SCENT for GLAM: a tool for giving meaning to professional controlled vocabularies

1. Diversity and evolution of professional practices in the cultural sector

Cultural institutions are digitising their objects and collections intensively over the last twenty years. If the first challenge of digitisation of heritage was the long-term conservation, the development of the Web in the 2000's has made access to the widest audience of this digital heritage another major challenge.
The Information Technology and Communication (ICT) and the growing amount of scanned documents and objects have resulted in cultural professionals new documentary practices for indexing and cataloging of the scanned items and their management in databases that are mainly open for a an internal use first. These databases were linked to collections management tools, if that is still the case for most institutions, some preferred to separate the internal management of their collections and their dissemination and valorisation via databases open to the general public.

The cultural sector may look uniform and homogeneous but in reality each cultural field has its own practices and objects. Galleries, libraries, archives and museums, that is called the GLAM sector, have their own professionnal practices for organising and managing their objects and collections. They have different professional practices since they deal with different object types. Libraries have to describe precisely books and manuscripts for example whereas museums deal with a variety of objects that may be from archeological objects to paintings from the 16th century. Each field has defined and is still defining its own standards and norms to describe these objects and collections.

Cultural institutions when they are digitizing their collections and objects, produce a digital representation of this object / collection but also information that describe the object / collection. This information is what we call associated metadata. This metadata can be rudimentary giving basic information such as title, author, period or subject of the object, but they can also be very complex by providing information on the work itself but also on its acquisition mode by the institution or biographical information about the author, for example. Many institutions use controlled vocabularies, ie predefined term lists, structured and unstructured, which allow to normalise some of these metadata according to the choice of the institution or the choices and constraints of the collection management system

2. Europeana, the European digital library

In 2008, the European Commission has massively supported the building of a European digital library Europeana[1] with the vocation to offer a broad and open access to European cultural heritage to all European citizens. Europeana gathered initially the major European

1 http://www.europeana.eu

national libraries, including the National Library of France. Europeana has subsequently broadened its scope by displaying a willingness to aggregate the content of all types of cultural institutions and not only European libraries. Thus, the contents from archives, audiovisual institutes or centers, museums or private organisations with heritage collections began to converge to Europeana. Number of specific European projects for the different fields - archives, libraries, museums - have been funded by the European Commission in order to support the institutions in their digitisation process and provision of content to Europeana. The term "content" means the metadata associated with an object or a heritage collection. The object or collection in itself is not aggregated by Europeana insofar as the rights associated with these objects or collection may restrain their dissemination and reuse. Europeana brings together cultural metadata from all fields, languages and European countries and then redirects the user to the institution's website that holds the object or collection.

Europeana and the strong impulse given by the European Commission for the digitisation of the cultural heritage and its promotion and dissemination to a large and European audience has deeply changed the mentalities among professionals from the different fields.
Indeed institutions that were focusing so far to share their metadata at national level via local databases had to take into account the European level but also the fact that their metadata will not be only shared to a professional audience. Europeana as a European portal for cultural heritage is intended to the professionals, the students but also the general public.
In order to be part of this European portal, cultural institutions that were used to work within their own standards had to get close to the standards of their field.
Europeana first implemented a metadata schema called ESE[2], which stands for Europeana Simple Elements, to fit the basic metadata of cultural institutions from all fields. ESE was an improved Dublin Core format that was convenient in the beginning but quickly became very limited for the cultural institutions willing to promote their objects and collections. Considering that the development of the ICT technologies have significantly improved harvesting protocols, data transfer and storage systems, the definition of a more complex metadata schema has became a new priority for Europeana. In 2011, EDM[3], the Europeana Data Model, has been defined as the new metadata schema in use for the metadata aggregated by Europeana. This schema is not linear as ESE was but modular and based on classes and properties. This organisation in classes and properties allows richer and more complex descriptions for objects and collections and also fit the specific needs of all fields of the GLAM sector.

