A Reduced CRM-Compatible Form Ontology for the virtual Emigration Museum

Ricardo G. Martini[1], Cristiana Araújo[1], Giovani R. Librelotto[2], and Pedro R. Henriques[1]

[1] Algoritmi Research Centre, Department of Informatics
University of Minho - Gualtar - 4710-057, Braga, Portugal
[2] Programa de Pós-Graduação em Informática - Universidade Federal de Santa Maria - 97105-900 - Santa Maria, RS, Brazil
rgm@algoritmi.uminho.pt,decristianaaraujo@hotmail.com,librelotto@inf.ufsm.br,prh@di.uminho.pt

Abstract. In this paper we discuss the construction of a Reduced CRM-compatible form ontology for the virtual Emigration Museum based in the international standard for museum ontologies, CIDOC-CRM. To extract knowledge from the information of the virtual Emigration Museum when navigating through it, abstract data models should be used to conceptualize, the emigration documents stored in a relational database. In that way, resorting to an ontology (as abstract layer), the information contained in those documents can be accessed by the end-users (the museum visitors) to learn about the emigration phenomena. We also describe how we instantiate the ontology through a parser that automatically translates a plain text description of emigration data into RDF. Finally, we also discuss the choice of a triple storage system to save the RDF triples in order to enable the use of SPARQL to query the RDF data.

Keywords: Emigration Museum, Emigration Documents, Ontology, RDF, Triplestore, CIDOC-CRM

1 Introduction

The institutions holding cultural heritage – like libraries, museums, archives – are responsible for the technical treatment, preservation and dissemination of the documentary collections [1]. This kind of repositories is fundamental to comprehend History, contributing not just for the patrimonial wealth of a country but also for the council's history.

In the context of this work, we are particularly interested in bringing together the documentation about the emigration movement with the aim of creating a virtual Emigration Museum.

We discuss along the paper the digital preservation and exploitation of emigration documents via a web interface using an ontology.

In general it is known that libraries, archives, and museums hold many documents in paper. However, to consult and learn about the documents information

in paper becomes a hard task (since often the documents are old and handwritten). Moreover, the excessive paper handling causes a rapid degradation [1].

Those facts rise up the need to preserve all documents in digital repositories. Avoiding excessive paper handling and making its data available on the web, it is possible to consult, relate, and understand the information in an easy and interesting way. But to allow the data available on the web and extract some information of it, it is necessary to translate the data to a machine understandable format.

To handling this data, Semantic Web[3] technologies like Resource Description Framework (RDF) enable the creation of datasets on the web, so applications can consume the data and present them in an human-readable format [2].

The combination of different documental fonds related to the emigration movement like biographies, almanacs, passport application forms, passports, ship information, can bring a significant value to the existent data, and this bring to us the necessity of represent data and information (all kind of emigration documents) in RDF, because it facilitates, even in different schemas, the data merging [3].

The work here reported is a crucial part of a doctoral project that focus on the goal of automating the creation of web-based virtual Learning Spaces using ontologies and Domain Specific Languages to specify the virtual environment.

The main goal of the work reported in this paper is the creation, automatic instantiation through the translation of plain text to RDF notation, and exploitation of an ontology for the emigration phenomena in Portugal. We adopt the international standard for museum ontologies called International Committee for Documentation - Conceptual Reference Model (CIDOC-CRM[4]). Having the ontological view of the fond described in RDF, we consult the information using SPARQL Protocol and RDF Query Language (SPARQL).

In Section 2 is presented the Emigration Museum and its documental assets. In Section 3, CIDOC-CRM standard for ontology description is introduced and a Reduced CRM-compatible form ontology, to describe the emigration domain, is discussed. Section 4 discusses data representation alternatives in order to simplify the storage and query of the ontological triples (subject, property, and object). The development of a parser to convert the triples from a plain text description into RDF notation is described in Section 5. Finally, in Section 6 the paper is summarized, conclusions are drawn, and directions for future works are proposed.

