Web Metadata Editor: a Web Application to Build a Knowledge Base Based on Linked Data

Angelica Lo Duca, Andrea Marchetti

Institute of Informatics and Telematics
National Research Council
via Moruzzi 1, 56124 Pisa, Italy
email: [name].[surname]@iit.cnr.it

Abstract. Within the field of Digital Humanities, a great effort has been made to digitize documents and collections in order to build catalogs and exhibitions on the Web. In this paper, we present WeME, a Web application for building a knowledge base, which can be used to describe digital documents. WeME can be used by different categories of users: archivists/librarians and scholars. WeME extracts information from some well-known Linked Data nodes, i.e. DBpedia and GeoNames, as well as traditional Web sources, i.e. VIAF. As use case of WeME, we describe the knowledge base related to the Christopher Clavius's correspondence. Clavius was a mathematician and an astronomer of the XVI Century. He wrote more than 300 letters, most of which are owned by the Historical Archives of the Pontifical Gregorian University (APUG) in Rome. The built knowledge base contains 139 links to DBpedia, 83 links to GeoNames and 129 links to VIAF.

Keywords: Metadata Editor, Linked Data, Knowledge Base, Digital Humanities

1 Introduction

Over the last years, a great effort has been made in the field of Digital Humanities to digitize documents and collections in different formats, such as PDF, XML, plain texts and images. All these documents are often stored either in digital libraries or big digital repositories in the form of books and catalogs (e.g. the Oxford Digital Library,[1] the Library of Congress,[2] and the Perseus Digital Library.[3]). Sometimes, projects are developed to annotate a subset of texts and images, such as the Clavius on The Web project,[4] [9, 1] where the idea behind the work presented in this paper was originated. Other projects include the Digital Vercelli Book[5] and Burckhardtsource.[6]

[1] http://www.odl.ox.ac.uk
[2] https://www.loc.gov
[3] http://www.perseus.tufts.edu
[4] http://claviusontheweb.it
[5] http://vbd.humnet.unipi.it/beta2/
[6] http://burckhardtsource.org

The process of cataloging requires also the creation of a knowledge base, which contains contextual resources associated to documents of the catalog, such as the authors of the documents and places where documents were written. Information contained in the knowledge base can be used to enrich documents details, i.e. metadata associated to documents. Most of the existing tools for catalog creation allows to build the knowledge base manually, in the sense that the user must insert each piece of information (metadata) one by one. This process is often tedious, because it consists in editing well-known information about a document, such as the author's name and date of birth. In addition, this process is repetitive, because many documents are written by the same author and in the same place thus requiring to write the same information twice or more. In general this manual effort produces three main disadvantages: a) the probability of introducing errors increases, b) the whole process is slowed down because it is not automatic, c) inserted information is isolated, i.e. not connected to the rest of the Web.

In this paper we present the Web Metadata Editor (WeME), a Web application which provides users with a user-friendly interface to build a knowledge base associated to a collection. WeME helps archivists to enrich their catalogs with resources extracted from two kinds of Web sources: Linked Data [3] and traditional Web sources. WeME mitigates the three described disadvantages, produced by manual effort, by extracting well-known metadata from some Linked Data nodes (e.g. DBpedia,[7] [2] GeoNames[8]) and other traditional Web sources (VIAF[9]). WeME exploits semantic and traditional Web to extract information, through the construction of SPARQL [7] and RESTful APIs queries to the Web, in a way totally transparent to the user. In fact, in the Web interface, the user must specify only the name of the resource to be searched. WeME then retrieves information from the Web and shows them to the user, who can decide whether or not to accept, edit or discard them. Through this automatic search of metadata, the process of metadata insertion is acccelerated and the probability of introducing errors is reduced. The advantages derived from WeME are essentially two: firstly WeME eases the task of building a knowledge base; secondly, WeME establishes new relations both among documents within the same catalog and with documents belonging to Web sources.

