Linked Heritage: a collaborative terminology management platform for a network of multilingual thesauri and controlled vocabularies

Marie-Veronique Leroi

Context

The semantic web and linked data are now well known principles of the Web. Applying the rules of the linked data has been defined as a priority for the development of Europeana, the European Digital Library. Although the normalisation of metadata is a work that has been investigated for years, the importance of terminologies to understand and exploit these metadata in a structured way has risen to the foreground thanks to the linked data and semantic web.

The semantic web is defined by Tim Berners-Lee as "the Web of data with meaning in the sense that a computer program can learn enough about what the data means to process it". The Web is not about documents anymore but it is about data. Therefore Linked data can be explained as "The semantic web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data". The Web of Data makes

sense only if the data are linked.

Many European projects are contributing specific content to Europeana. Cultural heritage is rich thanks to its diversity. Galleries, Libraries , Archives and Museums, also known as the GLAM sector, have a different perception of cultural objects. Besides the diversity of the content itself, there is also a huge diversity of languages and expert terminologies. There is therefore a need for the harmonisation of terminologies at European level to enable a better understanding of the content available at European level.

The Michael project[1] was one of the first European project to take into account the issue of multilingualism in a European context. Indeed this portal offers a multilingual access to digital cultural heritage in 12 languages. The Michael Culture association that has been created after the end of the project in order to sustain the activity of the European portal has been involved in the European projects contributing to Europeana.

Athena has been one of the major projects contributing content to Europeana since almost 1,8 million objects descriptions were provided via this initiative. We present in this paper how the work on terminologies initiated within the Athena Project is now reused and implemented within the Linked Heritage project.

Athena

The Athena Project that started in November 2008 and ended in April 2011, aimed at providing content from European museums to Europeana. The Michael Culture Association has been involved in the Athena Project as workpackage leader for the one dedicated to terminology and multilingualism (WP4). The main objectives of this

---

[1]Michael: http://www.michael-culture.org
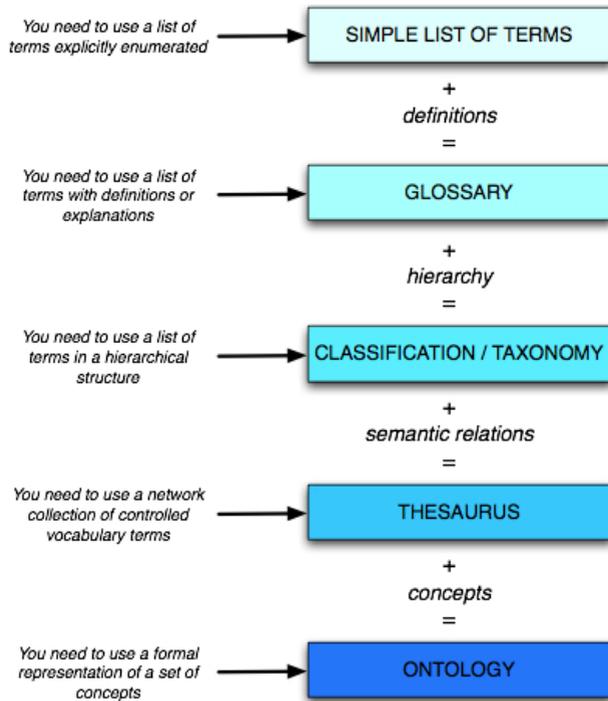
Figure 1: Kinds of terminology resources

workpackage were to provide the European museums with recommendations for terminology management. Therefore we proceeded with a definition and presentation of the different kinds of terminology resources. Indeed as many kinds of terminology resources can be hidden behind the «controlled vocabulary»phrase, we made a synthetic view in order to guide the institutions for answering our survey.

The survey launched among the Athena Partners was intended to
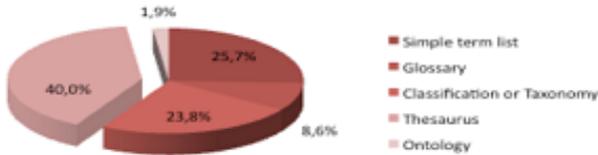
Figure 2: Kinds of terminology

achieve a complete state of the art and give us a precise overview of the terminology use and methods in the European museums. All the results of the state of the art could be found on the Athena and Linked Heritage wiki and on the deliverable of the project.[2]

The results of the survey shown us that most of the cultural institutions, 40% (on 105 total answers) use a thesaurus-type terminology. Indeed most of the cultural institutions use a structured controlled vocabulary. Thesaurus offers both hierarchical and associative relations between the descriptors and it is then a very powerful and simple tool for indexing and cataloguing.