2 Emigration Museum and its Documental Fond

International Council of Museums (ICOM[5]) defines virtual museums as "a logically related collection of digital objects composed in a variety of media which, because of its capacity to provide connectedness and various points of access,

[3] http://www.w3.org/standards/semanticweb/
[4] http://www.cidoc-crm.org/
[5] http://icom.museum/

lends itself to transcending traditional methods of communicating and interacting with visitors; it has no real place or space, its objects and the related information can be disseminated all over the world" [4].

The Emigration Museum is not an exception. It has cultural information extracted from a collection of documents inherited from a municipal archive to be exhibited to the public. Besides the backoffice archive, this museum has some thematic exhibition rooms [5].

In this paper we focus on the archive, because it holds the relevant documents to study and understand the emigration phenomena. The structure and content of the documents considered in this work were detailed in the paper [6].

3 CIDOC-CRM Ontology

The objective of CIDOC-CRM is to promote a shared understanding of the Cultural Heritage domain by providing a common and extensible semantic framework that any Cultural Heritage information can be mapped to. In this way, it can provide the semantic glue needed to mediate between different sources of information, such as that published by museums, libraries and archives [7][8].

To understand how CIDOC-CRM is organized, Figure 1 presents its core, showing the main entities and relations.

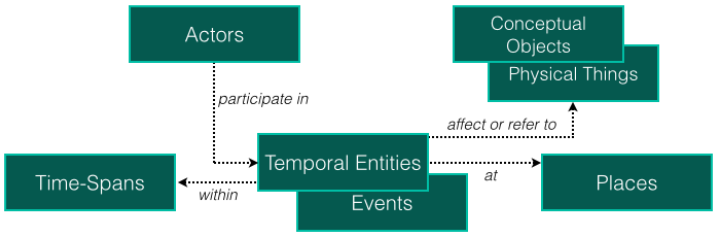

Fig. 1. Core Structure of CIDOC-CRM (adaptated from [8])

CIDOC-CRM is an event-based ontology where the main entities are related to *Temporal Entities* as depicted in Figure 1. As their name implies, *Temporal Entities* are concepts related to events in the past and because of this, they are related to a temporal length of events (period), so they can have date and time associated to the *Time-Spans* entity. The *Actors*, *Conceptual Objects*, *Physical Thing* and *Places* classes can not be directly linked to time (*Time-Spans*), so they need to be associated to events (*Temporal Entities*).

A *Place* can be anything that describes a location (geographical or e.g., in the bank of the Douro River or on top of Eiffel Tower).

Actors are entities that hold a legal liability. An actor can be an individual or a group; the first one is related to a person and the second one can be associated

to a company, for example. Actors interact with things (*Conceptual Objects* and Physical Things) through events.

A *Physical Thing* is something that can be physically destroyed and, case some part is preserved, it can be turned into something new. By other hand, *Conceptual Objects* can not be crashed. For instance, a physical thing like a smartphone, or a magazine can be destroyed, but the information (content) related to that physical thing can not. To destroy a *Conceptual Object* it is necessary to extinguish the source, i.e., anything that represents that concept, including people.

Things in CIDOC-CRM can have *Appellations*. They can be a name, an identification number, etc. Furthermore, different organizations have distinct classification types. In CIDOC-CRM, these classifications are called *Types* and they classify things. For instance, events can have diverse types like birth, marriage, race, earthquake, flood, war, etc. Both *Appellations* and *Types* can be related to any entity.

Besides, the CIDOC-CRM ontology has name conventions that should be followed. Any concept starts its name with the capital letter "E" (of Entity) followed by a numerical code (e.g. E39 Actor, E53 Place, etc.). The relations are no different, they start their names with the capital letter "P" (of Property) followed by a numerical code (e.g. P89 falls within, P131 is identified by, etc.).

Section 3.1 details how the emigration documents – held by Municipal Archive of Fafe – were described in CIDOC-CRM.

