Semantic Web 1 (0) 1–5
IOS Press



# A Strategy for Archives Metadata Representation on CIDOC-CRM and Knowledge Discovery

Dora Melo [a,c,*,**], Irene Pimenta Rodrigues [b,c] and Davide Varagnolo [b]

[a] *Coimbra Business School|ISCAC, Polytechnic Institute of Coimbra, Portugal*
*E-mail: dmelo@iscac.pt*
[b] *Department of Informatics, University of Évora, PT Portugal*
*E-mails: ipr@uevora.pt, d.varagnolo@studenti.unipi.it*
[c] *NOVA Laboratory for Computer Science and Informatics, NOVA LINCS , PT Portugal*

**Editors:** First Editor, University or Company name, Country; Second Editor, University or Company name, Country
**Solicited reviews:** First Solicited Reviewer, University or Company name, Country; Second Solicited Reviewer, University or Company name, Country
**Open reviews:** First Open Reviewer, University or Company name, Country; Second Open Reviewer, University or Company name, Country

**Abstract.**
This paper presents a strategy for the semantic migration of Portuguese National Archives records into CIDOC-CRM standard, an ontology developed for museums, within the context of EPISA project. The approach to automatically populate the CIDOC-CRM is based on Mapping Description Rules to semantically translate the archives descriptive information into CIDOC-CRM representation. The compliance of the CIDOC-CRM model recommendations guarantees that the populated CIDOC-CRM ontology of archives descriptive information verifies interoperability, and could be linked and integrated with other populated CIDOC-CRM ontologies. In the information modelling, requirements on the mapping representation, due to the intent of interpreting natural language text to automatically extract information of metadata text fields and to interpret natural language queries, are taken into account. To automatically interpret the Mapping Description Rules, `OWL API` was used to obtain the set of assertions that represents the information in the target ontology and two datasets are available with some migration examples. The exploration of the knowledge representation is done through some Description Logic queries to highlight the advantages of having this new representation of the National Archives. The evaluation of the resulting representation can be done automatically proving its correctness for the metadata that has a direct representation in CIDOC-CRM.

Keywords: Knowledge Representation and Reasoning, Natural Language Processing, Archives Ontology, Semantic Migration, CIDOC-CRM, Linked Data Science

## 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,

*Corresponding author. E-mail: dmelo@iscac.pt.
**This work is financed by National Funds through the Portuguese funding agency, FCT (Fundação para a Ciência e a Tecnologia) within I&D Projects with identification EPISA (DSAIPA/DS/0023/2018) and NOVA LINCS (UIDB/04516/2020).

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 [1], the archive national system that uses well-established description standards, namely the ISAD(G) (General International Standard Archival Description) [2] and ISAAR(CPF) (International Standard Archival Authority Record for Corporate Bodies, Persons and Families) [3] with a hierarchical structure adapted to the nature of archival assets.

The data model and description vocabularies adopted are built upon the CIDOC-CRM (Conceptual Reference Model) standard [4], an ontology developed for museums by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM) [5, 6].

The aim of this paper is to introduce an approach to automatically populate the CIDOC-CRM with the Portuguese National Archives metadata that obey ISAD(G) and ISAAR recommendations. The methodology is based on Mapping Description Rules to semantically translate the archives descriptive information into CIDOC-CRM ontology representation. The Mapping Description Rules, as defined, can be easily adapted to the use of other ontologies. The compliance of the CIDOC-CRM model recommendations guarantees that the populated CIDOC-CRM ontology of archives descriptive information verifies interoperability, and could be linked and integrated with other populated ontologies using CIDOC-CRM representation.

The remainder of this paper is divided into the following sections. Section 2 presents the norms and formats to universally describe archives metadata, proposals for mapping ISAD(G) into ontologies such as CIDOC-CRM, natural language interpretation of queries and of raw text to automatic populate an ontology and a brief resume of current related work with CIDOC-CRM representation and interfaces to query OWL2 knowledge base.

The representation of ISAD(G) and ISAAR(CPF) Archives Metadata in CIDOC-CRM is presented in Section 3. This section introduces the methodology based on Mapping Description Rules for automatizing the migration process, presents the CIDOC-CRM recommendations for modelling information, to guarantee the effectiveness and the consistency of the final populated ontology, as well as some requirements on the mapping representation due to the intend of interpreting natural language text to automatically extract information of metadata text fields and to interpret natural language queries.

Section 4 presents the architecture of the migration process from DigitArq HTML records into CIDOC-CRM and describes in detail each one of its steps. Some illustrative examples are presented for clarification and better understanding.

An evaluation on the results of the migration process is presented in Section 5. A set of questions performed over the knowledge base are presented in order to approve that the CIDOC-CRM Ontology representation of the DigitArq metadata is well-performed. To help and facilitate the task of querying the knowledge base, an application program interface was also developed and it is also presented in this section.

In Section 6, it is presented a set of open problems that arose from occurred issues while developing and implementing the Mapping Description Rules, together with the analysis of different examples.

Finally, in Section 7, it is drawn the conclusions, as well as further work and a future evaluation.

## 2. The Archival Description Scenario

The International Council of Archives[1](ICA) defines archives as "the documentary by-product of human activity retained for their long-term value." They are characterized as contemporary records created by individuals and organisations about their business, providing information on past events. These records can be of a wide range of formats including written, photographic, moving image, sound, digital and analogue.

The aim of the ICA is to promote the management and use of records and archives, and the preservation of the archival heritage of humanity around the world. The sharing of experiences, research and ideas on professional archival, records management, as well as on the management and organisation of archival institutions, are part of the strategy for their success.

In this follow-up, the ICA Committee Description Standards developed the General International Standard Archival Description (ISAD(G))[2], which provides general guidance for creating descriptions of archival materials, establishing a model based on the principle of *respect des fonds* within a multi-level description. ISAD(G) defines 26 elements that may be

[1] https://www.ica.org/en

combined in seven areas to constitute the description of an archival entity. These areas are identified by Identity Statement, Context, Content and Structure, Condition of Access and Use, Allied Materials, Note, and Description Control, and provides general content guidelines. The structure and content of the information in each of those elements should be formulated in accordance with applicable national rules. As general rules, these are intended to be broadly applicable to descriptions of archives regardless of the nature or extent of the unit of description (subsequently identified also as just 'unit').

The International Standard Archival Authority Record for Corporate Bodies, Persons and Families (ISAAR (CPF))[3], also developed by the same ICA Committee Descriptions Standards, provides guidance for preparing archival authority records which introduce descriptions of entities, such as corporate bodies, persons and families, associated with the creation and maintenance of archives.

The ISAD(G) content model, along with the ISAAR (CPF), serves as the basis for the development of the guidance document for the standardization of Portuguese archival descriptions [7]. This document was developed by the General Directorate for Book, Archives, and Libraries (DGLAB)[2], through an archival description standardization working group. Besides the introduction of the ISAD(G) elements and ISAAR (CPF) descriptions, the Portuguese guidance has two main purposes: first, the inclusion of the detailed perspective of the lower levels of description, such as installation unit, compound document and simple document (named as Item); and second, the addition of a unifying view of the description that included coherent description of documents in electronic form.

The need for a mean to facilitate the archival work with the generation of indexes, lists, inventories catalogues and transference guides about Historical documents, as well as a coherent finding aids on helping users and archivists attain the artefacts they seek, resulted in the development of the DigitArq [1] platform. DigitArq is characterized by a common digital format based on an international standard and an archival management software to maintain all information, supported by a centralised repository to store all the collected material.

The development of the semantic migration process of the DigitArq metadata uses the CIDOC-CRM ontology as a data model and description vocabularies. One of the main reasons for this choice is that semantic integration and interoperability can be achieved trough the use of CIDOC-CRM, since there are many platforms available to access the information in CIDOC-CRM representation for several domains. The semantic mapping of archival metadata into the CIDOC-CRM Ontology can be straightforward for some elements [8–10].

The first approach to present a set of mapping rules was a study to explore the representation expressiveness of CIDOC-CRM into archival metadata domain [11]. This approach presents a set of rules which allows to map Encoded Archival Description (EAD) into CIDOC-CRM representation. EAD is a XML language designed to represent the ISAD(G) elements in XML syntax and is maintained by the standards initiative of the Library of Congress, and a rigorous mappings between EAD and ISAD(G) and vice-versa are maintained [12]. More recent, this first study was extended with a set of mapping rules and a language to write them [8]. Using this mapping rules, a conceptual ontology for Archival Knowledge Model was proposed in [9], with the purpose of querying archival or historical knowledge bases, and where natural language queries are translated to the CIDOC-CRM and appropriate extensions.

The semantic integration of CIDOC-CRM with other standards has been a recurring goal [13]. An example of an effort in this regard is the proposal for semantic integration of collection description illustrated with Dublin Core and CIDOC-CRM [10].

The importance of the migration process lies not only in the direct translation of the ISAD(G) elements, but also in the possibility of adding information to the knowledge base that can be extracted and inferred from the textual elements. In fact, there are elements of ISAD(G) descriptions whose content is free text about the record itself and for which there are no general mapping rules available. This content must be interpreted in the CIDOC-CRM ontology context in order to represent the entities, events, locals, dates, relations and properties in the ontology. This process is achieved by applying Natural Languages Processing (NLP) techniques. OntoPrima [14] is a NLP-based Ontology Population that extracts instances of concepts and instances of relations from text in order to pop-

[2]The DGLAB (http://dglab.gov.pt/) is a public body under the Portuguese Ministry of Culture's responsibility, is a central service of the direct administration of the State, endowed with administrative autonomy, whose mission is to ensure the coordination of the national archives system and the implementation of an integrated policy for non-school books, libraries and reading.

ulate a given ontology based on NLP techniques for language processing, semantic web techniques (RDFS, RDF, Jena APIs) for knowledge modeling and representation, and on domain expert's intervention for validating extracted instances. This topic is explored in other works such as [15–17].

In the past few years, some interfaces were developed for CIDOC-CRM knowledge bases, mainly in the cultural heritage domain, such as OpenArcheo [18], that allows the users to create complex query with a user's friendly GUI and facilitates the task of searching for information that users seek to find, or even Arches heritage inventory and management system [19] and ONTOME a collaborative ontology management environment [20, 21]. An example of a differentiation tool is the interface for manipulating narratives, Narrative Building and Visualisation Tool [22], that allows the users to add new narratives and visualizes information about them. All these platforms are a mean to integrate different domain knowledge bases for interoperability.

## 3. Representing ISAD(G) and ISAAR(CPF) Archives Metadata in CIDOC-CRM

As mentioned before, ISAD(G) content model is based on the principle of *respect des fonds* within a multi-level description. This principle has as practical consequence that archival description proceeds from the general to the specific, in order to represent the context and hierarchical structure of the fonds and its component parts. Generically, this means that each level of description can be subdivided into the sub-levels considered necessary to mirror the different documentary realities. In addition, the multi-level description model also complies with the following rules: Set information relevant to the level of description, with the aim of accurately representing the context and content of the unit of description; the existence of a link between descriptions, in order to make explicit the position of the unit of description in the hierarchy; and no repetition of information, in order to avoid redundancy of information in hierarchically related archival descriptions. Figure 1 presents the model of the levels of arrangement of a fonds.

Concerning to each unit at some level of description, all 26 information elements provided for in ISAD(G) can be considered, in their entirety, at any level of description, according to the desired degree of completeness. However, just the following elements are considered essential for international exchange of descriptive information [2, 7]: reference code; title; creator; date(s); extent of the unit of description; and level of description. The reference code is the information that allows to identify uniquely the unit of description and to provide a link to the description that represents it. The creator of the unit identifies the corporate body, family or person that created, accumulated and/or maintained records in the conduct of personal or corporate activity. The date(s) identifies and records the date(s) of the unit of description, such as date range or creation date. The extent of the unit of description is the information that allows to identify and describe the physical or logical extent and the medium of the unit of description. Finally, the level of description is the position of the unit of description in the hierarchy of the fonds (Figure 1).

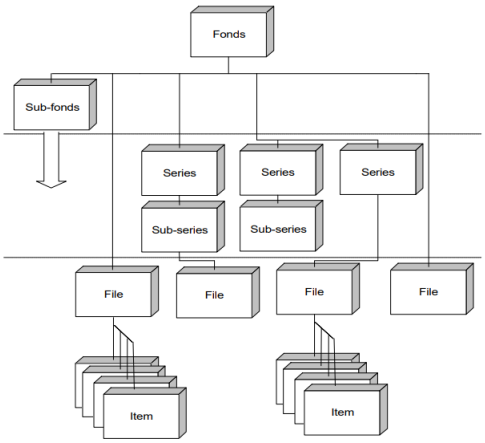



Fig. 1. Model of the levels of arrangement of a fonds [2, p. 36]

Taking this knowledge into account, it was necessary to establish the principles of information representation to ensure that the migration process of the archives' metadata into the CIDOC-CRM ontology is successfully completed. Therefore, the representation of the archives' metadata in CIDOC-CRM uses the criteria explained in the following subsections.

