Representing and Aligning Thesauri for an Integrated Access to Cultural Heritage Resources

Antoine Isaac
Vrije Universiteit Amsterdam
aisaac@few.vu.nl
Henk Matthezing
Koninklijke Bibliotheek, Den Haag
Henk.Matthezing@kb.nl

Abstract: In this paper, we show how Semantic Web techniques can help to solve semantic interoperability issues in the cultural heritage domain. In particular, these techniques can enable integrated access to heterogeneous collections by representing their controlled description vocabularies (e.g. thesauri) in a standardized format – Simple Knowledge Organization System (SKOS). We also present existing automatic alignment procedures that can assist cultural heritage practitioners to connect such vocabularies at the semantic level, building similarity links between the concepts they contain.

Keywords Semantic Web, thesaurus alignment, SKOS, semantic interoperability

1. Introduction: the Semantic Interoperability Problem

Currently, efforts are being put into building portals gathering collections from different institutions and containing heterogeneous material.[2] A crucial point for the success of these efforts is their ability to provide a seamless, unified access to different collections from a semantic perspective: when querying the portal for a given subject, a user should get *all* the material matching this subject, independently of its form(at) or location.

A way to achieve this is to rely on one single conceptual vocabulary for querying or browsing the objects contained in different collections. In this case integrated portals would be able to re-use the approaches and tools developed for individual collections, where objects are described and accessed by means of a dedicated knowledge organization system (KOS) – i.e. a thesaurus, a classification scheme or a subject heading list.

Still, difficulties remain with objects from different collections being most of the time described (indexed) using different vocabularies. These descriptions are not *interoperable at the semantic level*: when searching for objects showing "landscape with ruins" one will only retrieve objects that were indexed using this specific description of subject. This is of course not optimal if the portal contains objects indexed as "classical ruins", which is a conceptually similar subject description, but comes from another controlled vocabulary. If no help is provided by the system, the user has to cope with this heterogeneity problem, and has to query the database twice, using the two descriptions for getting objects from the two collections.

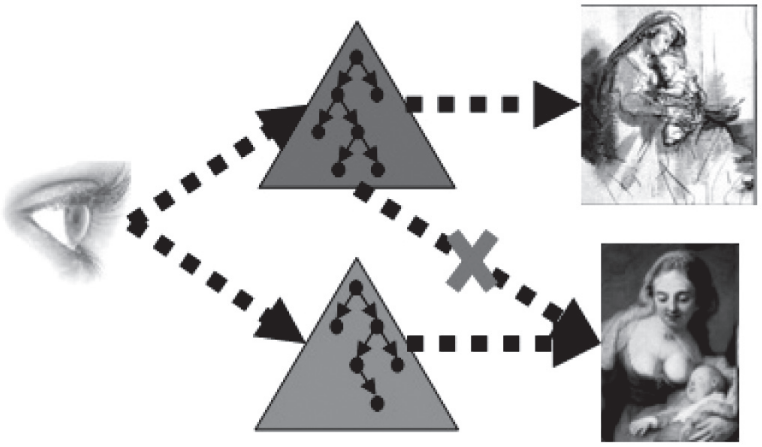

Figure 1 The semantic interoperability problem for collection access

In this paper, we insist on the need to solve two heterogeneity problems in order to enhance the interoperability of controlled vocabularies and, hence, of the systems and collections that use them:

- *representation heterogeneity*: vocabularies often come in models and formats that are not directly compatible, either because they mirror different general information needs (*e.g.* thesauri *vs.* classification schemes) or institutional and technical concerns.
- *conceptual heterogeneity*: vocabularies may contain concepts that have similar meanings (they can be just equivalent, or one can be more general than the other). Such similarity links have to be determined and exploited so that the final system can provide users with seamless access to content described by several vocabularies.

We will show how these problems can be addressed using solutions that are currently being investigated in the context of the Semantic Web research domain. In section 2 we describe some elements of the Semantic Web infrastructure, and illustrate how one can use the SKOS model to represent different vocabularies homogeneously. In section 3, we show how the representation of the different vocabularies can be semantically aligned to enable semantic integration of different collections.

