ARKIVO: an Ontology for Describing Archival Resources

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

Abstract. In this paper we present ARKIVO, an ontology designed to accommodate the archival description of historical document collections. The aim of ARKIVO is to provide a reference schema for a rich representation of data elements in digital historical archives. This paper briefly reports design and implementation of ARKIVO, as well as its application on a real world case study, namely the Józef Piłsudski Institute of America digitized collections.

1 Context & Motivation

Information technologies changed the way of doing archival research. The real change happened in the 1990s, after the advent of the Web, when a great number of historical documents were published online stored in digital archives, by providing the users the possibility to have a direct access to millions of documents in a rapid and easy way.

Recently, digital archives are facing new challenges in order to overcome traditional data management and information browsing. In this context, Semantic Web (SW) [1] technologies can improve digital archives by facilitating archival metadata storage and adding semantic capabilities, which increase the quality of the information retrieval process. In particular, ontologies, which are defined as a formal specification of domain knowledge conceptualization [2], play a key role in several aspects, e.g., resources description by means of taxonomies and vocabularies in order to promote interoperability and consistency between different sources [3].

In the last decade, there has been a great amount of effort in designing vocabularies and metadata standards to catalogue documents and collections, such as Functional Requirements for Bibliographic Records (FBRB) [1], MAchine-Readable Cataloging (MARC) [2], Metadata Object Description Schema (MODS) [3], and Encoded Archival Description (EAD) [4], just to cite a few well-known examples. Metadata standards such as FRBR, EAD and MODS seem to be more

[1] http://www.cidoc-crm.org/frbroo/
[2] https://www.loc.gov/marc/
[3] http://www.loc.gov/standards/mods/
[4] https://www.loc.gov/ead/

devoted to human consumption rather than machine processing [4], while concerning MARC some experts experienced that it is not suitable neither for machine processable nor for actionable metadata [5, 6]. 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.

In this paper we present ARKIVO, 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 [7], but also to describe the relevant elements contained in these documents.

The paper is organized as follows. In Section 2, we describe the design process of ARKIVO and we give some details about the implementation. In Section 3, we describe the application of ARKIVO to the digitized collections of the Józef Piłsudski Institute of America, and we conclude the paper in discussing future work.

2 ARKIVO Ontology

In this Section, we describe the design process of ARKIVO, in order to define key concepts and relations. We also give some insights about its implementation.

As a starting point, we considered the archival management practices and the most common methods used by archives for storing and cataloging materials. Archival resources are typically organized following a hierarchy composed of different layers (i.e., fonds, file, series, item), in which *fonds* represent the highest level of that structure, while *item* the lowest.

In the design phase, we considered the best practices used by archive workers in the metadata collection process. Usually, this tasks is conducted by selecting items and trying to locate, in addition to a title and abstract, key persons, places, dates and events mentioned in the item. These data are especially important in describing an historical document. One can sometimes identify the author, or document creation date, in which case the assignment of the role can be more specific. In most cases, however, one can only note that the name, place or date was mentioned in the document. We decided to model these categories of information into the ontology, in order to give all the useful elements to support the annotation process.

Next, we examined some existing standards and metadata vocabularies potentially compatible with the considered domain, in order to re-use them as core ontologies. In the following, we report core ontologies and vocabularies used in ARKIVO. `Dublin Core` [8] and `schema.org` [9] represent the most used vocabularies for describing and cataloging both Web and physical resources, such as

Web pages and books. Due to the generic nature of their terms, we also refer to `BIBO`[5] ontology in order to have a detailed and exhaustive document classification. Concerning the description of personal information, we decided to use `FOAF`[6], a widely used vocabulary to describe persons as well as organizations. ARKIVO uses also `Geonames`[7] ontology for linking a place name to its geographical location and `LODE` [10] ontology which is one of the most often-used model for publishing events as Linked Data.

ARKIVO has been developed using the OWL 2 DL language [11]. In the following, we report some implementation details of the ontology[8]. The main classes in ARKIVO are `Collection`, which represents the set of documents or collections, and `Item`, which is the smallest indivisible unit of an archive. In order to describe the structure of the archive, different subclasses of the class `Collection` are modeled, namely the subclasses *Fonds*, *File* and *Series*. Using existential quantification property restriction (`owl:someValuesFrom`), we defined that the class `Item` as the class of individuals that are linked to individuals in the class `Fonds` by the `isPartOf` property, as shown below using the DL syntax [12]

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

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.

Moreover, we defined some 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:

$$CreativeThing \equiv Collection \sqcup HistoricalEvent \sqcup Item$$

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

$$NamedThing \equiv Place \sqcup Date \sqcup Agent$$

It refers to things, such as date, place and agent, that are related to individuals in the `CreativeThing` by the object property *isMentionedIn*, and it includes individuals that belong to at least one of the classes `Place`, `Date` or `Agent`.

