Exploring Semantic Archival Collections: The Case of Piłsudski Institute of America

Laura Pandolfo[1](✉), Luca Pulina[1], and Marek Zieliński[2]

[1] Dipartimento di Chimica e Farmacia, Università di Sassari,
via Vienna 2, 07100 Sassari, Italy
{laura.pandolfo,lpulina}@uniss.it

[2] Piłsudski Institute of America, 138 Greenpoint Avenue, Brooklyn, NY 11222, USA
MZielinski@pilsudski.org

Abstract. Over the last decades, a huge amount of available digital collections have been published on the Web, opening up new possibilities for solving old questions and posing new ones. However, finding pertinent information in archives often is not an easy task. Semantic Web technologies are rapidly changing the archival research by providing a way to formally describe archival documents.

In this paper, we present the activities employed in building the semantic layer of the Piłsudski Institute of America digital archive. In order to accommodate the description of archival documents as well as historical references contained in these, we used the ARKIVO ontology, which aims at providing a reference schema for publishing Linked Data. Finally, we present some query examples that meet the domain experts' information needs.

Keywords: Semantic technologies for digital archives · Ontologies · Linked data

1 Introduction

Semantic Web (SW) technologies [1,2] have been offering new opportunities and perspectives in their use in historical research and, more in general, in the humanities. Until recently, the overall historical documents was scattered among different archives, each of which holds a specific and unique collection of information separated from the others. Historians and scholars had to physically visit archive repositories each time they wanted to consult primary sources, try to get relevant information and then manually assemble cross-references. Today, the huge amount of available digital collections, usually converted in interchangeable formats, makes it feasible to access resources from any place at any time, by offering users the possibility to have direct access to information regardless of where they are physically located. Also, the publication of several datasets on the Web provide a comprehensive picture of historical and social patterns

P. Manghi et al. (Eds.): IRCDL 2019, CCIS 988, pp. 107–121, 2019.
https://doi.org/10.1007/978-3-030-11226-4_9

by allowing historians to explore unknown interactions between data that could reveal important new knowledge about the past.

In the last decades, there has been a great amount of effort in designing vocabularies and metadata formats to catalogue documents and collections, such as Dublin Core (DC)[1], Functional Requirements for Bibliographic Records (FBRB)[2], MAchine-Readable Cataloging (MARC)[3], Metadata Object Description Schema (MODS)[4], and Encoded Archival Description (EAD)[5], just to cite a few well-known examples. While DC is particularly suited for enabling searches of library catalogs of digital collections, metadata such as FRBR, EAD and MODS seem to be more devoted to human consumption rather than machine processing [3]. Concerning MARC, some experts experienced that it is not suitable neither for machine processable nor for actionable metadata [4,5]. Also, MODS is focused on objects such as books, and EAD, even reflecting the hierarchy of an archive, is focused on finding aids and the support for digitized objects is limited.

Despite the wide range of metadata standards, there is an ongoing lack of clarity regarding the use of these resources, which leads to the conclusion that in absence of a standardized vocabulary or ontology, different institutions will continue to use their own distinct systems and different metadata schemas. Notice that both vocabulary and ontology describe the way human beings refer to things in the real world, but they are different in a number of aspects. For instance, vocabularies may have no formal semantics, no defined interrelationships between different terms, and consequently no automatic reasoning technique can be exploited. On the contrary, ontologies are usually specified using the Web Ontology Language OWL [6] language, which has its logical grounding in Description Logics (DLs) [7]. Their formal semantics allows humans and computer systems to exchange information without ambiguity as to their meaning, and also makes it possible to infer additional information from the facts stated explicitly in an ontology [8].

In this paper, we present the activities employed in building the semantic layer of the Piłsudski Institute of America digital archive. In order to accommodate the description of archival documents, supporting archive workers by encompassing both the hierarchical structure of archival collections and rich metadata created during the digitization process, we used the ARKIVO ontology for modeling the Piłsudski archival collections. ARKIVO aims at providing a reference schema for publishing Linked Data [9] about archival documents as well as to describe historical elements referenced in these documents, by giving the opportunity to represent meaningful relationships between data.

The paper is organized as follows. Section 2 includes some preliminaries that will be used in the rest of the paper, and some relevant work in the field of digital

[1] http://dublincore.org/.