### *3.1. CIDOC-CRM recomendations*

The translation of the Archival metadata into the CIDOC-CRM representation follows the main principles of the CIDOC-CRM model [4, 23].

1. The introduction of a new class should comply with the minimality modelling principle of CIDOC-CRM:
   "A class is not declared unless it is required as

the domain or range of a property not appropriate to its superclass, or it is a key concept in the practical scope"
Regarding properties, a new one only should be added if "it is a key concept in the practical scope".

2. The representation of terms that declare that an object belongs to a particular category of items follows the CIDOC-CRM specific modeling constructs 'about types'.
The class '$E_{55}$ Type' comprises such terms from thesauri and controlled vocabularies used to characterize and classify instances of CIDOC-CRM classes. Therefore, instances of '$E_{55}$ Type' represent concepts (universals) in contrast to instances of '$E_{41}$ Appellation', which are used to name instances of CIDOC-CRM classes.
In addition, the property '$P_2$ has type (is type of)' provides the mechanism to specialize the classification of CIDOC-CRM instances to any level of detail, by linking to external vocabulary sources, thesauri, classification schemas or ontologies.
3. The cases in which categorization is established in the relationship (property) between two individuals, i.e., stating the role of a relation between individuals, their representation also follows the CIDOC-CRM specific modeling constructs 'about types'.
With an analogous purpose of the '$P_2$ has type (is type of)' property, some properties of the CIDOC-CRM are associated with an additional property and are numbered with a '.1' extension. The range of these properties of properties always falls under '$E_{55}$ Type'. The purpose of a property of a property is to provide an alternative mechanism to specialize its domain property through the use of property subtypes declared as instances of '$E_{55}$ Type'.

### 3.2. *Information Useful for Natural Language Interpretation of Text or Queries.*

- In every appellation (or identifier) introduce a type (instance of class '$E_{55}$ Type') to explicitly classify the relation between the name and the entity. As an example, consider the reference code of a unit. The reference code is the type of the reference code value, identifier of the unit.
- In the case of one concept represented as a subclass of other concept, the representation is established by defining the first one as having as type the second one. As an example, consider the unit with a description level fond. The description level is the type of fond.

Consider the following natural language sentence "The description level of a unit is Fonds."

In its natural language interpretation, if 'Description level' was considered a class of CIDOC-CRM the matching would be straightforward, and merely classify 'Fonds' as an instance of such class, see Figure 2. However and according to the previous described CIDOC-CRM recommendations rules, the 'Description level' should not be a class. It can be omitted[3] or it can be a new instance of type. Therefore, instead of having 'Fonds' has an instance of the class 'Description Level', the choice is to have ´Fonds' as a type defined with the type 'Description level'. This mechanism is suitable for natural language interpretation, since it is possible to match the term 'Description level' with the instance of '$E_{55}$ Type', the term 'Fonds' with another instance of '$E_{55}$ Type', the term 'unit' to '$E_{31}$ Document', the preposition to the property '$P_2$ has type' and the verb 'is' to the same property '$P_2$ has type'. The matching between the sentences terms (nouns, adjectives, prepositions, verbs, named entities) and class, properties and instances of an ontology is a common step in natural language interpretation for querying an ontology or mining text to populate an ontology [24, 25].

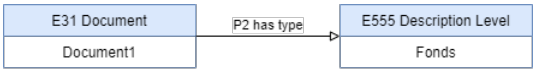



Fig. 2. 'Fonds' representation as an individual of the new class entity.

A little more elaborated example is the following sentence, where the tags and the syntactic tree are obtained with the Stanford parser.
Sentence:
"The description level of the unit with reference code xxx is Fonds"

Tagging:
The/DT description/NN level/NN of/IN the/DT unit/NN with/IN reference/NN code/NN xxx/NN

[3]One unit belongs only to one description level and their values are from a controlled vocabulary, Fond, Sub-fond, Series, Subseries, ... (see Figure 1)

After searching for equivalent terms in the CIDOC-CRM representation, it is possible to obtain new assertions as the ones following:
Sentence$_a$:
"The Description_level of the unit with Reference_code xxx is Fonds"

Tagging$_a$:
`The/DT Description_level/NNP of/IN the/DT unit/NN with/IN Reference_code/NNP xxx/NNP is/VBZ Fonds/NNP`

Parser tree$_a$:
```
S(NP(NP (DT The) (NNP Description_level)) (PP (IN of) (NP (NP (DT the) (NN unit)) (PP (IN with) (NP (NNP Reference_code) (NNP xxx))))))
(VP (VBZ is) (NP (NNP Fonds)))
```

By quering the CIDOC-CRM with the CIDOC assertions and the migration rules (introduced later on), the matching between unit and $E_{31}$*Document* is done by the interpretation of 'PP (IN with)' as '$P_1$ is identified by xxx' and 'PP (IN of)' as '$P_2$ has type Fond'.

### *3.3. Mapping Description Rules*

As mentioned before, each unit, at some level of description, has a well-known structure of information defined by the ISAD(G) elements. In order to define the representation of each unit, the elements can be grouped, according to their content and what they refer to, and associated with three concepts, namely the object itself that belongs to the physical archive; the digital registration that describes the object; and the language properties associated to the object (when they exist). These three concepts are mapped into the following CIDOC-CRM classes, respectively '$E_{22}$ Human-Made Object'; '$E_{31}$ Document'; and '$E_{33}$ Linguistic Object'. Whereby, each unit is itself mapped into an instance of '$E_{31}$ Document', that documents (property '$P_{70}$ documents') a physical object (an instance of '$E_{22}$ Human-Made Object) and refers to (property '$P_{67}$ refers to') a conceptual object (an instance of '$E_{33}$ Linguistic Object').

The hierarchical structure of the archives is represented using the relation '$P_{106}$ is composed of' between the '$E_{31}$ Document' and their sub-documents (also represented as individuals of the class '$E_{31}$ Document').

These representations follows the CIDOC-CRM recommendations, and similar approaches for representing archives and collections are presented in [8, 10, 11]. The representation of the archival description units in the CIDOC-CRM Ontology is done through rules that express the metadata mapping into the ontology entities. These rules define the Mapping Description Rules that establish the basis for the automatic migration process. Table 1 presents some of the Mapping Description Rules.

Therefore, the representation of the unit explained before is translated into the rule No. 1 and the hierarchy of the archive is captured in rule No. 17.

The Mapping Description Rules use the notation defined in [8] and widely used in the context of CIDOC-CRM mappings. They have been extended with the use of:

- $\langle$`expression`$\rangle$, where the 'expression' is added to the knowledge base only if the 'expression query' does not succeed. For instance, the expression query $ID_E \rightarrow \langle P_P \rangle \rightarrow \langle E_E \rangle$ will add the property $P_P$ between the instance $ID_E$ and a new instance of $E_E$ if there is no property $P_P$ between instance $ID_E$ and any instance of $E_E$.
- $\|$`expression`$\|$, where 'expression' is ignored if the 'expression query' is not part of the knowledge base. For instance, the expression query $\|ID_{E_1}\| \rightarrow \|P_P\| \rightarrow \|ID_{E_2}\|$ will only be considered if the property $P_P$ already exists between the existing instances $ID_{E_1}$ and $ID_{E_1}$, otherwise is ignored and nothing is added to the knowledge base.

This expressiveness is important for the automatic interpretation of the mapping description rules that will allow to obtain the representation of the Portuguese National Archives in CIDOC-CRM representation. Each rule of the Table 1 gives rise to a sequence of assertions. These assertions are OWL2 facts in CIDOC-CRM.

In order to establish the adequate sequence of assertions for each rule, the expressiveness of the rules takes into account the following commands:

**NewInst**($I_{Id}$, $C_{Id}$) - Creates a new instance of $C_{Id}$ with value $I_{Id}$. If the instance already exists, raises an exception.

**Inst**($I_{Id}$, $C_{Id}$, **Value**) - If there is an instance of $C_{Id}$, $I_{Id}$ = Value; else creates a new instance $I_{Id}$ of $C_{Id}$ with value Value.

**InstS**($I_{Id}$, $C_{Id}$, **Value**) - If there is an instance of the $C_{Id}$, $I_{Id}$ = Value; else raises an exception.

**NewProp**($I_{Id_1}$, $P_{Id}$, $I_{Id_2}$) - Creates a new instance of a object property $P_{Id}$ with domain $I_{Id_1}$ and range

Table 1
Mapping Description Language Rules

| Rule No. | Left part (rec[attribute value list]) | Right part CIDOC-CMR |
|---|---|---|
| 1 | DigitArq(Rec) | $E_{31}\{= ID_{E_{31}}\} \rightarrow P_{70} \rightarrow E_{22}\{= ID_{E_{22}}\} \rightarrow P_{67} \rightarrow E_{33}\{= ID_{E_{33}}\}$ |
| 2 | ['Description level', V] | $\$ID_{E_{31}} \rightarrow P_2 \rightarrow (\langle E_{55}\{= V\}\rangle \rightarrow \langle P_2\rangle \rightarrow \langle E_{55}\{^{= \text{'Description level'}}\}\rangle)$ |
| 3 | ['Reference code', V] | $\$ID_{E_{31}} \rightarrow P_1 \rightarrow (E_{42}\{= V\} \rightarrow P_2 \rightarrow \langle E_{55}\{^{= \text{'Reference code'}}\}\rangle)$ |
| 4 | ['Language of the material', V] | $\$ID_{E_{33}} \rightarrow P_{72} \rightarrow \langle E_{56}\{= V\}\rangle$ |
| 5 | ['Original numbering', V] | $\$ID_{E_{22}} \rightarrow \langle P_{55}\rangle \rightarrow (\langle E_{53}\rangle \rightarrow P_{89} \rightarrow (E_{53} \rightarrow P_1 \rightarrow \langle E_{41}\{= V\}\rangle \rightarrow P_2 \rightarrow \langle E_{55}\{^{= \text{'Original numbering'}}\}\rangle))$ |
| 6 | ['Scope and content',V] | $\$ID_{E_{31}} \rightarrow P_{01i} \rightarrow (PC_3 \rightarrow P_{02} \rightarrow \langle E_{62}\{= V\}\rangle \rightarrow P_{3.1} \rightarrow \langle E_{55}\{^{= \text{'Scope and content'}}\}\rangle)$ |
| 7 | ['Recipient', V]' | $\$ID_{E_{31}} \rightarrow P_{01i} \rightarrow (PC_{129} \rightarrow P_{02} \rightarrow (E_{21} \rightarrow P_1 \rightarrow \langle E_{41}\{= V\}\rangle) \rightarrow P_{129.1} \rightarrow \langle E_{55}\{= \text{'Recipient'}\}\rangle)$ |
| 8 | ['Title and Type',($T_i$, $T_y$)] | $\$ID_{E_{31}} \rightarrow P_{01i} \rightarrow (PC_{102} \rightarrow P_{02} \rightarrow \langle E_{35}\{= T_i\}\rangle) \rightarrow P_{102.1} \rightarrow \langle E_{55}\{= T_y\}\rangle)$ |
| 9 | ['Creation date', V] | $\$ID_{E_{31}} \rightarrow \langle P_{94}\rangle \rightarrow (E_{65} \rightarrow P_4 \rightarrow (E_{52} \rightarrow P_2 \rightarrow \langle E_{55}\{^{= \text{'Creation date'}}\}\rangle \rightarrow P_{170i} \rightarrow \langle E_{61}\{= V\}\rangle))$ |
| 10 | ['Modification date',V] | $\$ID_{E_{31}} \rightarrow \langle P_{94}\rangle \rightarrow (E_{65} \rightarrow P_4 \rightarrow (E_{52} \rightarrow P_2 \rightarrow \langle E_{55}\{^{= \text{'Modification date'}}\}\rangle) \rightarrow P_{170i} \rightarrow \langle E_{61}\{= V\}\rangle))$ |
| 11 | ['Date Range', (I,F,C)] | $\$ID_{E_{22}} \rightarrow \langle P_{108i}\rangle \rightarrow (E_{12} \rightarrow P_4 \rightarrow (E_{52} \rightarrow P_2 \rightarrow \langle E_{55}\{= C\}\rangle \rightarrow P_{79} \rightarrow \langle E_{61}\{= I\}\rangle \rightarrow P_{80} \rightarrow \langle E_{61}\{= F\}\rangle \rightarrow P_2 \rightarrow \langle E_{55}\{^{= \text{'Production Date'}}\}\rangle))$ |
| 12 | ['Dimension and support', ($V_s$, $D_u$, $V_d$)] | $\$ID_{E_{22}} \rightarrow P_{45} \rightarrow \langle E_{57}\{= V_s\}\rangle \rightarrow P_{43} \rightarrow (E_{54} \rightarrow P_{91} \rightarrow \langle E_{58}\{= D_u\}\rangle \rightarrow P_{90} \rightarrow \langle E_{60}\{= V_d\}\rangle)$ |
| 13 | ['Current keeper', ($V_{abv}$,$V_n$)] | $\$ID_{E_{22}} \rightarrow P_{50} \rightarrow (E_{74} \rightarrow P_{01i} \rightarrow (PC_1 \rightarrow P_{02} \rightarrow \langle E_{41}\{= V_{abv}\}\rangle \rightarrow P_{1.1} \rightarrow \langle E_{55}\{\{^{= \text{'Institution abbreviation'}}\}\}\rangle)P_{01i} \rightarrow (PC_1 \rightarrow P_{02} \rightarrow \langle E_{41}\{= V_n\}\rangle \rightarrow P_{1.1} \rightarrow \langle E_{55}\{^{= \text{'Institution name'}}\}\}\rangle))$ |
| 14 | ['Country', ($V_{abv}$, $V_n$)] | $\$ID_{E_{22}} \rightarrow P_{55} \rightarrow \langle P_{55}\rangle \rightarrow (\langle E_{53}\rangle \rightarrow P_{89} \rightarrow (E_{53} \rightarrow P_2 \rightarrow \langle E_{55}\{= \text{'Country'}\}\rangle \rightarrow P_{01i} \rightarrow (PC_1 \rightarrow P_{02} \rightarrow \langle E_{41}\{= V_{abv}\}\rangle \rightarrow P_{1.1} \rightarrow \langle E_{55}\{\{^{= \text{'Country abbreviation'}}\}\}\rangle)P_{01i} \rightarrow (PC_1 \rightarrow P_{02} \rightarrow \langle E_{41}\{= V_n\}\rangle \rightarrow P_{1.1} \rightarrow \langle E_{55}\{^{= \text{'Country name'}}\}\}\rangle)))$ |
| 15 | ['Producer', ($V_{abv}$,$V_n$)] | $\$ID_{E_{22}} \rightarrow \langle P_{108i}\rangle \rightarrow (\langle E_{12}\rangle \rightarrow P_{14} \rightarrow (E_{74} \rightarrow P_{01i} \rightarrow (PC_1 \rightarrow P_{02} \rightarrow \langle E_{41}\{= V_{abv}\}\rangle \rightarrow P_{1.1} \rightarrow \langle E_{55}\{\{^{= \text{'Institution abbreviation'}}\}\}\rangle)P_{01i} \rightarrow (PC_1 \rightarrow P_{02} \rightarrow \langle E_{41}\{= V_n\}\rangle \rightarrow P_{1.1} \rightarrow \langle E_{55}\{^{= \text{'Institution name'}}\}\}\rangle))$ |
| 16 | ['Producer Type', ($T_y$)] | $\$ID_{E_{22}} \rightarrow \langle P_{108i}\rangle \rightarrow (\langle E_{12}\rangle \rightarrow \langle P_{14}\rangle \rightarrow (\langle E_{74}\rangle \rightarrow P_2 \rightarrow \langle E_{55}\{= T_y\}\rangle$ |
| 17 | ['Hierarchy',(Root$_{ref}$1,Son$_{ref}$2)] | $\|\|E_{31}\|\| \rightarrow \|\|P_1\|\| \rightarrow \|\|E_{42}\{= \text{Root}_{ref}\}\|\| \rightarrow P_{106} \rightarrow (\|\|E_{31}\|\| \rightarrow \|\|P_1\|\| \rightarrow (\|\|E_{42}\{= \text{Son}_{ref}\}\|\|)$ |
| 18 | other rules | $\|\|E_{53}\|\| \rightarrow \|\|P_{89i}\|\| \rightarrow (\|\|E_{53}\|\| \rightarrow \|\|P_{55i}\|\| \rightarrow \$ID_{E_{22}} \rightarrow \|\|P_2\|\| \rightarrow \|\|E_{55}\{^{= \text{'Original numbering'}}\}\|\|) \rightarrow P_{89} \rightarrow (\|\|E_{53}\|\| \rightarrow \|\|P_2\|\| \rightarrow \|\|E_{55}\{= \text{'Country'}\}\|\|\|P_{89i}\|\| \rightarrow \|\|E_{53}\|\| \rightarrow \|\|P_{55i}\|\| \rightarrow \$ID_{E_{22}})$ |

