

Integration of distributed terminology resources to facilitate subject cross-browsing for library portal systems

Libo Eric Si, Ann O'Brien, Steve Proberts
Department of Information Science, Loughborough University.

Abstract

Purpose: To develop a prototype middleware framework between different terminology resources in order to provide a subject cross-browsing service for library portal systems.

Design/methodology/approach: Nine terminology experts were interviewed to collect appropriate knowledge to support the development of a theoretical framework for the research. Based on this, a simplified software-based prototype system was constructed incorporating the knowledge acquired. The prototype involved mappings between the computer science schedule of the Dewey Decimal Classification (which acted as a spine) and two controlled vocabularies UKAT and ACM Computing Classification. Subsequently, six further experts in the field were invited to evaluate the prototype system and provide feedback to improve the framework.

Findings: The major findings showed that given the large variety of terminology resources distributed on the web, the proposed middleware service is essential to integrate technically and semantically the different terminology resources in order to facilitate subject cross-browsing. A set of recommendations are also made outlining the important approaches and features that support such a cross browsing middleware service.

Originality/value: Cross browsing features are lacking in current library portal meta-search systems. Users are therefore deprived of this valuable retrieval provision. This research investigated the case for such a system and developed a prototype to fill this gap.

Keywords: mapping, semantic interoperability, knowledge organisation systems, KOS, library portal, and cross-browsing.

Classification: research paper.

1. Introduction

A number of library portal systems have been developed and applied in libraries, some of which offer meta-search engines. The basic capability of these meta-search engines is to allow users to enter queries and get results returned from heterogeneous resources. In order to develop these meta-search services, a wide range of mappings between different metadata schemes used by different services have been established. However, because different knowledge organisation systems (KOS) are used to describe the different metadata records within different metadata repositories, and these KOS differ in their subject areas, degree of pre-coordination/post-coordination, level of granularity, language, etc., the development of meta-search services has been impeded by the heterogeneity of different KOS. With the exponential increase in

scholarly information resources, information is often indexed using different vocabularies and browsed using different subject structures. End-users have to switch mental models between different KOS, and re-familiarise themselves with different terminologies. In addition, due to the lack of established conceptual mapping between KOS, most of these meta-search services do not provide subject cross-browsing services. Subject cross-browsing is particularly helpful for inexperienced users or for users not familiar with a subject, its structure and terminology.

This research aimed to develop a framework to improve the interoperability between different KOS used by different metadata repositories, and facilitate subject cross-browsing for library meta-search services, such as Ex Libris MetaLib, and MuseGlobal SingleSearch. The paper is organised as follows: Based on reviewing relevant literature, Section 2 outlines suitable methods to improve the interoperability between different KOS and points out the necessary requirements for the development of such a framework for subject cross-browsing services. Section 3 introduces the research methodology applied. Section 4 describes the findings and discusses the rationale and principles used to develop the theoretical framework for facilitating subject cross-browsing. Section 5 presents some recommendations for further development.

2. Literature Review

Prior work has been undertaken in exploring theories to improve interoperability between different KOS (BS8723-Part4, Chan and Zeng 2004, Koch *et al* 2001, Chaplan 1995, McCulloch *et al* 2005). Among these theories, the main method focuses on establishing terminology mapping between different KOS. The basic idea of establishing concept mapping consists of identifying different mapping relationships between concepts from different KOS (Miles and Brickley 2004, Chaplan 1995, Vizine-Goetz *et al* 2004 and Koch *et al* 2001), and establishing the equivalence between two or more concepts (Chan and Zeng 2004). A number of relevant factors challenging semantic mapping work needed to be considered before conducting the mapping work. These include:

- structural models for mapping (BS8723-Part4)
- the direction of the mapping (BS8723-Part4)
- methods to distribute a huge amount of mapping work to different participants (Koch *et al* 2001a)
- how compound concepts are handled (BS8723-Part4)
- automatic mapping solutions (Issac *et al* 2007)
- the top-level metadata schemes to describe different KOS (Zeng 2008, and Golub and Tudhope 2008).

There are also a number of technical factors that challenge compatibility between different KOS (Si 2007, Tudhope and Binding 2008, Tuominen *et al* 2008, and OCLC Research 2008). These technical factors include:

- different concept identification mechanisms (URI, local identification mechanisms, subject-based identification systems, etc)
- knowledge representation formats to encode KOS content (e.g. SKOS, Zthes XML Schema, DD-8723, MARC21, etc.)
- protocols to access KOS content (e.g. SRW, SRU, Z39.50,)

- and database systems where different KOS are located (RDF triple stores, relational databases, etc).