[2] http://www.cidoc-crm.org/frbroo/.

[3] https://www.loc.gov/marc/.

[4] http://www.loc.gov/standards/mods/.

[5] https://www.loc.gov/ead/.

archives. Section 3 describes all the aspects related to the use case, including the ARKIVO ontology for modeling resources and the connection to external datasets. Section 4 is dedicated to present some query examples that meet the domain experts' information needs. Finally, Sect. 5 concludes the paper with some final remarks and future work.

2 Background and Related Work

2.1 Preliminaries

An ontology is usually defined as a formal specification of domain knowledge conceptualization [10]. Ontologies can be defined by ontology languages such as the Resource Description Framework Schema (RDFS) [11] and the Web Ontology Language OWL [6]. While RDF is suitable for modeling simple ontologies, OWL is considered a more expressive representation language. The latest version of OWL is OWL 2, which addresses the complexity issue by defining profiles [12], namely fragments for which at least some reasoning tasks are tractable. OWL 2 DL, the version of the Web Ontology Language we focus on, is defined based on Description Logics (DL), which is a family of formal knowledge representation languages that models concepts, roles, individuals and their relationships. In DL, a database is called a knowledge base. In particular, if $\mathcal{K} = \langle \mathcal{T}, \mathcal{A} \rangle$ is a knowledge base, then the Tbox $\mathcal{T}$ is a set of *inclusion assertions*, i.e., concept descriptions in $\mathcal{AL}$ or some of its extensions, whereas the Abox is a set of *membership assertions* of the form $A(x)$ and $R(x, y)$ where A is some atomic concept, R is some atomic role and x, y are objects of a domain. Some OWL 2 constructors with the corresponding DL syntax are listed in Table 1.

Table 1. OWL 2 constructors and the corresponding DL syntax.

Constructor	DL syntax
Class	C
SubClassOf (C D)	$C \sqsubseteq D$
EquivalentClasses ($C_1 ... C_n$)	$C_1 \equiv ... \equiv C_n$
DisjointClasses ($C_1 ... C_n$)	$C_i \sqcap C_j \sqsubset \bot$ $i \neq j \; i, j \in \{1, ..., n\}$
DisjointUnion (C $C_1 ... C_n$)	$C = C_1 \sqcup ... \sqcup C_n$
objectProperty	R
datatypeProperty	T
ObjectInverseOf (R)	R^-
ObjectSomeValuesFrom (R C)	$\exists R.C$

RDFS is considered as the basic representation format for developing the SW. It represents sentences in the form of triples, which consist of a subject, a predicate and an object. Triples can also be represented using the Turtle notation [13], which provides a slightly more readable RDFS serialization. As shown in the example below, Turtle syntax separates the data into two parts: a list of URIs and their abbreviated prefixes, and a list of the triples. In this example, the three triples express that the subject in these is a book, identified by a specific URI, which has its title and an author.

```
@prefix ex:  <http://example.com> .
@prefix book:<http://books.com> .
 book:uri rdf:type   ex:Book ;
          ex:title   "The Whale" ;
          ex:author  "Herman Melville".
```

The standard query language is SPARQL [14], whose latest version, namely SPARQL 1.1, includes many new language features such as aggregates, subqueries, a new suite of built-in functions, and path expressions. SPARQL queries typically consist of various clauses and blocks, which specify basic graph patterns to be matched along with keywords that join, filter and extend the solution sequences to these patterns.

```
PREFIX dbp: <http://dbpedia.org>
PREFIX dcterms: <http://purl.org/dc/terms/>
SELECT ?title ?author ?publisher ?date
WHERE  {
         ?book dbp:title     ?title .
         ?book dbp:author    ?author .
         ?book dcterms:publisher ?publisher .
         OPTIONAL {?book dbp:releaseDate ?date}
        }
```

Considering the SPARQL query example above, the keyword `PREFIX` declares a namespace prefix, similar to @prefix in Turtle. The keyword `SELECT` determines the general result format, while the statements after `SELECT` refer to the remainder of the query. The listed names are identifiers of variables for which return values are to be retrieved. The above query returns all values for the variables `?title`, `?author`, `?publisher`. The actual query is initiated by the keyword `WHERE`. It is followed by a simple graph pattern, enclosed in curly braces. Intuitively speaking, these identifiers represent possible concrete values that are to be obtained in the process of answering the query. Finally, the keyword `OPTIONAL` refers to the subsequent group pattern. Optional patterns are not required to occur in all retrieved results, but if they are found, may produce bindings of variables – in our case `?date` – and thus extend the query result [2].