Since it has been acknowledged as a W3C recommendation in August 2009 and since it is the format expected by Europeana, SKOS (Simplified Knowledge Organisation System) is used more and more to bring all the terminology resources into a interoperable format. Very few of the institutions who answered the survey had already SKOSifified terminology, e.g. terminology converted in theSimplified Knowledge Organisation System (SKOS) format. Most of the museums manage their vocabularies within their collections management system. These tools are generally proprietary and often allow for an XML export and in some cases for a SKOS export. Besides the iden-

---

[2]D4.1: identification of terminology resources in museums: http://www.athenaeurope.org/getFile.php?id=398

tification of terminology resources in use in the European museums, the second task of the workpackage was to define some guidelines and a tutorial for SKOS.

Recommendations

The first step is about the conception of your terminology. So to say, at this stage an institution manage its terminology "internally" in order to make a thesaurus in a "human" perspective. We gave the different key-steps for an institution that is about to create on its own a new terminology or adapt a terminology already in use in order to optimize your digital resources descriptions on Europeana. These operations have to be done in priority since they determine the two other steps. In this step, the institution has to think of the domain(s) the terminology will cover, who will be the users expected to use it, what will be the languages the terminology will be available in. The terms and their organisation within the thesaurus structure is defined at this level.

Then the second step consists in making the terminology interoperable. Now it is about rising the terminology out of the museum. Indeed when a terminology is SKOSified, it takes into account the machine perspective. With this second step, we are in the perspective of the semantic web. In this second step, the institution has to evaluate if SKOS is the relevant format for the kind of terminologies it manages. SKOS, for example, won't be the appropriate format in the case of authors/persons' names. After this evaluation, the main object of this step is to proceed with the conversion of the terminology into SKOS. Some tools are available for validating the SKOS output.

Finally we address the cultural institutions our last recommendations as they concern the networking of their terminology with

Figure 3: Benchmark - workflow

others. At this third and last stage, for an institution, it is about being visible in Europe in a network perspective by integrating their terminology in a network of SKOSified terminology. Define some metadata to describe the administrative details of the terminology is a necessary task within this step. On the basis of the metadata provided by other terminologies, terminology resources that can be mapped with the one of the institution will be identified. With this last step, we are in the perspective of the linked data.

From Athena to Linked Heritage

The WP3 of Linked Heritage relies on the legacy of the WP4 of the Athena Project and has been organised to tackle in the best way the following two aspects: content management and technical developments for terminology management. Within the Athena WP4, by gathering experiences from museums and other heritage domains, we identified the most logical process and functional needs related to the management, semantic interoperability and enrichment of terminologies. In doing this, the project identified some use cases and set up a benchmark. The work on the use cases made it clear that first of all a workflow specification was needed for a collaborative production and moderation of cultural heritage terminologies. With the help of the expert working group on terminologies and the uses cases defined earlier on, the following workflow was considered:

This defined workflow helped to evaluate existing tools, interfaces and methods that might be suitable for this work. Unfortunately none of the evaluated tools could actually manage the entire process as shown above. Some of the tools came close, offering search and navigation, as well as semantic mapping and enrichment in a collaborative environment. But there was no complete software environment available offering the possibility for cultural heritage institutions to upload, register and SKOSify the terminologies first, before proceeding to the next steps. This was exactly the point of failure in the workflow of existing tools, because most cultural heritage institutions use own in-house reference terminologies and haven't got the available resources for managing them in a standard interoperable format such as SKOS. This technological lack has been fully considered in the framework of the Linked Heritage WP3 as a complete software will be developed in order to tackle all the steps and process identified within Athena.

From theory to practice: Terminology Management Platform (TMP)

The Terminology Management Platform (TMP) will be this complete software for terminology management and is meant to follow the same structural workflow presented above. Linked Heritage goes beyond the work done in Athena with the purpose to develop a prototype of a tool able to deal with the different steps of the defined workflow and thus to lessen the economical efforts the institution has to make when wanting to share the terminology in an exchangeable format to the community and Europeana.

