Koch, I., Pires, C., Teixeira Lopes, C., Ribeiro, C., Nunes, S. (2023). From ISAD(G) to Linked Data Archival Descriptions. In: Alonso, O., Cousijn, H., Silvello, G., Marrero, M., Teixeira Lopes, C., Marchesin, S. (eds) Linking Theory and Practice of Digital Libraries. TPDL 2023. Lecture Notes in Computer Science, vol 14241. Springer, Cham. https://doi.org/10.1007/978-3-031-43849-3_27

From ISAD(G) to Linked Data Archival Descriptions

Inês Koch(✉), Catarina Pires(✉), Carla Teixeira Lopes(✉), Cristina Ribeiro(✉), and Sérgio Nunes(✉)

INESC TEC and Faculty of Engineering, University of Porto, Porto, Portugal
{ines.koch,catarina.o.pires}@inesctec.pt, {ctl,mcr,ssn}@fe.up.pt

Abstract. Archives preserve materials that allow us to understand and interpret the past and think about the future. With the evolution of the information society, archives must take advantage of technological innovations and adapt to changes in the kind and volume of the information created. Semantic Web representations are appropriate for structuring archival data and linking them to external sources, allowing versatile access by multiple applications. ArchOnto is a new Linked Data Model based on CIDOC CRM to describe archival objects. ArchOnto combines specific aspects of archiving with the CIDOC CRM standard. In this work, we analyze the ArchOnto representation of a set of archival records from the Portuguese National Archives and compare it to their CIDOC CRM representation. As a result of ArchOnto's representation, we observe an increase in the number of classes used, from 20 in CIDOC CRM to 28 in ArchOnto, and in the number of properties, from 25 in CIDOC CRM to 28 in ArchOnto. This growth stems from the refinement of object types and their relationships, favouring the use of controlled vocabularies. ArchOnto provides higher readability for the information in archival records, keeping it in line with current standards.

Keywords: Archival Description · CIDOC CRM · ArchOnto

1 Introduction

Archives play a central role in understanding and interpreting the past. They are a resource from which we reflect and attempt to revisit what has already transpired [11]. The content of public archives is part of humanity's knowledge heritage for present and future generations. It is essential to safeguard and ensure the continued accessibility of archives [10]. As the information society moves forward, archives face new challenges, among which is the increase in the amount of information produced, specifically information from the digital world. Most documents today are created electronically [7].

The change in information access habits increased the need for digitally available archives. However, access is only one requirement when people explore an extensive collection, such as public archives. Archives should also follow the other FAIR Principles [15], which include findability, interoperability, and reusability.

O. Alonso et al. (Eds.): TPDL 2023, LNCS 14241, pp. 303–309, 2023.
https://doi.org/10.1007/978-3-031-43849-3_27

This work aims to analyze the representation of existing archival records in the Portuguese National Archives using ArchOnto. ArchOnto is a modular ontology developed within the scope of the EPISA Project that introduces a set of specific classes and properties.

To understand the impact of using a Linked Data Model to represent archival records, we compare the representations of a sample of documents in CIDOC CRM and ArchOnto. The impact is measured in terms of applicability in archives.

2 Background

The Portuguese National Archives curate a unique collection of historical and contemporary objects accumulated since the 9th century, distributed among the various institutions that compose the archives. The National Archives comprise two national archives and 16 regional archives at the district level. It curates over 3,5 million records described through a combination of the various standards for archival description developed by the International Council on Archives (ICA), namely the ISAD(G) – General International Standard Archival Description [3], and the ISAAR(CPF) – International Standard Archival Authority Record for Corporate Bodies, Persons and Families [4].

Among the assets held by this institution are a large number of Fonds, which are organized in groups, namely Central and Local Administration; Collections; Companies; Judicial; Monastics; Notaries; Parish; and Personal. Collections include records from previous political systems; records from contemporary, ecclesiastical, monastic, and conventual institutions; records of archives of individuals, families, associations, companies, commissions, and congresses; and records of photographic archives [8].

In the archival domain, efforts have been made to develop a data model to represent the archival assets. ICA is developing the RiC-CM (Records in Context Conceptual Model) and RiC-O (RiC-Ontology) to illustrate archival concepts, considering the main descriptive entities [5]. As this model was still preliminary when this work started, CIDOC CRM emerged as the model to use.

The CIDOC Conceptual Reference Model (CRM) is a formal ontology developed in the scope of museums by the International Committee for Documentation (CIDOC) of the International Council of Museums (ICOM). It intends to facilitate the integration, mediation and interchange of heterogeneous cultural heritage information and similar information from other domains [2]. It is under active development by the CIDOC CRM Special Interest Group. It is the only ontology in the Cultural Heritage domain accepted as an ISO standard (ISO 21127:2014) [6]. It has events as a central concept and provides a detailed description of people, places, and periods [1].