One of the Europeana related project, ATHENA[4], which started in 2008 simultaneously with the official launch of Europeana, was focusing on the agregation of metadata from European museums. This European project has developed a standard and a tool for aggregating metadata from European museums. The standard, LIDO[5], which stands for Lightweight In
Formation Describing Objects, is itself based on two Museum existing standards CDWA lite and the museumdat metadata schemas. LIDO is a rich metadata schema that allows describing in a modular way an object as it is not a linear model such as Dublin Core. Specific classes with precise properties describe each event that might occurr in the lifecycle of a Museum object like its discovery, its acquisition, its proofing or even its use in

[2] ESE : http://pro.europeana.eu/share-your-data/data-guidelines/ese-documentation
[3] EDM : http://pro.europeana.eu/share-your-data/data-guidelines/edm-documentation
[4] Athena : http://www.athenaeurope.org
[5] http://www.lido-schema.org/schema/v1.0/lido-v1.0-specification.pdf

a exhibition.
The aggregation tool developed in the framework of the ATHENA project, MINT[6], is based on the LIDO schema. LIDO is used in MINT as a harvesting pivot format towards the metadata schema used by Europeana, EDM.

The principle defined within the ATHENA project was that each European museum prepares its set of objects, e.g. thumbnails and associated metadata, then upload it in the aggregating tool MINT. Once the set of metadata is imported into MINT, the institution has to make a mapping of its own metadata schema with the LIDO datamodel in order not to lose any information contained in the institution's metadata. MINT automatically makes the conversion from the LIDO model to the EDM model and the metadata is then transferred to the Europeana portal.

Among all the workpackages of the ATHENA project, one was dedicated to terminologies and multilingualism. Indeed a key challenge of Europeana is to provide access to the digital cultural heritage in Europe in a multingual way. The principle of aggregation is fostered by the thematic European projects but it relies mostly on the national organisation of the cultural institutions. Unless it is prescribed by national policies like in Belgium where French, Dutch and German are considered as official languages, institutions used to index and catalogue their collections only in the national language and eventually English in some cases. The consequence at Europeana level is to have metadata from a variety of countries in a variety of languages. Terminologies are the most efficient way to tackle the issue of multilingualism.

3. The use of terminologies to enrich the cultural metadata

Terminologies are used complementarily to the metadata. If we consider that the metadata schema is the grammar, the terminologies are the vocabulary and this is the combination of both that provides meaningful metadata.

A state of the art on the terminologies in use in European museums[7] has been conducted within the dedicated workpackage of the Athena project. This state of the art was based on a survey led among the project partners and a benchmark of the existing reference terminologies.

As an introduction to this state of the art, a clear distinction has been made between « terminology » considered by the linguists as a discipline aiming at identifying all the terms proper to a specific domain and « terminology » as a neutral designation for any type of controlled vocabulary.

[6] MINT : http://mint.image.ece.ntua.gr/redmine/projects/mint/wiki/Introduction_to_MINT
[7] Athena D4.1 : Leroi, Marie-Véronique, Holland, Johann, 2009. *Identification of existing terminology resources in museums*.