A wide range of terminology resources using different techniques are currently distributed on the web. These terminology resources include:

- Terminology services, such as OCLC Terminology Service, HILT Terminology Service, etc., which were developed as shared services that allow other services to access their terminological data
- Controlled vocabularies, which are used to index important collections. They are represented in particular encoding formats, and published on the Web
- Mapping sets between different controlled vocabularies which are represented in encoding formats, and published on the Web
- Local vocabularies which are used by library portal systems for local subject indexing and cataloguing.

It is worth noting that in most current cases, one terminology resources only holds a small amount of terminological data, and it may be impossible to have a terminology resource that includes all the KOS required in a library portal. For example, the HILT terminology service, as the largest terminology service in the UK, only includes thirteen controlled vocabularies. A number of fairly widely-used vocabularies, such as UDC, LCC, BLISS, UKAT, are not covered.

It would therefore, be useful to develop a middleware platform to integrate technically and semantically different terminology resources, and then provide subject cross-browsing services to different library portal systems. The next section will introduce the specific research methods to develop such a platform.

3. Methodology

From the literature review, a number of essential elements to improve the semantic and technical interoperability between different terminology resources were identified. However, it is still not clear how elements, such as technologies, standards, and semantic methods need to be combined to make up a middleware platform, and there is a lack of developmental effort into terminology mapping research in the real world. Most previous research has been context-dependent, and may not be within the scope of this research. As a result, the findings of the literature review could not be directly translated into a real system implementation.

For this reason, it was decided to collect in-depth ideas from a number of experts, who were involved in different terminology mapping projects, and match these ideas to the objectives of this research. Nine expert interviews were conducted, in which the interviewer introduced the research context to the experts and encouraged them to discuss the issues of importance to the development of the framework. Much valuable data were gained from the interviews. These data were analysed to form a basis to develop a theoretical framework.

Before using appropriate technologies, standards, and semantic methods to develop the framework, it was necessary to investigate a number of KOS used in different collections. Without a clear insight into the different characteristics of various KOS, it is impossible to make the right decision on using appropriate methods to establish the mappings between these KOS. For this reason, an investigation was conducted to

review a variety of KOS and their characteristics, and the findings formed a basis to establish the mappings.

Subsequently, a Design Research approach (Hevner *et al* 2004) was undertaken, and a prototype system was developed based on the guidelines and theories. The prototype was used as a basic platform for evaluating the theories developed. Dewey Decimal Classification (DDC) was used as the switch language. Mappings were created within the computer science section between DDC (which acted as a spine) and two controlled vocabularies UKAT and ACM Computing Classification. A range of heuristics were developed for expert evaluation. These heuristics were translated into a set of questions. A number of user interaction tasks were designed to enable six expert consultants (not those who were previously interviewed) to walk through the prototype system. The experts were then asked to answer the relevant questions based on the developed heuristics, and provide suggestions to improve the system. The evaluation findings became the foundation to develop the final theories.

4. Research discussion

The following discussion reports the findings of each stage of the research. Nine expert interviewees and six consultant evaluators were in agreement with the different stages. This section outlines the overall findings resulting from the initial interviews and proposed amendments from the evaluation.

4.1 Structural model of terminology mapping

Establishing ‘many-to-many direct mappings’¹ between different KOS is a very precise method to facilitate subject cross-browsing and cross-searching. Based on this approach, a user can be navigated by any vocabulary to get directly mapped terms from all other KOS without interacting with a mediator. However, this approach is very labour-intensive and time-consuming. In a large information environment where there are also a large number of KOS, this approach is not suitable, because establishing direct mappings is very costly. It was therefore considered appropriate to apply or create a switch language to exchange terminological information between a number of KOS. A number of requirements of the switch language are identified below:

Requirement 1: It is important to use a switch language that has great granularity and covers most subject areas.

Requirement 2: A switch language should be well-known across different communities.

Requirement 3: A switch language should be encoded in a well-defined interchange format.

Requirement 4: A switch language should have excellent concept synthesis capability.

Figure 1 represents these four requirements, and how different vocabularies are mapped into each category.

