Towards an Ontology for Describing Archival Resources

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

[1] POLCOMING, Università di Sassari, Viale Mancini n. 5 – 07100 Sassari – Italy
laura.pandolfo@uniss.it, lpulina@uniss.it
[2] Pilsudski Institute of America, 138 Greenpoint Avenue, Brooklyn, NY 11222 – USA
MZielinski@pilsudski.org

Abstract. Several digital libraries and archives are emerging around the world due to the need to store, organize and make available on the Web a lot of resource collections. However, managing this information poses new challenges in order to overcome traditional data management and information browsing. Semantic Web technologies can improve digital libraries and archives by facilitating metadata storage and adding semantic capabilities, which increase the quality of the information retrieval process. In this paper we present ARKIVO, an ontology designed to model the archival description of historical document collections.

1 Introduction

The Web changed the way people can search and discover information providing them the opportunity to have direct access to millions of documents easily. Online repositories, such as digital libraries, support users' exploration of large document collections and, as in the case of digital historical archives, also facilitate access to original and rare documents. Recently, digital archives are facing new challenges in order to overcome traditional data management and information browsing. The Semantic Web (SW) [2] technologies provide ways to address these challenges by offering valuable solutions to represent, organize, and retrieve such kind of data. In particular, ontologies play a key role providing a common shared vocabulary that can be used to describe domains, annotate documents and promote interoperability and consistency between different sources [9, 10].

In the context of digital libraries and archives, some of the most used metadata and ontologies include Dublin Core Metadata Initiative (DCMI) [17], CIDOC Conceptual Reference Model (CRM) [4], MAchine-Readable Cataloging (MARC), Metadata Object Description Schema (MODS) [8], and Encoded Archival Description (EAD) [14]. However, none of these can exhaustively support both the representation of the archival arrangement structure and the annotation of historical data embedded within the documents – the importance of which has been highlighted in, e.g., [1].

To address these needs, in this paper we introduce ARKIVO[1], an ontology designed to accommodate the description of historical archival documents, support-

[1] Arkivo is the translation of "Archive" in Esperanto.

ing archive workers by encompassing both the hierarchical structure of archival collections and rich metadata created during archive digitization, such as historical elements. The aim of ARKIVO is not only to provide a reference schema for publishing Linked Data [3] about historical archival documents, but also to describe the historical elements contained in these documents, e.g., giving the opportunity to represent useful relationships between people, places, and events. In this paper, we also describe the usage of ARKIVO in the context of the historical archive stored by the Józef Piłsudski Institute of America, which houses a rich collection of historical sources covering the period from the 1863 to the present day.

The paper is organized as follows. In Section 2, we briefly the ARKIVO ontology and its design process, while in Section 3 we show the usage of ARKIVO in the context of the digitized collections of the Józef Piłsudski Institute of America. We conclude the paper in Section 4 with some final remarks and future work.

2 The ARKIVO Ontology

The ontology development process can be characterized by different strategies and methodologies – see, e.g., [16, 7]. ARKIVO has been developed 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. In the following, we report the main phases of the development process of ARKIVO, which have been carried out with the support of the experts.

Requirements Specification and Knowledge Acquisition. In this phase, we considered different scenarios, use-cases and end-users, focusing on the archival management practices and the most common methods used by archives for storing and cataloging materials. Moreover, we analyzed the best practices used by archive workers in the metadata collection process. This phase allowed us to detect the main concepts useful to represent the domain of interest.

Conceptualization and Formalization. In the light of the knowledge gained in the previous phase, we have drawn up a glossary of terms that identify the proper terminology used in the archival domain. The aim of the conceptualization resulting from this activity was intended to structure the domain knowledge, in terms of concepts and relations, in order to meet the pre-established requirements. In particular, we compute a taxonomy for describing the archival arrangement levels, from the concept of *collection*, which can contain items or other collections as *fonds*, to the concept of single *item*, which typically is the smallest indivisible unit.

Integration. Some of the concepts resulting from the conceptualization phase can be represented by reusing existing standard metadata and vocabularies. For this purpose, we integrated ARKIVO with the several core ontologies and vocabularies. In details, DCMI, `FOAF`[2], and `schema.org` [13] were used to model some

[2] `http://www.foaf-project.org/`

general information related to documents, organizations, places and persons. We also referred to `BIBO`[3] ontology in order to have a detailed classification of documents. In order to link a place name to its current geographical location, we used `Geonames`[4]. Finally, we integrated `LODE` [15] ontology to model events and their properties.

Implementation. ARKIVO has been developed using the OWL2 language [6] with the PROTÉGÉ [5] editor. The ontology is composed of 43 classes, 24 object properties, and 34 data properties. In the following, we pinpoint some of the main classes and properties of ARKIVO ontology. Notice that we include the core ontologies prefixes, namely `dc` for Dublin Core, `foaf` for FOAF, `schema` for schema.org, and `bibo` for BIBO ontology. Finally, the empty prefix is used for original classes and properties of `arkivo`.

