Date : 20/10/2008 (2nd Version)

Semantic web and vocabularies interoperability: an experiment with illuminations collections

Anila Angjeli
Bibliothèque nationale de France
anila.angjeli@bnf.fr

Antoine Isaac
National Library of the Netherlands and Vrije Universiteit Amsterdam
aisaac@few.vu.nl

Translation by:
Nadia PAZOLIS-GABRIEL

Meeting: **129. Classification and Indexing**
Simultaneous Interpretation: English, Arabic, Chinese, French, German, Russian and Spanish

WORLD LIBRARY AND INFORMATION CONGRESS: 74TH IFLA GENERAL CONFERENCE AND COUNCIL
10-14 August 2008, Québec, Canada
http://www.ifla.org/IV/ifla74/index.htm

The experiment described here was the subject of an internal report at the Bibliothèque nationale de France (BnF), co-authored by Thierry Cloarec and Frédéric Martin, of the Département de la bibliothèque numérique, in close cooperation with Antoine Isaac and Lourens van der Meij, researchers at the National Library of the Netherlands (Koninklijke Bibliotheek, KB) and at the Vrije Universiteit, Amsterdam.

Abstract

During the years 2006 and 2007, the BnF has collaborated with the National Library of the Netherlands within the framework of the Dutch project STITCH. This project, through concrete experiments, investigates semantic interoperability, especially in relation to searching. How can we conduct semantic searches across several digital heritage collections? The metadata related to content analysis are often heterogeneous. Beyond using manual mapping of semantically

similar entities, STITCH explores the techniques of the semantic web, particularly ontology mapping.

This paper is about an experiment made on two digital iconographic collections: Mandragore, iconographic database of the Manuscript Department of the BnF, and the Medieval Illuminated manuscripts collection of the KB.

While the content of these two collections is similar, they have been processed differently and the vocabularies used to index their content is very different. Vocabularies in Mandragore and Iconclass are both controlled and hierarchical but they do not have the same semantic and structure. This difference is of particular interest to the STITCH project, as it aims to study automatic alignment of two vocabularies.

The collaborative experiment started with a precise analysis of each of the vocabularies; that included concepts and their representation, lexical properties of the terms used, semantic relationships, etc. The team of Dutch researchers then studied and implemented mechanisms of alignment of the two vocabularies. The initial models being different, there had to be a common standard in order to enable procedures of alignment. RDF and SKOS were selected for that. The experiment lead to building a prototype that allows for querying in both databases at the same time through a single interface. The descriptors of each vocabulary are used as search terms for all images regardless of the collection they belong to.

This experiment is only one step in the search for solutions that aim at making navigation easier between heritage collections that have heterogeneous metadata.

1. Context

Heritage collections are more and more present on the World Wide Web. The institutions who hold these collections have adapted by progressively integrating functions and technologies of the web, moving from simple showcase to digital representation of the collections after referencing them online. Today we face new challenges.

It is time for interconnection and intelligent interaction between different collections. The frame for presence and action is getting wider; the collections are no longer considered only as strongly attached to their holding institutions, but as part of a European treasure, even a worldwide one,

that needs to be interconnected. The institutions widen their ambitions from an institutional to a cross-institutional scale, from a national to an international scale.

The technologies of the semantic web open up new perspectives to achieve these ambitions and heritage collections are ideal candidates for that. The goal is to reuse the tools already created by professionals in various domains, to "release" the meaning contained within the existing metadata, and connect the "intelligence" they capitalize on a long-term basis. Highly professional and specialized knowledge organization systems have been developed for the description and organization of heritage collections. The metadata linked to these collections are traditionally rich, precise and structured; they are based on tools that are regularly updated. This information is added value; it builds up and accumulates as time goes by. Thanks to the techniques of the semantic web, we will be able to use it as a way to interconnect the collections.

However, this is not without difficulties. The challenges to meet in order to interconnect and make the collections interoperable include: the wealth and the variety of these collections, the different cultural contexts in which they are produced, the different ways to process the collections and to create their metadata.