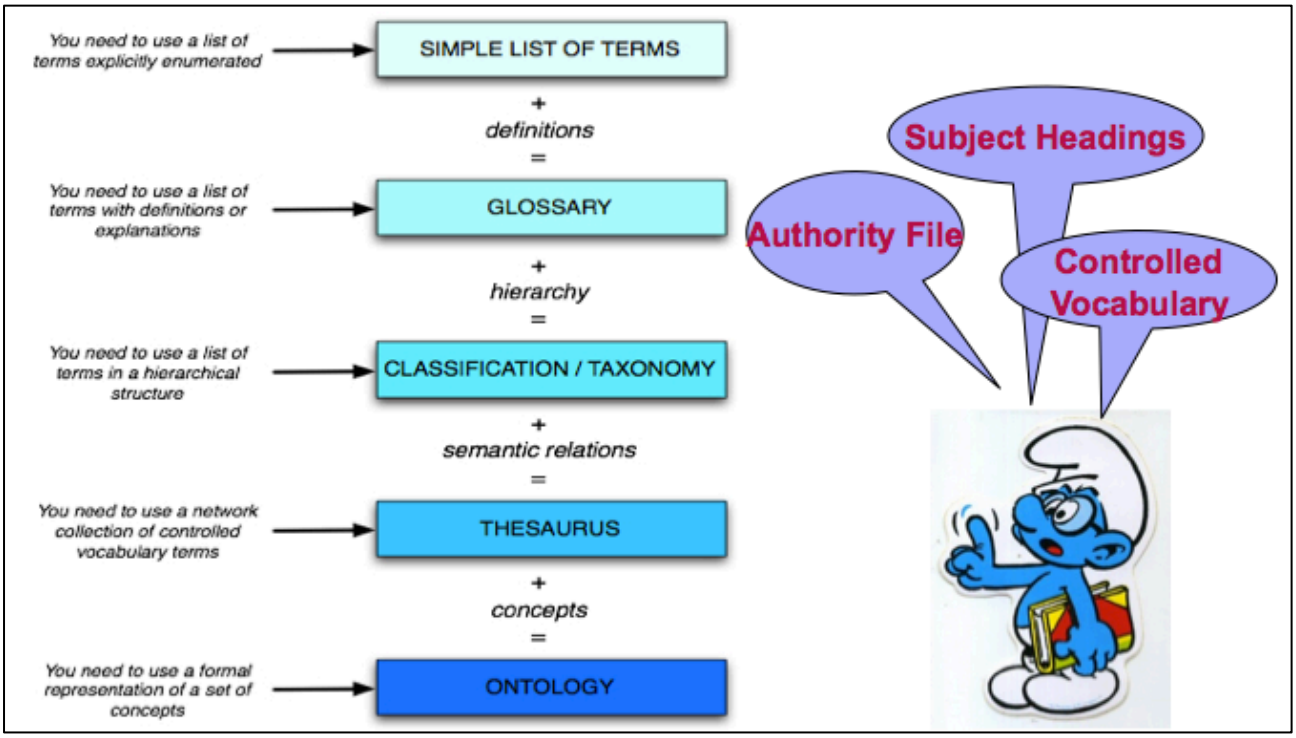

Figure 1 : Types of terminology

The figure above shows the different types of terminology that can be used by many disciplines and especially cultural professionals. Professionals usually speak about subject headings, authority lists or controlled vocabulary.

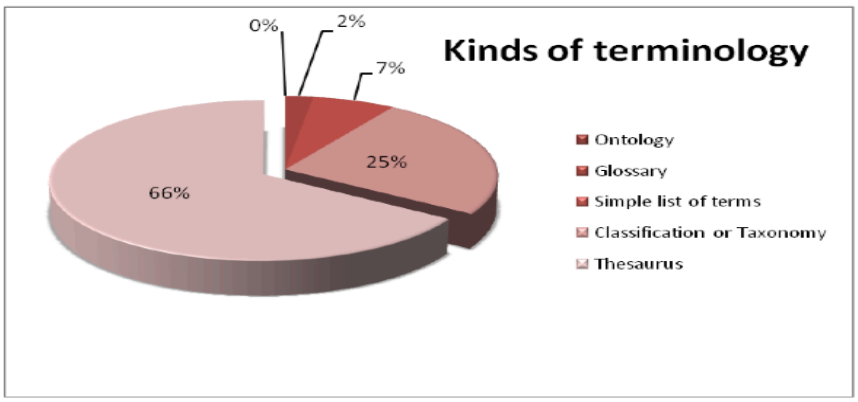

Figure 2 : Results of the survey on the use of terminology in the Athena project

The figure above shows that most of the Europeana museums use thesaurus type of terminologies. A thesaurus is a documentary resource different from a simple list of terms, a glossary or a classification. As shown in the figure x, a simple list of terms does not provide any hierarchical information or definitions. A glossary is a simple list of terms generally organised alphabetically that provides definitions. A classification is a list of terms organised hierarchically and that might provide some definitions. The thesaurus combines hierarchical and associative information and can also provide definitions and documentary notes.