`bibo:Collection` is the class that represents set of documents or collections. This class has several sub-classes, including `:File` and `:Fonds`. The former is the class devoted to describe a file, namely an organized unit of items grouped together, while the latter relates the whole of the records organically created and/or accumulated by a particular person, family, or corporate body in the course of that creator's activities and functions.

`:Date` is the class containing dates mentioned in an item.

`foaf:Organization` is used to describe an organization related to bibliographic items or to events.

`foaf:Person` represents people related to a bibliographic item or to a specific event.

`:Item` represents the archival item, in other words the smallest intellectually indivisible archival unit. This class contains several sub-classes, such as `bibo:Article`, `:Document` and `bibo:Letter`.

`dct:creator` is the relationship that shows who has created a specific item, connecting individuals in `:Item` class to individuals in `foaf:Agent` class.

`dc:created` indicates the date when it was created an individual of the class `Item`.

`schema:isPartOf` indicates that an individual in the class `:Item` is part of a collection, by linking that individual to another in the class `bibo:Collection`.

`schema:mentions` is useful to indicate that an instance of `foaf:Person` and/or `schema:Place` is mentioned in an individual of the class `bibo:Collection`.

`:isSectionOf` connects instances of `:File` to instances of `:Fonds`.

`:repository` connects instances of `foaf:Organization` class to instances of `bibo:Collection` class, in order to describe that an organization can be a repository of collections or items.

ARKIVO ontology is licensed under a Creative Commons Attribution 3.0 Unported License and it can be downloaded at `http://purl.org/arkivo`. For more details about the full list of classes and properties see also the documentation at `https://github.com/ArkivoTeam/ARKIVO`.

[3] `http://bibliontology.com/`

[4] `http://www.geonames.org/ontology`

3 Case Study: the Józef Piłsudski Digital Archival Collections

The Józef Piłsudski Institute of America[5] was established 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. The Polish State was re-established in 1918 in the aftermath of the Great War and after several regional wars and uprisings, the borders were settled in 1922. Soon after 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. Today, the Institute has some 240 linear meters, namely 2 million pages, of archives covering mostly the Polish, European and American history of late 19th and 20th century. The collection includes documents, photographs, films, posters, periodicals, books, personal memoirs of diplomats, and political and military leaders, as well as collection of paintings by Polish and European masters. For the last nine years, the archival collections are being digitized, and gradually put online.

The main objective of the historical research is to understand the past through the study of historical sources, such as documents stored in archives. In this context, researches are mainly interested in detecting facts (e.g., people, places, events) cited in the documents in order to analyze them, discover relationships and draw inferences. ARKIVO ontology, unlike, e.g., EAD, provides elements to represent both the hierarchical structure of archival documents and the historical data expressed in them.

As an example, in the following we report the description (in Turtle language) of one of the document stored in the Józef Piłsudski Institute archive, namely the "Letter to comrades in London". Such document has been wrote by Piłsudski in 1898, and it contains a mention of different people and places, as depicted in Figure 1.

```
:LetterToComradesInLondon a bibo:Letter .
:A701.001.012 a :File .
:A701.001 a :Fonds .
:PilsudskiInstitute a foaf:Organization .
:PilsudskiJosef a foaf:Person .
:JedrzejowskiBoleslaw a foaf:Person .
:MalinowskiAleksander a foaf:Person .
:Sachalin a schema:Place .
:Bialystok a schema:Place .
:LetterToComradesInLondon schema:isPartOf :A701.001.012 .
:A701.001.012 schema:isSectionOf :A701.001 .
:A701.001 :repository :PilsudskiInstitute .
:LetterToComradesInLondon dc:creator :PilsudskiJosef .
```