SPARQL 1.1 allows to use many other operators, such as, e.g, `FILTER` to restrict the set of matching results and `SERVICE` to support queries that merge data distributed across the Web. For the full list of operators, please refer to [14].

2.2 Semantic Digital Archives: A Brief Survey

Although digital archives and digital libraries may appear similar, a number of distinctions can be identified. The main difference lies in the focus on history in archives work, since they usually house collections of unique and unpublished materials not available anywhere else. Moreover, digital archive catalogues have to reproduce in some way their hierarchical structure composed of different layers, in which *fonds* represent the highest level of that structure, while *item* the lowest. Notwithstanding these differences, both are facing new challenges in order to overcome traditional data management and information browsing. In this context, SW technologies can improve the annotation metadata process by adding semantic capabilities, which increase the quality of the information retrieval process. Moreover, the use of shared ontologies can enable interoperability and promote consistency between different systems [15,16].

Since the late 1990s, after a great digitization effort, several digital archives on the Web have been developed with the aim of facilitating the document storage and retrieval processes. Europeana[6] represents the most visible effort to link cultural heritage resources and their metadata across several cultural institutions throughout Europe. The data are represented in the Europeana Data Model (EDM) [17], which is based on SW languages. EDM ensures a suitable level of interoperability between the datasets from different institutions, but the automatic conversion process into new data formats causes loss of the original metadata [18].

Pergamos[7] is a Web-based digital library implemented by the University of Athens that offers a uniform platform for managing, documenting, preserving and publishing heterogeneous digital collections. Pergamos provides even old and rare historical material in PDF format and the users can browse this collection and also visualize biographic and bibliographic information. A beta version of its open data platform has been published recently.

The Franklin D. Roosevelt (FDR) Library and Digital Archives[8] represent one of the most significant collections of historical material concerning the history of America and the world in the 20th century. In the context of this digital library, the FDR/Pearl Harbour project was intended to develop means to support enhanced search and retrieval from a specific set of documents included in the FDR library which refer to the history of the Pearl Harbour attack. The main goal of the FDR/Pearl Harbor project was to provide new search methods for historians based on an ontology in the background, which supported the retrieval process not only on the basis of specific names and events but also by category and/or role [19].

Another example of digital archive is the Józef Piłsudski Archive created by the Piłsudski Institute of America[9], which has been used as use case in this paper and it will be described in details in the next Section.

[6] https://www.europeana.eu/portal/en.

[7] https://pergamos.lib.uoa.gr/uoa/dl/frontend/en/index.html.

[8] http://www.fdrlibrary.marist.edu/archives/collections.html.

[9] http://www.pilsudski.org/portal/en/about-us/history.

3 The Piłsudski Institute of America Digital Collections

The Józef Piłsudski Institute of America was set up in 1943 in New York City for the purpose of continuing the work of the Institute for Research of Modern History of Poland, established in Warsaw in 1923. After Poland regained its independence at the end of the Great War, a group of historians and officers begun to travel around the country to collect archival documentation. At the beginning of World War II, part of the archives were evacuated and landed in Washington, eventually creating the seed of the Institute archival collections, which grew in time by donations from politicians, officers and organizations of prewar Poland and Polish diaspora. From its establishment, the Institute was to be a cultural and historical research center that would gather archives in order to disseminate the history of Poland [20].

Today, the Institute is devoted to collecting, safe-keeping and preserving the documents and other historical memorabilia as well as to make these resources accessible to researchers and visitors by providing support to scholars during archival queries on site. To give some idea of the range of archival material, the collections occupy about 240 linear meters, namely 2 million pages of documents covering mostly the Polish, European and American history of late 19th and 20th century. The majority of archival documents are written in Polish, but the amount of documents in other languages (e.g., Italian, English, Russian, French, Portuguese, and others) is not trivial. The international character of the archival resources draws the attention of a large number of experts and visitors coming from different countries. The collections include not only archival documents but also photographs, films, posters, periodicals, books, personal memoirs of diplomats as well as collection of paintings by Polish and European masters. In the last years, the archival collections have been annotated, digitized, full-text indexed, and gradually put online on the website of the Institute. The annotation process has been carried out in two steps. In the first step, archive workers have manually annotated each document with relevant entities (e.g., title, author, date of creation, etc.). In the second step, the annotations has been regularly validated and stored in a database. Considering the maturity reached in the development of Semantic Web technologies and Linked Data applications, the Piłsudski Institute of America started to link the archival data to external resources.