The thesaurus is defined by several ISO norms : the ISO 2788 and the ISO 5964

respectively give guidelines a recommendations for the design of monolingual and multilingual thesauri. The ISO 2788 was first published in 1974 and revised in 1986. The ISO 5964 which extends the scope of the ISO 2788 with multilingualism was published in 1985. These two norms have been strongly revised in 2011 with the publication of a new norm the ISO 25964.

This new norm has been published in two parts: the first one titled « Thesauri for information retrieval » has been published in August 2011 and the second part titled « Interoperability with other vocabularies » was published in March 2013. This norm adapts the two previous norms to the actual technological context and its impact on Knowledge Organisation System.

There is in the cultural field several terminologies that are considered as reference resources: it is the case of lists of authorities from the Library of Congress (LCHS: Library of Congress' Subject Headings)[8], terminologies from the Getty[9] (AAT : Art and Architecture Thesaurus, TGN: Thesaurus for Geographic Names and very recently ULAN : Union List of Artist Names) or the RAMEAU authority list coordinated by the National Library of France.

The study showed that the majority of European museums is using either in-house terminologies created from scratch or adapted, translated or customized versions of these reference terminologies. Each of these terminologies is of great value for understanding the metadata associated with an object or a museum collection. This is all the more true in a necessarily multilingual European context.

4. Semantic Web and Linked Data: new Web, new practices?

Semantic Web and Linked Data, evolutions of the Web of documents introduced by Sir Tim Berners-Lee has strongly widened the possibilities and impacts in the use of KOS in documentary systems.

We mentioned earlier the updated norm on thesaurus, the ISO 25964, this norm takes especially into account the SKOS format, recommendation from the W3C since 2009. The SKOS format is one of the KOS that implements the principles of the Semantic Web and follows the classical structure of a thesaurus.

Aggregating metadata from cultural institutions in the context of Europeana requires multilingualism and the Semantic Web principles and technologies offer to cultural institutions an economic and efficient way to enable multilingualism thanks to the semantic enrichment of metadata.

One of the objectives of the Terminology workpackage in the ATHENA was to sensitise and train European museums on these principles and technologies and also convince them that moving towards a format such as SKOS is the key to provide semantically enriched metadata.

The main difficulties in this process were essentially the partners' professional practices which are tightly connected to the collection management system (CMS) in use in the

[8] LCSH : http://id.loc.gov/authorities/subjects.html
[9] Getty : http://www.getty.edu/research/tools/vocabularies/lod/index.html

institution. In most of the cases, these CMS are proprietary tools and institutions pay an annual fee for the licenses. Apply Semantic Web technologies might imply additional costs to update or change of CMS and/or associated tools. So a first difficulty is financial.

In some cases, possibility of imports and exports in the CMS are limited and/or outdated and the professionals are dependent on the software company developing the tool for any modification. A second difficulty is technical.

Another difficulty consists in the practices themselves since professionals fear the automatic processes and possibilities implied by the Semantic Web technologies. There is a fear that the quality of metadata could be impacted.

A last difficulty that is unfortunately common to everything related to the cultural sector is the intellectual property rights issue and the principle of open data. Indeed there is a fear that the intellectual work provided by curators and professionals might be just given away and perverted by the collaborative use that might be made of it.

In order to answer to the reluctances of the professionals, a set of recommendations has been defined within the Terminology workpackage of the Athena Project. These recommendations have been published as a booklet addressing all cultural institutions[10].

