Enriching Archival Linked Data Descriptions with Information from Wikidata and DBpedia

Inês Koch[1,2(✉)], Cristina Ribeiro[1,2], María Poveda-Villalón[3], Mariano Rico[3], and Carla Teixeira Lopes[1,2]

[1] Faculty of Engineering of the University of Porto, Porto, Portugal
[2] INESC TEC, Porto, Portugal
ines.koch@inesctec.pt
[3] Ontology Engineering Group, Universidad Politécnica de Madrid, Madrid, Spain

Abstract. Various sectors within the heritage domain have developed linked data models to describe their cultural artefacts comprehensively. Within the archival domain, ArchOnto, a data model rooted in CIDOC CRM, uses linked data to open archival information to new uses through the prism of linked data. This paper seeks to investigate the potential to use information in archival records in a larger context. It aims to leverage classes and properties sourced from repositories deemed informal due to their crowd-sourcing nature and the possibility of inconsistencies or lack of precision in the data but rich in content, such as the cases of Wikidata and DBpedia. The anticipated outcome is attaining a more comprehensive and expressive archival description, fostering enhanced understanding and assimilation of archival information among domain specialists and lay users. To achieve this, we first analyse existing archive records currently described under the ISAD(G) standard to discern the typologies of entities involved. Subsequently, we map these entities within the ArchOnto ontology and establish correspondences with alternative models. We observed that entities associated with people, places, and events benefited the most from integrating properties sourced from Wikidata and DBpedia. This integration enhanced their comprehensibility and enriched them at a semantic level.

Keywords: Archival description · ArchOnto · DBpedia · Wikidata · Data propagation

1 Introduction

The Web has moved from a read to a read-write model in which users have shifted from passive information consumers to active contributors, a transformation emblematic of the Web 2.0 era.

More recently, the Web evolved into a structure understandable to humans and machines. This iteration, Web 3.0, introduced the concept of linked data,

A. Antonacopoulos et al. (Eds.): TPDL 2024, LNCS 15177, pp. 396–412, 2024.
https://doi.org/10.1007/978-3-031-72437-4_23

Koch, I., Ribeiro, C., Poveda-Villalón, M., Rico, M., Teixeira Lopes, C. (2024). Enriching Archival Linked Data Descriptions with Information from Wikidata and DBpedia. In: Antonacopoulos, A., et al. Linking Theory and Practice of Digital Libraries. TPDL 2024. Lecture Notes in Computer Science, vol 15177. Springer, Cham. https://doi.org/10.1007/978-3-031-72437-4_23

marking a pivotal advancement in web architecture [5]. Consequently, this digital landscape has witnessed an exponential proliferation of information online and in digital formats. This growth has prompted a reevaluation of knowledge dissemination practices, leading Cultural Heritage sectors to explore novel approaches to describe, preserve, and disseminate cultural artefacts, increasingly in digital formats.

As foundational repositories of collective knowledge, archives assume primary importance in the Cultural Heritage context, necessitating adaptive frameworks that preserve semantic integrity amidst evolving informational formats. In Portugal, more specifically at Directorate of Books, Archives and Libraries (DGLAB), where archival description standards are keeping pace with international evolution, initiatives such as the EPISA Project – Entity and Property Inferences for Semantic Archives – have emerged, proposing a novel data model, ArchOnto[1], to enhance archival record characterisation and accessibility [7].

Archival information often requires specialised knowledge, often accessible only to professionals like historians or researchers due to its specificities. For instance, mentions of titles like the cognomen of a monarch, such as *The Pious King* in the context of King D. João III of Portugal, require a nuanced understanding. By establishing relations between the descriptive elements of ArchOnto and external repositories of information, specifically Wikidata and DBpedia, arises the possibility to augment and streamline the informational content, thereby facilitating comprehension for lay users. Furthermore, enriching archive records will enhance their accessibility not only solely through archive platforms but also across other types of platforms, such as Wikidata and DBpedia, thereby expanding their reach to a broader audience.

To analyse the potential for enriching archival records using Wikidata and DBpedia, we studied the equivalences between the classes and properties of the three models under study to determine how they could be linked. With the established equivalences, we could identify the properties that can be used to extend the ArchOnto ontology and enrich the archival description.

This paper is organised as follows. Section 2 will present the background of this work and the data models used to develop the work. Section 3 outlines the methods employed in this study, while Sect. 4 delves into the results obtained in the work carried out and in the discussion. Lastly, Sect. 5 presents the conclusions drawn from the study.

2 Background

The Torre do Tombo National Archive (ANTT) is one of the oldest institutions in Portugal and one of the oldest national archives in the world. It maintains a distinctive collection of historical and contemporary objects accumulated since the 9th century, which are described with the ISAD(G) – General International Standard for Archival Description, and ISAAR(CPF) – International Standard Archival Authority Record for Corporate Bodies, Persons and Families [7].

[1] Ontology available at: https://purl.org/episa/archonto.

In the last few years, alongside the restructuring of the Portuguese National Archives, the Directorate of Books, Archives and Libraries (DGLAB) has rethought the existing archival system, starting to work on the transition from ISAD(G) multilevel description to a linked data model used to represent information archives in an open source knowledge platform prototype[2]. The data model resulting from this work is called ArchOnto [7]. With ArchOnto and the awareness of the semantic features inherent in linked data, the prospect of interconnecting the entities delineated within this ontology with external entities emerges.