WeME was used to build the knowledge base related to the Christoper Clavius correspondence. Clavius wrote and received more than 300 letters to and from other scientists of the same period. Among them, Galileo Galilei and Tycho Brahe. Most of these letters are hosted by APUG. Around this correspondence, the Clavius on the Web project (CoW) was started in 2013 and lasted four years.

The remainder of the paper is organized as follows: Section 2 illustrates some related work. In Section 3 we describe the approach employed in this paper, while Section 4 illustrates the Web application. In Section 5 we describe the use case

[7] http://dbpedia.org

[8] http://www.geonames.org

[9] http://viaf.org

of WeME to the Clavius' domain. Finally, in Section 6, we give our conclusions and future work.

2 Related Work

In this section firstly we review the current literature on tools and projects which exploit Linked Data to build knowledge bases and then we briefly illustrate some tools for cataloging.

DaCura [5] is a framework which provides tools to collect and curate high quality linked datasets. DaCura is not thought for digital libraries or digital repositories. However, it covers more aspects that are important in the context of digital humanities, such as data provenance, data quality, etc. Another important initiative is the CULTURA project [6], which develops a metadata-driven personalization environment to navigate collections. In addition, it supports different categories of users, such as professional researchers and simple users. A more recent initative is the FREME project,[10] developed by the group behind DBpedia. FREME provides an interactive editor to identify and annotate entities in texts in an interactive editor. Users are even able to manage the entities discovered. The FREME tool suite furthermore discovers people, place and events.

With respect to the existing tools, frameworks and projects, WeME provides a simple Web application, which does not require any specific skill. In fact, WeME can be used by any kind of user, e.g. scholars and archivists/librarians, as well as students. In addition, WeME can be easy installed and run within a Web server, without any specific configuration. Finally, its source code can be downloaded as open source from the GitHub platform, as described later in the paper.

2.1 Tools for cataloging

Many software tools have emerged recently, making it possible to catalog and manage digital collections. Among proprietary tools, the most famous is CONTENTdm,[11], created by OCLC. CONTENTdm is a digital collection management tool that permits to upload, describe, manage and access digital collections. It is a very powerful tool with an easy-to-use interface. However, its cost is prohibitive for many no profit organizations, i.e. entry level license options start at $4,300 annually.

Open-source software tools include: Omeka,[12] which provides a unified application for the Web interface and back-end cataloging system; Collective Access,[13], whose main focus is on cataloging and multiple metadata schemas; CollectionSpace,[14] which does not permit to create digital collections, but it enables

[10] http://www.freme-project.eu/
[11] http://www.contentdm.org/
[12] http://omeka.org/
[13] http://collectiveaccess.org/
[14] http://www.collectionspace.org/

users to connect with other existing open-source applications; Open Exhibits.[15] a multitouch, multi-user tool, whose main aim is to develop online and interactive exhibits of collections.

Existing tools provide very powerful interfaces to add, edit or delete metadata associated to digital documents, but all these information must be edited by the user, manually. Our tool, instead, exploits Web sources (Linked Data and RESTful APIs) to retrieve contextual resources automatically.

3 Approach

The core idea of this work consists in building a knowledge base which contains contextual resources connected to the documents of a collection, such as the authors and places of the documents. Allowed resources are a subset of the Europeana Data Model (EDM) [8] ontology: *person*, *place* and *cultural heritage object* (CHO). We choose EDM to represent our data because it defines relations among resources in a very efficient way: a CHO is related to a person, if the person is its author, as well as a place is related to a CHO, if the CHO was created in that place.

Every resource can be built through a simple Web interface, which gives the possibility to edit resources manually or by invoking Linked Data and Web RESTful APIs. The user formulates a simple query, based on the pair (name, surname) for people, and (name) for places. The application triggers a call to some remote Web services (e.g. DBpedia, VIAF and GeoNames) to retrieve information associated to the resource, such as the birth place and a description. The user is then free to accept, edit or discard retrieved information and save them to the knowledge base. Then, the user can view, edit and organize in collections her resources. A short video tutorial of WeME can be found at the following link: https://youtu.be/AqQS8N16OhY.