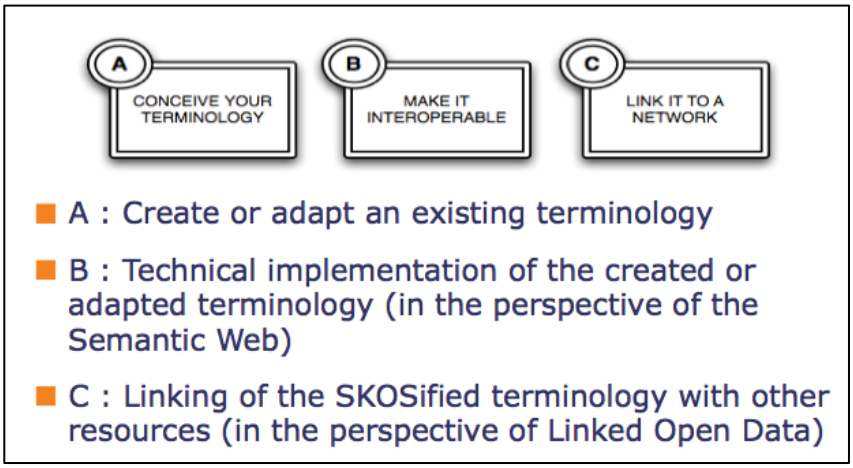

Figure 3 Recommendations for terminology management in the perspective of the Semantic Web

These recommendations have been formulated to address museums and more broadly cultural institutions to create and manage their terminologies in the context of the Semantic Web. These recommendations have been defined on the basis of the feedbacks and needs from the project partners and grouped under three main steps.

The figure above summarizes these recommendations and their organisation within the three consecutive steps.

The first set concerns the creation phase of a terminology resource. This first step aims at defining the scope of the terminology, the domain, the spatial and temporal coverage, the language(s) and the target audience of the terminology. At this stage the work is more

[10] Leroi, Marie-Véronique, Holland, Johann, 2011. *Your terminology as a part of the semantic web recommendations for design and management*

conceptual and institutionnal.

The second set of recommendations deals more techincally with how to make the terminology consistent with the Semantic Web technologies. This second phase introduces the SKOS datamodel and the possible bridges that can be made from a thesaurus structure to a SKOS terminology. This phase is more technical and needs some validation as the initial conceptual work should be technically implemented.

Finally the third and final set of recommendations stresses the provision of these terminologies and the need to establish links with external terminologies to semantically enrich and share them. This final step is clearly about linking concepts and terminologies together to enable multilingualism and richer KOS network. This step is technical and institutionnal since institutions should accept to open, share and publish their terminologies and connect them with external ones.

Through these recommendations, a major effort has been put to convince cultural institutions that they could really benefit from this change of practices.

5. SKOS, Something Kool Original and Sexy for a simple semantic interoperability

We already mentioned the SKOS format that has been published as a W3C recommendation in August 2009 and that is “Something Kool Original and Sex” for Alistar Miles, one of its creator. The acknowledgement of SKOS as a Semantic Web standard has been decisive in the “semantic” orientation that has been given to these recommendations. SKOS stands for Simple Knowlegde Organization System and is an exchange format which offers an interesting compromise to cultural institutions that wish to take the step of the semantic web. Compromise mainly because some might consider thtat SKOS is too simple a model. SKOS provides a simple model but rich enough to respect the structure of a thesaurus for example. Many classes and properties have been created since 2009 to meet the needs of institutions to represent more complex terminologies. SKOS has notably evolved to take into account the recommendations and best practices presented by the ISO 25964 standard for thesaurus. A format such as SKOS allows for a cultural institution to limit costs as such represent the creation of an ontology, it also changes their documentary practices without generating deep changes. Indeed, many collections management systems now offer an import or export SKOS. Using Semantic Web technologies could be an economic opportunity for cultural institutions that are facing lowering of their annual budget and staff.

The SKOS format has been chosen to represent the terminologies used by European museums in terminology management tool developed in the framework of the European projects Linked Heritage[11] and Athena Plus[12], SCENT for GLAM. Both projects follow the Athena project already cited in this article.