Next, we will present the linked data models used in this study in more detail, namely ArchOnto, DBpedia, and Wikidata.

2.1 ArchOnto

The ArchOnto[3] is a modular ontology designed to facilitate the organisation and description of archival records. Within this framework, elements of existing metadata standards applied to archives, notably ISAD(G) and ISAAR(CPF), are systematically mapped to combinations of classes and properties derived from the CIDOC CRM (Conceptual Reference Model).

Comprising five distinct ontologies, ArchOnto encompasses CIDOC CRM, DataObject, Link2DataObject, ISAD Ontology, and N-ary. Beyond incorporating elements from these foundational ontologies, ArchOnto introduces its classes and properties tailored to represent archive-specific entities absent in existing ontological frameworks. Each component within ArchOnto uniquely identifies through designated prefixes, with classes prefaced by *ARE* (**AR**chival **E**ntity) and properties by *ARP* (**AR**chival **P**roperty) [7].

The CIDOC CRM is the cornerstone of ArchOnto. The International Documentation Committee (CIDOC) of the International Council of Museums (ICOM) developed CIDOC CRM to facilitate the exchange, mediation and integration of diverse sources of information on Cultural Heritage [6]. The CRM Special Interest Group (SIG) is currently developing this model.

The DataObject is an auxiliary ontology that validates literal values utilised within ArchOnto properties, ensuring adherence to defined data types and maintaining semantic consistency across the ontology [7].

N-ary ontology, aligning with CIDOC CRM principles, facilitates the representation of tuples with an arity exceeding two, accommodating scenarios necessitating associations between multiple individuals [7].

The ISAD Ontology encapsulates elements of the ISAD(G) standard, preserving its structure without fragmentation. This ontology employs data properties to capture individual ISAD(G) elements, thereby safeguarding the integrity of original descriptions. Additionally, ISAD Ontology validates atomised contents to ascertain alignment with the source information in ISAD(G) [7].

[2] More information about this transition on the EPISA Project website – https://episa.inesctec.pt/.

[3] Ontology available at: https://purl.org/episa/archonto.

Lastly, Linked2DataObject enables seamless integration between DataObject and CIDOC CRM, facilitated by an object property "hasValue", fostering cohesive interconnectivity between these imported ontologies within the ArchOnto framework [7].

2.2 DBpedia

DBpedia [9] represents a collaborative endeavour to extract structured data from various Wikimedia platforms, notably Wikipedia, and disseminate it across the Linked Data Web. It was created in 2007 by the Free University of Berlin and Leipzig University in collaboration with OpenLink Software, extracting data from Wikipedia and building an RDF graph [1].

This initiative enables users to pose sophisticated queries against datasets derived from Wikipedia and establish connections between diverse datasets available online and Wikipedia-derived data. Despite the resource-intensive nature of the information extraction process, DBpedia stands as a central hub of linked data, interconnecting with numerous knowledge bases beyond Wikipedia, including YAGO, Wikidata, Freebase, and Cyc [2,10].

The construction of DBpedia entails extracting organised information from numerous Wikipedia editions spanning over 100 languages, with active participation from the user community pivotal in providing the semantics to Wikipedia data [9]. Central to DBpedia is its ontology, originally crafted through manual extraction from the infoboxes of the English Wikipedia edition in 2008. Evolving through contributions from the public, the DBpedia Ontology was transformed into a hierarchical structure that initially included 170 classes and mapped 2350 attributes to 720 ontology properties, covering a wide range of domains [4].

The DBpedia ontology currently encompasses 768 classes organised in a subsumption hierarchy and characterised by around 3000 properties, accommodating approximately 4,233,000 instances. Additionally, DBpedia offers a dynamic version, "DBpedia Live", synchronising with real-time updates from Wikipedia, thus ensuring the continual alignment of DBpedia with the latest information available on Wikipedia[4].

2.3 Wikidata

Wikidata [3] is an open and collaborative project hosted and supported by the Wikimedia Foundation whose main objectives are to be the central storage for structured data from its Wikimedia Foundation sister projects, including Wikipedia and Wikisource, among others, as well as providing data for other projects and third-party initiatives [1].

Since its public launch in late 2012, the site has amassed data on over 15 million entities, encompassing over 34 million statements, along with over 80 million labels and descriptions available in over 350 languages [3].

[4] Website DBpedia - Last consulted on 12/04/2024. Available at: https://www.dbpedia.org/resources/ontology/.

Wikidata is organised into entities, much like its data. Each subject for which Wikidata contains structured data is referred to as an entity, and each entity has its own page. Currently, the system distinguishes between entities: items and properties. According to semantic technologies, the former represents individuals and classes, while the latter is akin to RDF properties [3].

Wikidata items are not identified by language-specific tags but by opaque item identifiers, which are automatically assigned at the time of item creation and cannot be modified after that. Item identifiers consist of the letter *Q* followed by a number that identifies the resource that is being described [3].

3 Methodology

The Portuguese National Archive manages a vast collection of cultural objects organised according to their fonds. These fonds are categorised into groups covering central and local administration records, collections, companies, judiciary, monastic institutions, notarial activities, parishes, and personal archives [7].

Identifying elements suitable for expansion to enrich archive records with information from external sources, specifically Wikidata and DBpedia, was crucial to understanding the equivalences between the models used and which classes and properties could be used to make the extension.