One of the main issues while dealing with different sources regards resources disambiguation. In fact, it can happen that there is a conflict on a given field (e.g. birth date) between two or more sources. Currently, WeME leaves the user the task of performing resources disambiguation. However, as future work, we could organize the sources in to a hierarchy of importance (i.e. associate a score to each source). If a field is found in more than one source, the system could suggest to the user the field provided by the source with the highest priority.

Another aspect of WeME concerns the fact that the built knowledge base is completely self-contained, while still maintaining links to external sources. Another possible approach could consist in updating existing sources, such as DBpedia and GeoNames. However, we preferred to follow the self-contained strategy essentially for four reasons: a) users are able to claim their authorship on their work (i.e. towards academy or funding agencies), b) users can keep control of updates that could break their work, c) avoid delays and blockage in updating data due to validation processes, d) needed resources are too contextual to the

[15] http://openexhibits.org/

dataset and not of sufficient general interest to be accepted in an encyclopedic knowledge database.

4 WeME

The Web Metadata Editor (WeME) provides a Web editor to build a knowledge base, which contains contextual resources, related to digital documents. The application is envisaged for archivists/librarians, but in general it can be used by scholars, students and other people who want to build a knowledge base and connect it to the Web.

4.1 Users

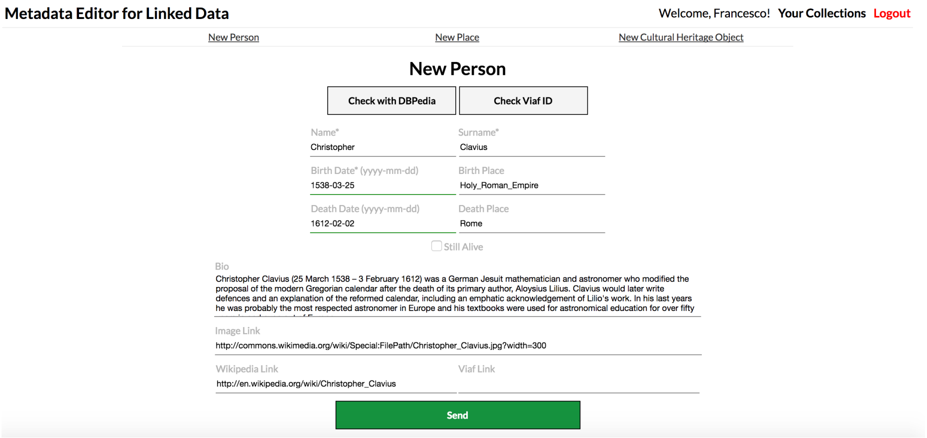
Although WeME can be used by various stakeholders, a distinction should be done between librarians, archivists and scholars [4]. From the point of view of WeME, librarians and archivists can be grouped in the same category. They own very specific skills to create a knowledge base for a collection of documents. Their main interest is capturing all reusable and relevant metadata to facilitate discovery, classification, exploration of catalogs. Scholars, instead, are concerned with compiling a knowledge base for answering their research questions. On the one hand, archivists/librarians may have an expertise in a specific field, for instance history, that facilitate their task. Scholars, on the other hand, do not necessary have this specific background.

WeME tries to satisfy needs of both archivists/librarians and scholars. From the point of view of archivists/librarians, WeME exploits Linked Data to capture common metadata, shared by different resources thus allowing resource reusage and common metadata classification. Regarding scholars, WeME provides a mechanism to link resources both to external sources, such as GeoNames and DBpedia and to internal sources, such as places and people within the same knowledge base. Given these relations, a scholar could execute some reasoning tools to extract new information. At the moment, WeME does not implement reasoning mechanisms. Anyway, it would be interesting to extend it to provide also this feature. WeME differs from the strategy adopted in [4], where two different knowledge bases are built, one for archivists/librarians and the other for scholars. In WeME, instead, only one knowledge base is built to satisfy both needs. In this way, the application is kept simple and there is no replication of information.