¹ This refers to establishing equivalence between two KOS.

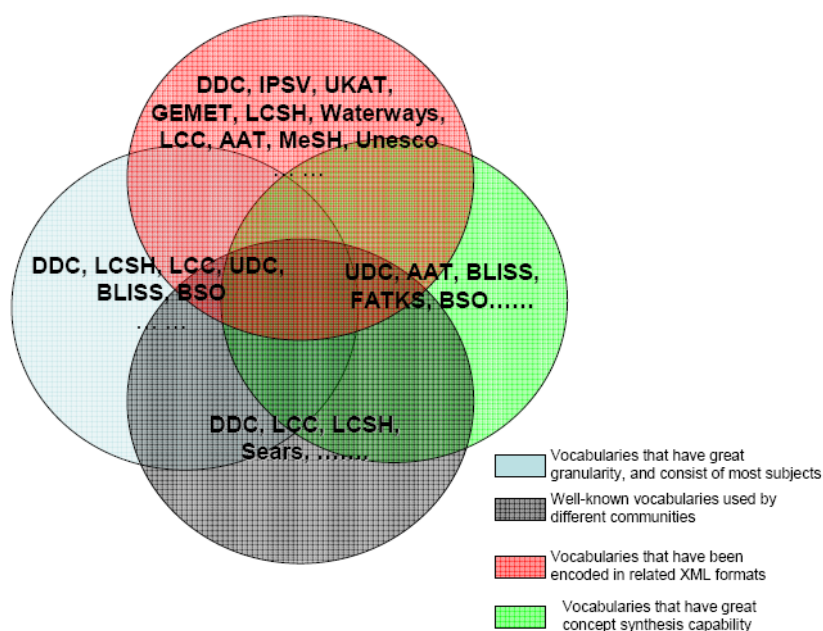


Figure 1: Four requirements that a switch language should possess

It was difficult to find an existing vocabulary to act as a switch language to achieve all the four requirements. However, some vocabularies, such as, DDC, can achieve three of these requirements.

A number of faceted classification schemes, such as UDC, BC2, etc., with their great notational capability and detailed subject coverage may have great potential to be powerful switch languages in the future. However, the disadvantages of these classification schemes are that they currently lack any encoding work that represents complex faceted classification data (e.g. UDC). It was therefore decided to use DDC as a switch language. This makes sense in the medium to long term because:

1. DDC can offer a limited capability of notational synthesis. For example, it is possible to combine the 026 (library) and 780 (music) into a compound concept 026.78 (music library).
2. DDC is not only a widely-used classification scheme used by many academic libraries throughout the world, but also has been applied as a switch language by a number of terminology services such as the HILT terminology service, OCLC terminology service, Renardus, etc.
3. DDC has been encoded in MARC21 XML data format
4. Many metadata records have been indexed not only by DDC, but also by other vocabularies.

4.2 Indirection problems caused by the use of DDC as a switch language

When a user interacts with a local vocabulary structure to exchange terminological data with other external KOS through a DDC spine, there is a two-step journey from the local taxonomy through DDC to the other KOS, this problem of indirection, which might cause loss of precision, was highlighted. For example, it is possible that a term “house cat” in a local KOS can only be mapped to DDC concept “cat”, but in another vocabulary, there is an exact term called “house cat”. In this situation, precision is lost.

Based on the findings, two possible solutions were identified. In the first, when the mapped terms returned through the DDC spine are not appropriate to a user's subject requirements, it might be necessary to give users the option to further refine the search term by local browsing or by expanding the mapped concepts within the vocabulary, and then reformulating the subject search before searching within collections. It would be helpful to develop some query expansion algorithms to return more terms considered semantically close to the mapped concepts, and show these expanded terms to the users. The users might then be able to find more appropriate terms to further refine their search. This solution encourages users to combine their own intelligence with machine intelligence (query expansion algorithms) to consider and compare these expanded terms, and make judgement on selecting the most appropriate subject terms to refine their searches.