This paper presents an experiment focusing on the interconnection of two heritage collections: the National Library of the Netherlands' illuminations and the Bibliothèque nationale de France's. In Part 2, we will consider this experiment within a general framework of problems related to interconnecting collections with heterogeneous data. Parts 3, 4, 5, and 6 deal with the characteristics of a concrete example, detailing the intellectual and technical options chosen to conduct the experiment in an appropriate way. In part 7, we present the prototype built for the experiment and our conclusion in Part 8 stresses the ways to re-use the results of the experiment.

2. Problems to be solved and state of the art of the issue

Semantic access and interconnection within heritage collections is the object of many ongoing research efforts. Often, the themes, topics and concepts represented in the resources are similar and constitute as many interconnection points to enable interoperability between these collections. However, the collections are usually indexed with specific vocabularies. These vocabularies are as many heterogeneous tools, developed according to different principles of representation and with their own indexing rules. Heterogeneity lies in the type of indexing tool and the way it is represented (thesaurus, classification system, authorities) as well as in the

variations found in the semantic scope of similar concepts. Heterogeneous linguistic environments add up to the complexity.

How can we enable access to documents belonging to two or several collections, each indexed with distinct vocabularies, by using search terms from either vocabulary? In other words, we would like to use search terms of a specific vocabulary to access documents that have no direct link with this vocabulary! In order to achieve this, the terms must be aligned.

Two major directions stand out in current studies. One is to establish manual links between different vocabularies used in the collections to be interconnected, as is done in the MACS project (Landry, 2007). The other one, more recent, uses the techniques of the semantic web, based on the fact that controlled vocabularies are true knowledge organization systems (KOS) and therefore correspond to the type or artifacts focused on in the semantic web vision.

That is to say, the techniques of the semantic web can give a new life to traditional tools as controlled vocabularies by using them with standard technologies developed for a networking environment. Controlled vocabularies can thus be used in wider spaces of connected resources.

3. A concrete collection alignment experiment

Dutch researchers who study these techniques have already carried out experiments using Dutch collections with the research program CATCH[1] which aims at finding innovating solutions to access heritage collections. STITCH (Semantic Interoperability To Access Cultural Heritage)[2] is one of the projects of this program, and precisely aims at solving the aformentioned interoperability issues using the technologies of ontology alignment (Shvaiko & Euzenat, 2005).

In order to explore a more multilingual context, an experiment has been conducted between the STITCH team and the BnF. It focused on two collections with similar content:

---

[1] CATCH: Continuous Access to Cultural Heritage http://www.nwo.nl/catch

[2] STITCH is financed by De Nederlandse Organisatie voor Wetenshappelijk Onderzoek (NWO), Dutch Organization for Scientific Research. We would like to thank the RKD Institute and Gerda Duijfjes-Vellekoop for letting us access Iconclass. Within the STITCH team, we would like to thank Frank van Harmelen, Henk Matthezing and Stefan Schlobach who contributed to this experiment through their support and many discussions. To learn more about STITCH, see http://www.cs.vu.nl/STITCH/

- The illuminated manuscripts database at the National Library of the Netherlands (KB)[3] – the most extensive iconographic medieval collection of the Netherlands, almost 11,000 miniatures from illuminated medieval manuscripts of the KB and the Meermanno Museum, digitized and accessible online.
- The Mandragore iconographic database from the Manuscripts Department of the BnF[4], hosting more than 140,000 miniatures from tens of thousands manuscripts at the BnF, some dating back to ancient Egypt, up to the contemporary era. More than 50,000 or the bibliographic descriptions are displayed along with the digital image online.

The experiment took place during the second half of 2006 and ended with the delivery of a demonstrator in January 2007. It aimed at studying and the aligning the Iconclass[5] and Mandragore[6] vocabularies that are used in the two collections respectively. This required first an analysis to compare the native models of each vocabulary.