The choice of the SKOS format, except its simplicity and the fact that it is a standard, is mainly due to the types of terminologies in use in the cultural institutions. SKOS perfectly fits the thesaurus structure and can be extended and customised with properties from

[11] Linked Heritage : www.linkedheritage.eu/
[12] AthenaPlus : www.athenaplus.eu/

other data models. Moreover since 2009, many software dedicated to terminology and collections management systems can handle SKOS import or exports.

Thesauri are used to organise descriptors that are the main terms connected to each others according to hierarchical or associative relations. Non-preferred terms (synonyms, alternative terms) are considered non-descriptors. The challenge for cultural institutions is to understand the need to move from the level of a descriptor (or preferred term) to the concept level.

The state of the art presented in the study of the Athena project and the recommendations addressed to cultural institutions to create and manage their terminologies have reinforced the need for these institutions to have a tool enabling them to inexpensively create and manage their terminologies in accordance with the standards and technologies of the Semantic Web for better intelligibility of their metadata into Europeana.

SCENT for GLAM: a complete software environment for terminology management

Many software and tools handling SKOS are available nowadays, some are proprietary tools and some other are open source.

The idea and need to develop a tool for terminology management emerged subsequently to the state of the art on the terminologies in use made in the ATHENA project. Considering the lack of means and strictly technical skills, providing to cultural institutions a list of tools that could be used for converting into SKOS their terminologies and another list of tools for publishing or sharing terminologies is not very realistic. This is why the need for a complete software environment presented itself as evidence.

A first prototype of terminology management tool was developed on the basis of the functional requirements expressed in the Athena project. Functional specifications and major technological choices were made in this context.

A more operational version of this terminology management tool, SCENT GLAM has been possible in the framework of the European project AthenaPlus which followed the European project Linked Heritage but is more a result of the Athena project in that most partners are museums. SCENT for GLAM stands for Semantic and Collaborative Environment for a Network of Terminology for Galleries, Libraries, Archives and Museums.

SCENT for GLAM offers a number of features that allow an institution to create or import terminology, to convert it into SKOS format, to edit it, make alignments with other terminologies from the same institution or another one and finally publish and share this terminology and alignments. The idea is to provide a repository of terminologies that will include all of the cultural sector concepts.

SCENT for GLAM fully supports any terminology formulated in SKOS format, but can also include properties from other formal languages (FOAF (Friend of a Friend) or OWL (Web Ontology Language), ...) as long as these are expressed in RDF (Resource Description Framework).

The terminologies that are created in SCENT support natively the SKOS data model.

The key features for a tool such as SCENT for GLAM are the collaborative ones which ensures a trust for the terminologies that are published and shared.
SCENT for GLAM provides URI for the terminologies that are published within the tool.

6. Conclusion

SCENT for GLAM is still under development in order to improve and take more and more into account the needs of professionals.
More than the tool itself, the major challenge is to get the cultural institutions accepting and participating in this new KOS configuration.

Using a format such SKOS implies some technical work and understanding but the benefits for institutions that give that efforts is economic since they could benefit from the expertise and knowledge from other institutions via the collaborative features. Translations obtained via the mapping links enable multilingualism in a most economic way.

This is also a way to rely more on standards and open source tools and finally get out of the proprietary tools' loop that offer a tool that will not evolve without a financial effort.

Last but not least, the major benefit and impact for the cultural institution adopting this new KOS is the trust and guarantee that it provides to its cultural content thanks to the openness and share of its terminologies but also the semantic enrichment of its metadata.

7. References

[Web document] Leroi, Marie-Véronique, Holland, Johann, 2009. *Identification of existing terminology resources in museums*. Athena D4.1 report. Available at <http://www.athenaeurope.org/getFile.php?id=398>

[Web document] Olensky, Marlies, Stiller, Juliane, and Drøge, Evelyn, 2012. *Poisonous India or the Importance of a Semantic and Multilingual Enrichment Strategy.* Available at <http://dm2e.eu/files/PoisonousIndia.pdf>

[Booklet] Leroi, Marie-Véronique, Holland, Johann, 2011. *Your terminology as a part of the semantic web recommendations for design and management*. Linked Heritage and Athena Booklet.