In the second solution, when the mappings between the local taxonomy and different KOS are established through DDC as a switch language, it would be possible to use the existing mappings to further develop the direct mappings between the local taxonomy and different KOS. This would be accomplished by those responsible for doing the mapping. For example, query expansion algorithms can be used to expand the mapped terms from the KOS used by different databases to return a number of concepts considered semantically close. This may enable the mapping workers to find more appropriate terms from the expanded concepts, and select some of the expanded concepts to establish more accurate direct mappings between the local taxonomy and different KOS. This solution aims to combine human mapping workers' intelligence with machine intelligence (query expansion algorithms) to consider and compare these expanded terms, and then create more accurate direct mappings. Figure 2 shows an example of a query expansion algorithm semantically expanding a mapped KOS concept to return a number of its narrower concepts, and the mapping workers selecting one of the narrower concepts to establish the direct mapping with the concept in the local taxonomy.

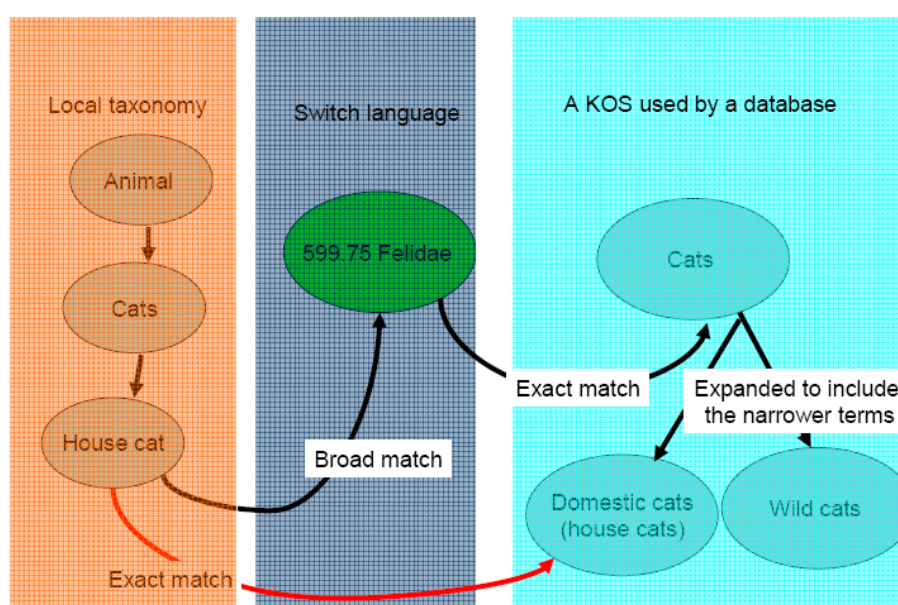


Figure 2: Discovering a direct mapping based on established mappings

In this figure, the black curved lines represent the mappings already established between the local taxonomy and switch language, and between the switch language

and an external KOS. The red curved line represents the mapping discovered by the mapping worker, based on analysing the established mappings between the local taxonomy and the switch language, and between the switch language and the external KOS.

4.3 Who should create the mappings?

Creating and maintaining the mappings is possibly the most important issue in any mapping-based solution. Based on the interview findings, it was found that a variety of participants, such as KOS owners, terminology service providers, local institutional subject librarians, etc., could potentially create the mappings for this middleware framework. Also, the importance of reusing existing mappings from other terminology services was highlighted by most experts. It is, therefore, important to ensure programmatic interfaces enabling access a variety of terminological data across the web. There eventually could be a number of databases storing different sets of mapping data. However, different terminology services may use different mapping strategies, such as provenance (source), methods (intellectual, co-occurrence, other automatic, etc), concept indicators, and so on. This might cause inconsistency between different mapping sets. In many cases, established mapping data might not be suitable for some particular use scenarios. If so, it is suggested that one central team with sufficient expertise should be formed to assess these distributed mapping data sets and should have the responsibility for improving the consistency and quality of distributed mapping resources.

4.4 Treatment of compound concepts

Based on the discussion in Section 4.1, DDC was selected as a switch language for mapping. It is a pre-coordinated controlled vocabulary that includes a number of compound concepts. When a post-coordinated vocabulary, in which most of the concepts are individual terms, is mapped against DDC, it is important to combine several relevant concepts in the post-coordinated vocabulary to map against one concept in DDC. For example, DDC concept “020.2854678 Internet—libraries” can be matched against the combination of the UKAT concepts “libraries” and “internet”. There are a variety of “connectors” that are able to combine different concepts from one vocabulary. These “connectors” might include Boolean Operators (and, or, not), facets (time, place, people, event, etc), and ontological relations. Interviews with experts showed that the best way to use these “connectors” for mapping is still an open issue, and that some concept connectors may make the mapping more complicated.