To enrich archival records with information from external sources, we undertook a study organised into five distinct stages, as presented in Fig. 1.

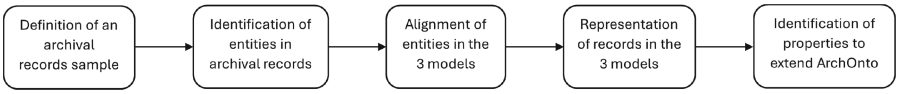

Fig. 1. Stages followed for the definition of properties to enrich archival description

In the initial phase of the methodology, we defined the archival record sample to take into account during the present work. The sample was curated by experts from the Portuguese National Archives, focusing on records abundant in information concerning individuals, locations, and events. These elements hold significant prominence in archival descriptions and are integral to CIDOC CRM, the foundational ontology of ArchOnto. Additionally, the potential for entities to have corresponding data outside the archival domain is considered high.

The selected fond for analysis is titled *Tribunal do Santo Ofício* and encompasses eight subfonds and 72 installation units. This fond pertains to the Portuguese Inquisition and contains information about various Inquisition courts across Portugal, including Lisbon, Porto, Coimbra, Évora, and Lamego (Table 1). It holds historical significance and documents the processes of individuals subject to judgment by the court of the Holy Office during its operational period. This richness in entities makes it conducive to extension using external properties. Details about the analysed documents can be accessed by describing the records and their digital representation.

Table 1. Sample for the first iteration

Title	Reference Code	LD*
Tribunal do Santo Ofício	PT/TT/TSO	Fond
Processo de João	PT/TT/TSO-IC/025/00001	File
Processo de Manuel Gonçalves	PT/TT/TSO-IE/021/00001	File
Processo de Helena Maria	PT/TT/TSO-IL/028/00002	File
Denúncia de Luís [Afonso] contra [...]	PT/TT/TSO-ILM/001/00003	File
Processo de Ana Cardoso	PT/TT/TSO-IP/002/0001/00001	File

*LD stands for Level of Description

In the second phase, we examine the sample of archival records to identify the various types of entities present. This identification was carried out by analysing the elements within the descriptions of the archival records, which were prepared using the ISAD(G) standard. We systematically read the fields of these descriptive elements, identifying all instances within the records and categorising them as corresponding entities. These values were associated with people, places, and events. This phase resulted in identifying entities that would be considered in the subsequent stages of this research.

In the third phase, we establish equivalent classes between ArchOnto, DBpedia and Wikidata to represent the identified entities. These equivalences were determined by mapping the entities in the three models and crossing between the classes identified to represent the entities. Among the various entities identified, it was possible to determine which could be defined in each data model, which were the existing equivalences and which classes could be used to represent them. It was also confirmed that several specific entities in the Portuguese historical panorama, such as ministries or government entities, still lack correspondence in Wikidata and DBpedia, such as *Ministério do Reino* (Table 2). Among the entities identified as not having correspondence in the three models used in the present study, more general classes were taken into account since, despite not having specific instances in the models, these can be represented through classes considered general.

The fourth phase involved manually representing the sample of archive records to ascertain which properties were used to describe the archives in each model. With this representation we wanted to understand which are the instances that each model is able to represent, which properties we need to take into account as possible properties to extend the model, and what information is available on the Wikidata and DBpedia that can enrich the archival representation done in ArchOnto. We observed that when the described documents pertain to processes related to a singular person, as in the *Processo de João* from the *Inquisição de Coimbra* (Table 1), it is not possible to obtain a correspondence of certain elements, such as a person. This is because only the first and last names of the people described are known, elements that become subjective and difficult to link to other sources.

In the fifth and final phase, we identified the properties capable of enriching the archive records with external information. This was achieved by systematically analysing multiple representations of the same type of entity, such as examining several individuals classified as persons.

4 Results and Discussion

In this section, we delineate the outcomes of our entity identification process and expose the correspondences between ArchOnto, DBpedia, and Wikidata classes for representing these entities. Subsequently, we will present a manual representation of an archive record across different data models to elucidate the construction of archive document representations and the utilised properties. Lastly, we will enumerate the properties suitable for enriching archive records with information sourced from DBpedia and Wikidata.

4.1 Identification of Entities and Properties

In the initial phase of the study, a comprehensive analysis of archival descriptions revealed a consistent presence of entities across all levels of description. Predominantly, these entities encompass people, places, dates, and events. Table 2, included herein, showcases a selection of identified entities chosen for their representativeness of the diverse range of entities encountered within archival records[5].

With the identified entities, we determined the classes that enable their representation in ArchOnto, Wikidata, and DBpedia. Table 2 showcases examples of correspondences within the archival description of the *Tribunal do Santo Ofício*. While certain entity instances may not be represented in one or more models, it's noteworthy that 70 entities (41,9%) can be described across all three models.

Table 2. Entities alignment

Entity	ArchOnto	DBpedia	Wikidata
1532	E52 Time-span	dbo:Year	year (Q577)
Afonso Gonçalves	E21 Person	dbo:Person	human (Q5)
Ministério do Reino	E74 Group	- - - - - -	- - - - - -
Pastor	ARE8 Role	dbo:Profession	profession (Q28640)
Paulo III	E21 Person	dbo:Pope	human (Q5)
Portugal	E53 Place	dbo:Place	country (Q6256)
Prisão	E5 Event	- - - - - -	arrest (Q1403016)
Rei	ARE8 Role	dbo:PersonFunction	monarch (Q116)
Rei Piedoso	E21 Person	dbo:Animal	human (Q5)
Roma	E53 Place	dbo:Capital_city	abolished municipality in Italy (Q3685476)
Tesoureiro	ARE8 Role	dbo:Person	profession (Q28640)
Tomar	E53 Place	dbo:City	municipality of Portugal (Q13217644)

[5] Due to space limitations, we only present some examples. The complete identification is available in the *Tribunal do Santo Ofício entities* file in the dataset [8].