2. Semantic Web techniques and controlled vocabulary representation

The Semantic Web (Berners-Lee, Hendler & Lassila, 2001) is a proposed extension of the existing Web, where information found on the web is augmented with machine-accessible knowledge[3]. The basic building blocks of the Semantic Web, as introduced by the Resource Description Format (RDF), are *resources* which denote any element that can be identified on (or even outside) the web. These resources are described by three-part *statements* that link them together. Each statement has a *subject* resource which is linked to an *object* resource via a *property* resource. Together, several such 'triplets' form a graph, such as the one represented in Figure 2. These graphs can contain:

- *factual knowledge: the third paragraph of the described document is about Amsterdam; the type of the described document is "Article";*

- *ontolo*gical knowledge: the Semantic Web is concerned about the way resources can be grouped in conceptual *classes*. These classes are introduced in *ontologies* that contain formally expressed knowledge about them. Here, Article is a class more specific than (a *subclass of*) Document[4].

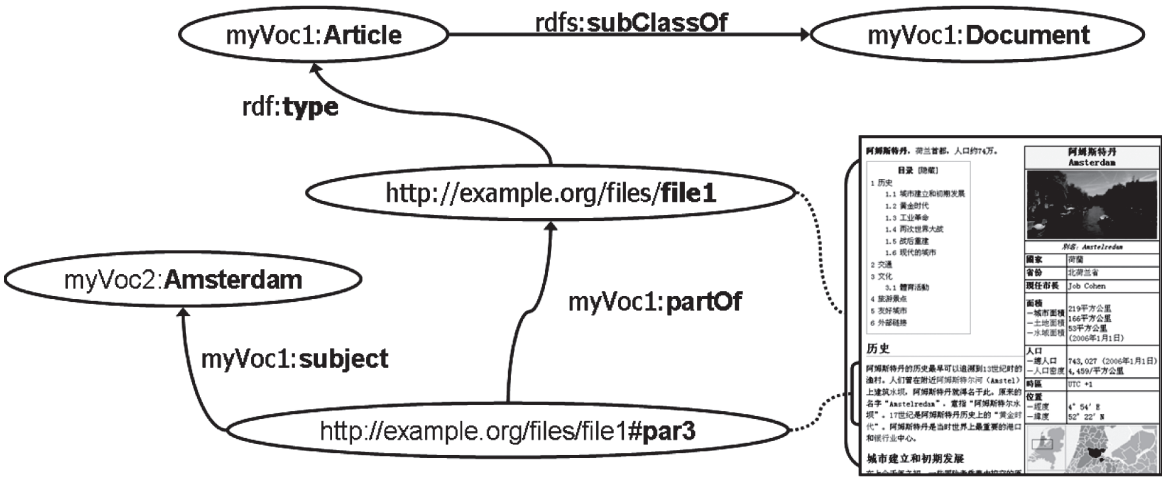

Figure 2 A Semantic Web RDF graph. "rdf:" namespace stands for "http://www.w3.org/1999/02/22-rdf-syntax-ns#", "rdfs:" for "http://www. w3.org/2000/01/rdf-schema#", "myVoc1:" for "http://example.org/voc1#", "myVoc2:" for "http://example.org/voc2#".

The semantic web research community is indeed well aware of the importance of having different sources of knowledge co-exist in a same space. Semantic Web data can merge resources coming from different information spaces. For example, the objects and links in figure 2 come from different *namespaces*, either user-defined (myVoc1:, myVoc2:) or proposed as standard (rdf:).

Semantic Web also promotes the porting of existing (meta-)data to semantic web formats like RDF. Repositories of concepts such as the KOS found in cultural heritage institutes or large companies are especially interesting, as they often provide consensual bases for exchanging information with shared and established meaning.