4.2 Layout

Figure 1 shows a snapshot of the interface. We defined a layout composed of three views:

1. *Person box*: the editor gives the possibility to add/edit a new person, by specifying the following fields: *name*, *surname*, *birth date*, *birth place*, *death*



Name*: Christopher
Surname*: Clavius
Birth Date* (yyyy-mm-dd): 1538-03-25
Birth Place: Holy_Roman_Empire
Death Date (yyyy-mm-dd): 1612-02-02
Death Place: Rome
□ Still Alive
Bio: Christopher Clavius (25 March 1538 – 3 February 1612) was a German Jesuit mathematician and astronomer who modified the proposal of the modern Gregorian calendar after the death of its primary author, Aloysius Lilius. Clavius would later write defences and an explanation of the reformed calendar, including an emphatic acknowledgement of Lilio's work. In his last years he was probably the most respected astronomer in Europe and his textbooks were used for astronomical education for over fifty
Image Link: http://commons.wikimedia.org/wiki/Special:FilePath/Christopher_Clavius.jpg?width=300
Wikipedia Link: http://en.wikipedia.org/wiki/Christopher_Clavius
Viaf Link:

Fig. 1. A snapshot of WeME.

date, *death place*, *image link*, *Wikipedia link*, *VIAF link*. There is also the checkbox *still alive*, which allows to specify whether the person is or not still alive. The user can edit all the fields, manually, or she can select the *check with DBpedia/check with VIAF* buttons, to populate, if available, the fields from DBpedia/VIAF. When the information is ready, the user can click the *send* button, to store the person in the knowledge base. If the person is already present in the knowledge base, the editor gives an alert.

2. *Place box*: the editor provides a form to add/edit a new place, by specifying the following fields: *original name*, *English name*, *country*, *region*, *population*, *latitude*, *longitude*, *description*, *image link*, *Wikipedia link* and *GeoNames link*. The user can edit all fields manually or she/he can use the button *check with DBpedia/check with GeoNames*, as specified in the case of the *add person box*.
3. *CHO box*: the editor allows the user to add a new cultural heritage object, such as a letter, a painting and so on, by specifying the following fields: *original title*, *English title*, *author*, *creation date*, *issue date*, *type* (text, video, sound, image, 3D), *language*, *description*, *image link* and *Wikipedia link*. All these fields, which follow the ontology defined by the Europeana Data Model, should be added by the user manually.

5 Use Case

WeME was used within the Clavius on the Web project, to help the construction of the knowledge base associated to Christopher Clavius's correspondence. Christopher Clavius (1538-1612) was a jesuit mathematician and astronomer and one of the most important characters in the scientific scene of the late 16th century. These manuscripts consist of two volumes of correspondence (about 330 letters) and seven volumes of works, some of which printed in those years and

Class	n.of instances	n.of links
Person	134	DBpedia: 55, VIAF: 129
Place	84	DBpedia: 84, GeoNames: 83
CHO	266	-

Table 1. Statistics about the knowledge base related to the Clavius's correspondence.

some still unpublished. The importance of the correspondence becomes clear just looking at the authors of the letters: Galileo Galilei, Tycho Brahe, Joseph Scaliger, Guido Ubaldo Dal Monte and many others.

The Clavius on the Web project (CoW) aimed at digitizing, annotating, enriching, exporting all this heritage to the Web and linking it to similar Web resources. One of the parts of the CoW project was the creation of a knowledge base of all people and places associated to the context of letters, such as people who wrote the letters and places where the letters were written. The idea was to link the APUG historical heritage to Web resources already contained on the Web, such as DBpedia and Wikipedia.

