A SKOS Core approach to implementing an M2M terminology mapping server

George Macgregor, Anu Joseph and Dennis Nicholson

Centre for Digital Library Research, Department of Computer & Information Sciences, University of Strathclyde, Glasgow, UK
{george.macgregor, anu.joseph, dennis.nicholson}
@cis.strath.ac.uk

Abstract: The proliferation of distributed digital libraries and repositories has increased the need for improved interoperability between terminologies in order to facilitate user access to the discrete heterogeneous digital objects held therein. The emergence of the Simple Knowledge Organization System (SKOS) Core is a useful development in this context. In this paper we describe a SKOS Core approach to implementing a web services (i.e. M2M) terminology server employing terminology mapping and using SKOS Core to wrap terminology responses. Aspects advantageous to this approach are explored, as are issues and areas for future research.

Keywords: terminologies, interoperability, Knowledge Organization Systems, SKOS Core, resource discovery, information retrieval

1 Introduction

Knowledge Organization Systems (KOS) encompasses a variety of disparate *terminologies* designed to present a systemised interpretation of knowledge (Zeng & Chan, 2004). KOS can include term lists (e.g. authority files, glossaries, gazetteers, etc.), classification schemes (e.g. bibliographic classification schemes, taxonomies, etc.) and relational vocabularies (e.g. thesauri, subject heading lists, etc.). The proliferation of digital libraries and repositories has increased the need for improved interoperability between terminologies in order to enhance user access to discrete heterogeneous digital objects (Chan & Zeng, 2002). This is particularly true within distributed resource discovery contexts in which digital objects are indexed and organized according to a variety of terminologies, perhaps deriving from disparate KOS. In such contexts it is impractical for users to query each repository individually or to acquaint themselves with the variety of terminologies in use. Simultaneous searching and browsing of multiple distributed repositories is therefore considered increasingly desirable and research exploring techniques designed to artificially or intellectually

In: *Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007), 21-23 February 2007, Bangalore, India.* Documentation Research & Training Centre, Indian Statistical Institute, Bangalore. pp.109-120.

augment cross-repository subject interoperability continues to be a significant area of study (e.g. Binding & Tudhope, 2004; Doerr, 2001; Koch, Neuroth & Day, 2003; Nicholson, Dawson & Shiri, 2006; Zeng & Chan, 2004).

The Simple Knowledge Organization System (SKOS) Core model (Miles & Brickley, 2005) offers a means of expressing the structure of 'concept schemes' on the web, facilitating the implementation of operational terminology services within a machine-to-machine (M2M) web services context. In this paper we describe and propose an approach to implementing a pilot M2M terminology server employing *terminology mapping* and using SKOS Core to mark-up terminology responses.

The remainder of the paper is structured as follows. We contextualise our work by defining terminology mapping and provide brief details of the underlying design of the terminology mapping server in section 2. SKOS Core is briefly introduced in section 3. The crux of the paper (section 4) describes the way in which SKOS Core is deployed in our system and explores the potential for server functions. Discussion, issues, areas for future research and development, and conclusions are addressed in section 5.

2 Terminology mapping

Before introducing SKOS Core, it is first necessary to define what is meant by terminology mapping. Mapping essentially involves relating equivalent terms, concept or hierarchical relationships, from one terminology to another (Doerr, 2001). The process of terminology mapping remains largely an intellectual process and is consequently heavily reliant upon human intervention. Within particular scenarios equivalence between terms can be derived via computational means (Vizine-Goetz, Hickey, Houghton & Thompson, 2004); however, most of these approaches still require significant human resources to verify and/or amend erroneous equivalences (McCulloch, Shiri & Nicholson, 2005). Research has therefore focussed on terminology *switching* to reduce the degree of human intervention required and to simplify the management of numerous terminology-to-terminology mappings.

Switching involves the use of a single terminology as an intermediary to translate requests from one scheme to another. For example, all the terminologies to be used within a retrieval system are mapped to a common terminology (X). This enables user queries entered using terminology A to be translated to X and then switched to the equivalent terms in terminology B.

In: *Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007), 21-23 February 2007, Bangalore, India.* Documentation Research & Training Centre, Indian Statistical Institute, Bangalore. pp.109-120.

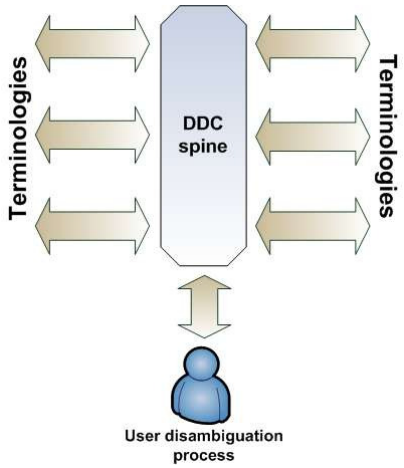

Figure 1: Diagram of the DDC spine-based model employing user 'disambiguation'.