In principle, any controlled vocabulary designer can create his own meta-vocabulary (ontology), fitting his specific representation need – *e.g.* for a standard thesaurus or a sophisticated classification scheme. However, to help the first conversion efforts as well as the future exchange of such vocabularies, the World Wide Web Consortium (W3C) has initiated the development of a standard that allows representing the basic features of knowledge organization systems: SKOS[5].

SKOS introduces a set of constructs for RDF. These mainly allow the description of *concepts* and *concept schemes* (Miles & Brickley, 2005).

Concept description

SKOS has chosen a concept-based approach for the representation of controlled vocabularies. As opposed to a term-based approach, where terms from natural language are the first-order elements of a KOS, SKOS describes abstract concepts that may

have different materialization in language (lexicalizations). SKOS introduces a special construct skos:Concept[6] to properly characterize the (web) resources that denote such KOS elements. To further specify these conceptual resources, SKOS features:

- *Labelling properties*, e.g. skos:prefLabel and skos:altLabel, to link a concept to the terms that represent it in language. The prefLabel value shall be a non-ambiguous term that uniquely identifies the concept, and can be used as a descriptor in an indexing system. altLabel is used to introduce alternative entries – synonyms, abbreviations, etc. SKOS allows concepts to be linked to prefLabels and altLabels in different languages. The represented concept can thus be used seamlessly in multilingual environments.
- *Semantic properties* are used to represent the structural relationships between concepts, which are usually at the core of controlled vocabularies like thesauri. We have skos:broader, which denotes the generalization link (BT in standard thesauri), skos:narrower, its reciprocal link, and skos:related, the associative relationship (RT).
- *Documentation properties*. Often, informal documentation plays an important role in a KOS. SKOS introduces explanatory notes – skos:scopeNote, skos:definition, skos:example – and management notes – skos:changeNote, skos:historyNote, etc.

Concept scheme description

A KOS as a whole also has to be represented and described. SKOS coins a skos: ConceptScheme construct for this. It also introduces specific properties to represent the links between different KOSs and the concepts they contain. skos:inScheme asserts that a given concept is part of a given concept scheme, while skos:hasTopConcept states that a KOS contains a concept as the root of (one of) its constituent hierarchical tree(s), that is a concept without broader concept.

As an example, the UDC class 512, "Algebra", identified by the (as yet fictitious) resource http://www.udcc.org/udc/class_512, could be partly represented by the graph in Fig. 3.

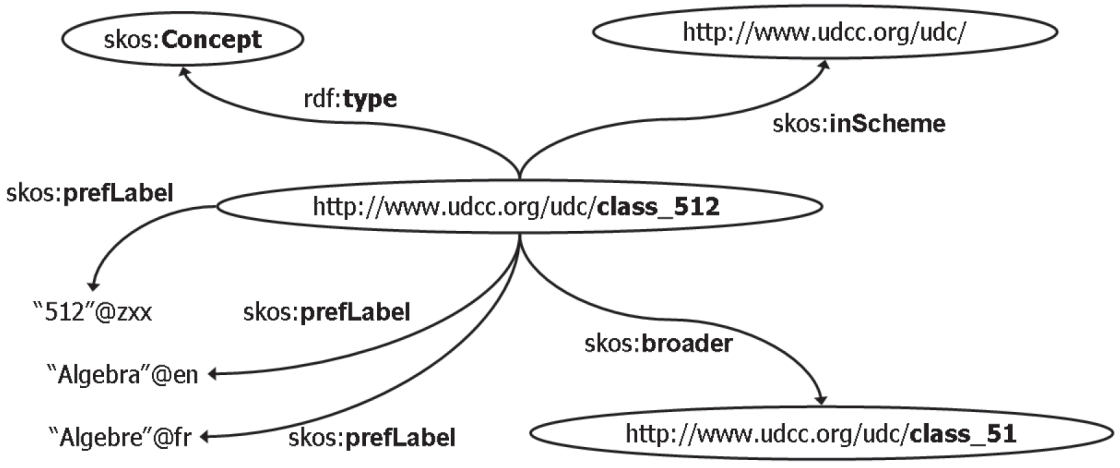

Figure 3 A SKOS graph partly representing the UDC class 512. "@" specifies the language of a label: "en" is the tag for "English", "fr" for "French" and "zxx" stands for any "artificial language".