Among the identified entities, we observed that 158 (94.6%) are represented in ArchOnto, which is justified since we are just starting to represent the entities using this model. Entities not defined by this ontology include those related to the age of the people described, as ArchOnto only represents this concept through the date of birth. Additionally, concepts such as "*bigamy*", "*crime*", "*death penalty*", and "*heretical propositions*" are also incapable of representation. They can only be illustrated by adding a note to the described document using the *P3 has note* property.

Regarding DBpedia, only 79 (47.3%) instances of the identified entities are represented. DBpedia demonstrates the lowest capacity for representing entities in selected archival descriptions among the three data models used in this study. This limitation arises because, although DBpedia can define classes, it lacks the specificity required, necessitating the use of higher-level and more general entities.

In Wikidata, 108 (64.7%) of the instances identified in archival entities can be represented. This higher representation may be attributed to the more active community on Wikidata compared to DBpedia, leading to a more significant number of available instances and classes in this information source.

By identifying classes, not only in terms of individuals but also regarding the archival record itself, we could determine the necessary properties to represent the archival description.

Table 3 presents the properties used to map the various ISAD(G) elements in ArchOnto, DBpedia, and Wikidata. Among these elements, the level of description stands out, represented in ArchOnto as a document (*E31 Document*) through the property *ARP12 has level of description*. Due to the specificity of this archival element, it is necessary to indicate its level using *dbo:type* in DBpedia. Conversely, Wikidata allows for the representation of this element through the property *description level (P6224)*. Another element to highlight is the Physical Location, which in DBpedia is identified through the *dcmi:identifier* property, a Dublin Core property used by DBpedia.

Table 3. Properties used to represent ISAD(G) elements

ISAD(G)	ArchOnto	DBpedia	Wikidata
Title	P102 has title	dbo:alternativetitle	title (P1476)
Level of Description	ARP12 has level of description	dbo:type	level of description (P6224)
Reference Code	P1 is identified by	dbo:documentNumber	identifies (P10476)
Production Date	P4 has time-span	dbo:productionStartDate dbo:productionEndDate	inception (P571)
Support	P45 consists of	dbo:material	made from material (P186)
Physical Location	P1 is identified by	dcmi:identifier	identifies (P10476)
Creation Date	P4 has time-span	dbo:publicationDate	publication date (P577)
Last Modification	P4 has time-span	dbo:dateLastUpdate	last update (P5017)

4.2 Manual Representation of "Processo de João"

After identifying the classes and properties, we manually represented the selected archive records for the sample in the three models, ArchOnto, Wikidata and DBpedia. This representation followed an iterative methodology, starting with the Fond representation and then representing five Files from distinct Inquisition Courts of Portugal. The primary goal of this representation was to analyse the equivalences and correspondences of the representation of the entities in the three linked data models, observing how the same sample of archival records is modelled in the three selected data models.

To elucidate the process of representing and selecting properties conducive to enriching archival descriptions through ArchOnto, this study focuses on the archival record titled "*Processo de João*" (João's Process) – Table 1. Designated as a Unit of Description at the File level, this record encapsulates many entities pertinent to the research scope, including individuals, locations, events, and temporal delineations. This description unit contains information concerning the inquisition process of João, who underwent trial for the crime of heretical propositions. It provides details about João, including his parents' names, age, place of birth, and information concerning his sentence.

Within this record, certain elements readily lend themselves to direct translation into ArchOnto classes and properties, notably the *Title*, *Reference Code*, and *Level of Description*. However, other components, such as the Scope and Content element, need atomisation before being integrated into the ArchOnto representation. This happens since *Scope and Content* is a descriptive element, which makes it rich in entities.

Figure 2 illustrates a part of the ArchOnto representation of the document under scrutiny[6]. An array of classes and properties requisite for encapsulating the content outlined in the ISAD(G) description is discernible within this depiction.

The representation of archive records in ArchOnto is consistently based on the description unit delineated by the *E31 Document*. This unit serves as the focal point for all elements related to the description of a record, encompassing aspects such as the title, level of description, identifiers, and the physical and conceptual attributes of the document. Physical attributes pertain to characteristics inherent to the object itself, including dimensions and support composition, along with the document's production date. On the other hand, conceptual attributes encompass aspects like the language in which the document is written, as well as dates associated with its archival description, such as the creation date and the date of the last update.

The *Level of Description* of a record is delineated through the *ARE1 Level of Description* class. The document title constitutes a supplied title, encapsulated within the *ARE3 Supplied Title* class. A complete representation of the title necessitates utilising the *DOE8 String* class from the DataObject ontology, facilitating the definition and validation of the precise sequence of characters within the title.

[6] Due to the loss of detail in the image, the full image of ArchOnto representation is available in the *Inquisição de Coimbra – ArchOnto* file in the dataset [8].