The mapping mechanism employed by the system documented in this paper (HILT: High-level Thesaurus: http://hilt.cdlr.strath.ac.uk/) is similar to switching but differs in that the switching terminology is also central to user *disambiguation* processes (Shiri, Nicholson & McCulloch, 2004). It should be noted that the process of disambiguation not only resolves the existence of homographs (as the term 'disambiguation' may suggest), but encompasses a variety of processes allowing users to qualify their search requirements (Figure 1). This so-called 'spine-based' approach uses the Dewey Decimal Classification (DDC) as a switching spine for searching and permits hierarchical browsing and the discovery of like terms within other terminologies. Although the primary purpose of the terminology mapping server is to enable improved cross-repository searching, it can also provide other terminological functions, such as terminology-based interactive query expansion to assist user query formulation. Such terminology-based techniques have been more formally defined by Efthimiadis (1996) as interactive query expansion based on collection independent knowledge structures.

3 SKOS Core

Simple Knowledge Organization System (SKOS) Core (Miles & Brickley, 2005) is an application of the Resource Description Framework (RDF) and a model proposed by the W3C Semantic Web Best Practices and Deployment Working Group (W3C, 2006). SKOS Core provides a flexible framework for expressing the structure and content of terminologies (or 'concept schemes'), thus enabling efficient machine processing. The framework is flexible enough to accommodate most KOS (as defined in section 1) and essentially consists of a series of RDF properties and RDF Schema (RDFS) classes to encode

In: *Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007), 21-23 February 2007, Bangalore, India.* Documentation Research & Training Centre, Indian Statistical Institute, Bangalore. pp.109-120.

terminologies' content and structural characteristics. For example, a thesaurus could be considered as a series of `skos:Concepts` containing preferred labels (`skos:prefLabel`) and non-preferred labels (`skos:altLabel`). It may also contain various broader terms (`skos:broader`) or related terms (`skos:related`), and so forth. Since the data is encoded in RDF it is inherently pliable and can be utilised or integrated with other RDF data via semantic web applications.

To complement SKOS Core, Miles and Brickley (2004) have proposed the SKOS Core Mapping Vocabulary Specification (MVS). The SKOS Core MVS allows the mapping of concepts between different terminologies using the SKOS Core framework. The properties proposed by SKOS MVS are: exactMatch, broadMatch, narrowMatch, majorMatch and minorMatch. Since like to like mappings are often rare, the MVS also supplements the match types with a series of classes (AND, OR, NOT) for combining or excluding concepts. For example, the 'AND' class is used to denote the intersection of two or more concepts. The term of *Education (United Kingdom)* in terminology A may therefore map to *Education* AND *United Kingdom* in terminology B.

4 Using SKOS Core for terminology services

4.1. SKOS Core: deployment context

The motivation behind the terminology mapping server is to ameliorate the limited terminological interoperability currently afforded between the federation of repositories, digital libraries and information services comprising the UK Joint Information Systems Committee (JISC) Information Environment (JISC, 2003). The expectation is that participant services in the federation will employ Search/Retrieve Web service (SRW) (http://www.loc.gov/standards/sru/srw/) clients to interact transparently with the SRW compliant terminology mapping server during normal service operation (Figure 2). Client requests made to the server will be sent to a database of terminology sets and associated mappings to DDC. Hits identified are then sent back to the server for onward communication to the SRW clients. Testing of this underlying architecture is currently being conducted in collaboration with 'GoGeo!' (http://www.gogeo.ac.uk/) hosted at EDINA (http://www.edina.ac.uk/). GoGeo! provides access to a variety of geospatial datasets, many indexed using disparate terminologies, and constitutes a sound test bed.