3.1 Modeling Archival Collections with ARKIVO Ontology

ARKIVO [21] is an ontology designed to accommodate the archival description, supporting archive workers by encompassing both the hierarchical structure of archives and the rich metadata attributes used during the annotation process. The strength of ARKIVO is not only to provide a reference schema for publishing Linked Data about archival documents, but also to describe the historical elements contained in these documents, e.g., giving the possibility to represent relationships between people, places, and events. We used ARKIVO ontology to model the Piłsudski digitized archival collections.

The ontology development process has been carried out according to a top-down strategy, which consists first in identifying the most abstract concepts of the domain and then in specializing the specific concepts. Given the specificity of the archival field, domain experts and archivists often used real scenarios to validate the design of ARKIVO ontology throughout its development process. In the light of reusability principle [22], we selected some existing standard metadata, such as, e.g., `Dublin Core` and `schema.org`[10] for describing and cataloguing both physical resources, `BIBO` ontology[11] in order to have a detailed and exhaustive document classification, `FOAF`[12] for describing agents, `Geonames`[13] for linking a place name to its geographical location and `LODE`[14] for representing events. Throughout this paper, prefixes, such as `dc` for Dublin Core, `foaf` for FOAF, `schema` for schema.org, and `bibo` for BIBO ontology, are used to abbreviate URIs. The empty prefix is used for `arkivo`.

ARKIVO has been developed using the OWL 2 DL profile [23]. In the following, we describe some of the classes, properties and axioms of the ontology[15]. Furthermore, a graphical representation of ARKIVO is shown in Fig. 1.

Some of the main classes in ARKIVO are `bibo:Collection`, which represents the set of documents or collections, and `:Item`, which is the smallest indivisible unit of an archive. In order to model the different categories of collections as well as describe the structure of the archive, different subclasses of the class `bibo:Collection` are asserted, as shown below using the DL syntax:

$$\begin{aligned} :Fonds &\sqsubseteq bibo:Collection \\ :File &\sqsubseteq bibo:Collection \\ :Series &\sqsubseteq bibo:Collection \end{aligned}$$

Using existential quantification property restriction (`owl:someValuesFrom`), we define that the individuals of class `:Item` must be linked to individuals of class `:Fonds` by the `schema:isPartOf` property:

$$:Item \sqsubseteq \exists\, schema:isPartOf.:Fonds$$

This means that there is an expectation that every instance of `:Item` is part of a collection, and that collection is a member of the class `:Fonds`. This is useful to capture incomplete knowledge. For example, if we know that the individual `:701.180/11884` is an item, we can infer that it is part at least of one collection.

[10] http://schema.org.

[11] http://bibliontology.com.

[12] http://www.foaf-project.org.

[13] http://www.geonames.org/ontology/documentation.html.

[14] http://linkedevents.org/ontology/.

[15] The full ARKIVO documentation is available at https://github.com/ArkivoTeam/ARKIVO.

We also define union of classes for those classes that perform a specific function on the ontology. In this case, we used `owl:unionOf` constructor to combine atomic classes to complex classes, as we describe in the following:

$$:CreativeThing \;\equiv\; bibo:Collection \;\sqcup\; :HistoricalEvent \;\sqcup\; :Item$$

This class denotes things created by agents and it includes individuals that are contained in at least one of the classes `bibo:Collection`, `:HistoricalEvent` or `:Item`.

$$:NamedThing \;\equiv\; schema:Place \;\sqcup\; :Date \;\sqcup\; foaf:Agent$$

It refers to things, such as date, place and agent, and it includes individuals that are contained in at least one of the classes `schema:Place`, `:Date` or `foaf:Agent`. Individuals of class `:NamedThing` are connected to individuals of class `:CreativeThing` using the `schema:mentions` object property.

$$:GlamThing \;\equiv\; bibo:Collection \;\sqcup\; :Item$$

GLAM is the acronym of Galleries, Libraries, Archives and Museums. This class denotes individuals that are or can be stored in a GLAM institution. It includes individuals that are contained in at least one of the classes `bibo:Collection` or `:Item`.