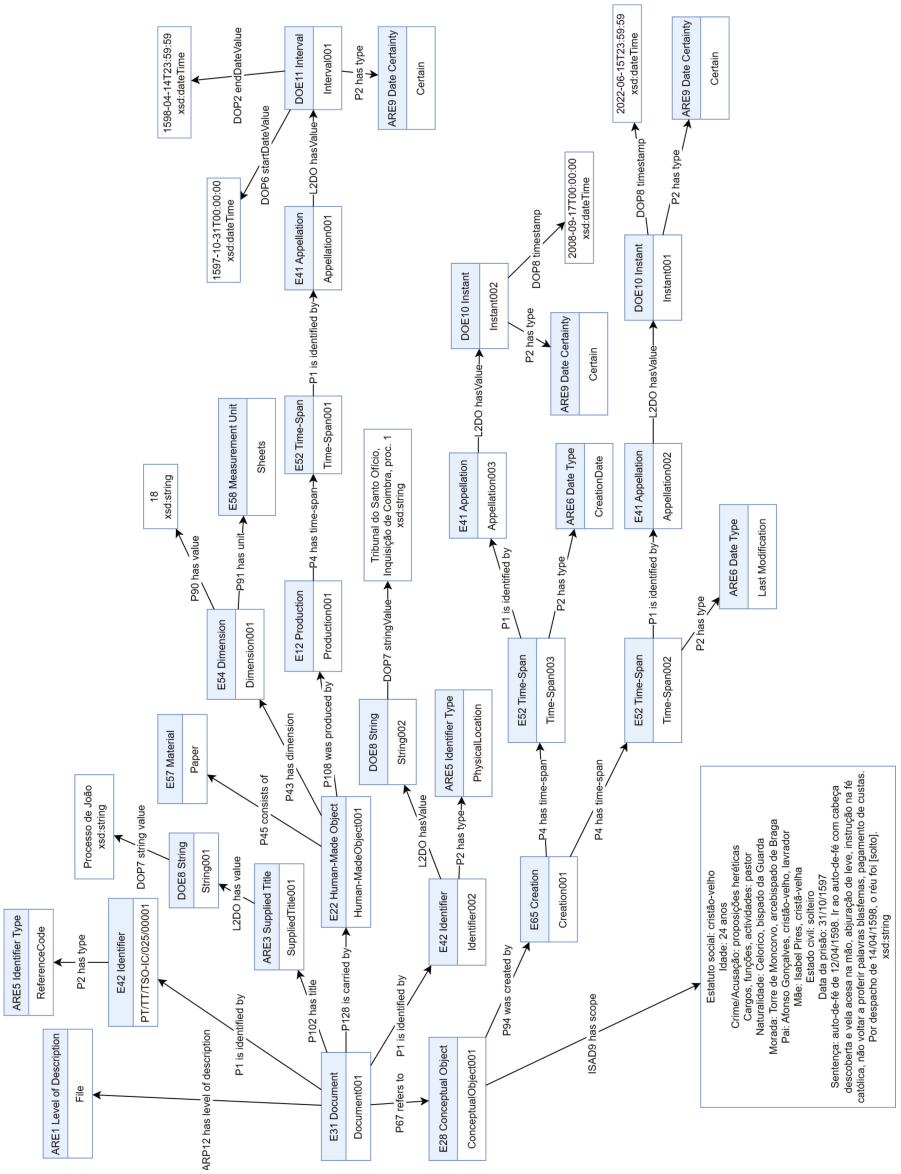

Fig. 2. Processo de João – ArchOnto Representation

In ArchOnto, the *Reference Code* and the document's *Physical Location* are uniformly represented by the *E42 Identifier* class, as can be seen in the representation of this record. Since a single document may have multiple types of identifiers, specifying the mapped identifier type becomes imperative, requiring using the *ARE5 Identifier Type* class.

Regarding the dimensions of a document, they are represented by the *E54 Dimension* and encompass two associated values: the dimension's *Measurement Unit (E58)* and its numerical value. Concerning the physical aspects of the record, the production date of the document is also addressed through a series of classes and properties linked to *E12 Production*.

Within Fig. 2, we depict the creation date of the archival document. The representation of the date commences with elucidating the archival document creation event (*E65 Creation*), followed by two distinct *E52 Time-Spans*, differentiated via the *ARE6 Date Type* class. The first time span denotes the document's creation, while the second denotes the last modification date.

Due to its lengthy nature and absence of direct mapping, numerous classes and properties are employed to represent Scope and Content. Primarily featured are those about individuals (*E21 Person*), locations (*E53 Place*), events (*E5 Event*), and chronological references (*E52 Time-span*).

In this representation, the ISAD Ontology comprehensively encapsulates all information within this file's ISAD(G) elements. An illustrative instance of applying this ontology is evident in the Scope and Content element representation, wherein the data property *ISAD9 scope* is employed to retain all textual components without any form of atomisation.

After having the mapping in ArchOnto, we moved on to mapping the same sample in DBpedia and Wikidata. The ISAD(G) descriptions made available by the Portuguese National Archives were considered to carry out these mappings. In Figs. 3 and 4, it is possible to see how the mappings of the "Processo de João" record are made for these two models[7].