ArchOnto [8,9] is a modular ontology developed in the scope of the EPISA Project. Its classes and properties capture concepts that contribute to specific aspects of an archival organization. The ontology also specializes CIDOC CRM to include controlled vocabularies used in the archives. It comprises five ontologies that can be imported whenever needed — CIDOC CRM, N-ary, DataObject,

Link2DataObject, and ISAD Ontology. A prefix identifies each ontology according to Semantic Web best practices.

The CIDOC CRM is the base ontology of ArchOnto and provides the concepts and properties to capture archival records' essential features, e.g., event, date, location, person, and group. N-ary systematically represents non-binary associations, i.e., those that connect more than two individuals. This ontology is based on the CIDOC's early proposal for representing tuples with an arity higher than two. DataObject is an ontology created to handle literal values and their validation. The goal is that each individual with a representation as a simple type, such as a date or a string, is validated against the corresponding DataObject class. Link2DataObject connects DataObject to CIDOC CRM.

Finally, to ensure the integrity of information when migrating data from the legacy description to ArchOnto, the ISAD Ontology was created to represent the entire description expressed with the elements of the ISAD(G) standard. This allows the structured contents in the Linked Data to be validated against the information in the original ISAD(G) record.

3 Linked Data Representation

A sample of archival documents from the Portuguese National Archives was selected to understand and discuss the applicability of ArchOnto and CIDOC CRM in the archival context. The selected sample [13] contains 1,318 records, including Groups of Fonds that stand out among those existing in this National Archive. Among the records represented, 102 relate to the Decentralized Central Administration (*"Registo de Passaportes Deferidos"* about passports from 1914 to 1918), and 1,216 are related to parish records (*"Registos de Baptismo"* about baptisms from 1644 to 1911). This sample considers one series, 16 installation units, and 1,301 items.

The *Baptismo* and *Passaportes* datasets [13] were automatically represented in CIDOC CRM from their ISAD(G) description. The translation is based on rules that map the archival descriptive information to the CIDOC CRM representation semantically [12]. Additionally, the translated information was subject to some refinements to correspond to the same CIDOC CRM version used in ArchOnto. The migration from CIDOC CRM to ArchOnto also followed an automatic approach. The two data models were aligned based on the ontologies, identifying the differences and developing SPARQL Update queries for the required transformations. The result is a valid ArchOnto representation [14].

4 Results and Discussion

We could identify and quantify the classes and properties used in each representation of the selected archival records in CIDOC CRM and ArchOnto. Table 1 show an excerpt[1] of the results obtained when representing all the documents

[1] Due to space limitations, we could only place the Top-5 from 2 of the 4 tables. The remaining tables are available in the *Statistics* folder of the dataset in [14].

that are part of the series *Registos de Baptismo* and *Registo de Passaportes Deferidos* in CIDOC CRM and ArchOnto. For each class, we include the number of individuals and, for each property, the number of assertions that use it.

Table 1. Top 5 Classes and Properties in *Registos de Baptismo*.

Ontology	Class	CIDOC	ArchOnto
CIDOC	E52 Time-Span	4,927	4,927
CIDOC	E21 Person	3,782	3,782
CIDOC	E67 Birth	3,367	3,367
CIDOC	E41 Appellation	1,203	2,801
CIDOC	E53 Place	2,539	2,539

Ontology	Property	CIDOC	ArchOnto
CIDOC	P1 is identified by	12,404	16,134
CIDOC	P2 has type	11,035	10,878
Link2DataObject	L2DO has value	0	5,236
CIDOC	P4 has time-span	3,730	3,730
CIDOC	P98 brought into life	3,367	3,367

By CIDOC we mean CIDOC CRM. Full Tables are available in the *Statistics* folder of the dataset in [14].

An automatic and systematic method was used to obtain the results presented in the tables. The results were obtained using a 2-step script. The first step consists of importing the baptism dataset using Apache Jena[2], namely its RDF API, to load the dataset file into a model. In the second step, we count the occurrences of classes and properties and export them to a file in tabular form.

The *Registo de Baptismo* and *Registo de Passaportes Deferidos* representations offer similar results, and the same conclusions can be drawn. As the *Passaportes* dataset is smaller when compared to the *Baptismo* dataset, the results discussed are based on the latter (see Table 1).

Taking into account their specificity, we verified that the ArchOnto representation uses, in total, eight more classes and three more properties than the representation in CIDOC CRM. From a more general perspective, regarding the number of statements in each representation, CIDOC CRM has 151,950 statements, whereas ArchOnto has 160,611 statements, an increase of 6%. Observing the representation of records in the sample, it was possible to see an increase in the number of classes (+16%) and properties (+13%) used in ArchOnto, relative to CIDOC CRM.