$I_{Id_2}$. $I_{Id_i}$ are class instances. If the data property instance already exists raises an exception.

**Prop($I_{Id_1}$, $C_{Id_1}$, $P_{Id}$, $I_{Id_2}$, $C_{Id_2}$)** - If there is an instance of an object property $P_{Id}$ with domain $I_{Id_1}$, instance of the class $C_{Id_1}$, and range $I_{Id_2}$, instance of the class $C_{Id_2}$; else creates a new instance of the object property $P_{Id}$ with domain $I_{Id_1}$ and range $I_{Id_2}$.

**PropS($I_{Id_1}$, $C_{Id_1}$, $P_{Id}$, $I_{Id_2}$, $C_{Id_2}$)** - If there is an instance of an object property $P_{Id}$ with domain $I_{Id_1}$, instance of the class $C_{Id_1}$, and range $I_{Id_2}$, instance of the class $C_{Id_2}$; else raises an exception.

Therefore, rule No. 1 will be translated into the following sequence of commands:

1 NewInst($ID_{E_{31}}$, '$E_{31}$ Document')
2 NewInst($ID_{E_{22}}$, '$E_{31}$ Man-Made Object')
3 NewInst($ID_{E_{33}}$, '$E_{33}$ Linguistic Object') (1)
4 NewProp($ID_{E_{31}}$, '$P_{70}$ documents', $ID_{E_{22}}$)
5 NewProp($ID_{E_{31}}$, '$P_{67}$ refers to', $ID_{E_{33}}$)

With a similar interpretation, rule No. 2 is translated into the following commands sequence:

$$
\begin{array}{ll}
1 & \text{Inst}(ID_{E55_1}, \text{'}E_{55}\text{ Type'}, \text{V}) \\
2 & \text{Inst}(ID_{E55_2}, \text{'}E_{55}\text{ Type'}, \text{'Description level'}) \\
3 & \text{Prop}(ID_{E55_1}, \text{'}E_{55}\text{ Type'}, \text{'}P_2\text{ has type'}, ID_{E55_2}, \\
 & \text{'}E_{55}\text{ Type'}) \\
4 & \text{NewProp}(ID_{E31}, \text{'}P_2\text{ has type'}, ID_{E55_1})
\end{array} \quad (2)
$$

Each commands sequence rule is then translated directly into Java instructions using the `OWL API` library (further details in Subsection 4.1).

Consider, for instance, the ISAD(G) element 'Reference code'. Each unit is uniquely identified by this code. The 'Reference code' can be represented as an instance of the class '$E_{42}$ Identifier', and to represent that the value of the reference code is the identification of the unit, it is possible to use the property '$P_1$ is identified by', resulting the statement '$E_{31}$ Document' '$P_1$ is identified by' '$E_{42}$ Identifier {PT/TT/...}', see Figure 3.

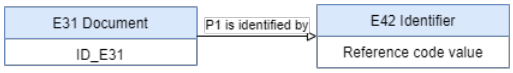



Fig. 3. Reference code representation as unique identifier of a document

However, in this representation, the information that the 'Reference code' is the identifier of the document is implicit. If this information needs to be explicit, then it is possible to apply a type to the identifier with the rule '$E_{42}$ Identifier {PT/TT/...}' '$P_2$ has type' '$E_{55}$ Type {Reference code}', as illustrated in Figure 4.

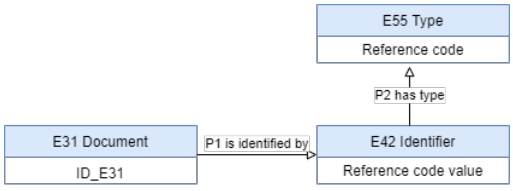



Fig. 4. Reference code representation as unique identifier of a document and a type

If the same 'reference code value' expression is intended to be used to identify other entities, then the identifier could have other types depending on the entity that it identifies, so the type of the identifier on the document should be placed on the relation '$P_1$ is identified by', as shown in Figure 5.

But OWL2 only allows the use of binary properties, so this representation should be done as presented in

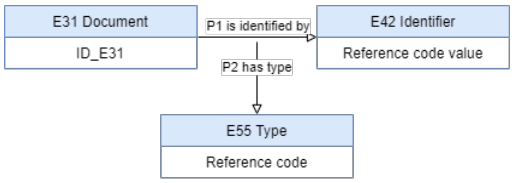



Fig. 5. Reference code representation with the type on the $P_1$ relation

Figure 6 and follows the recommendation of CIDOC-CRM: a subclass of '$PC_0$ CRM Property' is created with the name of the property that has a type, $PC_1$, '$PC_1$ is identified by' with a new data property '$P_{1.1}$ has type'. The properties $P_{01}$ and $P_{02}$ are already defined in CIDOC-CRM for class '$PC_0$ CRM Property' as its domain and '$E_1$ CRM Entity' as its range.

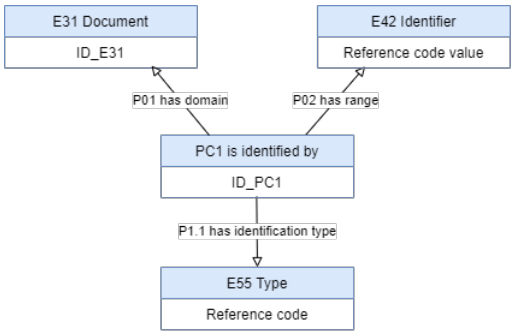



Fig. 6. Reference code representation with the type on the $P_1$ relation as binary relation

However and as explained before, the expression of 'reference code value' is a unique value and with the intention of allowing to search and retrieve information using the term 'reference code', the mapping description rule used for this element needs only to establish the type over the identifier, as illustrated in Figure 4, and is captured by the rule No. 3, shown in Table 1. The representation illustrated in Figure 6 is used when the identifier instance of '$E_{42}$ Identifier' can identify more then one instance and can have more then one type, e.g. 'PT' can be an instance of '$E_{41}$ Appellation' with type 'Country abbreviation' when it identifies a Place or an '$E_{41}$ Appellation' with type 'Person name abbreviation' when it identifies a Person. In this case the type must be on the relation as it happens in rules No. 6, No. 7, No. 8, No. 13, No. 14 and No. 15, from Table 1.

Consider now the element 'Description level' of a unit. Its value establishes the type of the unit, according to ISAD(G) model of the constituents description units of an archive (Figure 1), such as Fonds, Sub-

Fonds, Series, Sub-Series, File, Item, etc. As a result, it is considered that the 'Description level' is the only type property of the '$E_{31}$ Document' that represents the unit. So an instance of '$E_{55}$ Type' is created with the value of the unit type and this instance will have the type 'Description level', if those instances do not exist already. Additionally the property '$P_2$ has type' is added to link the '$E_{31}$ Document' to its type '$E_{55}$ Type'. This representation is adequate for the interpretation of natural language queries, since it makes explicit that the value is a 'Description level'. Note that if an unit could have more then one type, then the type of the type should be a ternary property '$P_{2.1}$ has type' (similar to the one explained before and showed in Figure 6). The Mapping Description Rule No. 2 captures the translation of the 'Reference code' element into CIDOC-CRM representation.

The Mapping Description Rule No. 18 states that the 'original numbering' place of a human-made object falls within the country place of the object unit. This rule must be triggered after the original numbering and country places were generated.

The remain Mapping Description Rules that are presented in Table 1 correspond to some of other interpretation ISAD(G) elements. All of these Mapping Description Rules are displayed, in Appendix A, in a diagram format for better understanding.

## 4. Automatic Migration of Archives Metadata to OWL2

The DigitArq platform, as mentioned before, is supported by a centralized repository (named DigitArq database, from now on), which allows to store all the collected material in a well-structured organization determined by the archival representation. The automatic migration of DigitArq records into CIDOC-CRM is then based on simple translation rules for the elements where there is a mapping between ISAD(G) and CIDOC-CRM. Some of these Mapping Description Rules were already introduced in the previous Section 3 and the Table 1 presents some of the rules established for the migration process. However, there are elements of the ISAD(G), such as 'scope and content', that provide a semi-structured text with additional information to the ones established by the translations of the elements themselves. For these texts there is no direct mapping rules between concepts and the representation must be made by building new mapping rules, according to their structure.

The complete migration process is done in three main steps: 1) DigitArq Metadata Extraction; 2) Migration Process; and 3) Ontology Knowledge Discovery. At first step, the metadata to be represented in CIDOC-CRM are extracted from the DigitArq database. The second step represents the effective mapping process between the ISAD(G) elements and the CIDOC-CRM representations, and is made using the introduced Mapping Representation Rules. Finally, the third step refers to the interpretation of some pieces of text provided by some ISAD(G) elements and that are not yet represented in the CIDOC-CRM Ontology. This last step is done entirely over the information already represented in CIDOC-CRM, and obtained in the second step. The objective of the third step is to map valuable information to the knowledge base, by applying Natural Language Processing techniques to extract the additional information. Figure 7 presents the architecture of the migration process from DigitArq HTML records into CIDOC-CRM, the main tasks of each module are explained in the following subsections.