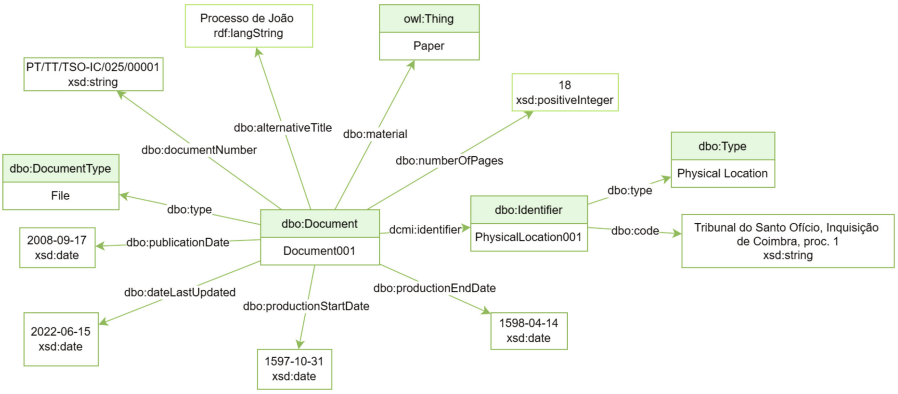

Fig. 3. Processo de João – DBpedia Representation

Concerning ISAD(G) elements specific to archives, it was observed that DBpedia and Wikidata do not always facilitate the direct representation of all the components. However, some properties enable direct mapping.

[7] Due to the loss of detail in the full images, the complete picture of these representations is available in the dataset [8].

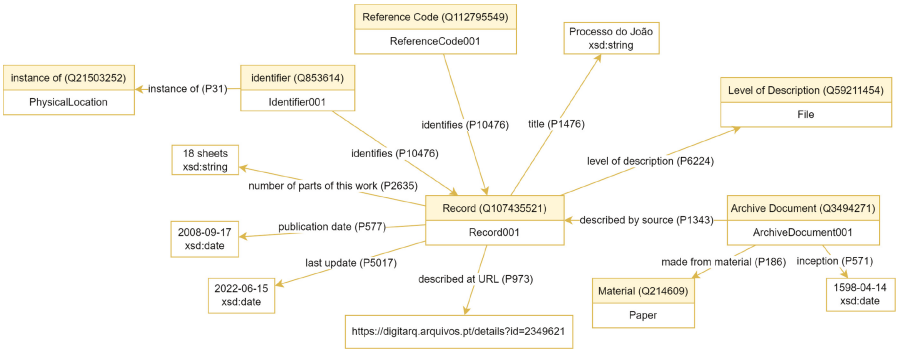

Fig. 4. Processo de João – Wikidata Representation

In terms of indicating the title and its type, given that titles in archives typically assume a formal or supplied type, it was observed that a more intricate representation is necessary for ArchOnto. This complexity arises from the requirement to utilise DataObject to represent the value present in the string. Conversely, in DBpedia and Wikidata, this mapping is straightforward, accomplished through the *dbo:alternativeTitle* or *dbo:title* properties in DBpedia, and the *title (P1476)* property in Wikidata. However, in Wikidata, it is not possible to distinguish the title type.

Other aspects that could be verified were that, in ArchOnto and Wikidata, it was possible to differentiate the physical part of the document and its registration. At ArchOnto, this is done through the *E22 Human-Made Object* and *E28 Conceptual Object* classes. On the other hand, in Wikidata, this differentiation is made through the use of the *Record (Q107435521)* and *Archive Document (Q3494271)* classes.

The simplicity of mapping these two data models is immediately evident, as they require fewer steps to represent certain concepts. An example is the representation of the archive record's creation and last modification dates. In DBpedia, these two concepts are represented through the *dbo:publicationDate* and *dbo:dateLastUpdate* properties, while in Wikidata, they are expressed through the *publication date (P577)* and *last update (P5017)* properties.

4.3 Properties to Extend the Model

With the previous stage, it was possible to determine how ArchOnto, Wikidata and DBpedia represent the existing archival records in the Portuguese National Archives. Given the understanding of the alignment between the classes of the three analysed data models, as well as the representation capabilities each offers for archive records, it became feasible to discern the ArchOnto entities that would benefit from increased expressiveness and completeness through enrichment with properties from DBpedia and Wikidata.

To extend the properties, we analysed the representations of each entity in Wikidata and DBpedia, selecting the properties suitable for use in archival descriptions.

It is discernible that Wikidata boasts a more extensive array of properties than DBpedia, thereby facilitating the enrichment of ArchOnto. This disparity can be attributed to the heightened activity level and collaborative engagement within the Wikidata community. Due to the more significant number of properties available in the representations of the selected sample, Wikidata was favoured as a source of information, as it has the potential to enrich the archive records further. However, among the various properties used, the different identifiers present in Wikidata were not considered, which could provide additional enrichment to the archival records from other external sources.

It is evident that the entities associated with the level of description (*ARE1 Level of Description*), places (*E53 Place*), and occupations (represented by the *E74 Group*) were subject to extension.

In Fig. 5, it is possible to observe an example of a place enriched with information from external sources to archives. In this case, the entity *E53 Place* is exclusively enriched with Wikidata properties and classes since DBpedia does not represent this specific place. Although this happens in this instance, it is important to note that places can also be further enriched with the incorporation of DBpedia properties.