---

[3] See www.kb.nl/manuscripts/

[4] See http://mandragore.bnf.fr/html/accueil.html

[5] See http://iconclass.nl

[6] The Mandragore vocabulary can be accessed on the search pages of the iconographic database Mandragore, http://mandragore.bnf.fr/html/accueil.html, from the descriptor field search in particular.

4. Vocabulary analysis

The table below briefly shows the general characteristics of the two vocabularies, Iconclass and Mandragore.

Iconclass	**Mandragore**
Classification system Designed in the 1970s by Dutch scholars for the description and indexing of images International scope	**Controlled vocabulary** Designed internally for illuminations indexing purposes at the Manuscripts Department of the BnF
Form Each term is made of: - a complex alphanumeric identifier (notation) - a definition (or descriptor in a textual form) - (associative) cross-references	**Form** Text descriptor. A numeric identifier is assigned to the descriptors for internal management purposes only. The descriptor record features also: - alternative forms - information notes
Structure Strong hierarchy: 10 levels Each notation inherits from the semantic of the superior levels	**Structure** Loosely structured in 2 parts: - a list of subject descriptors - a two-level hierarchy with about 150 classification elements (inspired by the Dewey classification) gathers the descriptors according to general topics
Language Multilingual (English, German, French, Italian, some Finnish and Norwegian)	**Language** French
Semantic cover Descriptors for objects, people, events and abstract ideas that may be the subject of an image	**Semantic cover** Descriptors for objects, people, events and abstract ideas that may be the subject of an image
Use As needed, to express the meaning of: - entire scenes - isolated elements contained within an image	**Use** As needed, to express the meaning of: - entire scenes - isolated elements contained within an image
Information objects that the vocabulary aims at analyzing Paintings, drawings, photographs, etc.	**Information objects that the vocabulary aims at analyzing** Images in manuscripts
Quantity 28,000 descriptors divided in 10 main classes Alphabetical index of 14,000 keywords used to find identifiers in the vocabulary and the textual descriptors 40,000 bibliographic references for books or articles of iconographic interest	**Quantity** 16,000 descriptors

The collaborative experiment started with a close analysis of each of these descriptive vocabularies: the concepts and their representation, the lexical properties of the terms, the semantic relationships, the syntax linking the concepts together, etc. The team of Dutch researchers then studied and implemented mechanisms to link both vocabularies.

5. Semantic web and vocabulary alignment

In this experiment we aimed at testing semantic web techniques on digital heritage collections, focusing on ontology alignment. This technique consists in automatically identifying the possible correspondences between the vocabularies terms. Even if the alignments are later manually readjusted by experts – the process is then semi-automatic – this allows saving up considerably on human effort.[7]

The first step for vocabulary alignment actually consists of reducing heterogeneity of vocabulary syntax and model. For this, we have used the common standard RDF for data representation in combination with SKOS. SKOS[8] (Simple Knowledge Organization System) is a simple and standard model for controlled vocabulary representation on the Web. It allows representing in an RDF format:

- Concepts `(Concept)`;
- The description of KOS themselves (here, Iconclass and Mandragore) as SKOS `Conceptscheme` objects;
- Lexical properties of descriptors `(prefLabel, altLabel)` including linguistic variations;
- Semantic relationships between descriptors `(broader, related)`;
- Information provided in notes `(scopeNote, definition)`;

[7] A STITCH publication (Gendt, M. van et al., 2006) studies the alignment techniques used in the project

[8] http://www.w3.org/2004/02/SKOS/ See also (Miles & Bechhofer, 2008) and (Isaac & Summers, 2008) for more information about this standard

The analysis of the Iconclass and Mandragore vocabularies presented in the previous section helped define the links between components of each vocabulary and appropriate elements of the SKOS model. For example, a preferred term in Mandragore is represented with the element `skos:prefLabel`; the concept corresponding to a descriptor is linked to the concept corresponding to a more general descriptor in the hierarchy with the element `skos:broader`; a scope note is represented by the element `skos:definition`, etc. It is to be noticed that SKOS, which aims to allow data portability to the semantic web, is a *simple* representation model and that the conversion of data always causes loss of information. This is why the native models retain their value *in their initial context*.