The work done in WP4 of the Athena Project resulted in the proposed solution to design and the implement an integrated software

environment for terminology management, enabling any institution to manage its terminology according to Europeana ingestion rules. In Linked Heritage this proposed solution will become reality with the development of a prototype of a Terminology Management Platform (TMP) for the cultural heritage sector to collaboratively create a network of interlinked multilingual terminologies in a Europeana compliant format (SKOS). In the development of this platform, the expertises of four technical partners are brought together in a combined effort to create an integrated environment for terminology management. This approach of combining expertises and integrating existing tools into a single web environment, allows us to select and combine best practice technological features in a time and cost efficient way. Considering the outcomes from the Athena Benchmark and the first tasks achieved within Linked Heritage WP3, here are the identified features for the Terminology Managment Platform (TMP):

- to be a web service: For collaborative work online;

- to have a user-friendly GUI: Adapted for a non-expert use in European museums, lirabries and archives;

- to combine open-source components: Such a service must stay independent of proprietary codes and formats;

- to be logically structured with an intuitive Workflow: The user must find which actions to do according to his/her needs;

- to be flexible enough to be adapted to new standards: What if SKOS is updated in a new version or evolving towards an ontology description?

As said above, four technical partners are involved in the development of the TMP. Each of them has its own experience of development and usability and this diversity is very enriching for the whole
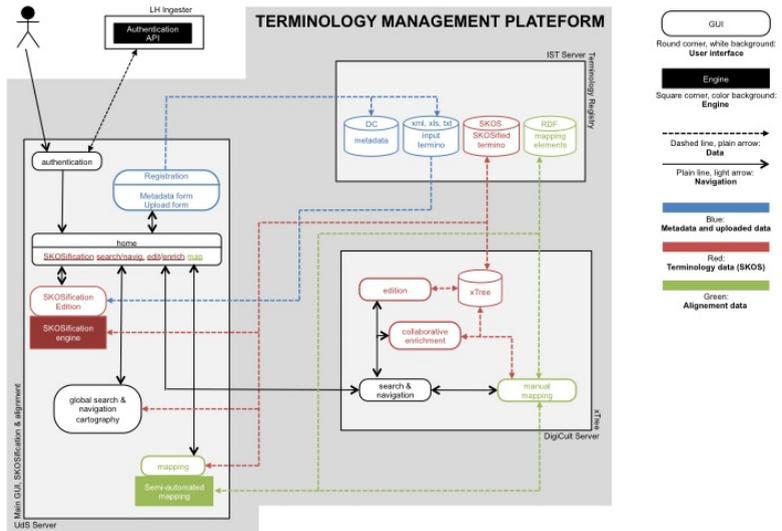
Figure 4: TMP Architecture

toolset. The schema in figure 4 represents the technical architecture of the TMP.

Digicult, the technical partner from Germany is already developing and maintaining an editing tool, xTree. The Institute of Science and Technology (IST) from Portugal has already been involved in projects contributing to Europeana and especially in the development of a metadata registry. This knowledge and experience is now applied for the Linked Heritage TMP since Istituto di Scienza e Tecnologia (IST) is in charge of developing the terminology registry of the TMP. The National Technical University of Athens was already involved in the Athena Project and is the creator of the Mint tool that is used for the ingestion of metadata. In order to guide the institutions who will use the TMP, we decided to use the same authentication process

than the one used in the Mint tool for ingestion so institutions won't need to register themselves in the TMP if they are already using the Ingester. Finally, the University of Savoie from France who has a strong experience in knowledge engineering is in charge of developing a tool for SKOSification and for visual navigation. But most of all, University of Savoie is working on bringing together the different technologies and modules from the rest of the technical partners.

Conclusion

Linked Heritage take a great benefit from the Athena Project as a legacy since it reuses the Mint tool for ingestion and the LIDO format. In the field of terminologies it even goes one step beyond putting into practice all the recommendations and workflow defined in Athena. The Terminology Management Platform will be a great deal for the institutions who will be then autonomous for the management and mapping of their terminologies. This will also enable a better awareness on the semantic web and linked data issues.

MARIE-VERONIQUE LEROI, Ministère de la culture et de la communication (France).

marie-veronique.leroi@culture.gouv.fr

ABSTRACT: Terminology and multilingualism have been one of the main focuses of the Athena Project. Linked Heritage as a legacy of this project also deals with terminology and bring theory to practice applying the recommendations given in the Athena Project. Linked Heritage as a direct follow-up of these recommendations on terminology and multilingualism is currently working on the development of a Terminology Management Platform (TMP). This platform will allow any cultural institution to register, SKOSify and manage its terminology in a collaborative way. This Terminology Management Platform will provide a network of multilingual and cross-domain terminologies.