### *4.1. DigitArq Metadata Extraction*

The DigitArq database contains a large and diverse amount of records, currently over 2 millions. As mentioned before, this database is structured, using a well-established standard archival description, with a hierarchical structure adapted to the nature of archival assets.

Along with the development of the DigitArq database, a web-based search engine (web service) 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 service[4] that for each unit shows the whole information needed to be considered in the migration process.

For this purpose, the `jsoup` library [5] is used to extract web page content from specific fields. `jsoup` is a Java HTML Parser that provides a very reliable, user-friendly, and easy configuration and parameter adjustments 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 [26, 27]. The first one presents a solution for querying Greek governmen-

[4] https://digitarq.arquivos.pt/
[5] http://jsoup.org

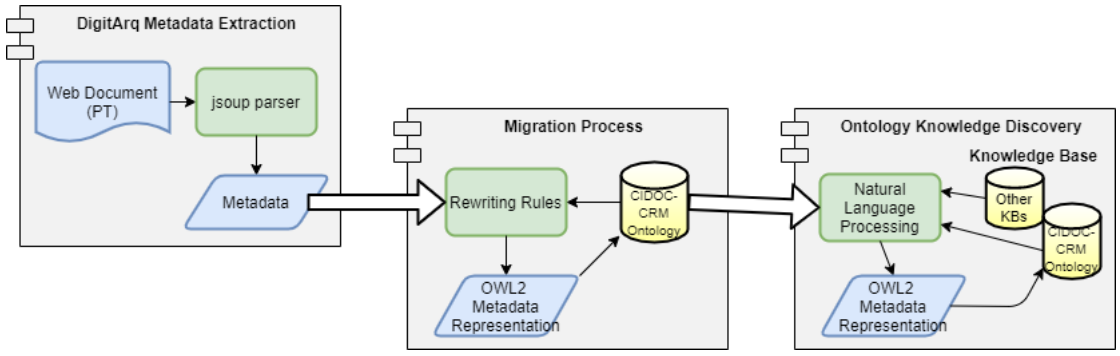



Fig. 7. Architecture for Automatic Migration of ISAD(G) Units into CIDOC-CRM

tal 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.

Each record's web page has a standardized scheme following the ISAD(G) and ISAAR(CPF) 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 the already introduced predefined Mapping Description Rules (summarized in Table 1).

As an illustration, consider the fonds record, named "PARÓQUIA DE ALDOAR"[6], which describes the set of archival documents that composes it, regardless of its form or support, and concerning to baptisms, weddings, deaths, usages and customs, legacies and obligations of religious masses and indices registered by the Parish of Aldoar, from Oporto district, Portugal.

Using a set of `jsoup` functions, it is possible to extract information like the title, and other ISAD(G) elements (fields), as well as their values, presented in the web page record. For instance, the function *title()* allows to extract the title of the record; the function *getElementsByClass()* allows to extract the information per elements, as well as to get the set of its child records; the function *f.select("span").first().html()* allows to extract the name of the each field and their values could be obtained using the function *text()*. `jsoup` also provides functions to connect and parse directly the web page source, such as *connect()* followed by *get()*, and *parse()*, respectively. The fragment of Java code Listing 1 illustrates how this is performed. Using such strategy to extract information, it is possible to consider other public webpage platforms as database that can offer the same or additional information, such as the "Archives Portal Europe"[7].

Listing 1: Metadata Extraction Java Code for the fonds record "PARÓQUIA DE ALDOAR"

```
Document record = Jsoup.connect("http://pesquisa.
    ↪ adporto.arquivos.pt/details?id=488455").get
    ↪ ();
String title = record.title());
Elements fields = record.getElementsByClass("Field"
    ↪ );
for(Element f: fields) {
  String field = f.select("span").first().html();
  String value = f.text().substring(field.length()
      ↪ +1);
}
(...)
Elements childs = record.getElementsByClass("AspNet
    ↪ -TreeView-Leaf␣AspNet-TreeView-
    ↪ ParentSelected");
```

Some of the fields and the corresponding values extracted from "PARÓQUIA DE ALDOAR"'s fonds record are presented in Table 2.

The information extracted is adequately analysed, where each fields' name and their values are identified, the adequate ontology representation is established, and the corresponding ontology entities, such as individuals and properties, are then generated. This process is made by applying the predefined set of Mapping Description Rules, some of them presented in Table 1 and introduced in the previous Section 3, and their implementation is explained with more details in the following subsections.

[6]http://pesquisa.adporto.arquivos.pt/details?id=488455, with dataset [28]

[7]The Archives Portal Europe provides access to information on archival material from different European countries as well as information on archival institutions throughout the continent. https://www.archivesportaleurope.net/home

Table 2
Example of some fields and the corresponding values extracted from "PARÓQUIA DE ALDOAR"'s fonds unit

| Fields | Values |
|---|---|
| Description level | Fonds |
| Title | PARÓQUIA DE ALDOAR |
| Reference code | PT/ADPRT/PRQ/PPRT01 |
| Title type | Formal |
| Date Range | 1640-05-15 to 1911-03-31 |
| Scope and content | Documentação relativa a baptismos, (...). |
| Creation date | 05/22/2012 00:00:00 |

### 4.2. *Migration Process from DigitArq into CIDOC-CRM*

The migration process consists in generate automatically CIDOC-CRM entities from DigitArq database records by applying the set of mapping description rules, already introduced in Subsection 3.3.

At this step, the `OWL API`[8] [29] and the `SPARQL-DL`[9] [30, 31] libraries are used to upload and model the CIDOC-CRM archival representation into a well-structured model for Java environment, to implement the mapping description rules, to update the mapping knowledge base, and also to reasoning over the knowledge base. The `OWL API` is a high level Application Programming Interface (API) for working with OWL ontologies, and is closely aligned with the OWL 2 structural specification[10]. It supports parsing and rendering in the syntaxes defined in the W3C specification, manipulation of ontological structures, and the use of reasoning engines. The `SPARQL-DL` is a Java query engine, settled on top of the `OWL API`, and it is fully aligned with the OWL2 standard and adds a `SPARQL-DL` interface to every `OWL API 3` reasoner.

Using the mentioned tools, the set of commands representing each mapping description rule is directly translated to Java instructions, which allows for automatically generate the CIDOC-CRM representation for each DigitArq record, and save it in OWL2 format.

As mentioned in the previous Subsection 4.1, each DigitArq database record, interpreted as a unit of description, has a well-known structure represented by a set of fields and their values, as well as their hierarchical relationship with other units, according to archival standards. As presented before, the migration process defines for each unit the application of:

1. Rule 1, Table 1 - The unit itself is mapped into an instance of '$E_{31}$ Document' that documents (property '$P_{70}$ documents') a physical object (instance of '$E_{22}$ Human-Made Object') and refers to (property '$P_{67}$ refers to') a linguistic object (an instance of '$E_{33}$ Linguistic Object'). The corresponding set of commands introduced in 1, Subsection 3.3, generates the set of Java instructions as shown in Listing 2.

Listing 2: "Java code translation of Rule No. 1"

```
String erlangen_crm = "http://erlangen-crm.org/200717/";
(...)
OWLIndividual IDE22 = newInst(erlangen_crm + "E22_Human-Made_Object");
add_prop(IDE31,erlangen_crm + "P70_documents",IDE22);
OWLIndividual instanceE33 = newInst(erlangen_crm + "E33_Linguistic_Object");
add_prop(IDE31,erlangen_crm + "P67_refers_to", IDE33);
```

2. Rule 17, Table 1 - If the unit is composed by a collection of other units then using the property '$P_{106}$ is composed of' allows to represent the hierarchical relationship between the unit and the units that are part of it. The hierarchical link between two units '$ID_{E31_1}$' and '$ID_{E31_2}$', instances of the '$E_{31}$ Document', is established by applying the following command
Prop($ID_{E31_1}$, '$E_{31}$ Document', '$P_{106}$ is composed of', $ID_{E31_2}$, '$E_{31}$ Document')
with the following corresponding Java instruction

```
Prop(IDE31_1,"http://erlangen-crm.org/200717/P106_is_composed_of", IDE31_2);
```

The unit $ID_{E31_2}$ is interpreted as a part of the unit $ID_{E31_1}$ and with a description level below in the hierarchical tree representation (Figure 1).
3. The remain rules, Table 1 - For each ISAD(G) element that are described in the unit, the corresponding rule is applied to map the information into CIDOC-CRM representation. For instance, considering rule No. 2, which maps the 'Description level' of the unit, the corresponding set of commands, presented in 2, Subsection 3.3, is translated to the Java instructions illustrated

[8] http://owlapi.sourceforge.net/
[9] https://www.derivo.de/en/resources/sparql-dl-api/
[10] https://www.w3.org/TR/owl2-syntax/

in Listing 3. The set of Mapping Representation Rules applied varies according to the information that is described in the unit.

Listing 3: "Java code translation of Rule No. 2"

```
String erlangen_crm = "http://erlangen-crm.org/200717/";
(...)
case "Description level":
  OWLIndividual ID_E55_1 = inst(erlangen_crm + "E55_Type", V);
  OWLIndividual ID_E55_2 = inst(erlangen_crm + "E55_Type", "
      Description level");
  prop(ID_E55_1, erlangen_crm + "P2_has_type", ID_E55_2);
  add_prop(instanceE31, erlangen_crm + "P2_has_type",ID_E55_1);
  break;
```

For better understanding, let us go back to the fonds unit "PARÓQUIA DE ALDOAR" and consider the elements 'Description level', 'Reference code', 'Language of the material', and 'Date range'. Figure 8 shows the solution obtained in the mapping process by applying respectively the rules No. 2, No. 3, No. 4 and No. 11 for the corresponding elements and their values. The mapping representation of the unit is obtained by applying the rule No. 1, and the mapping representation of the hierarchical relationship with other units is obtained by applying the rule No. 17. The fonds unit is composed by 8 other units (Figure 8 just presents 2 of them), each one with the classification of 'Series' as 'Description level' (and each one is composed by other units, according to the hierarchical model level in Figure 1). The size of the hierarchy depends on the type and composition of the fonds and on what is described in DIgitArq database. The complete representation of the fonds unit "PARÓQUIA DE ALDOAR" in CIDOC-CRM representation is obtained by applying the mapping process for each unit, belonging to the hierarchical tree of the fonds. Table 3 shows the mapping process metrics of the CIDOC-CRM representation of the fonds unit "PARÓQUIA DE ALDOAR", as well as its complete hierarchical composed units.

### 4.3. Ontology knowledge Discovery

The Ontology Knowledge Discovery step consists of, by applying Natural Languages Processing (NLP) techniques, finding the proper interpretation of some text fields, from instances of CIDOC-CRM class, such as '$E_{62}$ String', that contain strings with description information as their value, and therefore extract additional information. The 'Scope and content' is one of the ISAD(G) elements that is characterized by having additional information describing its unit, and it is in text format.

Table 3
Mapping Process Metrics of the complete CIDOC-CRM representation of the fonds "PARÓQUIA DE ALDOAR"

| | Total |
|---|---|
| Axiom | 146609 |
| Logical axiom | 106766 |
| Declaration axioms | 34058 |
| Class | 84 |
| Object property | 298 |
| Data property | 10 |
| Annotation Property | 4 |
| Individuals | 33666 |
| Object property assertion between Individuals | 72016 |

These texts, usually, have a structure that can be recognized, by using NLP tools, and giving as output a feature value list that will be the input of the migration sub-process. Enumeration is a structured pattern that is frequent in these text fields.

Consider as an illustration the fonds unit entitled "JUÍZO DA ÍNDIA E MINA"[11], which describes the set of archival documents that compose it, regardless of its form or support, and concerning civil and criminal processes registered under Portuguese discoveries root "Índia" and "Mina". Most of those processes are related to shipping damage, payment of soldiers, collection of freight, freight and unloading, qualification of heirs, processes of individuals who wanted to prove to be Portuguese and that their ships had made in Portuguese shipyards, with no foreigners interested in their cargo, and also about Corsair and piracy lawsuits. This fonds unit has the following piece of text withdrawn from its 'Scope and content' element.

> "Referem ainda o tipo de embarcações: navio, corveta, bergantim, galera, escuna, brigue, iate, caíque, nau, sumaca, barco, corsário, polaca"
>
> (They also mention the type of vessels: ship, corvette, brigantine, galley, schooner, brig, yacht, caique, ship, sumaca, boat, corsair, polish)

From this text using some NLP tools (e.g., tagger and lemmatization) and some grammar rules, it is obtained a list of Type-Value (vessel, name) elements.