Considering the data in the tables, we can see that several classes were used the same number of times. This is the case with generic classes such as *E52 Time-Span*, *E21 Person*, and *E67 Birth*. However, when archival records are represented using ArchOnto, there is a 40% increase in the number of classes used: from 20 classes in CIDOC CRM to 28 in ArchOnto.

Due to its more specific nature, ArchOnto allows a more detailed categorization of concepts related to the archival domain. These concepts consider the existing controlled vocabularies in this area of cultural heritage. Thus, the use of the *E55 Type* class decreased by 21%, going from 24 in the CIDOC CRM representation to 19 in ArchOnto. This results from using eight classes in ArchOnto that allow types specific to archival concepts rather than the more generic CIDOC CRM classes. This is the case with *ARE1 Level of Description* (for the archival hierarchical structure), *ARE2 Formal Title* and *ARE3 Supplied Title* (for titles), and *ARE5 Identifier Type*, *ARE6 Date Type*, *ARE8 Role*

[2] https://jena.apache.org.

Type, *ARE9 Date Certainty* and *ARE14 Place Type* (for more specific types). Most of these classes are subclasses of *E55 Type*, making them a specialization. Among these are the *ARE1 Level of Description*, *ARE5 Identifier Type*, *ARE6 Date Type*, *ARE8 Role Type*, *ARE9 Date Certainty* and *ARE14 Place Type*. On the other hand, the classes *ARE2 Formal Title* and *ARE3 Supplied Title* are subclasses of *E35 Title*, as the concept of *Title* is also present in CIDOC CRM.

Considering these classes, there is a decrease in the use of classes and properties used to represent the type of a title through a ternary relationship, which happens in the CIDOC CRM representation. With this, in the ArchOnto representation, the *E35 Title* class is no longer used, as well as the *PC102 has title*. Instead, the *ARE2 Formal Title* and *ARE3 Supplied Title* classes appear. Associated with these classes, properties used also differ, with a decrease in the use of *P01 has domain*, *P02 has range*, and *P102.1 has type*. Although these properties are present in CIDOC CRM, they are organized in the N-ary ontology in ArchOnto. They are therefore used in the same circumstances but taken from a different ontology.

In CIDOC CRM, the classes *E59 Primitive Value* and *E61 Time Primitive* represent time primitives, but information regarding their temporal extent is missing. In ArchOnto, on the other hand, it is possible to distinguish between an instant and a time interval with *DOE10 Instant* and *DOE11 Interval* classes, respectively. This means that the way dates are represented differs, and the classes and properties used are no longer those used in the CIDOC CRM representation. As a result, there is a decrease in the use of these classes and an increase in the expression through the DataObject ontology.

Furthermore, CIDOC CRM does not establish a sufficient distinction between literal values, whereas, in ArchOnto, it is also possible to differentiate strings with the help of DataObject. This is visible in the complete comparison of the left side of Table 1 with the *DOE8 String* and *DOE17 Person Name* classes, where people's names are separate from other strings.

The class that stood out the most was *E41 Appellation*, with an increase of 33%, from 1,203 occurrences in CIDOC CRM to 2,801 in ArchOnto. This happened since, in CIDOC CRM, the literal values were not considered an appellation, contrary to ArchOnto, particularly with DataObject.

As the number of classes increases, there is a subsequent increase in the number of properties used, as seen in the complete comparison of the right-side Table 1. There was an increase of approximately 12% in the present sample when represented in ArchOnto.

With the previously mentioned ArchOnto's temporal extent capabilities, it is possible to define a start and end date time for a time interval (*DOE11 Interval*) with *DOP6 start date value* and *DOP2 end date value* properties, respectively, and, for a time instant, to determine the timestamp associated with the *DOP8 timestamp* property.

The validation of all existing literal values in the ArchOnto representation resulted in an increase (+30%) in the use of the properties *P1 is identified by*

and the emergence of the use of the *L2DO has value*, a property that makes the connection of CIDOC CRM to the DataObject ontology.

To preserve the integrity of the original descriptions, the ISAD Ontology contains the property *ISAD18 has note*, where the description referring to the ISAD(G) notes is present. This property corresponds to the CIDOC CRM property *P3 has note*, used for informal notes. It is used in ArchOnto to make sure existing information from archival descriptions is kept throughout the migration.