3. Vocabulary alignment as a solution to the interoperability problem

Having unified and linkable representations of the concepts contained in different collections' vocabularies helps managing them in a single framework. However, this is not enough for solving the semantic interoperability problem mentioned in the introduction.
For this, one has still to find ways to determine semantic similarity links between the elements of the different vocabularies – to *align* them (Doerr, 2001). If a search engines knows that a concept C from a thesaurus T1 is semantically equivalent to a concept D from thesaurus T2, then it can return all the objects that were indexed against D for a query for objects described using C. We thus have an access to two different collections, using one single thesaurus.

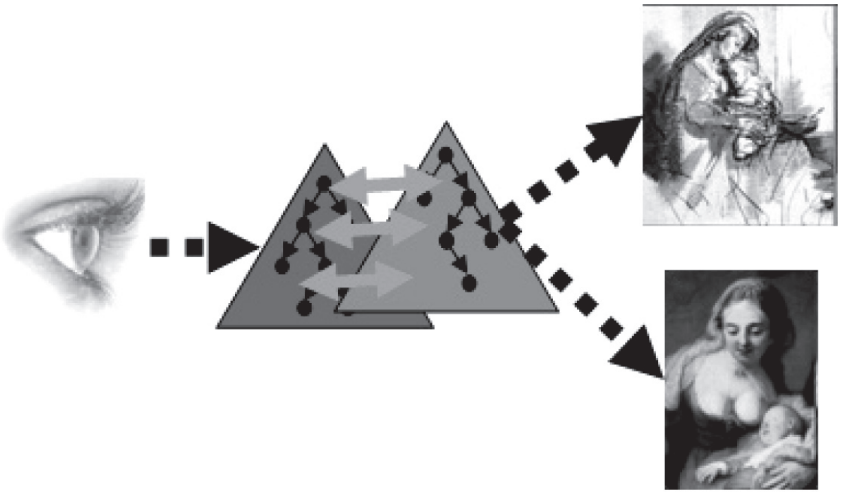

Figure 4 Using vocabulary alignment to for integrated access

To achieve this ideal situation requires a lot of work. Several projects in the past have tried to implement such a solution, like MACS[7] (Landry, 2004) and Renardus (Day, Koch & Neuroth, 2005). They have demonstrated very interesting results, but also highlighted the difficulty of manually aligning all the different vocabularies involved in practical cases. The vocabularies used in cultural heritage institutions, like the UDC, are generally very large - dozens of thousands of concepts. Furthermore, they evolve with time, which makes manual maintenance of alignment links an important problem.

In order to alleviate this problem, the creation of alignment tools has been proposed in order to provide for candidate mappings between two input vocabularies. Alignment would then become a (semi-)automatic task where human work is assisted. Recently, the Semantic Web community has produced a lot of alignment tools. These are aimed at the problem of formal ontology matching (Shvaiko & Euzenat, 2005) but the techniques they employ and the goals they advertise make them deployable in the context of more general KOS alignment – some of them actually originate from the older field of database research.

Most of the existing tools rely on sophisticated methods (Euzenat & Shvaiko, 2007). It is however possible to distinguish a number of basic techniques that use different materials, coming from the vocabularies or from other sources: lexical information attached to the concepts, structure of vocabularies, collection objects and external knowledge sources.

Lexical alignment techniques

In these techniques the lexical materializations of the concepts are compared together. If a significant similarity is found, then we can establish a semantic link between the concepts concerned. A straightforward example is when two concepts have the same label. But one can also search for string inclusion patterns or more complex techniques relying for instance on lemmatizers – getting normalized forms of labels – and syntactical analysis tools. A concept labelled "(map of) the North Pole" can be detected as a narrower concept of another which is labelled "Charts, maps". These lexical methods exploit the preferred labels of concepts, but they can also turn to their lexical variants or their associated definitions and scope notes. Of course, such approaches encounter the same problems as humans when dealing with words taken out of context. Polysemy and homonymy, for instance, are common sources of errors. This has to be compensated for by turning to contextual information.