With this in mind, one approach would be not to use these connectors at all. Instead, a number of relevant concepts could be put into a “bag”, and the bag is mapped to an equivalent DDC concept. The bag becomes a very abstract concept that may not have a clear meaning, but based on the evaluation findings, it was widely-agreed that using a bag to combine a number of concepts together is a good idea. Figure 3 shows how to use a bag to combine a number of concepts, and match the bag against a relevant compound concept.

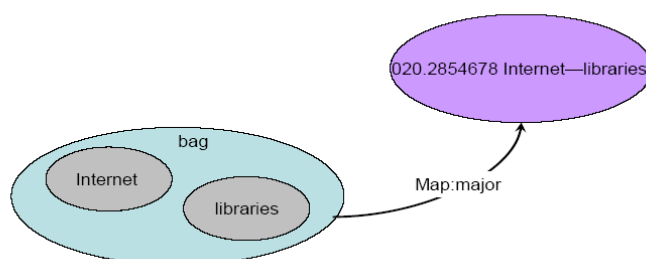


Figure 3: Treatment of a compound concept

In the prototype solution created in this research, when a number of individual concepts are returned from a bag, and listed to the users, users can select some terms, or use appropriate Boolean operators to combine some of the selected terms in the bag to further their searches. One evaluator suggested that when a number of mapped terms are returned and listed, many users might not realise that they could use Boolean operators to further combine these mapped terms. Another evaluator suggested that using a Google-styled “Do you mean...?” sentence plus all possible Boolean combinations of mapped terms might be an appropriate way to present these terms to the users. For example, it is possible to ask “Do you mean internet, libraries, internet and libraries, internet or libraries ?” to present all options for a user’s further subject search.

4.5 Technical Architecture

As mentioned in Section 2, there are a number of technical factors challenging the interoperability between different terminology resources. Based on the interview findings, it was impossible to identify one standard technology for all applications. For example, although it was felt that RDF data would become more and more widely-used in different communities, MARC21 XML in the library community is still the most appropriate, accurate, and cost-effective format to encode library data. This research, therefore, focused on developing a middleware platform to cross-access terminology data from terminology resources that use different formats, access protocols, identification mechanisms and that are located in different database systems. In order to cross-access heterogeneous terminology resources, a knowledge base was developed to store connectivity details of different terminology resources. The purpose of the knowledge base was to translate the users’ queries into appropriately structured queries that the different terminology resources could understand, and convert the returned terminological records into a consistent format. In the knowledge base, a number of appropriate APIs were employed to translate a user’s query, and a number of XSLT files were developed for data conversion. Because of the conversion programmes, the knowledge base can become a data converter that converts different terminological resources into the formats that different clients need, and converted data can be aggregated and stored into a new database system as a new terminology service. Figure 4, for example, indicates that through this framework, different terminological resources are “SKOSified”, and that “SKOSified” data are stored in a database as a terminology service for various web clients. In this example, the middleware system can be used as a “SKOSification” tool to convert different vocabularies into SKOS format.