[11] http://digitarq.arquivos.pt/details?id=4208377, with dataset [28]

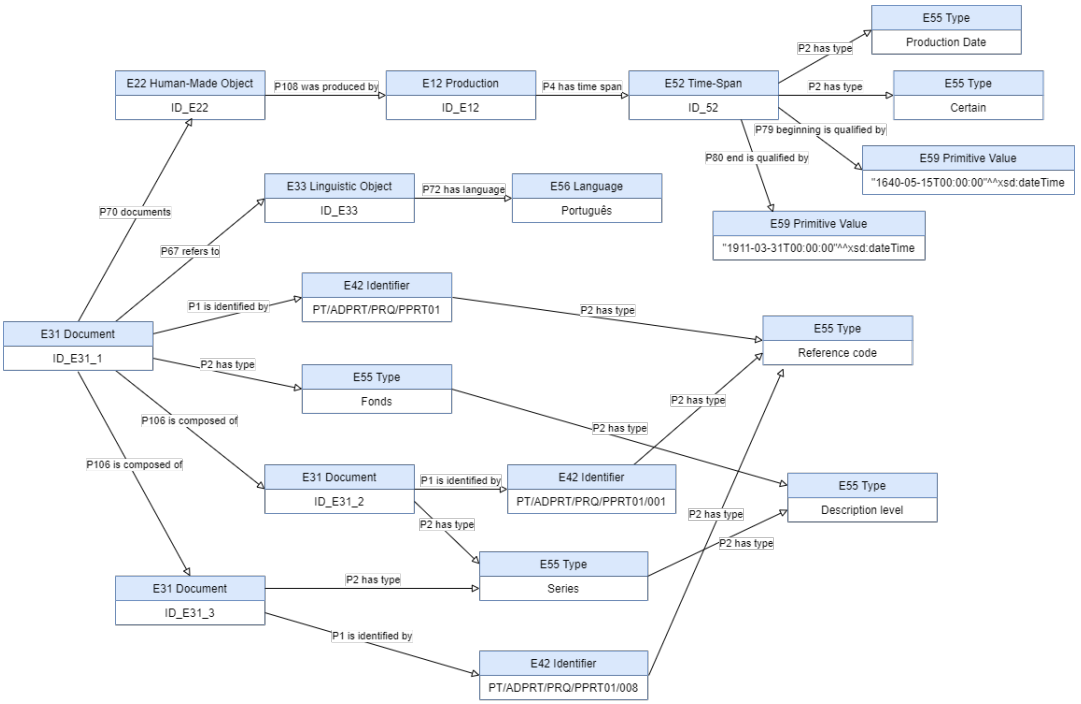



Fig. 8. The fonds unit “PARÓQUIA DE ALDOAR” partial mapping.

The representation of this information in CIDOC-CMR is given by rule No. 19, Table 5.

Each pair vessel-value gives rise to a new instance of ’$E_{22}$ Human-Made Object’ connected to the ‘$E_{31}$ Document’ of the fonds unit by the Object property ’$P_{129}$ is about’, the instance ’$E_{22}$ Human-Made Object’ has type value that has type ‘vessel’. Therefore, rules No. 19 to No. 21, from Table 5, can be automatically generated from the text structure in ISAD(G) properties to obtain the CIDOC-CRM representation of the corresponding information.

For this example, Table 4 presents the total of axioms generated to represent the information interpreted, which increases substantially the amount of entities in the knowledge base compared to the entities generated in the migration process of the fonds unit itself. More important, it is possible to retrieve such information and infer about it, both automatically.

Another structure pattern that is often found in ’Scope and content’ element is the identification of people and their relationship role with the ’Recipient’ when referring to events [32]. These events can be baptisms, weddings or deaths, and the information is organized in a list of names and tagged by the role of the relationship that connects to the ’Recipient’ of the unit. For instance, to illustrate this pattern, consider

Table 4

Populate Metrics of the Scope and content field of the “JUÍZO DA ÍNDIA E MINA”’s fond unit

| Individual Axioms | Scope and content | Migration Step | Total |
|---|---|---|---|
| Class assertion | 245 | 43 | 288 |
| Object property assertion | 381 | 44 | 425 |

the Item unit entitled "REGISTO DE BAPTISMO"[12], which refers to the baptism happening of the person named "Ana", ’Recipient’ of the unit, and its ’Scope and content’ element has the following text value:

> "Pais: Manuel de Oliveira e 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"
>
> ("Parents: Manuel de Oliveira and Rufina Maria
> Maternal grandparents: Manuel da Fonseca and Rosa da Silva
> Paternal grandparents: José de Oliveira and Jacinta de Oliveira
> Godparents: Manuel Martins Ramos and Maria Francisca
> Birthdate: 10th February, 1812")

[12]http://pesquisa.adporto.arquivos.pt/details?id=1374655

Table 5
Mapping Description Language Rules for the Ontology Knowledge Discovery Process

| Rule No. | Left part (rec[attribute value list]) | Right part CIDOC-CMR |
|---|---|---|
| 19 | [‘vessel’,V] | $\$ID_{E_{31}} \rightarrow P_{129} \rightarrow (E_{22} \rightarrow P_2 \rightarrow (\langle E_{55}\{= V\}\rangle \rightarrow \langle P_2\rangle \rightarrow \langle E_{55}\{= \texttt{'vessel'}\}\rangle))$ |
| 20 | [’products traded’,V] | $\$ID_{E_{31}} \rightarrow P_{129} \rightarrow (E_{22} \rightarrow P_2 \rightarrow \langle E_{55}\{= V\}\rangle \rightarrow \langle P_2\rangle \rightarrow \langle E_{55}\{= \texttt{'products traded'}\}\rangle))$ |
| 21 | [‘Vessel name’,V] | $\$ID_{E_{31}} \rightarrow P_{129} \rightarrow (E_{22} \rightarrow P_1 \rightarrow \langle E_{41}\{= V\}\rangle \rightarrow P_2 \rightarrow \langle E_{55}\{= \texttt{'vessel'}\}\rangle)$ |
| 22 | [’Baptism’, birth(Mother, Father, DBirth)] | $\$ID_{E_{31}} \rightarrow P_{67} \rightarrow (E_{67} \rightarrow P_{98} \rightarrow (\|E_{21}\| \rightarrow \|P_{02i}\| \rightarrow (\|PC_{129}\| \rightarrow \|P_{129.1}\| \rightarrow \|E_{55}\{\texttt{'Recipient'}\}\| \rightarrow \|P_{01}\| \rightarrow \$ID_{E_{31}})) \rightarrow P_{96} \rightarrow (E_{21} \rightarrow P_1 \rightarrow \langle E_{41}\{\texttt{Mother}\}\rangle) \rightarrow P_{97} \rightarrow (E_{21} \rightarrow P_1 \rightarrow \langle E_{41}\{\texttt{Father}\}\rangle) \rightarrow P_4 \rightarrow (E_{52} \rightarrow P_{170i} \rightarrow \langle E_{61}\{\texttt{DBirth}\}\rangle)$ |
| 23 | [’Baptism’, bapt(Godfather, Godmother, DBap)] | $\$ID_{E_{31}} \rightarrow P_{129} \rightarrow (E_5 \rightarrow P_2 \rightarrow \langle E_{55}\{\texttt{'Baptism'}\}\rangle \rightarrow P_{01i} \rightarrow (PC_{14} \rightarrow P_{02} \rightarrow (\|E_{21}\| \rightarrow P_{02i} \rightarrow (\|PC_{129}\| \rightarrow \|P_{129.1}\| \rightarrow \|E_{55}\{\texttt{'Recipient'}\}\| \rightarrow \|P_{02}\| \rightarrow \$ID_{E_{31}})) \rightarrow P_{14.1} \rightarrow \langle E_{55}\{\texttt{'Baptized'}\}\rangle) \rightarrow P_{01i} \rightarrow (PC_{14} \rightarrow P_{02} \rightarrow (E_{21} \rightarrow P_1 \rightarrow \langle E_{41}\{\texttt{Godmother}\}\rangle) \rightarrow P_{14.1} \rightarrow \langle E_{55}\{\texttt{'Godmother'}\}\rangle) \rightarrow P_{01i} \rightarrow (PC_{14} \rightarrow P_{02} \rightarrow (E_{21} \rightarrow P_1 \rightarrow \langle E_{41}\{\texttt{Godfather}\}\rangle) \rightarrow P_{14.1} \rightarrow \langle E_{55}\{\texttt{'Godfather'}\}\rangle) \rightarrow P_4 \rightarrow (E_{52} \rightarrow P_{170i} \rightarrow \langle E_{61}\{\texttt{DBap}\}\rangle)$ |
| 24 | [’grandparents’, mother(Mother, Father)] | $\$ID_{E_{31}} \rightarrow P_{67} \rightarrow (E_{67} \rightarrow P_{96} \rightarrow (E_{21} \rightarrow P_1 \rightarrow \langle E_{41}\{\texttt{Mother}\}\rangle) \rightarrow P_{97} \rightarrow (E_{21} \rightarrow P_1 \rightarrow \langle E_{41}\{\texttt{Father}\}\rangle) \rightarrow P_{98} \rightarrow (\|E_{21}\| \rightarrow \|P_{96i}\| \rightarrow (\|E_{67}\| \rightarrow \|P_{98i}\| \rightarrow (\|E_{21}\| \rightarrow \|P_{02i}\| \rightarrow (\|PC_{129}\| \rightarrow \|P_{129.1}\| \rightarrow \|E_{55}\{\texttt{'Recipient'}\}\| \rightarrow \|P_{02}\| \rightarrow \$ID_{E_{31}}))))$ |
| 25 | [’grandparents’, father(Mother, Father)] | $\$ID_{E_{31}} \rightarrow P_{67} \rightarrow (E_{67} \rightarrow P_{96} \rightarrow (E_{21} \rightarrow P_1 \rightarrow \langle E_{41}\{\texttt{Mother}\}\rangle) \rightarrow P_{97} \rightarrow (E_{21} \rightarrow P_1 \rightarrow \langle E_{41}\{\texttt{Father}\}\rangle) \rightarrow P_{98} \rightarrow (\|E_{21}\| \rightarrow \|P_{97i}\| \rightarrow (\|E_{67}\| \rightarrow \|P_{98i}\| \rightarrow (\|E_{21}\| \rightarrow \|P_{02i}\| \rightarrow (\|PC_{129}\| \rightarrow \|PC_{129.1}\| \rightarrow \|E_{55}\{\texttt{'Recipient'}\}\| \rightarrow \|P_{02}\| \rightarrow \$ID_{E_{31}}))))$ |
| 26 | other rules - about persons on events and births | |

Applying to this text some NLP tools (e.g. tagger and lemmatization) and some grammar rules, it is possible to extract the names of each person and the corresponding role of the relationship that links each person to "Ana" through the baptism event. In particular, this means that it is possible to identify "Ana"'s parents, grandparents from both sides and also her birthdate.

Unlike what happens with the birth event for which the CIDOC-CRM has the entity '$E_{67}$ Birth' to represent this concept, CIDOC-CRM model does not have a specific entity to represent the baptism event. In this case, any baptism event is established as an instance of the entity '$E_5$ Event' and to mark the event as a baptism, it is applied a type to the event, i.e., '$E_5$ Event' '$P_5$ has type' '$E_{55}$ Type'.

To establish parenting relationship through the birth event, CIDOC-CRM model has proper properties, like '$P_{96}$ by mother', '$P_{97}$ by father' and '$P_{98}$ brought into life', that allow to perfectly represent the roles of being a father or a mother of someone else. To set a name to a person, the CIDOC-CRM has the class '$E_{41}$ Appellation' whose instances values represent names and using the object property '$P_1$ is identified by'. For instance, to set 'Manuel de Oliveira' as the name of a person, the representation expression is '$E_{21}$ Person' '$P_1$ is identified by' '$E_{41}$\{Manuel de Oliveira\}'. The rule No. 22, in Table 5, expresses the complete mapping description for a birth event, described in a baptism record unit.

Additionally, for instance, using the information about the grandparents, it is also possible to identify who are the parents of "Ana"'s parents and state a birth event for each "Ana"'s parents. Rules No. 24 and No. 25, in Table 5, are the complete mapping description rules representing the grandparents relationships from both sides.

The role of being a godparent is established through the baptism event and the CIDOC-CRM model does not have a proper set of entities and properties that explicitly represent those concepts. The solution is to establish a ternary relation where the role of the relationship is expressed as the type of a property. The entity '$PC_{14}$ Carried Out By' represents the relationship, which has type ('$P_{14.1}$ has type') the role of the re-

lationship ('$E_{55}$ Type'), with domain the baptism '$E_5$ Event' and range the godparent '$E_{21}$ Person'. Rule 23, in Table 5 captures the complete mapping description for the godparents relationship.

Figure 9 shows the complete solution for the baptism example, obtained after the application of the Mapping Description Rules identified and expresses the axioms that are added to the knowledge base.

The Mapping Description Rules presented in Table 5, are also displayed, in Appendix A, in a diagram format for better understanding.

## 5. Querying CIDOC-CMR representation of Archives metadata

The result of the migration process can be evaluated by querying the knowledge base, consisting of CIDOC-CRM Ontology and the complete set of assertions obtained through the Migration Process and the Ontology Knowledge Discovery. The guarantee that the CIDOC-CRM Ontology representation of the DigitArq metadata is well-performed is established when questioning (searching) the knowledge base, it is possible to retrieve the original information.