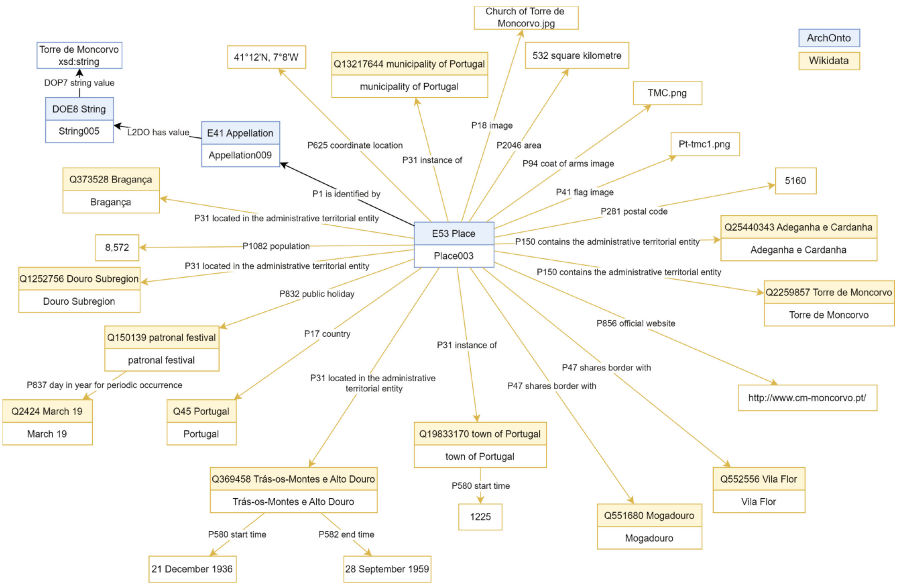

Fig. 5. Processo de João – Place enrichment

Table 4[8] left side illustrates the properties found while representing and analysing the different records of the Portuguese National Archives to enhance the description level and place representation.

Among the identified properties, two represent concepts present in ArchOnto but had not been mapped in the current registry. These properties relate to the organisation of the hierarchy of description records. In ArchOnto, indicating that a description level is above or below another is feasible by utilising the *ARP8 upper level* and *ARP9 lower level*. Conversely, this representation is made through the *next higher rank (P3730)* and *next lower rank (P3729)* properties, respectively.

Table 4. Wikidata properties' description and number that can be used to extend Level of Description and Place

Level of Description
described at URL (P973)
described by source (P1343)
different from (P1889)
has part(s) (P527)
next higher rank (P3730)
next lower rank (P3729)
part of (P361)

Place
area (P2046)
coat of arms image (P94)
contains the administrative territorial entity (P150)
coordinate location (P625)
country (P17)
flag image (P41)
image (P18)
located in the administrative territorial entity (P131)
official website (P856)
population (P1082)
postal code (P281)
public holiday (P832)
shares border with (P47)

On the right side of Table 4, properties relating to the specific location of the described places in the record are visible. These properties include information about the country, territorial entity, postal code, coordinates, and neighbouring places. Details about public holidays, population, area, and images depicting the location, flag and coat of arms are also presented. Concerning place properties, this table exclusively lists those utilised in the description of this record. The remaining identified properties are available for consultation in the INESC TEC repository [8].

The exploration of relevant information within DBpedia and Wikidata concerning occupation entities facilitated the enrichment of the mapping process, contributing to a clearer understanding of the represented entities. Table 5 outlines the properties employed to enhance the information in ArchOnto, which previously focused solely on the profession of each individual mentioned in the archival description.

[8] Due to space limitations, all tables are available in *Properties extension* file [8].

Wikidata properties facilitate the indication of the name's feminine and masculine forms, specifying their field of activity and identifying any coincidences or differences with other professions, as well as the patron saint and extra things used in this occupation, as an animal or a material. Furthermore, provides information about the source where the profession is described and includes images illustrating it. Conversely, DBpedia properties also allow for indicating the field of activity, its type, and the sector of activity.

In both Wikidata and DBpedia, it is possible to reference additional properties related to the occupations of a given person; however, these were not relevant to this representation.

Table 5. Wikidata and DBpedia properties to extend Occupation

Wikidata Properties	DBpedia Properties
described by source (P1343)	dbp:activitySector
different from (P1889)	dbp:employmentField
female form of label (P2521)	dbp:type
field of this occupation (P425)	
image (P18)	
male form of label (P3321)	
partially coincident with (P1382)	
patron saint (P41)	
uses (P2283)	

In addition to the entities presented in this archive record, it was also possible to observe other entities that could benefit from the enrichment of properties from Wikidata and DBpedia. These include people, events, and dates.

Regarding people, it was evident that archival documents often only refer to individuals' names without providing context regarding the time and place in which they exist or existed. Therefore, information from these tools can help elucidate such details for users. An example is King D.João III, referred to in the *Tribunal do Santo Ofício* record as the Pious King (Rei Piedoso). However, no further information is provided about this monarch. Figure 6 displays some of the properties that can be added to this entity[9].

[9] More properties can be consulted in the complete representation of this individual found in the *Person Extension* file available in the dataset [8].