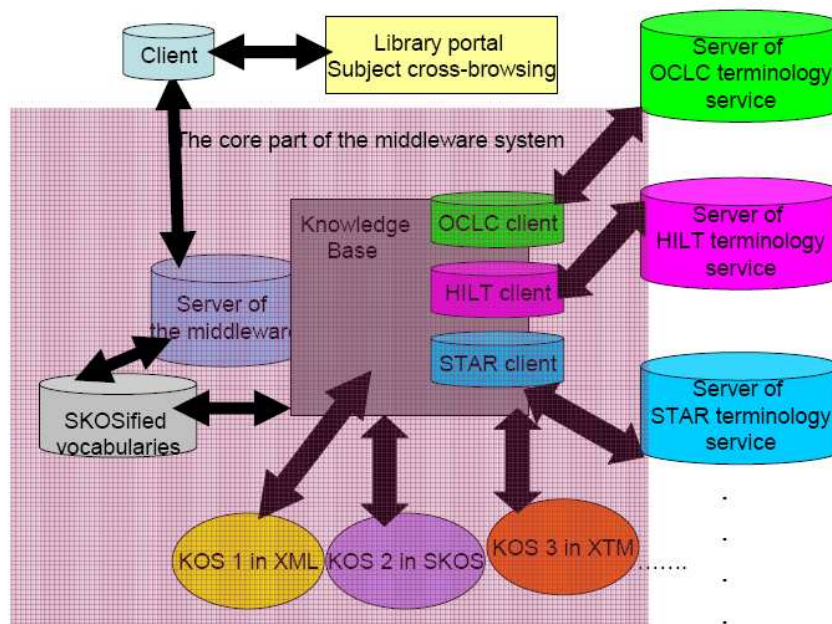


Figure 4: The middleware as “SKOSification” tool

In another example, it is possible that through this middleware system, different terminological resources could be converted into MARC21 XML format, and then different clients could use MARC21 XML data to support their subject cross-browsing. See Figure 5.

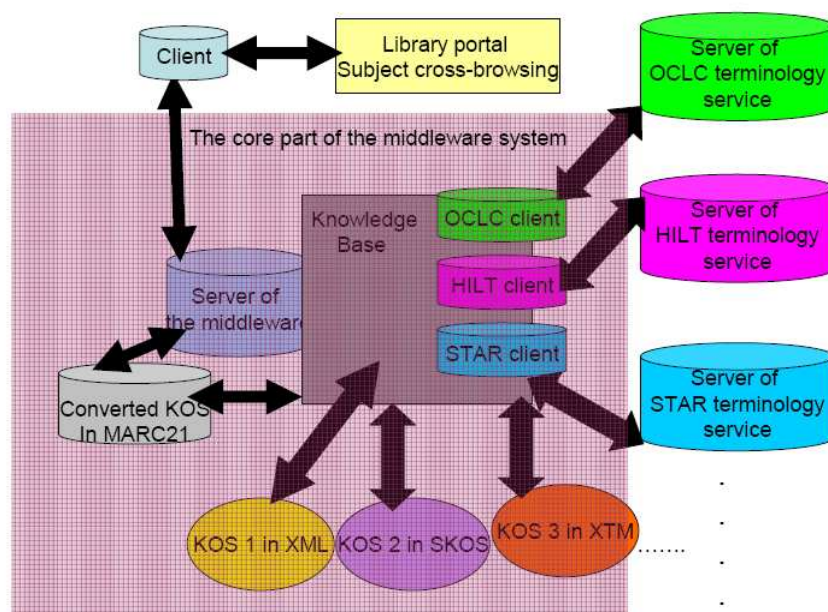


Figure 5: The middleware as a tool to convert data into MARC21 XML

4.6 Machine-to-Machine Interaction with Meta-search engines and collection registry

Based on reviewing a number of subject cross-browsing projects (e.g. HILT, Renardus), it was found that most subject cross-browsing services can return mapped

conceptual terms, but that end-users are more likely to be concerned with gaining the relevant metadata records through subject cross-browsing. The interview findings pointed out that it is important to enable a subject cross-browsing service to interact with meta-search services provided by library service vendors. When a user selects a mapped term returned from a particular KOS, the mapped conceptual terms could become queries against the specific databases that were indexed by this KOS. In this context, a database registry that records the usage of KOS in different databases should be developed. The purpose of this registry would be to enable the mapped conceptual terms from each particular KOS to become meta-search queries against the specific databases that are indexed by this KOS. See Figure 6 as an example. The specific steps are described below:

1. Users interact with a subject cross-browsing interface, and get a number of mapped conceptual terms from various KOS;
2. The database registry is used to enable the federated search service to use mapped terms to search against the relevant databases using these mapped terms. Thus, a number of item-level metadata results could be returned from different databases;
3. The item-level metadata results returned would be converted into a consistent format, and presented to the end-users;
4. A ranking algorithm should be developed based on the five different types of mapping relationships (exact, broad, narrow, related, and close).