### *5.1. Querying the Knowledge Base*

The process of retrieving the information about the archival units uses the Mapping Description rules, presented in Table 1 and Table 5, to define the Description Logic (DL) queries on the subject of a question.

The following examples illustrate queries to obtain some of the elements of an unit[13], such as: Reference code, Description level, Title, Title Type, Creator (Producer), Date Range.

1. The unit with 'Reference code' PT/ADPRT/PRQ/PPRT01/001/0004/00001
   $DLq_1$ = `P1 is identified by` *some* (`inverse 'P1 is identified by'` *some* (`'P1 is identified by'` *value* `PT/ADPRT/PRQ/PPRT01/001/0004/00001`) *and* `'P2 has type'` *value* `Reference code`
   Answer: E31_Document20
2. 'Description level' of the unit with 'Reference code' PT/ADPRT/PRQ/PPRT01/001/0004/00001
   $DLq_2$ = `inverse 'P2 has type'` *some* $DLq_1$ and `'P2 has type'` *value* `Description level`
   Answer: Item
3. 'Title' of the unit with 'Reference code' PT/ADPRT/PRQ/PPRT01/001/0004/00001
   $DLq_3$ = `inverse 'P02 has range'` *some* (`'PC102 has title'` *and* `'P01 has domain'` *some* $DLq_1$)
   Answer: Registo de Baptismo
4. 'Title type' of the unit with 'Reference code' PT/ADPRT/PRQ/PPRT01/001/0004/00001
   $DLq_4$ = `inverse 'P102.1 has type'` *some* (`'PC102 has title'` *and* `'P01 has domain'` *some* $DLq_1$)
   Answer: Formal
5. 'Creator (the producer)' of the unit with 'Reference code' PT/ADPRT/PRQ/PPRT01/001/0004/00001
   $DLq_5$ = `inverse 'P02 has range'` *some* (`inverse P01i is domain of'` *some* (`inverse 'P14 carried out by'` *some* (`'P108 has produced'` *some* (`P70i is documented in'` *some* $DLq_1$))))
   Answer: PPRT01 *and* Paróquia de Aldoár
6. 'Date range' of the unit with 'Reference code' PT/ADPRT/PRQ/PPRT01/001/0004/00001
   $DLq_6$ = `inverse 'P79 beginning is qualified by'` *some* (`inverse 'P4 has time-span'` *some* (`'P108 has produced'` *some* (`'P70i is documented in'` *some* $DLq_1$)))
   Answer: "1811-07-07T00:00:00"^^xsd:dateTime
7. 'Reference code' of units that mention 'corveta'
   $DLq_7$ = `inverse 'P1 is identified by'` *some* (`'P129 is about'` *some* (`'E22 Man-Made Object'` *and* `'P2 has type' value corveta`))
   Answer: `PT/TT/JIM`

To help the evaluation of the CIDOC-CRM representation of the migrated data, a web interface was developed, see Subsection 5.2 for further information. To explore the advantages of the archives OWL representation, the interface application enables smart queries, such as:

1. Instances identified by *Name*
   (`'PC1 is identified by'` *and* `'P02 has range'` *value* **Name**) *or* (`'P1 is identified by'` *value* **Name**)
2. What is the type of **Name**:
   `inverse 'P2 has type'` *some* (`inverse P01 has domain'` *some* ( `'PC1 is identified by'` *and* `'P02 has range'` *value* **Name**)) *or* (`inverse 'P2 has type'` *value* **Name**) *or* (`inverse 'P2 has type'` *some* (`'P1 is identified by'` *value* **Name**)) *or* (`inverse 'P0.1 has type'` *some* (`'P02 has range'` *some* (`'P1 is identifi-`

[13] These queries were done in Protegé with the reasoner Pellet over the dataset [33].

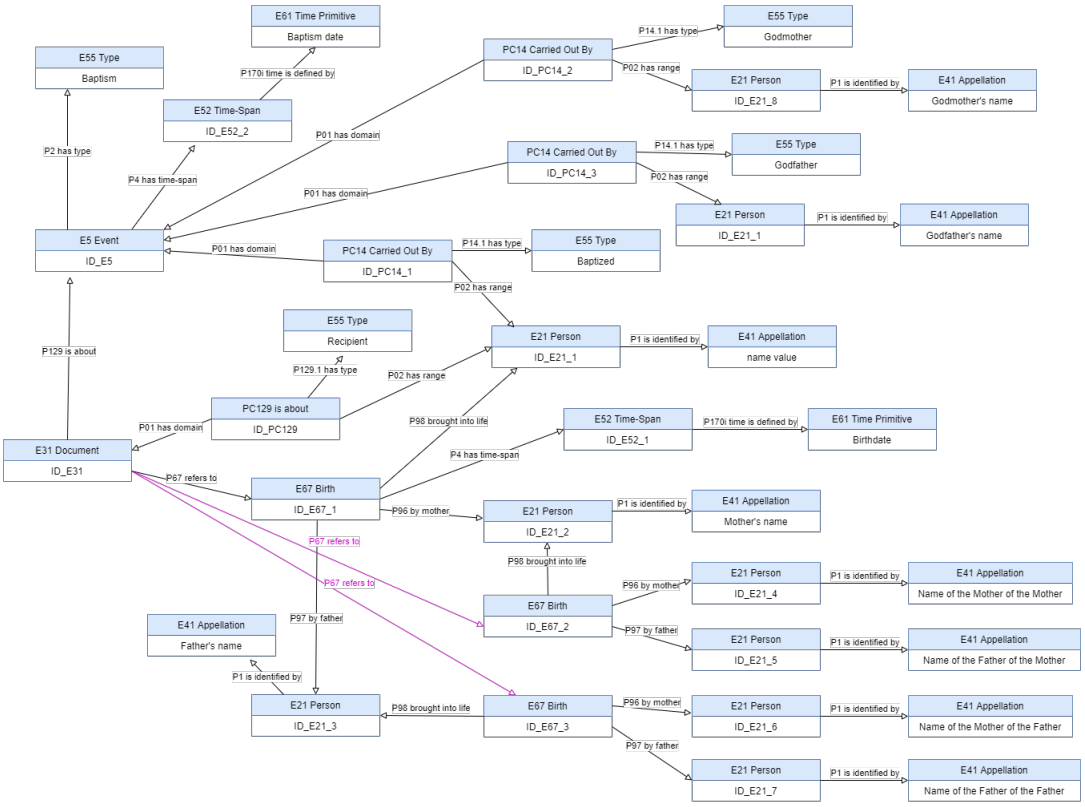



Fig. 9. The baptism event of the person "Ana" partial mapping.

ed by' *value* **Name**))) *or* (inverse 'P$_{0.1}$ has type' *some* (P$_{02}$ has range' *value* **Name**))

Examples of answers obtained when questioning the knowledge base about the type of a given name:

- Name=Ana
  'E$_5$5 Type': Recipient
- Name=Paróquia de Aldoar
  'E$_5$5 Type': Institution name
- Name=Portugal
  'E$_5$5 Type': Country name
- Name=PT
  'E$_5$5 Type': Country abbreviation

Note that Protegé does not allow to query for a property of an instance, but it is possible to do it with SPARQL-DL as in Listings 4. This query retrieves all the instances of 'E$_{31}$ Document' that its 'Description level' is 'Fonds' (Mapping Description Rule No. 2, Table 1).

Listing 4: "Units with 'Description level' 'Fonds'"

```
PREFIX erlangen-crm: <http://erlangen-crm.org
    ↪ /200717/>
SELECT ?x WHERE {
  Type(?x,erlangen-crm:E31_Document),
  PropertyValue(?x,erlangen-crm:P2_has_type,?
      ↪ instance),
  Type(?instance,erlangen-crm:E55_Type>),
  PropertyValue(?instance,erlangen-crm:P2_has_type,
      ↪ erlangen-crm:Description_level),
  SameAs(?instance,erlangen-crm:Fonds)
}
```

These 'smart' queries are useful in the interface application, not only for helping the users to explore the knowledge base, but also to use in the interpretation of natural language text and assign ontology terms to sentences tokens.

Another kind of query that is important for the natural language interpretation process is to obtain object properties that links instances from a class domain and instances from a class range. The query presented in Listings 5 is an example of an SPARQL-DL query to obtain the object properties that link instances of 'E$_{31}$ Document' to a 'Reference code'. This query can not be done in Protegé.

Listing 5: "properties that link an 'E$_{31}$ Document' to a 'Reference code'"

```
PREFIX erlangen-crm: <http://erlangen-crm.org/200717/>
SELECT DISTINCT ?p WHERE {
  Type(?x,erlangen-crm:E31_Document),
  PropertyValue(?y,erlangen-crm:P2_has_type,erlangen-crm:Reference_code>),
  ObjectProperty(?p),
  PropertyValue(?x,?p,?y)
}
```

### 5.2. Query Ontology Interface

The knowledge base querying process is supported by an application program interface (API), entitled Query Ontology Interface, that facilitates the interaction between regular users and the knowledge base. The main goals for the development of such API are to allow retrieving information from knowledge base without technically know how the information is represented in the ontology, as well as to make those question as near as possible to natural language text. The main target users, such as the librarians or archivists, are in general not able to make queries using SPARQL language, or even using description logic languages.

The Query Ontology Interface was developed using Spring Boot[14], a Java-based framework that allows to create a Graphical User Interface (GUI) and export the final API in a stand-alone application (originally a web-application). The SPARQL-DL Java query engine is used to search the knowledge base, and serves as a middle layer application between the GUI and the knowledge base. The question made by the user at the GUI level is translated to the corresponding CIDOC-CRM representation and the answer is retrieved using the SPARQL-DL engine and then presented at the GUI application level. The approach used to query the knowledge base is as much user-friendly as possible, besides the use of a GUI, it also uses Natural Language understanding mechanisms to help the users in the querying process. Other existing interfaces over CIDOC-CRM Ontology use similar approach, such as OpenArcheo [18].

The Query Ontology Interface is able to retrieve information about single individuals and about the structure of the whole knowledge base. For instance, it is possible to retrieve information based on the value of some key-entities, like '$E_{41}$ Appellation' or '$E_{42}$ Identifier', that are expressed in Natural Language and stored as xsd:string. It is also possible to define a constraint (or a joint of constraints) to retrieve the desired individuals. The result of such query is an individual (or a joint of individuals) with all the properties and other individuals linked to it. In addition, it was defined a set of predefined queries that work like filters, such as displaying all the class entities or all the individuals belonging to a class entity.. An example on how it works is shown in Figure 10, where the search is made using the 'Reference code' value, see Figure 10a, and the result is the information of the corresponding unit elements, see Figure 10b.

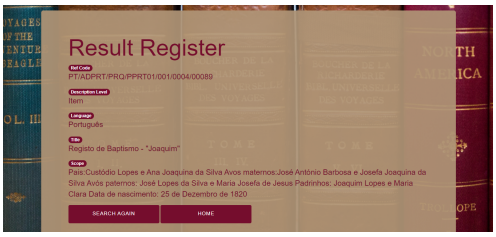
[14] https://spring.io/

## 6. Open Problems

The development of the Mapping Description Rules and their implementation process, together with the analysis of different examples, allowed to notice some issues that led to a set of open problems.

In the Ontology Knowledge Discovery process, two of the major problems identified are, first, to know exactly the information available in the text elements and, second, what is possible and important to infer from them. The text fields are free text, but depending on what they are about, what event or subject they are describing, it is possible to identify some structure which allows to define proper mapping description rules for their representation. For instance, consider the example of "Ana"'s baptism unit presented in Subsection 4.3 and the semi-structured text of its 'Scope and content' element. The text format happens to be equal for all the units referring to baptism events. After knowing the subject type of the unit, the information available in the text fields can be represented by applying the mapping description rules established for the corresponding semi-structured text, as explained before. For the "Ana"'s baptism unit and its 'Scope and content' value example, the axioms generated by the Mapping Descriptions Rules from No. 22 to No. 26, from Table 5, depends on the information available. For instance, if the information about the grandparents is not available, it is not possible to infer the representation of the birth events for both "Ana"'s parents, then the corresponding axioms are not added to the knowledge base. However, if the information to generate the axioms, even being complete, is not correctly interpreted and identified, it may lead to inaccurate representation. Therefore, proper NLP techniques are necessary to make the adequate interpretation and identification of the information available, allowing,


(a) Query the ontology about a specific 'Reference code' value




(b) The query result for a specific 'Reference code' value.

Fig. 10. Query Ontology Interface example

beyond the application of the correct mapping description rules, to generate accurate information representation.

Looking in particular to this example, some other questions occurred, beyond the simple interpretation of the text. For instance, when a new person shares some properties with a known person, should it be considered that it is the same person? when a person has the same names for its parents and grand parents of a known person in the knowledge base, are they siblings? should that relationship role be considered and added to the knowledge base?