Structural alignment techniques

A first kind of context is provided by the vocabulary itself, because of the hierarchical and associative links between the concepts it contains. These links, especially the hierarchical generalization and specialization ones, are useful to constrain a concept's natural interpretation: "bank" will be understood differently if it is a narrower term of "finance" or "geography". Some tools will analyze this semantic context, either to check similarities obtained by other techniques or to derive new similarities from existing ones. If two concepts from different vocabularies are semantically equivalent, their children in these vocabularies might also be semantically related.

Extensional alignment techniques

A second kind of context comes from the actual usage of the concepts in applications. For instance, a class from a classification scheme will be used to categorize a number of objects – *e.g.*, books – in a collection. Accessing this information will give an extensional characterization of the class' intended meaning – akin to its *literary warrant*. When documents are described against two different vocabularies[8], it is possible to use statistical techniques to compare the sets of documents described by the concepts from these vocabularies (Figure 5). A high degree of overlap between these sets of documents will yield a high similarity between corresponding concepts.

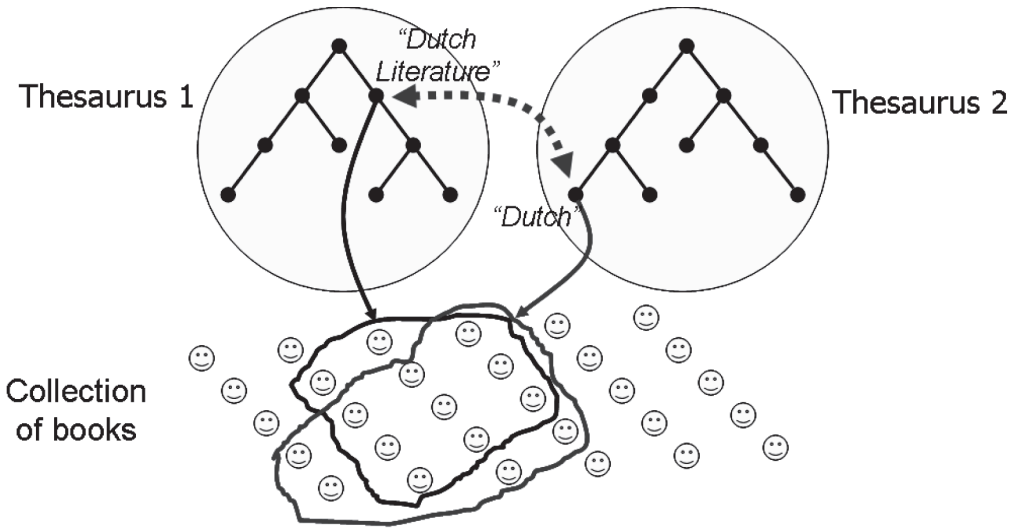

Figure 5 Using object-level information to align vocabularies [adapted from (Harmelen, 2005)]

Background knowledge-based alignment techniques

A final group of alignment methods rely on knowledge sources that are external to the application and the vocabularies being considered. These sources can be of a different nature, from synonym dictionaries to semantic networks like Wordnet (Miller, 1995). They are used as extra knowledge for other techniques, compensating for the lack of lexical information or semantic structure in the vocabularies to be aligned. For example a concept "calendar" from one thesaurus can be aligned to the more general "publication" from another thesaurus, using the hypernymy relation that stands between the two corresponding terms in Wordnet.