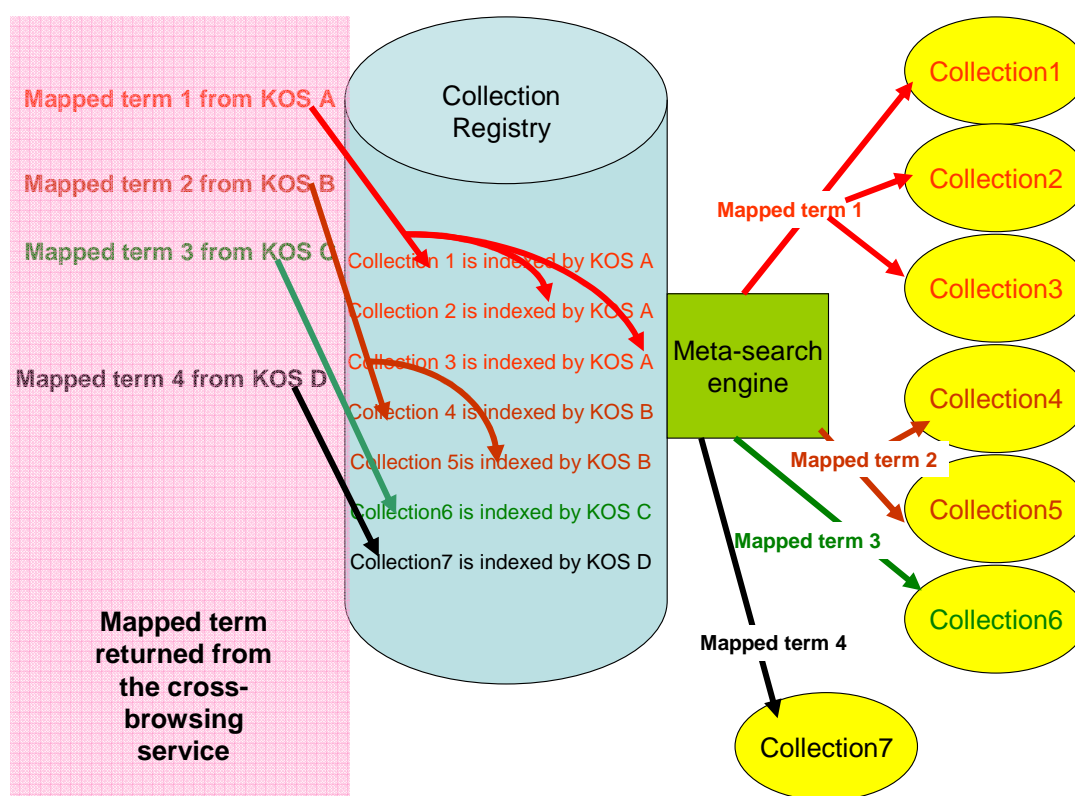


Figure 6: M2M interaction between the framework and a meta-search engine

In Figure 6, the collection registry is responsible for splitting these mapped terms from different KOS into different query terms for searching against the databases indexed by the relevant KOS.

5. Recommendations

Based on a combination of the literature review, findings from the research, and self-reflection, several recommendations are presented as follows:

Recommendation 1—The structural model for the development of such a framework

There are various advantages and disadvantages to mapping different vocabularies to a DDC spine. A number of terminology mapping projects (HILT, Renardus, OCLC TS, etc.) are committed to this approach. Given the nature of many other academic databases which include DDC data, this still makes sense in the medium term.

However, with the further development of faceted classification schemes with great notational synthesis capability, such as BLISS, BSO, etc., these might be a better option than DDC. Thus, it is recommended to further explore a switch language with better notational synthesis capability by employing advanced faceted classification theories, and explore the methods to encode these classifications in semantic web-enabled formats to improve the reusability of these classifications.

Recommendation 2—An approach to improving the consistency of the mappings from different terminology resources

A variety of mapping initiatives have been proposed and developed based on different mapping strategies. The mappings from different initiatives may vary in different features, such as their provenances (source), methods (intellectual, co-occurrence, other automatic, etc), subject indicators, encoding formats. The terminology services holding the mappings also vary. For this reason, it is recommended to develop a metadata application profile to characterise these features, and a centralised team with mapping expertise should be formed to investigate the different characteristics of these mappings. They should focus their intellectual effort on enhancing the consistency and quality of the mapping data from different sources.

Recommendation 3—Technical architecture

Because there have been a number of existing terminology mapping services that use different representation formats, access protocols, API functions, and query languages, it is recommended to develop a knowledge base, which is able to record functionality provided by the different terminology services, translate the users' query into different forms of the queries that different terminology services can accept, and convert different results into a consistent format.

Recommendation 4—The use of bag to combine relevant individual concepts

When creating a mapping between several individual concepts and a compound concept, it is recommended to use a bag to combine these individual concepts, and map this bag against the compound concept. When presenting the mapped bag of concepts to end-users, it is recommended to use a Google-styled "Do you mean...?" plus all possible Boolean combinations of mapped terms. Thus, users can easily select different combinations of mapped terms relevant to their subject needs.