5 Conclusions

CIDOC CRM is one of the most mature ontologies regarding the representation of cultural objects in Linked Data. Based on events, this model can represent several concepts essential to heritage, such as people, places, and dates. However, the model showed limitations in representing critical elements in archival records. With the use of CIDOC CRM in the archives, it was possible to observe that very distinct concepts had to be mapped to the same class. The most obvious case was *E55 Type*. The examples made it clear that, in ArchOnto, it is possible to distinguish the various "types" with classes that enable the use of specific controlled vocabularies. Therefore, ArchOnto provides a more straightforward application of Linked Data in the archival domain.

With the migration of a collection of real-world records to CIDOC CRM and ArchOnto, it was possible to verify that the more specific "types" provide an appropriate range of classes and properties to be used. ArchOnto also added the validation of simple types using the DataObject ontology.

Representations in ArchOnto provide easier access to individuals associated with the specific types. In CIDOC CRM, this would require following extra relationships. For example, to retrieve all people's names, in ArchOnto, it is only necessary to search for individuals of type *DOE17 Person Name*. In contrast, in CIDOC CRM, we need to search for individuals whose type is *E21 Person* and then follow the respective link to arrive at that person's name.

We found that the records migrated from ISAD(G) considered in this work had a very similar structure, making the results less expressive than expected. However, the values obtained allowed us to conclude that ArchOnto provides greater granularity than CIDOC CRM alone. We conclude that ArchOnto is more expressive than CIDOC CRM, as it supersets the latter and generally favours the use of more specific classes and properties.

In the future, we plan to expand this study by considering a more diverse set of archival records to verify whether other description elements can be extracted at a more specific level. It will be interesting, for example, to extract statistics related to a single document or documents according to their description level.

Acknowledgements. National Funds finance this work through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia within project DSAIPA/DS/0023/2018. Inês Koch is also financed by National Funds through the Portuguese funding agency, FCT, within the research grant 2020.08755.BD.

References

1. Bruseker, G., Carboni, N., Guillem, A.: Cultural heritage data management: the role of formal ontology and CIDOC CRM. In: Vincent, M.L., López-Menchero Bendicho, V.M., Ioannides, M., Levy, T.E. (eds.) Heritage and Archaeology in the Digital Age. QMHSS, pp. 93–131. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-65370-9_6
2. ICOM/CIDOC CRM special interest group: definition of the CIDOC conceptual reference Model. ICOM, 7.1.2 edn. (2022)
3. International Council on Archives: ISAD(G): General International Standard Archival Description - 2nd edn. International Council on Archives (2000)
4. International Council on Archives: ISAAR (CPF): International Standard Archival Authority Record for Corporate Bodies, Persons and Families, 2nd edn. International Council on Archives (2004)
5. International Council on Archives Expert Group on Archival Description: Records in Context - Conceptual Model. International Council on Archives (2021)
6. ISO Central Secretary: Information and documentation - A reference ontology for the interchange of cultural heritage information. Standard ISO 21127:2014, International Organization for Standardization (2014), https://www.iso.org/standard/57832.html
7. Kampffmeyer, U.: Document life-cycle management for the European public sector industry white papers. In: Proceedings of the DLM-Forum, pp. 52–63 (2002)
8. Koch, I., Lopes, C.T., Ribeiro, C.: Moving from ISAD(G) to a CIDOC CRM based linked data model in the Portuguese archives. J. Comput. Cult. Heritage (JOCCH) (2023)
9. Koch, I., Ribeiro, C., Teixeira Lopes, C.: ArchOnto, a CIDOC-CRM-based linked data model for the Portuguese archives. In: Hall, M., Merčun, T., Risse, T., Duchateau, F. (eds.) TPDL 2020. LNCS, vol. 12246, pp. 133–146. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-54956-5_10
10. Liikanen, U.: The memory of the information society. In: Proceedings of the DLM-Forum, pp. 21–26 (2002)
11. Lyons, B.: Writing archives/crafting order: a critique on the longstanding archival practices of arrangement and description (2009)
12. Melo, D., Rodrigues, I.P., Varagnolo, D.: A strategy for archives metadata representation on CIDOC-CRM and knowledge discovery. Semant. Web **1**, 1–32 (2022). https://doi.org/10.3233/sw-222798
13. Melo, D., Rodrigues, I.P., Varagnolo, D.: CIDOC-CRM ontology representation of the Portuguese archival description unit obtained from the semantic migration process of DigitArq. Dataset, INESC TEC (2022). https://doi.org/10.25747/BSW1-TQ51
14. Pires, C., Koch, I., Nunes, S.: ArchOnto ontology representation of Portuguese archival description units (baptism records and passports). Dataset, INESC TEC (2023). https://doi.org/10.25747/x78e-1a27
15. Wilkinson, M.D., et al.: The FAIR guiding principles for scientific data management and stewardship. Sci. Data **3**, 1–9 (2016). https://doi.org/10.1038/sdata.2016.18