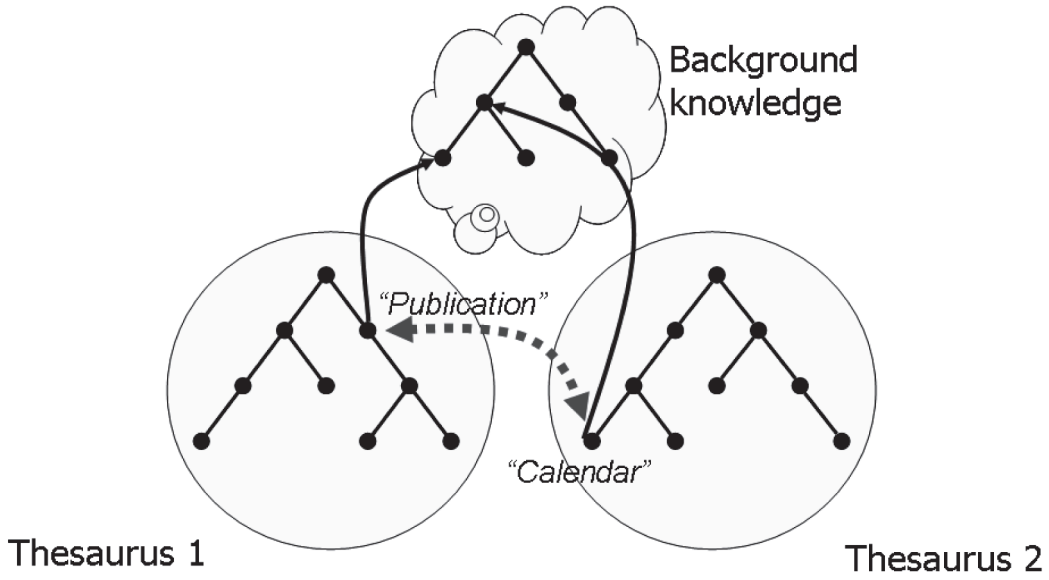

Figure 6 Using background knowledge to align vocabularies [adapted from (Harmelen, 2005)]

4. Discussion and conclusion

Existing alignment tools have been reported to perform relatively poorly on real-world cases like cultural heritage thesaurus alignment (Gendt et al., 2006). Alignment is still a difficult research problem, as no single technique among the ones mentioned in the previous section gives an ideal solution. Indeed, different techniques have to be selected and combined, depending on the characteristics of the case at hand, like the richness of the semantic structures of vocabularies, their lexical coverage or the existence of collections described by several vocabularies at the same time. However, the continuous development of new tools leads to significant improvements, as certified by the regular evaluation campaigns organized by the research community (Euzenat et al., 2006).

The Semantic Web-inspired methods and tools presented in this paper still require further experimentation in the context of real-world applications, and of course more accessible vocabularies. Nevertheless, the availability of current representation and alignment techniques already allows the building of demonstrators showing their potential value for integrating collections at the semantic level, leading from separate islands of cultural heritage knowledge to better connected networks of collections and vocabularies.

One such demonstrator is the (pilot) browser developed in the context of the STITCH project[9]. This browser enables a unified access to two collections of illuminated manuscripts,

using the description vocabulary used in the first collection, Mandragore,[10] or the one used by the second, Iconclass.[11] Other examples of Web portals illustrating the use of Semantic Web techniques in the cultural heritage domain can be seen on the websites of the *MuseumFinland*[12] and *eCulture*[13] projects.

In such experiments, porting vocabularies to SKOS plays an important role: it enables the reuse of off-the-shelf semantic web tools, including alignment ones. As an open, web-compatible standard, SKOS is also expected to improve the way controlled vocabularies are shared and re-used in many different applications, making most of the efforts invested in their design.

UDC can fit well in such a vision. As a multilingual vocabulary that is compatible with common information needs, it is an ideal pivot language for accessing different information sources. In a first category of scenarios, UDC can be used as background knowledge for alignment, as explained in section 3: it has a very good coverage of general subjects, as well as a rich and clean semantic structure. Here, access to collections is done via their original description languages, but UDC provides the ground for the alignment process. In a second category of scenarios, different vocabularies will be aligned to UDC, which will then be used to search and browse the collections described with these vocabularies. This is for example the approach chosen for the MSAC project (Balikova, 2005). Here, the access to collections is done directly by selecting relevant UDC classes.