Recommendation 5—The use of query expansion algorithm to expand the mapped term

When the mapped terms are returned from the middleware platform, and they are not suitable for the users' subject needs, users might be frustrated by the returned results. In this case, it is recommended to use a query expansion algorithm to expand the mapped terms to produce a number of terms considered semantically close to the

mapped terms. In this context, the users could use their intelligence to re-formulate their subject queries.

This research highlighted that given the variety of terminology resources on the web, it is important to consider the integration of these resources on both technical and semantic levels. A number of recommendations and guidelines were outlined. Hopefully, these recommendations and guidelines will be able to provide fertile ground and act as incentives for the development of subject cross-browsing services in library portal systems.

Bibliography

BS8723-Part4 (2008), *Structured vocabularies for information retrieval, Part 4: Interoperability between vocabularies*, London: British Standards Institution.

Chan, L.M. and Zeng, M. (2004), "Trends and issues in establishing interoperability among knowledge organization systems", *Journal of the American Society for information science and technology*, Vol.55 No.5, pp. 377-395.

Chaplan, M.A. (1995), "Mapping laborline thesaurus terms to Library of Congress Subject Headings: implications for vocabulary switching", *Library Quarterly*, Vol.65, No.1, pp.39-61.

Golub, K. and Tudhope, D. (2008), "*Terminology registry scoping study (TRSS): Excerpt on metadata*", Available at: <http://www.ukoln.ac.uk/projects/trss/dissemination/metadata.pdf> (accessed 09.12.08).

Hevner, A., March, S., Park, J. and Ram, S. (2004), "Design science in information systems research", *MIS Quarterly*, Vol. 28, No.1, pp.75-105.

Isaac, A., *et al* (2007), "Integrated access to cultural heritage resources through representation and alignment of controlled vocabularies", *Library Review*, Vol. 57, No.3, pp.187-199.

Koch, T., Neuroth, H., and Day, M. (2001), "Renardus: cross-browsing European subject gateways via a common classification system (DDC)", available at: <http://www.ukoln.ac.uk/metadata/renardus/papers/ifla-satellite/ifla-satellite.pdf> (accessed 09.09.08).

Koch, T., Neuroth, H. and Day, M. (2001a), "*DDC mapping report, Renardus D7.4.*", available at: <http://renardus.sub.uni-goettingen.de/wp7/d7.4> (accessed 18 May 2009).

McCulloch, E., Shiri, A. and Nicholson, D. (2005), "Challenges and issues in terminology mapping: a digital library perspective", *Electronic library*, Vol.23, No.6, pp.671-677.

Miles, A. and Brickley, D. (2004), "*SKOS mapping vocabulary specification*", available at: <http://www.w3.org/2004/02/skos/mapping/spec/> (accessed 9 March 2009).

OCLC Research, (2008), “*Terminology services: experimental services for controlled vocabularies*”, available at: <http://tspilot.oclc.org/resources/overview.pdf> (accessed 20 May 2009)

Si, L. (2007), “Encoding formats and consideration of requirements for mapping”, paper presented in NKOS 2007, Budapest, available at: <http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2007/papers/abstracts.html#si> (accessed 2 March 2009).

Tudhope, D., and Binding, C. (2008), “*Machine understandable knowledge organisation system*”, available at: <http://hypermedia.research.glam.ac.uk/media/files/documents/2008-07-05/SIEDL08-Tudhope-v3.pdf> (accessed 9 November 2008).

Tuominen, Jouni *et al*, (2008), “*ONKI-SKOS – Publishing and utilizing thesauri in the semanticweb*”, available at: <http://www.seco.tkk.fi/publications/2008/tuominen-et-al-onki-skos-2008.pdf> (accessed 1 March 2009).

Vizine-Goetz, D., *et al.*, (2004), “Vocabulary mapping for terminology services”, *Journal of digital information*, Vol.4, No.4, available at: <http://jodi.tamu.edu/Articles/v04/i04/Vizine-Goetz/> (accessed 21 May 2009).

Zeng, L. (2008), “Registry requirements and issues”, paper presented in *NKOS 2008*, Aarhus, Denmark, available at : <http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2008/programme.html> (accessed 20 May 2009).