3.2 Piłsudski Archival Collections to Linked Data

In order to support data integration process of combining data residing at different sources, we have used external identifiers. In this way, the resources of Piłsudski Digital Archival Collections have been linked to external datasets of the Linked Data in order to enrich the information provided with each resource. We have selected, among others, the Wikidata[16] and VIAF (Virtual International Authority File)[17], as the most common source of identifiers of people, organizations and historical events. In particular, VIAF is a system that is managed by the Online Computer Library Center (OCLC) with the goal to increase the utility of library authority files by matching and linking widely-used authority files and making that information available on the Web. These datasets appear to be stable, suggesting longevity, and data rich, which increase a chance of finding resources.

For example, we can express in Turtle notation that individuals Pius V (person), Józef Piłsudski Institute of America (organization) and Kampania

[16] https://www.wikidata.org/wiki/Wikidata:Main_Page.

[17] https://viaf.org.

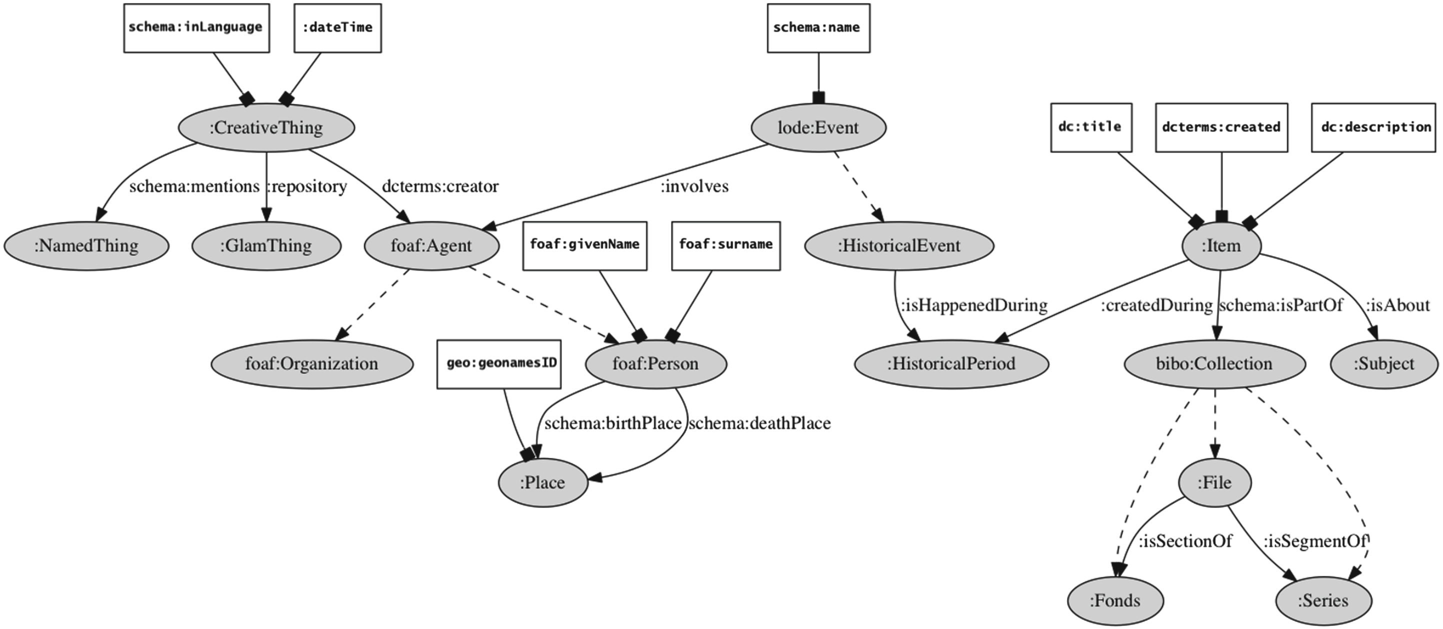

Fig. 1. Graph representation depicts some of the main classes and properties of the ARKIVO ontology. The classes are drawn as labeled ellipses and object properties between classes are shown as labeled edges, while dashed edges represent "is a" relationships. Finally, boxes represent data properties.