3.1 Onto ME, an Ontology for the Emigration Museum

After a CIDOC-CRM in-depth analysis, it was possible to correlate the compatible entities of the ontology with the emigration documental fond.

Thus, a compatible CIDOC-CRM based ontology was instantiated and reused. When an ontology is in accordance with certain rules imposed by the standard ontology, it is called *Reduced CRM-Compatible Form* [7].

To demonstrate how the emigration documents, that belong to the Municipal Archive of Fafe, fit in CIDOC-CRM, we show in Figure 2 an example of the previous ontology fragment instantiated with the information collected about the emigration movement of a person.

As can be seen in Figure 2, the main event is *E9 Move*, which refers to the emigration document that reflects a passport application form identified by the number *161*. *E9 Move* has four relations describing:

when the movement has occurred: described by *E52 Time-Span* named 'TS1', which in this case *(P78) is identified by* '1963-05-21', an *E50 Date*;

where the emigrant moved to: described by *E53 Place* named 'PL1', which in this case *(P87) is identified by* 'França', an *E44 Place Appellation*;

who emigrated: described by *E21 Person* named '2828624', an *E21 Person*, which in this case *(P131) is identified by* an *E82 Actor Appellation* 'José Carlos Magalhães'. *E21 Person* has a type to identify its role in *E9 Move*. So person '2828624' *(P2) has type* 'Emigrant';

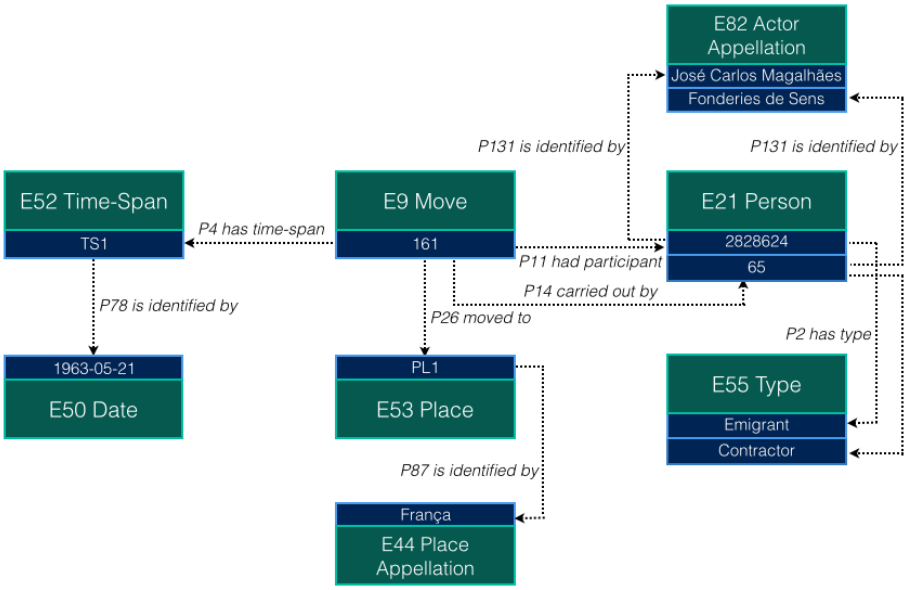

Fig. 2. Reduced CRM-Compatible Form instantiation example

who carried out [6]**:** described by *E21 Person* named '65', an *E21 Person*, which in this case *(P131) is identified by* an *E82 Actor Appellation* 'Fonderies de Sens'. *E21 Person* has a type to identify its role in *E9 Move*. So person '65' *(P2) has type* 'Contractor'. Notice that it is not possible to determine, from the sources, whether the contractor is a person or a company (*E74 Group*). So, it is always described as an *E21 Person*.

After describing the emigration documents using the CIDOC-CRM compatible form ontology, the need to represent the data in a machine-understandable format rose up. Section 4 presents the data representation formats used.