In the process of interpreting and representing enumeration lists, the following issues were identified and need to be taken into account:

- Synonyms - in enumerations the Type or the value can be a word or phrase that means exactly or nearly the same as another word or phrase that was already introduced as a new 'E55 Type´, e.g. vessel and ship.
- Names - in the same document or same enumeration list a name can appear more than once.

With regard to the Mapping Description Rules, from Table 1, there are some exceptions, presented below, that need to be considered in the process of extracting the information necessary for the migration process.

The Mapping Description Rule No. 7 correspond to the 'Recipient' representation, main entity to which the unit refers to. The 'Recipient' element could not be explicitly presented in the description of the unit and, when it happens, the ISAD(G) element that could provide this information is the 'Title' element of the unit. For these cases, the 'Title' element should be properly interpreted by applying NLP rules that allow to identify the title and the recipient of the corresponding unit. For instance, the unit with reference code 'PT/ADPRT/PRQ/PPRT04/001/0054/000013' does not have a 'Recipient' element, but its title 'REGISTO DE BATISMO DE ANA' (ANA BAPTISM REGISTRATION) includes the recipient name 'Ana'.

The Mapping Description Rules No. 13 and No. 14 are applied to explicitly represent, respectively, the current keeper and the current location country of the physical object described in a unit. When this information is not explicitly represented with an adequate unit element, it is possible to extract the information required from the 'Reference code' value of the unit. As mentioned before, the 'Reference code' element identifies uniquely the unit of description. To provide an accurate link to the information of the unit, the following conditions are taken into account when creating the 'Reference code' value of an unit: first, the country code in accordance with the latest version of ISO 3166 Codes for the representation of countries names; second, the repository code in accordance with the national repository code standard or other unique location identifier; and third, a specific local reference code, control number, or other unique identifier.

The Mapping Description Rules No. 15 and No. 16 capture the representation of the 'Creator' of the unit and its type. When these information values are not explicitly available to be interpreted, the 'Reference code' value of the unit also provides the information required to be interpreted.

As an example, consider the reference code 'PT/ADPRT/PRQ/PPRT04/001/0054/000013', where:

**PT** is the country abbreviation of Portugal.
**ADAVR** is the keeper abbreviation name of Arquivo Distrital do Porto.
**PRQ** is the abbreviation of the institution type, Paróquia (parish), of the producer.
**PPRT04** is the abbreviation of Paróquia de Cedofeita (Cedofeita parish), the producer or creator of the unit.

An adequate interpretation of the 'Reference code' value a unit will allow to extract and then represent the information about the current keeper and current location of the corresponding unit, as well as its creator and the type of the creator.

The Mapping Description Rule No. 18 expresses the relationship between the 'Original numbering' identification and the location country, i.e, the first place 'falls within' the second place. This interpretation only occurs when the unit explicitly presents the 'Original numbering' value and can only be set after the current location country of the physical object of the unit is already represented.

The Mapping Description Rule No. 14 associates a country to a human-made object. The object property '$P_{55}$ current location' has the constrain of max. 1, implying that there is at most one place. Therefore, to avoid the unification of different instances of place, whenever a country place is used, a new instance of '$E_{53}$ Place', the second in this rule, is created with the same linked properties apart the link to the first place. With this rule, the migration of a fond with hundreds of units will give rise to hundreds of '$E_{53}$ Place' to represent the (same) country. It is possible to avoid this proliferation of similar instances by correcting the Mapping Description Rule No. 14 or add a new Mapping Description Rule similar to rule No. 18 that will infer that those instances are the same.

The expressiveness power of the Mapping Description Rules with the proposed extensions is enough to deal with this kind of issues.

## 7. Conclusions and Future Work

The experience results show that the use of Mapping Description Rules with the proposed extensions has the expressiveness power necessary to define the representation of structured information, such as archives, in an OWL2 ontology such as CIDOC-CMR. These Mapping Description Rules can be automatically interpreted using an environment, such as `OWL API`, to obtain the set of assertions that represents the information in the target ontology.

The task of representing the information, such as an archive, in an ontology requires the study of the ontology and their recommendations in order to achieve interoperability sharing and to use information already represented in the ontology, as well as the use of platforms to explore the information represented. The use of CIDOC-CRM model is a guaranty that, on the one hand, there are already many information available in the area of cultural patrimony that can be used to integrate and linked with, and on the other, there are also many platforms available that can be used to explore the information migrated.

Another important issue when representing information in an ontology is to take into account the need of interpreting natural language text, to automatically obtain its ontology representation. Like in this subject domain, archives information, free text appears in a variety of metadata fields of other domains. Interpreting natural language text can condition the representations in the ontology as presented in regard to this work.

Some examples were presented about the migration of the metadata information within text fields, but currently this task is under development in order to achieve the automatic migration of events, persons, institutions, places, etc.

Regarding the migration process evaluation there are two sub processes, the set of mapping description rules presented in Table 1 and the set of rules from Table 5. For the first one, the result migration either is correct or not, if the information retrieved from OWL2 representation is successfully matched with the initial records, then it is correct. Otherwise, it is necessary to identify the problems in mapping representation and then they should be fixed. This evaluation can be made automatically, but an application interface as the one presented is helpful to debug the problems that can occur. For the second set of rules, the evaluation is more complex and requires human intervention to decide if the information extracted from the text fields is well represented and relevant. This evaluation is not done yet and, at this moment, the interface application only retrieve information represented for each unit, obtained in the first step of the migration process. This task is set as future work.

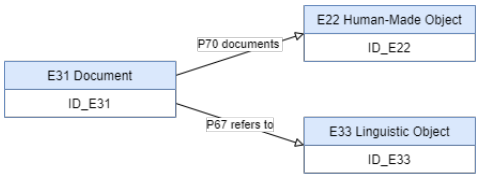
## Acknowledgement

This work is financed by National Funds through the Portuguese funding agency, FCT (Fundação para a Ciência e a Tecnologia) within I&D Projects with identification EPISA DSAIPA/DS/0023/2018. and NOVA LINCS (UIDB/04516/2020)

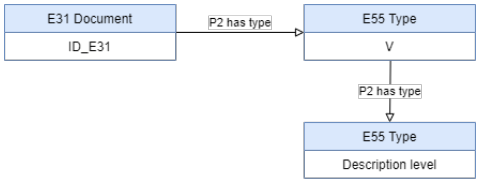
## Appendix A. Mapping Description Rules

The current Appendix is used to display the Mapping Description Rules, presented in Table 1 and in Table 5, in a diagram format for better understanding.