wrzesniowa (historical event) are linked to the corresponding Wikidata and VIAF resources as follows:

```
:P11373 a foaf:Person;
         schema:name "Pius V"^^xsd:string;
         owl:sameAs wikidata:Q131945;
         owl:sameAs viaf:76309925.

:ORG01 a foaf:Organization;
                              schema:name "Józef Pilsudski Institute of
  America"^^xsd:string;
        owl:sameAs wikidata:Q6320631;
        owl:sameAs viaf:151002901.

:R10001 a :HistoricalEvent;
         schema:name "Kampania wrzesniowa"^^xsd:string;
         owl:sameAs wikidata:Q150812.
```

Concerning geographical places, other than Wikidata and VIAF, we link the resources to Geonames[18], an open-license geographical database that provides RDF description about eight million locations and exposes them as Linked Data. In the following, we report as the place Warszawa is linked to external data:

```
:G10003 a schema:Place;
         schema:name "Warszawa"^^xsd:string;
         owl:sameAs geonames:756135;
         owl:sameAs wikidata:Q270;
         owl:sameAs viaf:146267734.
```

The process of linking the Piłsudski resources to other datasets is still ongoing and we are aiming to enrich our data with more external data sources. The triple store has been implemented using Stardog 5 Community edition [24], and currently the semantic archival collection of the Institute consist of more than 300 K triples.

4 Use Case Queries

In this Section, we describe the exploration of the Piłsduski archival collection with typical domain experts and historians information needs. Concerning the main information needs in querying archival collections, we can identify the following categories:

Collection. One of the main challenges is to detect the type of collections, such as fonds, file, and series, which are relevant to their research. In this context, also the collections' characteristics can provide meaningful cues. Moreover, relationships among records represent a primary issue.

Agent. People, organizations and authors are part of this category. Domain experts are interested, e.g., in people mentioned in documents, authors of documents as well as organizations that hold specific archival collection.

[18] http://www.geonames.org/ontology/.

Repository. The provenance of an archival collection plays a central role when accessing archival materials. The creator or the institution holding the collection, together with the name of the creator, are among the most common information used in archival queries.

Date and Place. Date and places may indicate specific references to the document. This data can be related with collections, e.g., date of creation, place mentioned in documents, or with biographical data.

Topic Information. Topics provide a way to find more content about a subject and do targeted searching in archival collections. In this regard, abstracts and descriptions of collections are used to reduce the amount of time spent in research.

We collected a set of queries formulated by domain experts. In the following, we illustrate some queries and their results. Notice that for each query in natural language, we wrote a corresponding query in SPARQL language (Table 2).

Query 01. Find all items, created between 1950 and 1955, that mention people with the surname “Churchill” and the place“Polska”. Thus, return the title of the item, the title of the collection to which it belongs and possibly the name of its author.

```
SELECT distinct ?itemTitle ?collectionTitle ?author
WHERE {
        ?item a :Item .
    ?item dc:title ?itemTitle .
    ?item schema:isPartOf ?collection .
    ?collection dc:title ?collectionTitle .
    ?item dcterms:created ?d .
         FILTER (?d >"1950"^^xsd:gYear && ?d < "1955"^^xsd:gYear) .
    ?item schema:mentions ?p .
    ?p foaf:surname "Churchill" .
    ?item schema:mentions ?place .
    ?place a schema:Place .
    ?place schema:name "Polska" .
             OPTIONAL {?item dcterms:creator ?name .
                   ?name schema:name ?author } .
     }
```

Table 2. SPARQL query **01** results

?itemTitle	?collectionTitle	?author
Wycinek z Orła Białego: Słowa papieża o Polsce	Wycinki prasowe dotyczace Watykanu	
15 rocznica śmierci Józefa Piłsudskiego	Artykuły prasowe na temat Józefa Piłsudskiego	
Sprawozdanie prezesa Rady Ministrów	Projekt zjazdu dyplomatów	Odzierzyński, Roman