Acknowledgements

The authors gratefully acknowledge support and ideas from the members of the STITCH project: Frank van Harmelen, who provided some excellent graphics, Lourens van der Meij, Stefan Schlobach, Shenghui Wang, Peter Wittenburg and Claus Zinn. STITCH is a project funded by NWO, the Dutch organization for scientific research.

Notes

1 Examples of such projects are The European Library – http://www.theeuropeanlibrary.org – and the Memory of the Netherlands – http://www.geheugenvannederland.nl.
2 The following is of course a very simplified introduction to the Semantic Web. For further detail, the reader is encouraged to read the Semantic Web Primer (Antoniou & Harmelen, 2004).
3 The information contained in ontologies is important, since it provides material for automated reasoning on the resources which populate the classes. For example, from the information found in Figure 2 for file1, Report and Document, an appropriate reasoning engine can infer that file1 is also an instance of the Document class, which will help to answer more queries.
4 SKOS stands for Simple Knowledge Organisation System. It is currently under scrutiny by the W3C Semantic Web Deployment Working Group – in which the author is involved – and is planned to be published as a W3C Proposed Recommendation in 2008. See http://www.w3.org/2004/02/skos/.
5 In the following "skos:" stands for http://www.w3.org/2004/02/skos/core#.
6 http://macs.cenl.org.

7 This also applies to the more general case when the similarity between objects from two collections described by their own vocabularies can be assessed, applying for example text similarity measures on textual documents.
8 http://stitch.cs.vu.nl/demos.html.
9 http://mandragore.bnf.fr.
10 http://www.iconclass.nl.
11 http://www.museosuomi.fi/.
12 http://e-culture.multimedian.nl/.

References

Antoniou, G., Harmelen, F. van (2004) Semantic Web Primer. Cambridge, MA: MIT Press.

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.

Berners-Lee, T., Hendler, J., Lassila, O. (2001) The Semantic Web. *Scientific American*, 284 (5), pp. 34-43. Available at: http://www.sciam.com/2001/0501issue/0501berners-lee.html.

Day, M., Koch, T., Neuroth, H. (2005) Searching and browsing multiple subject gateways in the Renardus service. In *Proceedings of the RC33 Sixth International Conference on Social Science Methodology*, Amsterdam, 2005.

Doerr, M. (2001) Semantic Problems of Thesaurus Mapping. *Journal of Digital Information*, 1 (8). Article No. 52. Available at: http://jodi.tamu.edu/Articles/v01/i08/Doerr/.

Euzenat, J. et al. (2006) Results of the Ontology Alignment Evaluation Initiative 2006. *International Workshop on Ontology Matching, 5th International Semantic Web Conference (ISWC 2006)*, Athens, Georgia, USA. Available at: http://www.dit.unitn.it/~p2p/OM-2006/7-oaei2006.pdf.

Euzenat, J., Shvaiko, P. (2007) *Ontology Matching*. Berlin, Heidelberg: Springer.

Gendt, M. van et al. (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), pp. 426-437.

Harmelen, F. van (2005) Ontology Mapping: A Way Out of the Medical Tower of Babel? In: S. Miksch, J. Hunter and E. Keravnou (Eds.). *Artificial Intelligence in Medicine: proceedings of the 10th Conference on Artificial Intelligence in Medicine*, AIME 2005, Aberdeen, UK. Berlin, Heidelberg: Springer. (Lecture notes in computer science, 3581), pp. 3-6.

Landry, P. (2004) Multilingual Subject Access: The Linking Approach of MACS. *Cataloging & Classification Quarterly*, 37 (3-4), pp. 177-191.

Miles, A., Brickley, D. (2005) *SKOS Core Guide*. W3C Working Draft. Work in progress, latest version available at http://www.w3.org/TR/swbp-skos-core-guide/.

Miller, G. (1995) Wordnet: a lexical database for English. *Communications of the ACM*,38 (11), pp. 39-41.

Shvaiko, P., Euzenat, J. (2005) Ontology Matching. *D-Lib Magazine*, 11 (12), In Brief. Available at: http://www.dlib.org/dlib/december05/12inbrief.html.

[End]