The Clavius knowledge base is composed of three main classes: person, place and cultural heritage object (CHO). A person is a historical character who wrote a letter to Cristopher Clavius; a place is a location where a letter was written; a CHO corresponds to a physical letter sent to Christopher Clavius from one of the people described before. Some persons had a related page in DBpedia or VIAF, thus WeME retrieved their related information. Other persons, instead, such as *Ilario Altobelli*, were not present in DBpedia, thus they were added to the knowledge base manually. The same was done for places. Table 1 resumes how many people and places were added to the knowledge base and how many links we found.

6 Conclusion and Future Work

In this paper we have illustrated WeME, a user-friendly Web editor of metadata based on semantic Web technologies, whose main design goal is to help archivists and scholars enter metadata of cultural-heritage objects while building a catalog. In addition, we have described the Clavius' knowledge base, which was built around the Clavius's correspondence of about 300 letters. The source code of WeME will be available for download in the GitHub portal.[16]

As future work, we are planning to extend WeME with the following features: a) exporting the knowledge base in different formats, i.e. RDF, XML and CSV; b) managing different ontologies, such as bibo;[17] c) supporting other classes, such as events. In addition we are planning to start a campaign among different categories of users to test the accessibility and usability of the interface, as well as the quality of the produced information. Finally, we are going to make

[16] https://github.com/alod83/metadata_editor

[17] http://bibliontology.com/specification

WeMe more configurable, thus it will be simple to customize it to deal with different scenarios, datasets, and criteria to match named entities with Linked Data objects.

Acknowledgment

The authors would like to thank Mr. Francesco Sacchini, who implemented WeME.

References

1. Abrate, M., Del Grosso, A., Giovannetti, E., Lo Duca, A., Luzzi, D., Mancini, L., Marchetti, A., Pedretti, I., Piccini, S.: Sharing cultural heritage: the clavius on the web project. In: LREC (2014)
2. Auer, S., Bizer, C., Kobilarov, G., Lehmann, J., Cyganiak, R., Ives, Z.: DBpedia: A Nucleus for a Web of Open Data, pp. 722–735. Springer Berlin Heidelberg, Berlin, Heidelberg (2007)
3. Bizer, C.: Evolving the Web into a Global Data Space, pp. 1–1. Springer Berlin Heidelberg, Berlin, Heidelberg (2011)
4. Debruyne, C., Beyan, O.D., Grant, R., Collins, S., Decker, S., Harrower, N.: A semantic architecture for preserving and interpreting the information contained in irish historical vital records. International Journal on Digital Libraries 17(3), 159–174 (2016)
5. Feeney, K.C., O'Sullivan, D., Tai, W., Brennan, R.: Improving curated web-data quality with structured harvesting and assessment. Int. J. Semantic Web Inf. Syst. 10(2), 35–62 (Apr 2014)
6. Hampson, C., Lawless, S., Eoin, B., Yogev, S., Zwerdling, N., Carmel, D., Conlan, O., O'Connor, A., Wade, V.: CULTURA: A Metadata-Rich Environment to Support the Enhanced Interrogation of Cultural Collections, pp. 227–238 (2012)
7. Harris, S., Seaborne, A., Prudhommeaux, E.: Sparql 1.1 query language. W3C recommendation 21(10) (2013)
8. Isaac, A., et al.: Europeana data model primer (2013)
9. Pedretti, I., Del Grosso, A., Giovannetti, E., Mancini, L., Piccini, S., Abrate, M., Lo Duca, A., Marchetti, A.: The clavius on the web project: Digitization, annotation and visualization of early modern manuscripts. In: Proceedings of the Third AIUCD Annual Conference on Humanities and Their Methods in the Digital Ecosystem. pp. 11:1–11:7. AIUCD '14 (2015)