Query 02. Return the number of items that belong to collection focused on the historical event “Kampania wrześniowa” (Invasion of Poland, in English). Return also the collection title and the name of the institution which houses that collection (Table 3).

```
SELECT ?title ?collection ?archive (count(distinct ?item) as ?triple)
WHERE {
    ?item a :Item .
    ?item schema:isPartOf ?collection .
    ?collection a :CreativeThing .
    ?collection :isAbout ?event .
    ?collection dc:title ?title .
    ?event a :HistoricalEvent .
    ?event schema:name "Kampania wrzeniowa" .
    ?organization :repository ?collection .
    ?organization schema:name ?archive .
     }
GROUP BY ?title ?collection ?archive
```

Table 3. SPARQL query **02** results

?title	?collection	?archive	?triple
Wojny Polskie	http://pilsudski.org/resources/A701.025	Piłsudski Institute of America	184

Query 03. Find all items created before or in 1936, which belong to a file collection and mention both Benito Mussolini and Adolf Hitler. Moreover, from the external dataset of the Italian Chamber of Deputies find information about the Premier office held by Mussolini and eventually some data pertaining his bio. Thus, return the title of the items, their creation date, the file resources, information about Mussolini’s Premier office and his biography data (Table 4).

```
PREFIX ocd: <http://dati.camera.it/ocd/>
SELECT ?title ?date ?file ?office ?bio
WHERE {
    ?item a :Item .
    ?item schema:isPartOf ?file .
    ?file a :File .
    ?item dcterms:created ?date .
        FILTER (?date <= ’1936’^^xsd:gYear) .
    ?item dc:title ?title .
    ?item schema:mentions ?p .
    ?p schema:name ?person1 .
    ?item schema:mentions ?otherperson .
    ?otherperson schema:name ?person2 .
        FILTER (?person1 = "Mussolini, Benito" && ?person2 = "Hitler, Adolf").
SERVICE <http://dati.camera.it/sparql> {
     ?s owl:sameAs <http://dbpedia.org/resource/Benito_Mussolini> .
     ?s ocd:rif_presidenteConsiglioMinistri ?off .
     ?off dc:title ?office .
         OPTIONAL {?s dc:description ?bio }.
       }
      }
```

Table 4. SPARQL query **03** results

?title	?date	?file	?office	?bio
Kariera Józefa Piłsudskiego	1935-05-13	A701.001.087	Presidente del Consiglio dei Ministri dal 31.10.1922 al 25.07.1943	Insegnante di scuole superiori, Pubblicista / Giornalista

5 Conclusions and Future Work

In this paper, we have presented our work in building the semantic layer of the Piłsudski Institute of America digital archive, which included the development of ARKIVO ontology, the representation of the Piłsudski archival collections with ARKIVO, the connection of these resources with external datasets and finally some of the domain experts' queries.

It is well-known that in order to be successfully and efficiently used both from end-users and service providers in practical cases, digital archives entail the need to deal with some critical issues. The first issue is related to the problem of maintaining updated an ontology-based application. In fact, manual ontology population is a time consuming task and it requires professional expertise for detecting extractable data. As future work, we will investigate methodologies for the automatizing of the ontology population process exploiting the techniques presented in [25,26]. The second issue deals with query answering over large ontology-enriched datasets. Query answering is not a simple retrieval procedure of explicit facts but involves some inference mechanism capable of discovering new information. In this context, our aim will be to provide fast and efficient query answering over large knowledge bases and thus allow user to build complex queries. The third issue concerns the ability to access, retrieve and use data by non-expert users, namely those who lack technical or domain knowledge skills. In this respect, we are planning to provide some visual tools for querying semantic data that would support users to easily explore all the interesting relationships that arise from encountering a single document in the archive.

References

1. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Sci. Am. **284**(5), 28–37 (2001)
2. Hitzler, P., Krotzsch, M., Rudolph, S.: Foundations of Semantic Web Technologies. CRC Press, Boca Raton (2009)
3. Alemu, G., Stevens, B., Ross, P., Chandler, J.: Linked data for libraries: benefits of a conceptual shift from library-specific record structures to rdf-based data models. New Libr. World **113**(11/12), 549–570 (2012)
4. Coyle, K., Hillmann, D.: Resource description and access (RDA): cataloging rules for the 20th century. D-Lib. **13**(1/2) (2007)
5. Tennant, R.: MARC must die. Libr. J. New York **127**(17), 26–27 (2002)
6. Antoniou, G., Van Harmelen, F.: Web ontology language: owl. In: Handbook on Ontologies, pp. 91–110. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-24750-0_4