4 Data Representation

To define and use an ontology, an explicit representation should be adopted. There are several representation languages that can be used for that purpose, like eXtensible Markup Language (XML), RDF, Web Ontology Language (OWL), among others. They vary in expressiveness [9].

CIDOC-CRM ontology can be described in such languages, but usually RDF is the one chosen by the museum community. The creators of CIDOC-CRM have chosen RDF aiming at an easy understanding by both computer experts

[6] Notice that exist other objects related to this same property, with the difference in the *E55 Type*. In this project there are types like: who intermediates the emigration movement (the intermedary); who is calling the emigrant (the caller); etc.

and non-experts [7]. So in this work RDF is used to describe the knowledge present in the emigration documents.

As aforementioned in Section 3.1, it was crucial to verify if the ontology created adequately represents our documental fond. In a first step, presented in Section 4.1, to specify the triples representing the emigration assets, a description in plain text was created.

The flow depicted in Figure 3 shows the way to query in SPARQL the triples database starting from an ontology *described in* a specific notation (RDF).

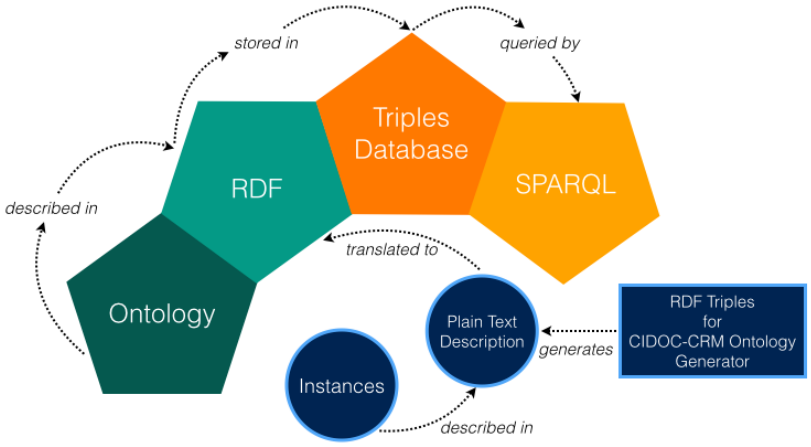

Fig. 3. Data Representation's schema

According to the flow of Figure 3, the ontology triples must be translated from the textual representation to RDF. This task was carried manually in a second step, as detailed in Section 4.2. This RDF description was stored in a triple database (Apache Jena TDB) to be queried through a SPARQL end-point (Apache Jena Fuseki).

Finally, and after understanding this manual process, it was automated by a compiler that translates plain text triples to RDF. The compiler was generated by ANother Tool for Language Recognition (ANTLR[7]) version 4; the process of translating the input (text triples) into RDF notation can be seen in Section 5.

Next sections describe the textual and RDF specifications used before storing the triples in a database appropriate to be queried by SPARQL.

4.1 Plain Text Triples

As mentioned in Section 4, to understand the structure and content of the emigration documents, we created a plain text description that follows the triples

[7] `http://www.antlr.org/`

concept (subject, predicate, object). An example of this plain text description can be seen in Listing 4-1.

Listing 4-1. Triples specified in the plain text description

```
1  161:E9 Move
2  P4 has time-span:TS1
3  P26 moved to:PL1
4  P11 had participant:2828624
5  P14 carried out by:65 .
6
7  TS1:E52 Time-Span
8  P78 is identified by:1963-05-21 .
9  . . .
```

This example describes exactly the same piece of knownledge previously shown in Figure 2; '161' is an *E9 Move* concept that has some relations described between lines 2 and 5. *E52 Time-Span* has relations specified starting at line 8. Notice that the "..." sign at line 9 indicates that exist other specifications (not listed) to describe the entire example of Figure 2.