The knowledge about vocabularies obtained during the analysis step also allowed choosing the most appropriate strategy to build the alignment algorithms, according to the project's constraints.

The semantic web community addresses the issue of ontology alignment through various approaches, each focused on a different kind of information found in the given controlled vocabularies. *Lexical approaches* use different types of linguistic information found in these vocabularies (preferred terms, see reference terms, definitions, etc.). *Structural approaches* use the hierarchical and associative architecture that link the concepts of a vocabulary. *Extensional approaches* widen the search beyond vocabularies. They use information from the metadata of the documents indexed against the vocabularies. For instance, the occurrences for a given concept in the indexed documents can be used as relevance criteria to test the validity of the alignment with a concept from another vocabulary – if a proper means to compare these documents with the document indexed against the target concept is available. *Background knowledge-based* approaches exploit external resources to compensate for deficiency of a given vocabulary (external dictionaries for lexical questions, external ontologies for matters regarding structure, etc.).

6. Strategies

For our experiment we focused on testing and combining several strategies of lexical alignment. As the architecture of the two vocabularies had significant differences (e.g. discrepancies in

structure, principle of inheritance used in Iconclass but inexistent in Mandragore), the structural approach was considered pointless. Implementation of the other approaches required more means and more time than the research team had.

Below are notes on the strategies that were accepted while implementing the lexical approach to establish alignments. Semantic links (equivalence or “broader”) between concepts were established when a lexical similarity was observed:

- between preferred forms, for example, between the preferred form `Grange` in Iconclass and its perfect corresponding term `Grange` in Mandragore, where it is also a preferred form;

- between a preferred form in Iconclass and its corresponding lexical term within a rejected form in Mandragore. For example, the term `Enterrement` as a preferred form in Iconclass and the term `Inhumation` as a rejected form in Mandragore;

- between a part of a preferred form in one vocabulary and a preferred term in another. For example, `Hercule est découvert par Junon et Minerve, celle-ci le met au sein de Junon` (Iconclass) and the term `Junon` (Mandragore);

- between `Zoologie (généralités)` (Mandragore) and the term `Zoologie` (Iconclass) that has a wider semantic scope than in Mandragore;

- between a definition or a note in Mandragore and a preferred form in Iconclass. In Mandragore, the definitions provide indications that help link a concept to other more general concepts. For example, an equivalence link was established between the term `Manche`, present in a definition in Mandragore, and the term `Mer`, a preferred form in Iconclass.

Some of these strategies provide lower quality results. However they were not neglected, as a readjustment could be made afterwards by combining the results of different strategies.

Other problems were not solved. The original models of the vocabularies being complex and not standard, some information that was important for the alignment process could not be converted in SKOS and therefore could not be used for the alignments.

7. Integrated access to multilingual collections: the demonstrator

The following step was to build a demonstrator[9] that provides simultaneous access to the collections of illuminations of the KB and the BnF with searches using terms from the Iconclass or the Mandragore vocabulary.

As the demonstrator is a simple prototype, it contains only a sample of the original collections, 2,170 records from the collection of illuminations of the BnF and 3,987 records of the KB's collection.

The demonstrator was inspired by the faceted search interface Flamenco,[10] developed at the University of California Berkeley. It gives the possibility to narrow down search criteria and to combine them. Users are guided through the whole search process. The indexed terms are displayed in their hierarchical environment in tree lists. Users make a search by clicking on one of the terms. They can unfold the hierarchy and narrow or broaden their search. Results are automatically recalculated.

It is important to note that the alignments are made from terms in French in both vocabularies, as Iconclass is a multilingual vocabulary. An indirect consequence of the vocabularies alignment is therefore the opportunity for Mandragore to benefit from this multilingualism. Searches made with Iconclass terms in English or in German will also retrieve documents from Mandragore, even though they are indexed in French only.