7. Baader, F., Lutz, C.: Description logic. In: Studies in Logic and Practical Reasoning, vol. 3, pp. 757–819. Elsevier, Amsterdam (2007)
8. Krötzsch, M., Simancik, F., Horrocks, I.: A description logic primer. arXiv preprint arXiv:1201.4089 (2012)
9. Bizer, C., Heath, T., Berners-Lee, T.: Linked data - the story so far. In: Semantic Services, Interoperability and Web Applications: Emerging concepts, pp. 205–227 (2009)
10. Guarino, N., Oberle, D., Staab, S.: What is an ontology? In: Staab, S., Studer, R. (eds.) Handbook on Ontologies. IHIS, pp. 1–17. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-92673-3_0
11. Brickley, D., Guha, R.V., McBride, B.: RDF schema 1.1. W3C Recommendation **25**, 2004–2014 (2014)
12. Motik, B., Grau, B.C., Horrocks, I., Wu, Z., Fokoue, A., Lutz, C., et al.: Owl 2 web ontology language profiles. W3C Recommendation **27**, 61 (2009)
13. Beckett, D., Berners-Lee, T., Prud'hommeaux, E., Carothers, G.: RDF 1.1. turtle-terse RDF triple language. W3C Recommendation (2014)
14. Harris, S., Seaborne, A., Prud'hommeaux, E.: SPARQL 1.1 query language. W3C Recommendation **21**(10), 806 (2013)
15. Kruk, S., Haslhofer, B., Piotr, P., Westerski, A., Woroniecki, T.: The role of ontologies in semantic digital libraries. In: European Networked Knowledge Organization Systems (NKOS) Workshop (2006)
16. Kruk, S.R., McDaniel, B.: Semantic Digital Libraries. Springer, Heidelberg (2009). https://doi.org/10.1007/978-3-540-85434-0
17. Doerr, M., Gradmann, S., Hennicke, S., Isaac, A., Meghini, C., van de Sompel, H.: The europeana data model (EDM). In: World Library and Information Congress: 76th IFLA General Conference and Assembly, pp. 10–15 (2010)
18. De Boer, V., et al.: Supporting linked data production for cultural heritage institutes: the Amsterdam museum case study. In: Simperl, E., Cimiano, P., Polleres, A., Corcho, O., Presutti, V. (eds.) ESWC 2012. LNCS, vol. 7295, pp. 733–747. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-30284-8_56
19. Ide, N., Woolner, D.: Exploiting semantic web technologies for intelligent access to historical documents. In: LREC, Citeseer (2004)
20. Pietrzyk, P.: A brief history of the mission and collections of the piłsudski institute of America for research in the modern history of Poland. Pol. Am. Stud. **60**, 91–98 (2003)
21. Pandolfo, L., Pulina, L., Zielinski, M.: Towards an ontology for describing archival resources. In: Proceedings of the Second Workshop on Humanities in the Semantic Web (WHiSe II) Co-located with 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, pp. 111–116, 22 October 2017
22. Gangemi, A., Presutti, V.: Ontology design patterns. In: Handbook on Ontologies, pp. 221–243. Springer, Heidelberg (2009). https://doi.org/10.1007/11574620_21
23. Grau, B.C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U.: Owl 2: the next step for owl. Web Semant. Sci. Serv. Agents World Wide Web **6**(4), 309–322 (2008)
24. Inc., C.: Stardog 5: The manual (2017). http://docs.stardog.com/. Accessed June 2018

25. Pandolfo, L., Pulina, L., Adorni, G.: A framework for automatic population of ontology-based digital libraries. In: Adorni, G., Cagnoni, S., Gori, M., Maratea, M. (eds.) AI*IA 2016. LNCS (LNAI), vol. 10037, pp. 406–417. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-49130-1_30
26. Pandolfo, L., Pulina, L.: ADnOTO: a self-adaptive system for automatic ontology-based annotation of unstructured documents. In: Benferhat, S., Tabia, K., Ali, M. (eds.) IEA/AIE 2017. LNCS (LNAI), vol. 10350, pp. 495–501. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-60042-0_54