This representation step is illustrated in Figure 3 by the node *instances* that are *described in* the *plain text description*. The plain text specification can be created in two ways: (1) simply using a text editor; or (2) using the *RDF Triples for CIDOC-CRM Ontology Generator* web application – which was developed by us and aids in the specification of the text triples (also depicted in Figure 3).

Having the plain text triples specified and the structure of the documents understood, it is necessary to describe them in a machine-readable format, so SPARQL can handle with it.

4.2 RDF Triples

Resources in RDF are identified by Uniform Resource Identifiers (URIs) and described with properties and property values, where: (1) Resources are subjects in RDF with a URI; (2) a property (a.k.a predicate) is used to describe the relations (e.g. 'is identified by', 'moved to', etc.) between the subject and the property values; and (3) a property value (a.k.a object) is an object that can be another subject or a literal. The aggregation of a Resource, a Property, and a Property value is known as a triple (subject, predicate, object), as already mentioned.

Thus, from the plain text representation, we can derive RDF triples to create our domain ontology. Listing 4-2 shows an excerpt of the RDF specification manually created to describe the example of Figure 2.

Listing 4-2. Triples specified in the RDF notation

```
1  <rdf:RDF . . .
2  <rdf:Description rdf:about="161">
3          <rdf:type rdf:resource="E9_Move"/>
```

```
4           <P4_has_time-span rdf:resource="TS1"/>
5           <P26_moved_to rdf:resource="PL1"/>
6           <P11_had_participant rdf:resource="2828624"/>
7           <P14_carried_out_by rdf:resource="65"/>
8   </rdf:Description>
9
10  <rdf:Description rdf:about="TS1">
11          <rdf:type rdf:resource="E52_Time-Span"/>
12          <P78_is_identified_by rdf:resource
                ="1963-05-21"/>
13  </rdf:Description>
14  . . .
15  </rdf:RDF>
```

Notice that the RDF file is describing exactly the same triples created in the plain text description (Listing 4-1).

5 Plain Text Description to RDF CIDOC-CRM (TXT2CIDOC)

Once understood how triples (both plain text and RDF) are specified, and the difficulty of manually describe the domain as a triple dataset in RDF, we feel the need to automate this process. With that in mind, we build a compiler TXT2CIDOC.

The first step in this case was to create a grammar in ANTLR that recognizes the language used by us to write the plain text description. A snippet of the grammar can be seen in Listing 5-3.

Listing 5-3. txt2rdfcidoc Grammar

```
1  grammar txt2rdfcidoc;
2
3  txt2rdfcidoc : (objectConcept NEWLINE (relationObject
       NEWLINE)* relationObject endStat)+ ;
4  objectConcept: object ':' concept ;
5  relationObject: relation ':' object ;
6  object       : OBJECT ;
7  concept      : 'E21 Person' | 'E53 Place' | 'E9 Move'... ;
8  relation     : 'P131 is identified by' | 'P2 has type'
9               | 'P26 moved to' | 'P4 has time-span'... ;
10 endStat      : '.' ;
```

For the sake of space, notice that at lines 7 and 9 there is a "..." sign indicating that exist other alternatives to identify a concept or a relation in CIDOC-CRM.

So, the main contribute here is the translation of the plain text to RDF notation. This translation process is represented by a *translated to* relation displayed

in Figure 3, that is used to illustrate that the compiler gets as input the textual description to be recognized according to the grammar and generates RDF triples.

This process is made by listening events thrown from a Java parse-tree walker. As shown in Listing 4-1, it is necessary to override some methods automatically created by ANTLR and used to visit each production of the grammar. These methods are called listeners and they work when the walker enters and exits a parse-tree node. Listing 5-4 shows the code of a listener to illustrate the approach (we use again the same running example).

