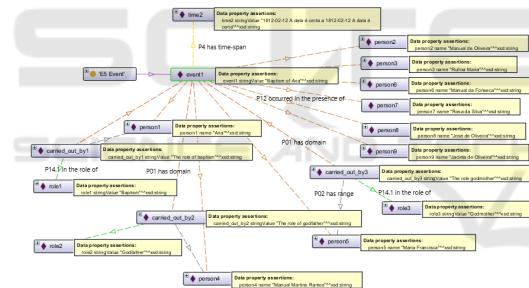


Figure 6: The baptism event.

3.6 Document Information

As mentioned before, the DigitArq has a great diversity of records, describing a huge amount of different subjects in the archival field. However, the information is structured in an homogeneously way and, regardless of the subject, the records contain information about reference code, title, dates, dimension and support, language, and physical location.

Identifiable immaterial items that make assertions about reality are represented by the class E31 Document, subclass of E71 Human-Made Thing. Therefore, a baptism record is classified as an instance of E31 Document.

The title of a document is intended to be a formal title and is represented in the ontology by the class ARE2 Formal Title, subclass of E35 Title.

The property P102 has title, with domain E71 Human-Made Thing and range E35 Title, assigns a formal title to a document.

Information like reference code is represented by the class E42 Identifier, subclass of E41 Appellation, and the property P1 is identified by, with domain E1 CRM Entity and range E41 Appellation, makes possible to assign a reference code to its document.

Information that is possible to extract beyond the structured is interpreted as information that is referred (directly or indirectly) by the record information. The CIDOC-CRM makes available the property P67 refers to, with domain E89 Propositional Object (superclass of E31 Document) and range E1 CRM Entity, which allows to assign that information was stated by a particular document.

Therefore, the "Ana"'s baptism record is an instance of the class E31 Document, has a formal title, a reference code, documents the baptism event, and other structured information, but also refers to the persons identified before, like "Ana", her parents, grandparents and godparents, as well as the births, events dates, and so on. Figure 7 shows the representation of "Ana"'s baptism record as an individual of the ontology and some of its relations with other individuals.

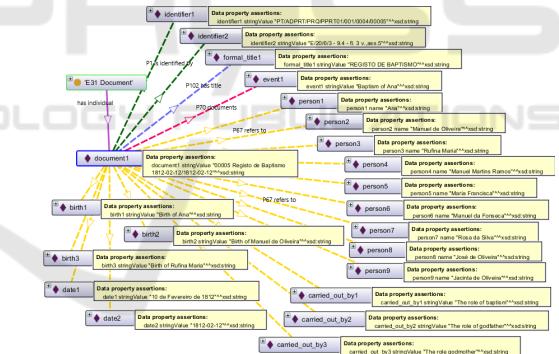


Figure 7: Ana's baptism record.

4 CONCLUSIONS AND FUTURE WORK

We have presented an automatic semantic migration prototype based on Knowledge Discovery from Digital Archive Data for ontology population in the domain of Archives metadata, ISAD(G). Our prototype uses Natural Language Processing (NLP) techniques for language processing and Semantic Web techniques for querying and updating the Ontology ArchOnto. We describe the migration process detailing the text fields morphological and syntactic analy-

sis as inputs for the process of Ontology Knowledge Discovery, showing how we query the ontology to finding the concepts and properties represented in ArchOnto that are related to the terms found at morphological and syntactic analysis level. The representation of different CIDOC-CMR concepts such as persons, events, times, locations and documents is illustrated highlighting the migration process of baptism document.

Future work includes the evaluation of the ArchOnto population quality. Traditionally this evaluation is done recurring to *precision* and *recall* measures as in Information Retrieval. However, for ontology's population the binary classification, yes/no, does not take into account cases where the information is partially captured, methods like Balanced Distance Metric(Maynard et al., 2008) are more adequate to our evaluation task. The EPISA project has the human resources that will enable us to embrace this task. The extension of the prototype to other documents types will bring new challenges that still need good solutions to be solved, for instance when a new person shares some proprieties with a known person, shall we consider that it is the same person? when a person has the same names for its parents and grand parents of a person known in ArchOnto, are they brothers? should that relation be considered as a extension of ArchOnto since it does not exists in CIDOC-CRM?

ACKNOWLEDGEMENTS

This work is financed by National Funds through the Portuguese funding agency, FCT (Fundação para a Ciência e a Tecnologia) within project DSAIPA/DS/0023/2018.

REFERENCES

- Bick, E. (2014). Palavras, a constraint grammarbased parsing system for portuguese. *Working with portuguese corpora*, pages 279–302.
- Bountouri, L. and Gergatsoulis, M. (2011). The semantic mapping of archival metadata to the cidoc crm ontology. *Journal of Archival Organization*, 9(3-4):174–207.
- Cavalcanti, M. C., Pereira, F. D., Fusco, E., and Mucheroni, M. L. (2017). Model of data extraction in the innovation environments of the state of são paulo based on semantic technologies. In *International Conference on Information Systems & Technology Management*.
- de Almeida, M. J. and Runa, L. (2018). ICON Project: Content integration in Portuguese National Archives using CIDOC-CRM. In *2018 CIDOC Annual Conference*.
- 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.
- Fragkou, P., Kritikos, N., and Galiotou, E. (2016). Querying greek governmental site using sparql. In *Proceedings of the 20th Pan-Hellenic Conference on Informatics*, PCI '16, New York, NY, USA. Association for Computing Machinery.
- ICOM/CIDOC-CRM Special Interest Group (2019). *Definition of the CIDOC Conceptual Reference Model*. ICOM, 6.2.7 edition.
- International Council on Archives (2000). *ISAD(G) Second Edition*. International Council on Archives.
- 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.
- Koch, I., Ribeiro, C., and Lopes, C. T. (2020). Archonto, a cidoc-crm-based linked data model for the portuguese archives. In *24nd International Conference on Theory and Practice of Digital Libraries*, TPDL.
- Makki, J. (2017). Ontoprima: A prototype for automating ontology population. *International Journal of Web/Semantic Technology* (IJWestT), 8.
- Makki, J., Alquier, A.-M., and Prince, V. (2008). An nlp-based ontology population for a risk management generic structure. In *Proceedings of the 5th International Conference on Soft Computing as Transdisciplinary Science and Technology*, CSTST '08, page 350–355, New York, NY, USA. Association for Computing Machinery.
- Maynard, D., Li, Y., and Peters, W. (2008). Nlp techniques for term extraction and ontology population.
- 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.
- Oldman, D. (2014). *The CIDOC Conceptual Reference Model (CIDOC-CRM): PRIMER*. CRM Labs.
- Ramalho, J. C. and Ferreira, J. C. (2004). 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, Brasilia, Brazil, June 23-26, 2004. Proceedings*.
- Scifleet, P. (2001). International standard archival description isad (g).
- 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.