[5] http://www.pilsudski.org/

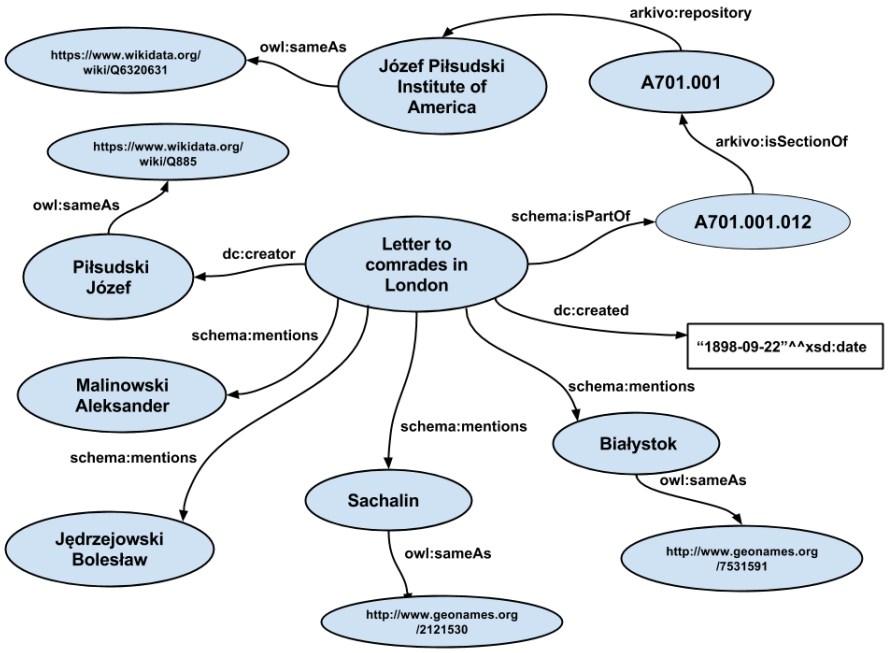

Fig. 1. A graphical example of entities and relationships in Piłsudski digitized collections using ARKIVO.

```
:LetterToComradesInLondon schema:mentions :JedrzejowskiBoleslaw .
:LetterToComradesInLondon schema:mentions :MalinowskiAleksander .
:LetterToComradesInLondon schema:mentions :Sachalin .
:LetterToComradesInLondon schema:mentions :Bialystok .
```

Finally we report that, actually, in the version of ARKIVO used for the Józef Piłsudski archival collections are stored about 270,000 triples, and it is populated by more than 130,000 individuals. In detail, there are 13,326 individuals related to items, 15,678 titles, 6,458 authors, 29,280 persons mentioned, 47,185 places, and 28,039 dates.

4 Conclusion and Future Work

In this paper we briefly presented ARKIVO, an ontology designed to model the archival description of historical document collections. In the paper we also show the current usage of ARKIVO in the context of the historical archive of the Józef Piłsudski Institute of America. Currently, we are working on the realization of an ontology-based digital archive.

Future work will include the implementation of automated and adaptive ontology population processes exploiting the techniques presented in [12, 11], as

well as the investigation of user interfaces aimed at providing the user with a rich interface to explore interesting relationships that arise from encountering a single item or file in the archive. It should help the users to find the unexpected and hidden knowledge accumulated both in the archive and in the Web.

References

1. Giovanni Adorni, Marco Maratea, Laura Pandolfo, and Luca Pulina. An ontology for historical research documents. In *International Conference on Web Reasoning and Rule Systems*, pages 11–18. Springer, 2015.
2. Tim Berners-Lee, James Hendler, Ora Lassila, et al. The semantic web. *Scientific american*, 284(5):28–37, 2001.
3. Christian Bizer, Tom Heath, and Tim Berners-Lee. Linked data – the story so far. *Semantic services, interoperability and web applications: emerging concepts*, pages 205–227, 2009.
4. Martin Doerr. The cidoc conceptual reference module: an ontological approach to semantic interoperability of metadata. *AI magazine*, 24(3):75, 2003.
5. John H Gennari, Mark A Musen, Ray W Fergerson, William E Grosso, Monica Crubézy, Henrik Eriksson, Natalya F Noy, and Samson W Tu. The evolution of protégé: an environment for knowledge-based systems development. *International Journal of Human-computer studies*, 58(1):89–123, 2003.
6. Bernardo Cuenca Grau, Ian Horrocks, Boris Motik, Bijan Parsia, Peter Patel-Schneider, and Ulrike Sattler. Owl 2: The next step for owl. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(4):309–322, 2008.
7. Stephan Grimm, Andreas Abecker, Johanna Völker, and Rudi Studer. Ontologies and the semantic web. In *Handbook of Semantic Web Technologies*, pages 507–579. Springer, 2011.
8. Rebecca S Guenther. Mods: the metadata object description schema. *Portal: libraries and the academy*, 3(1):137–150, 2003.
9. Sebastian Kruk, Bernhard Haslhofer, P Piotr, Adam Westerski, and Tomasz Woroniecki. The role of ontologies in semantic digital libraries. In *European Networked Knowledge Organization Systems (NKOS) Workshop*, 2006.
10. Sebastian Ryszard Kruk and Bill McDaniel. *Semantic digital libraries*. Springer, 2009.
11. Laura Pandolfo and Luca Pulina. 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.
12. Laura Pandolfo, Luca Pulina, and Giovanni Adorni. A framework for automatic population of ontology-based digital libraries. In *AI* IA 2016 Advances in Artificial Intelligence*, pages 406–417. Springer, 2016.
13. Peter F Patel-Schneider. Analyzing schema.org. In *International Semantic Web Conference*, pages 261–276. Springer, 2014.
14. Daniel V Pitti. Encoded archival description: An introduction and overview. 1999.
15. Ryan Shaw, Raphaël Troncy, and Lynda Hardman. Lode: Linking open descriptions of events. In *Asian Semantic Web Conference*, pages 153–167. Springer, 2009.
16. Mike Uschold and Michael Gruninger. Ontologies: Principles, methods and applications. *The knowledge engineering review*, 11(02):93–136, 1996.
17. Stuart L Weibel and Traugott Koch. The dublin core metadata initiative. *D-lib magazine*, 6(12):1082–9873, 2000.