Listing 5-4. A listener fired on entry in the node for the `objectConcept` production

```
1  @Override public void enterObjectConcept(
       txt2rdfcidocParser.ObjectConceptContext ctx) {
2          String object = ctx.object().getText();
3          object = object.replace(" ", "_");
4          instances += "\n<rdf:Description rdf:about=\""
               + object + "\">";
5          String concept = ctx.concept().getText();
6          concept = concept.replace(" ", "_");
7          instances += "\n\t<rdf:type rdf:resource=\""+
               concept +"\"/>";
8  }
```

The entry method of Listing 5-4 gets the text associated with the symbols `object` and `concept` and replaces any white space by the underline "_" character. After that, it concatenates the `object` and `concept` texts to the `instance` String, which should contain, after visiting all the tree nodes corresponding to the recognized grammar rules, the entire RDF description to be written in the final RDF file. This file is created when the Java walker visits the node for `txt2rdfcidoc` production by the last time (this is, on exit).

So, when the `txt2rdfcidoc` exit listener executes, the RDF file (as shown in Listing 4-2) is created.

After having the RDF file created, it should be stored and available in a triple database to be queried by SPARQL. In this work, Apache Jena Framework[8] was used to implement these tasks.

6 Conclusion

This paper presented a task that corresponds to one of the working phases of a bigger project that focus on the idea of creating virtual Learning Spaces to impart knowledge of cultural heritage information. The project here discussed aims at describing a documental fond of the Municipal Archive of Fafe (composed of emigration documents) in a machine-readable format in order to be possible

[8] Triple store (TDB): `https://jena.apache.org/documentation/tdb/` and SPARQL server (Fuseki end-point): `https://jena.apache.org/documentation/fuseki2/`

to extract the information about the emigration phenomena to build a virtual Emigration Museum. The basilar layer of the doctoral project – development of the digital repository – was already published in [6].

To achieve this, it was necessary to create an ontology based in CIDOC Conceptual Reference Model and automatically instantiate it through the translation of plain text to RDF. This compiler based translation was an important task, because manually creating the RDF triples is a time consuming and error-prone activity taking into account the RDF syntax.

As future work, this project can be extended to bring together the information not only about the passport application form, but another sources like biographies, letters, ships' route, etc. Thus, the final virtual Emigration Museum can be enriched with more knowledge about the emigration phenomena.

More information about this work can be seen in the TXT2CIDOC website at `http://www4.di.uminho.pt/~gepl/txt2cidoc/`.

Acknowledgements: This work has been supported by FCT – Fundação para a Ciência e Tecnologia within the Project Scope: UID/CEC/00319/2013. The work of Ricardo G. Martini is supported by CNPq, grant 201772/2014-0.

References

1. Young, S.F., of Illinois at Chicago. Library. Special Collections Department, U.: Don't Throw it Away!: Documenting and Preserving Organizational History. Special Collections Department, University Library, University of Illinois at Chicago (1995)
2. Shadbolt, N., Berners-Lee, T., Hall, W.: The semantic web revisited. IEEE Intelligent Systems **21** (2006) 96–101
3. : RDF 1.1 primer. Technical report, World Wide Web Consortium (2014)
4. Schweibenz, W.: The development of virtual museums. In: Virtual Museums. Volume 57(3). ICOM (2004)
5. Monteiro, M.: O museu da emigração e os "brasileiros" do rio: o público e o privado na construção de modernidade em portugal. Revista da Faculdade de Letras História **8** (2007)
6. Martini, R., Guimarães, M., Librelotto, G., Henriques, P.: Storing archival emigration documents to create virtual exhibition rooms. In Rocha, A., Correia, A.M., Costanzo, S., Reis, L.P., eds.: New Contributions in Information Systems and Technologies. Volume 353 of Advances in Intelligent Systems and Computing. Springer International Publishing (2015) 403–409
7. ICOM/CIDOC: Definition of the cidoc conceptual reference model. Technical report, ICOM/CIDOC (2013)
8. Oldman, D., Labs, C.: The CIDOC Conceptual Reference Model (CIDOC-CRM): PRIMER. International Council of Museums (ICOM) **1** (2014)
9. McGuinness, D.L., Harmelen, F.V.: Owl web ontology language overview (2004)