3 Case Study: the Józef Piłsudski Archival Collections

We applied the ARKIVO ontology to describe the archival holdings, partially digitized, of the Piłsudski Institute of America. In order to support data integration process of combining data residing at different sources, we used external identifiers. In this way, the resources of Piłsudski Digital Archival Collections have been linked to external

[5] `http://bibliontology.com/`
[6] `http://www.foaf-project.org/`
[7] `http://www.geonames.org/ontology`
[8] The full ARKIVO documentation is available at `https://github.com/ArkivoTeam/ARKIVO`

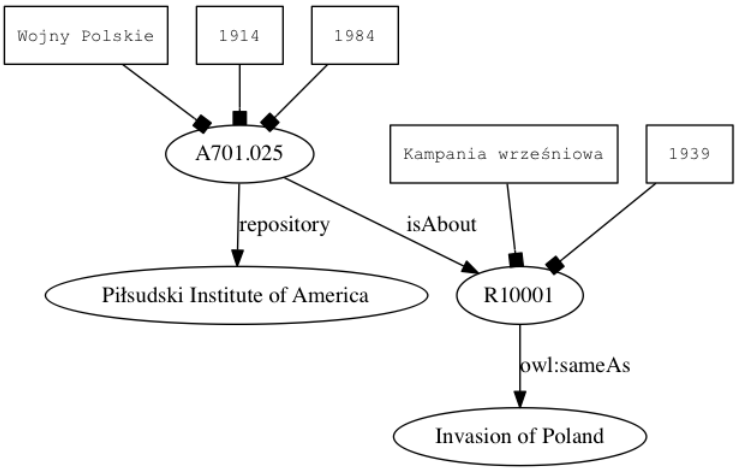

Fig. 1. Graphical representation of archival collections data using ARKIVO.

datasets of the Linked Data Cloud in order to enrich the information provided with each resource. We selected, among others, Wikidata[9], DBpedia[10], and VIAF (Virtual International Authority File)[11], as the most common source of identifiers of people, organizations and historical events.

In Figure 1, we report an example of individuals and properties stored in the Piłsudski digital archive, and how these data can be connected to resources belonging to external datasets. The individuals are drawn as labeled ellipses, object properties between individuals are shown as labeled edges, while boxes represent data properties related to individuals. Looking at this figure, we can see that the fonds "A701.025" is stored by the organization "Piłsudski Institute of America" and it has different data properties, such as its title, i.e., Wojny Polskie, its creation start date, i.e., 1914, and end date, i.e., 1984. The object property `isAbout` reveals the main topic addressed in this fonds, represented by the individual "R10001", which has its title, i.e., Kampania wrześniowa, and its date, i.e.,1939. The individual "R10001", which belongs to both the class `HistoricalEvent` and the class `Subject`, is in relationship to the DBpedia's instance "Invasion of Poland" through the property `owl:sameAs`, which indicates that these two URI references actually refer to the same thing. The integration with external dataset, such as DBpedia, makes it possible to discover and share information.

The archival collections of the Piłsudski Institute of America can be effectively explored and queried at `http://stardog.vuotto.tech/#/databases/arkivo_x3`[12]. The triple store has been implemented using Stardog 5 Community edition [13].

[9] `https://www.wikidata.org/wiki/Wikidata:Main_Page`
[10] `https://wiki.dbpedia.org`
[11] `https://viaf.org`
[12] We provided a test user with username `user_test` and password `test4jpi`.

4 Conclusion and Future Work

In this paper we briefly presented ARKIVO, an ontology designed to accommodate the archival description of historical document collections. We reported the current usage of ARKIVO in the context of the historical archive of the Józef Piłsudski Institute of America. In our work on the Józef Piłsudski archival collections, we collected approximately 300 thousand triples and its knowledge base is populated, among others, by 14,199 authors, 12,848 collections, 28,8644 items, 1,572 places and 6,615 dates.

Currently, we are working on the realization of the ontology-based digital archive of the Piłsudski Institute. As future work, we will investigate methodologies for the automatization of the ontology population process exploiting the techniques presented in [14, 15].

References

1. Berners-Lee, T., Hendler, J., Lassila, O., et al.: The semantic web. Scientific american **284**(5) (2001) 28–37
2. Guarino, N., Oberle, D., Staab, S.: What is an ontology? In: Handbook on ontologies. Springer (2009) 1–17
3. Kruk, S.R., McDaniel, B.: Semantic Digital Libraries. Springer (2009)
4. 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 Library World **113**(11/12) (2012) 549–570
5. Coyle, K., Hillmann, D.: Resource description and access (rda): Cataloging rules for the 20th century. D-Lib **13**(1/2) (2007)
6. Tennant, R.: Marc must die. LIBRARY JOURNAL-NEW YORK- **127**(17) (2002) 26–27
7. Bizer, C., Heath, T., Berners-Lee, T.: Linked data – the story so far. Semantic services, interoperability and web applications: emerging concepts (2009) 205–227
8. Weibel, S.L., Koch, T.: The dublin core metadata initiative. D-lib magazine **6**(12) (2000) 1082–9873
9. Patel-Schneider, P.F.: Analyzing schema.org. In: International Semantic Web Conference, Springer (2014) 261–276
10. Shaw, R., Troncy, R., Hardman, L.: Lode: Linking open descriptions of events. In: Asian Semantic Web Conference, Springer (2009) 153–167
11. Grau, B.C., Horrocks, I., Motik, B., Parsia, B., Patel-Schneider, P., Sattler, U.: Owl 2: The next step for owl. Web Semantics: Science, Services and Agents on the World Wide Web **6**(4) (2008) 309–322
12. Baader, F., Lutz, C.: Description logic. In: Studies in Logic and Practical Reasoning. Volume 3. Elsevier (2007) 757–819
13. Inc., C.: Stardog 5: The manual. Available online: http://docs. stardog.com/. Last accessed on June 2018 (2017)
14. Pandolfo, L., Pulina, L., Adorni, G.: A framework for automatic population of ontology-based digital libraries. In: AI* IA 2016 Advances in Artificial Intelligence. Springer (2016) 406–417
15. Pandolfo, L., Pulina, L.: Adnoto: A self-adaptive system for automatic ontology-based annotation of unstructured documents. In: To appear in Proc. of the 30th International Conference on Industrial, Engineering, Other Applications of Applied Intelligent Systems. Springer (2017)