The demonstrator offers two types of access to the collections: a single access, with the terms of one vocabulary; a combined access, with terms from both vocabularies. These two types of access branch out:

- access through terms from the Mandragore vocabulary;
- access through terms from Iconclass (in French, English, or German);
- access through a combination of Mandragore and Iconclass vocabularies;

[9] The demonstrator can be accessed at http://www.cs.vu.nl/STITCH/BNF_KB_demo.html

[10] The Flamenco faceted interface is explained on the Berkeley website at http://flamenco.berkeley.edu/index.html

- restrict display to one of the two collections;

The results of the query are displayed on the screen as a set of thumbnails that give direct access to the digital illuminations and their descriptive records.

8. Conclusion

This experiment is only a basis for reflection in the search for solutions aiming at facilitating navigation in heritage collections with heterogeneous metadata. As such, it is one of the numerous use cases the W3C workgroup "Semantic Web Deployment" (Isaac, Phipps & Rubin, 2007) is working on.

The information gathered throughout this experiment, together with information and findings from other experiments, will be useful in other projects and will lead to substantial improvements of the existing techniques. One can mention as an illustration the "Improve access to the collections" objective of the TELplus project.[11] Among other things, this objective aims at exploring the techniques of vocabulary automatic alignment for a set of European collections, paving the way for creating a continuum within European heritage collections.

References

Balikova, M. (2005) Multilingual Subject Access to Catalogues of National Libraries (MSAC) Czech Republic's collaboration with Slovakia, Slovenia, Croatia, Macedonia, Lithuania and Latvia. Paper presented at the 71th IFLA General Conference and Council "Libraries - A voyage of discovery", August 14th - 18th 2005, Oslo, Norway. Available at: http://www.ifla.org/IV/ifla71/papers/044e-Balikova.pdf

Gendt, M. van, Isaac, A., Meij, L. van der, Schlobach, S. (2006) Semantic Web Techniques for Multiple Views on Heterogeneous Collections: a Case Study. In: Julio Gonzalo et al. (Eds.). Research and advanced technology for digital libraries: proceedings of the10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006), Alicante, Spain, September 17-22 2006. Berlin, Heidelberg: Springer. (Lecture Notes in Computer Science, 4172), 426-437. Available at : http://www.few.vu.nl/~aisaac/papers/STITCH-ECDL06.pdf

[11] Portal of the TELplus project: http://www.theeuropeanlibrary.org/portal/organisation/cooperation/telplus

Isaac, Antoine, Phipps, Jon and Rubin, Daniel (Editors) (2007). SKOS Use Cases. W3C working draft, 17 May 2007. Last update available at: http://www.w3.org/TR/skos-ucr/.

Isaac, Antoine and Summers, Ed (Editors) (2008). SKOS simple knowledge organization system primer. W3C working draft, 21 February 2008. Last update available at: http://www.w3.org/TR/skos-primer/.

Landry, P. (2007) Multilingual Access to Subjects Access: mise à jour du projet. Présentation à la BnF le 19 janvier 2007. Available at : http://rameau.bnf.fr/informations/pdf/MACS-bnf-2007.pdf

Miles, Alistair and Bechhofer, Sean (Editors) (2008). SKOS simple knowledge organization system reference. W3C working draft, 25 January 2008. Last update available at: http://www.w3.org/TR/skos-reference/.

Sémantique et interopérabilité. Journée d'étude BnF / AFNOR CG46 Référentiels, données d'autorité, thésaurus, ontologies, taxonomies... Pour en savoir plus ! Bibliothèque Nationale de France, Paris, 28 mars 2008. Available at : http://www.bnf.fr/pages/infopro/journeespro/pdf/AFNOR2008/Isaac.pdf

Shvaiko, P and Euzenat, J (2005) Ontology Matching. D-Lib Magazine, In Brief, 11(12). December 2005