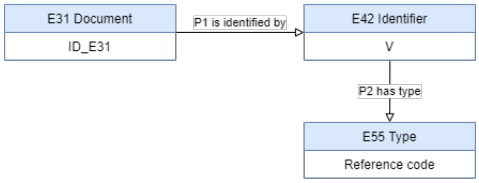
**Rule No. 1 -** Unit of Description



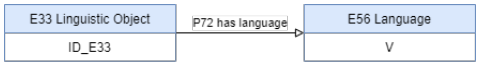

**Rule No. 2 -** Description level



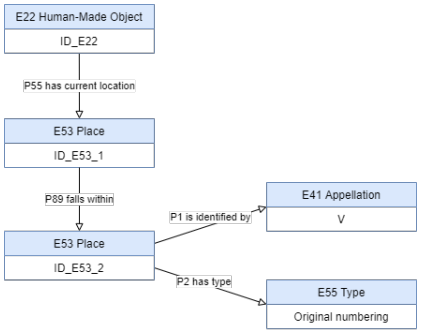

**Rule No. 3 -** Reference code



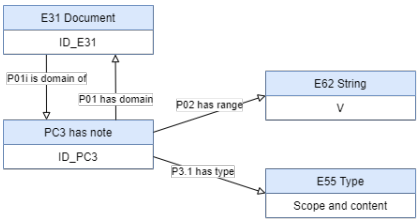

**Rule No. 4 -** Language of material



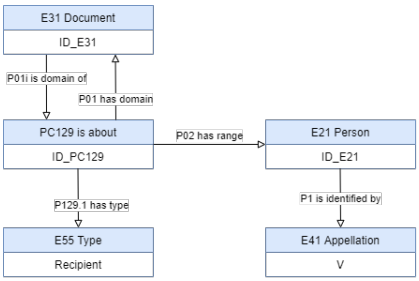

**Rule No. 5 -** Original numbering



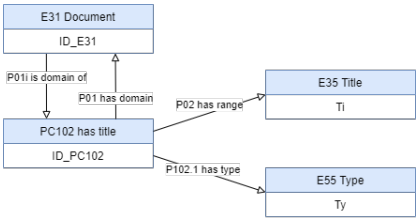

**Rule No. 6 -** Scope and content



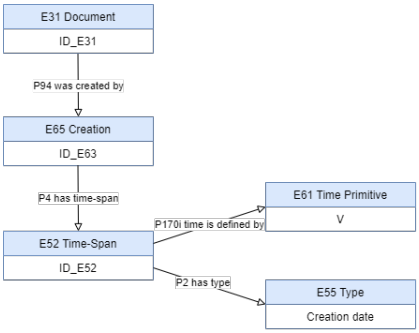

**Rule No. 7 -** Recipient



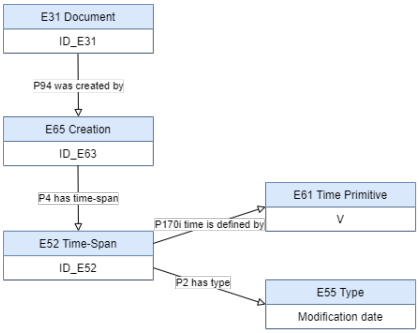

**Rule No. 8 -** Title and type



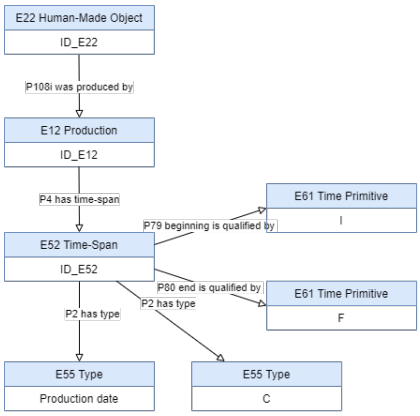

**Rule No. 9 -** Creation date



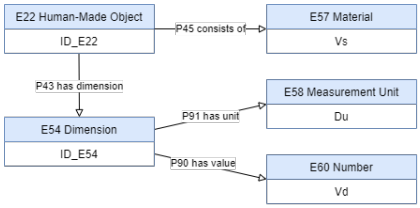

**Rule No. 10 -** Modification date



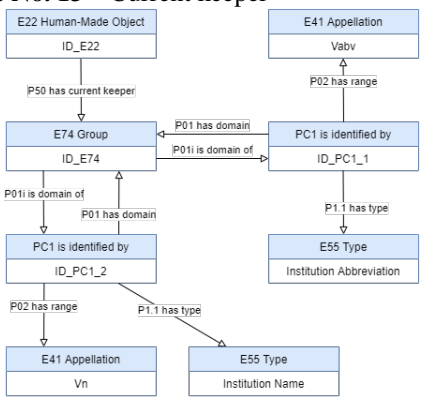
**Rule No. 11 -** Date range



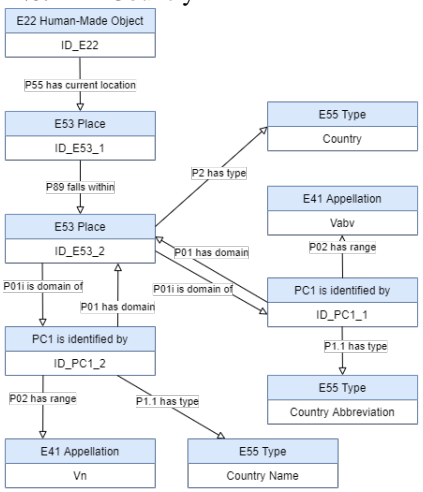

**Rule No. 12 -** Dimension and support



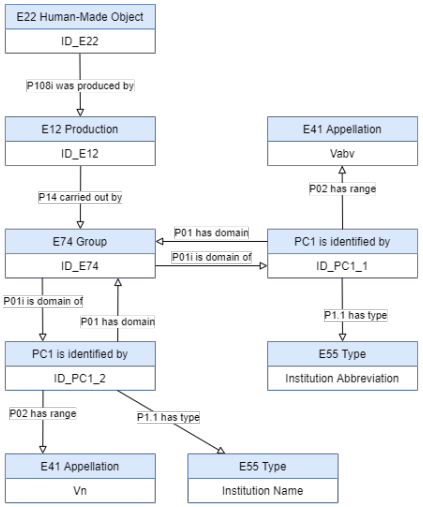

**Rule No. 13 -** Current keeper



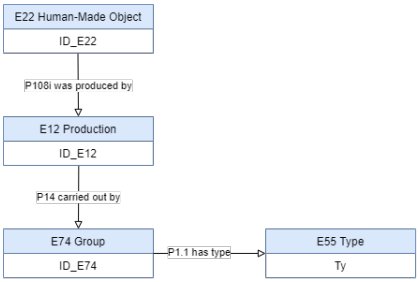

**Rule No. 14 -** Country



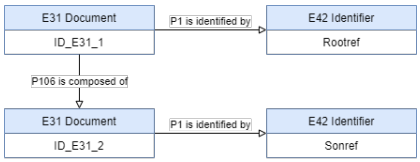

**Rule No. 15 -** Producer



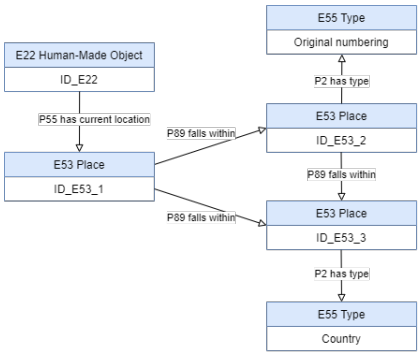

**Rule No. 16 -** Producer Type



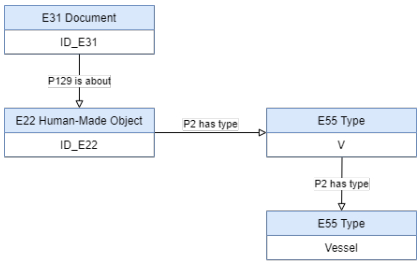

**Rule No. 17 -** Hierarchy



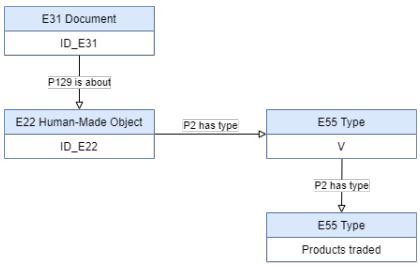

**Rule No. 18 -** Other Rules



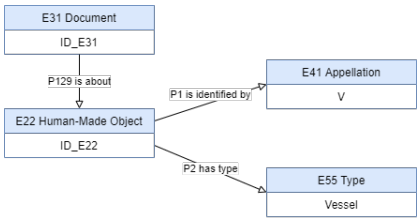

**Rule No. 19 -** Vessel



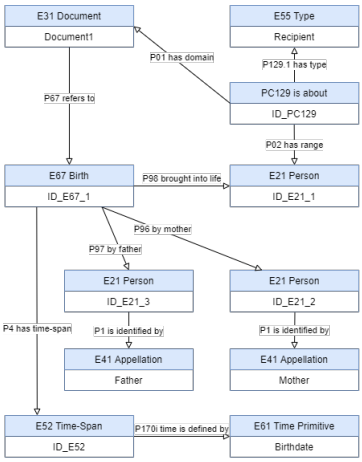

**Rule No. 20 -** Products traded



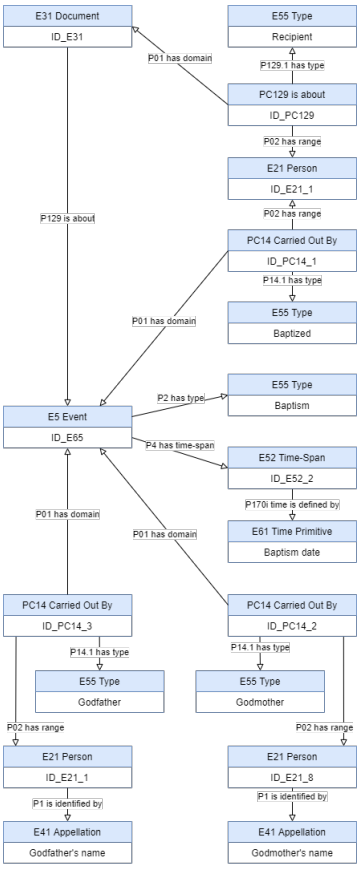

**Rule No. 21 -** Vessel name



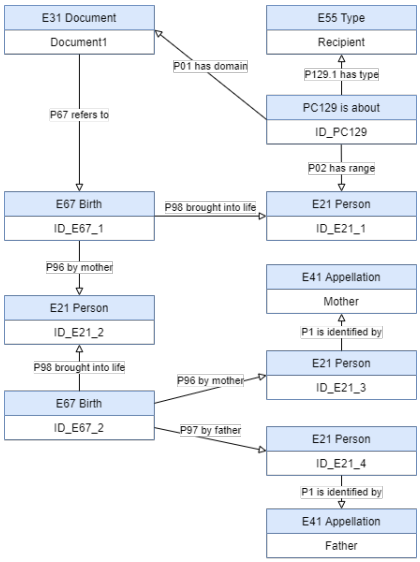

**Rule No. 22 -** Birth Event of the Recipient, of a Baptism unit



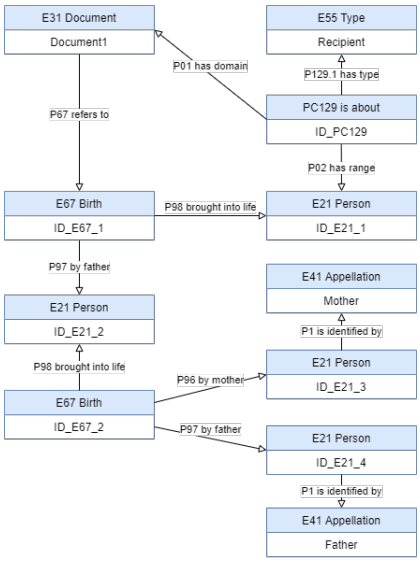

**Rule No. 23 -** Godparents of the Recipient, of a Baptism unit




**Rule No. 24 -** Grandparents from mother's side of the Recipient, of a Baptism unit




**Rule No. 25 -** Grandparents from father's side of the Recipient, of a Baptism unit




## References


[1] J.C. Ramalho and J.C. Ferreira, 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*, 2004.

[2] I.C. on Archives, *ISAD(G): general international standard archival description, Second Edition*, Springer Nature BV, 2011. ISBN 0-9696035-5-X.

[3] S. Vitali, 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) (2004), 185–199.

[4] C.b.t.C.S.I.G. ICOM/CIDOC Documentation Standards Group, *Definition of the CIDOC Conceptual Reference Model*, 7.0.1 edn, ICOM/CRM Special Interest Group, 2020.

[5] C. Meghini and M. Doerr, A first-order logic expression of the CIDOC conceptual reference model, *International Journal of Metadata, Semantics and Ontologies* **13**(2) (2018), 131–149.

[6] C.b.t.C.S.I.G. ICOM/CIDOC Documentation Standards Group, *Definition of the CIDOC Conceptual Reference Model*, 7.0.1 edn, ICOM, 2020.

[7] G.d.T.d.N.d.D. Diração-Geral de Arquivos, Orientações para a Descrição Arquivística, 3rd edn, 2011.

[8] L. Bountouri and M. Gergatsoulis, The Semantic Mapping of Archival Metadata to the CIDOC CRM Ontology, *Journal of Archival Organization* **9**(3–4) (2011), 174–207. doi:10.1080/15332748.2011.650124.

[9] S. Hennicke, Representation of Archival User Needs using CIDOC CRM, in: *Conference Proceedings TPDL: International Conference on Theory and Practice of Digital Libraries, Selected Workshops*, Valletta, Malta, 2013.

[10] I. Lourdi, C. Papatheodorou and M. Doerr, Semantic Integration of Collection Description: Combining CIDOC-CRM and Dublin Core Collections Application Profile, *D-lib Magazine - DLIB* **15** (2009). doi:10.1045/july2009-papatheodorou.

[11] M. Theodoridou and M. Doerr, Mapping of the Encoded Archival Description DTD Element Set to the CIDOC CRM, *Technical Report FORTHICS/TR-289. FORTH* (2001).

[12] T.T.S. for Encoded Archival Description of the Society of American Archivists, Encoded Archival Description Tag Library, Version 2002, 2002. https://www.loc.gov/ead/tglib/appendix_a.html#foot4.

[13] M. Doerr, The CIDOC Conceptual Reference Module: An Ontological Approach to Semantic Interoperability of Metadata, *AI Magazine* **24**(3) (2003), 75. doi:10.1609/aimag.v24i3.1720. https://www.aaai.org/ojs/index.php/aimagazine/article/view/1720.

[14] J. Makki, OntoPRiMa: A prototype for automating ontology population, *International Journal of Web/Semantic Technology (IJWesT)* **8** (2017).

[15] M.P. di Buono, M. Monteleone and A. Elia, How to Populate Ontologies, in: *Natural Language Processing and Information Systems*, E. Métais, M. Roche and M. Teisseire, eds, Springer International Publishing, Cham, 2014, pp. 55–58.

[16] J. Makki, A.-M. Alquier and V. Prince, 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, Association for Computing Machinery, New York, NY, USA, 2008, pp. 350–355–. ISBN 9781605580463. doi:10.1145/1456223.1456296.

[17] D. Maynard, Y. Li and W. Peters, NLP Techniques for Term Extraction and Ontology Population., 2008.

[18] O. Marlet, T. Francart, B. Markhoff and X. Rodier, OpenArchaeo for Usable Semantic Interoperability, in: *ODOCH 2019 @CAiSE 2019*, Rome, Italy, 2019. https://hal.archives-ouvertes.fr/hal-02389929.

[19] D. Myers, M.S. Quintero, A. Dalgity and I. Avramides, The Arches heritage inventory and management system: a platform for the heritage field, *Journal of Cultural Heritage Management and Sustainable Development* (2016).

[20] F. Beretta and V. Alamercery, Workflow for communal ontology management: aligning data models with OntoME, in: *APOLLONIS Workshop «Historical content metadata»*, 2019.

[21] F. Beretta, A challenge for historical research: making data FAIR using a collaborative ontology management environment (OntoME), *Semantic Web* (2020), 1–16.

[22] D. Metilli, V. Bartalesi and C. Meghini, A Wikidata-based tool for building and visualising narratives, *International Journal on Digital Libraries* **20** (2019), 417–432–. doi:10.1007/s00799-019-00266-3.

[23] M.D. Maria Theodoridou George Bruseker and M. Doerr, Methodological tips for mappings to CIDOC CRM, in: *Proceedings of CAA 2016 - 44th Computer Applications and Quantitative Methods in Archaeology Conference*, CAA 2016 OSLO Exploring oceans of data, 29 March—2 April, 2016.

[24] D. Melo, I.P. Rodrigues and V.B. Nogueira, Using a Dialogue Manager to Improve Semantic Web Search, *International Journal on Semantic Web and Information Systems (IJSWIS)* **12**(1) (2016). doi:10.4018/IJSWIS.2016010104.

[25] D. Melo, I.P. Rodrigues and V.B. Nogueira, Semantic Web Search Through Natural Language Dialogues, in: *In Innovations, Developments, and Applications of Semantic Web and Information Systems*, D.L. Miltiadis, N. Aljohani, E. Damiani and K.T. Chui, eds, IGI Global, Hershey, 2018, pp. 329–349. doi:10.4018/978-1-5225-5042-6.ch012.

[26] P. Fragkou, N. Kritikos and E. Galiotou, Querying Greek Governmental Site Using SPARQL, in: *Proceedings of the 20th Pan-Hellenic Conference on Informatics*, PCI '16, Association for Computing Machinery, New York, NY, USA, 2016. ISBN 9781450347891. doi:10.1145/3003733.3003807.

[27] M.C. Cavalcanti, F.D. Pereira, E. Fusco and M.L. Mucheroni, Model of data extraction in the innovation environments of the state of São Paulo based on semantic technologies, in: *CONTECSI - 14th International Conference on Information Systems & Technology Management*, CONTECSI USP, São Paulo, Brazil, 2017. ISSN 2448-1041. doi:10.5748/9788599693131-14CONTECSI/PS-4762.

[28] D. Melo, I.P. Rodrigues and D. Varagnolo, Semantic Migration of Fonds Examples, Mendeley Data, V1, 2020. doi:10.17632/knx84z3463.1.

[29] M. Horridge and S. Bechhofer, The OWL API: A Java API for OWL Ontologies, *Semant. Web* **2**(1) (2011), 11–21–.

[30] E. Sirin and B. Parsia, SPARQL-DL: SPARQL Query for OWL-DL, in: *3rd Workshop on OWL: Experiences and Directions*, Vol. 258 of CEUR Workshop Proceedings, 2007. CEUR-WS.org.

[31] P. Kremen and E. Sirin, SPARQL-DL implementation experience, in: *Proceedings of the 4th OWLED Workshop on OWL: Experiences and Directions Washington*, Vol. 496 of CEUR Workshop Proceedings, 2007. CEUR-WS.org.

[32] D. Melo, I.P. Rodrigues and I. Koch, 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, Knowledge Engineering and Knowledge Management - Volume 2: KEOD*, SciTePress, 2020, pp. 197–204, INSTICC. ISBN 978-989-758-474-9. doi:10.5220/0010134101970204.

[33] D. Melo, I.P. Rodrigues and D. Varagnolo, Installation Unit - REGISTOS DE BAPTISMOS, Mendeley Data, V2, 2020. doi:10.17632/wx7v7rmg7h.2.