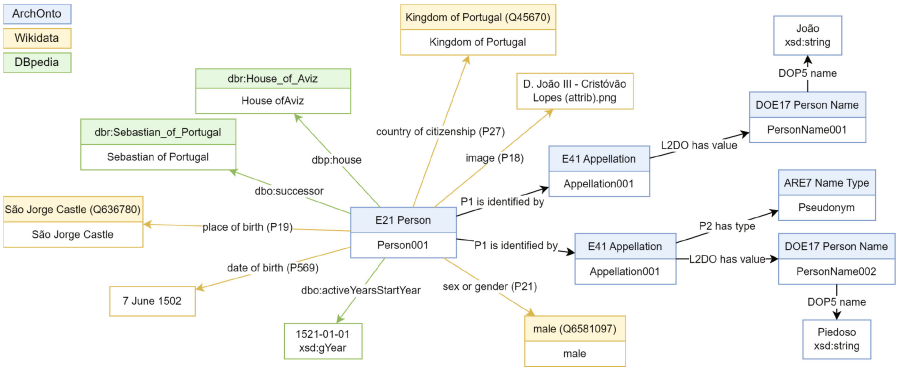

Fig. 6. King D. João III - Person enrichment

Concerning events, they can also benefit from the enrichment with properties from DBpedia and Wikidata, encompassing properties related to the event's name, location, start and end dates, participants, subject, and nature of the event.

5 Conclusions

In this work, we examined the potential of Wikidata and DBpedia to enrich ArchOnto, an ontology adapted for describing archival records. This research addressed the challenge of making archival records more accessible and understandable to a broader audience beyond domain experts by expanding existing and new archival records with information from other, less formal sources.

Through mapping archival records and identifying analogous classes and properties, it became evident that Wikidata and DBpedia offer a vast repository of information relating to entities found in records that could benefit from expansion for a better understanding of the documents. Leveraging this capacity allows for increased specificity of records, thereby enhancing their accessibility and comprehensibility for specialists and non-specialists alike.

Both Wikidata and DBpedia enable the representation of elements within archival descriptions. However, Wikidata generally contains a broader array of instances than DBpedia. This distinction is significant, as although both platforms can represent archival records, DBpedia often lacks depth of detail. DBpedia's representations are frequently confined to classes rather than instances, thereby not contributing significant additional information beyond what is already included in the archival description created in ArchOnto.

Upon comparative analysis between the two data models and subsequent mapping of archive records, it was established that Wikidata possesses a broader scope for representing information capable of enriching and expanding archival data. Nevertheless, DBpedia also contributes valuable properties that enrich records, fostering greater completeness and understanding across users.

Another notable finding is that DBpedia contains more descriptive information than Wikidata. Consequently, some properties may not be directly accessible, as the data is embedded within more descriptive fields, such as *dbo:abstract*.

Ultimately, the investigation demonstrated that the entities most poised to benefit from such enrichment are those deemed central within archival records and, by extension, within ArchOnto. These pivotal entities encompass individuals, events, and places whose comprehensive representation facilitates a deeper understanding of archival content and fosters greater user accessibility.

Acknowledgements. Thanks to the Ontology Engineering Group for allowing us to develop this work, especially to María Poveda Villalón and Mariano Rico for their support during the collaboration period and during Inês Koch's stay in Madrid. Inês Koch is financed by National Funds through the Portuguese funding agency, FCT – Fundação para a Ciência e a Tecnologia, within the research grant 2020.08755.BD - DOI: 10.54499/2020.08755.BD.

References

1. Abián, D., Guerra, F., Martínez-Romanos, J., Trillo-Lado, R.: Wikidata and DBpedia: a comparative study. In: Szymański, J., Velegrakis, Y. (eds.) IKC 2017. LNCS, vol. 10546, pp. 142–154. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-74497-1_14
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: a nucleus for a web of open data. In: Aberer, K., et al. (eds.) ASWC/ISWC -2007. LNCS, vol. 4825, pp. 722–735. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-76298-0_52
3. Erxleben, F., Günther, M., Krötzsch, M., Mendez, J., Vrandečić, D.: Introducing Wikidata to the linked data web. In: Mika, P., et al. (eds.) ISWC 2014. LNCS, vol. 8796, pp. 50–65. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-11964-9_4
4. Hellmann, S., Stadler, C., Lehmann, J., Auer, S.: DBpedia live extraction. In: Meersman, R., Dillon, T., Herrero, P. (eds.) OTM 2009. LNCS, vol. 5871, pp. 1209–1223. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-642-05151-7_33
5. Hiremath, B.K., Kenchakkanavar, A.Y.: An alteration of the web 1.0, web 2.0 and web 3.0: a comparative study. Imperial J. Interdisc. Res. **2**(4), 705–710 (2016)
6. ICOM/CIDOC CRM Special Interest Group: Definition of the CIDOC conceptual reference model, 7.1.3 edition. ICOM (2024). https://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_7.1.3.pdf
7. Koch, I., Teixeira Lopes, C., Ribeiro, C.: Moving from ISAD(G) to a CIDOC CRM-based linked data model in the Portuguese archives. J. Comput. Cult. Herit. **16**(4) (2023). https://doi.org/10.1145/3605910
8. Koch, I.: Tribunal do Santo Ofício in ArchOnto – extension of archival records through Wikidata and DBpedia properties. Dataset, INESCTEC (2024). https://doi.org/10.25747/SRYA-8115
9. Lehmann, J., et al.: DBpedia - a large-scale, multilingual knowledge base extracted from Wikipedia. Seman. Web **6**(2), 167–195 (2015). https://doi.org/10.3233/SW-140134
10. Piuri, V., Balas, V.E., Borah, S., Syed Ahmad, S.S. (eds.): Intelligent and Interactive Computing. LNNS, vol. 67. Springer, Singapore (2019). https://doi.org/10.1007/978-981-13-6031-2