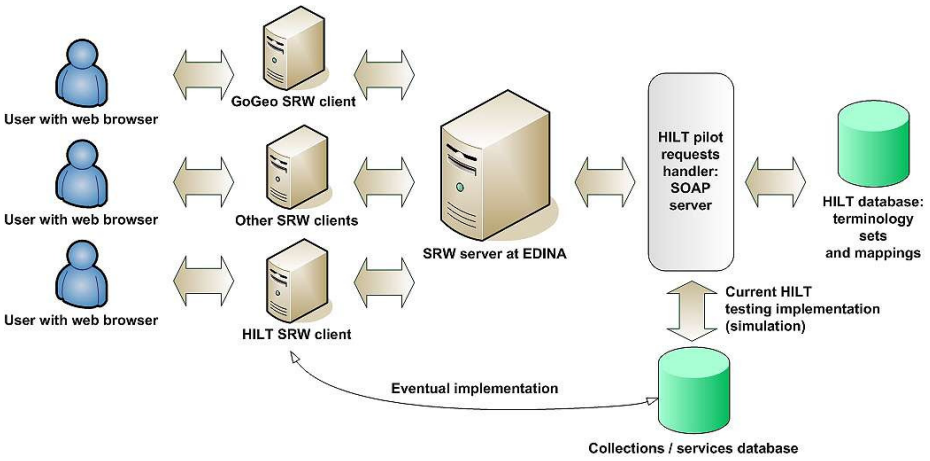

Figure 2: Underlying design of the M2M terminology mapping server

4.2 Wrapping terminologies using SKOS Core

Whilst it is acknowledged that SKOS Core may be deployed in novel and unanticipated ways (Miles, Matthews, Wilson & Brickley, 2005), the main objective of SKOS Core is to facilitate the publication of terminologies for the semantic web, not necessarily for dynamic client/server interactions as described above. However, we propose the use of SKOS Core for the terminology mapping server in order to facilitate meaningful communication with SRW clients regarding the structural nature of the terminological data requested and/or found in the database.

Within the context defined in 4.1, results identified in the database (e.g. scheme information, mapped terms, etc.) are 'wrapped' (i.e. marked-up) in SKOS Core by the SOAP (http://www.w3.org/TR/soap/) server (Figure 2). Since SRW requests (made in Common Query Language (CQL)) are handled using XML over HTTP via the SOAP protocol, terminological data marked-up in SKOS Core can easily be embedded within a SOAP XML envelope for messaging to clients. While such a technical approach potentially permits the future addition of further layers of abstraction, the use of SKOS Core for wrapping is advantageous for several reasons:

- It can accurately model and maintain the structural and semantic properties of the terminological data requested, thus facilitating flexible and reliable re-use by clients in local systems. It is worth noting that such re-use may entail the generation of innovative user interfaces or browsing structures, perhaps displaying results as RDF graphs or providing users with result displays that accurately reflect the hierarchical or semantic structure of the terminological data requested and/or found.

- Although issues exist (to be discussed in section 5), SKOS Core can accommodate the representation of terminology mappings.
- SKOS Core offers opportunities for enhanced interaction with the terminology mapping server, facilitating added functionality such as terminology-based interactive query expansion.

4.3 Server functions in SKOS Core

The M2M pilot terminology mapping server has currently been developed to offer six distinct terminological functions (Nicholson, 2006). These are overly detailed to address in this paper, therefore discussion will concentrate on two (`Get_filtered_set` and `Get_non_DDC_records`).

One purpose of `Get_filtered_set` is to enable the enrichment of users' search vocabulary, provide user feedback and allow limited interactive query expansion. The filtered search can consequently provide (where they exist) related terms (RT), broader terms (BT), narrower terms (NT), preferred terms (PT), and non-preferred terms (NPT). Scope notes may also be provided, depending on the characteristics of the terminology. Although it is possible to envisage such a function being deployed in a variety of user searching scenarios, it is expected that the `Get_filtered_set` function will be of most use to information services that wish to enhance the searching of their local service for users (i.e. enriching users' searching vocabulary to aid query formulation).

For example, a filtered query set to the UK Integrated Public Sector Vocabulary (IPSV) using the term 'Arboriculture' would return a SKOS Core record (Figure 3) providing details necessary to process and invoke a variety of local searching functionality and allowing users to re-interrogate (e.g. using BT, NT, RT, etc. to familiarise users with the topic area within the chosen terminology to aid the subsequent reformulation of search queries, and so forth).

Recall that the primary purpose of the server is to provide terminology mappings using DDC as a spine. There are several functions relating to terminology mapping; one is the `Get_non_DDC_records`. Subsequent to identification of a DDC number by the user via the disambiguation process (as discussed in section 2), `Get_non_DDC_records` provides the client with details of any non-DDC record that includes a mapping to the DDC number sent. Figure 4 provides the SKOS Core output from the terminology mapping server in response to a `Get_non_DDC_records` request for the DDC number 363.34 ('Disasters').

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<SOAP-ENV:Envelope SOAP-
ENV:encodingStyle="http://schemas.xmlsoap.org/soap/encoding/" xmlns:SOAP-
ENV="http://schemas.xmlsoap.org/soap/envelope/"
xmlns:xsd="http://www.w3.org/2001/XMLSchema"
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:SOAP-ENC="http://schemas.xmlsoap.org/soap/encoding/">
<SOAP-ENV:Body>
<ns1:get_filtered_setResponse xmlns:ns1="http://tempuri.org">
<return xsi:type="SOAP-ENC:Array" SOAP-ENC:arrayType="xsd:string[1]">
<item xsi:type="xsd:string">
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:Concept rdf:about="#2715">
          <skos:prefLabel xml:lang="en">Arboriculture</skos:prefLabel>
          <skos:broader rdf:resource="#504"/>
          <skos:narrower rdf:resource="#2633"/>
          <skos:related rdf:resource="#1566"/>
          <skos:related rdf:resource="#15"/>
          <skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/schemes/IPSV.xml"/>
</skos:Concept>
<skos:concept rdf:about="#504">
          <skos:prefLabel xml:lang="en">Horticulture</skos:prefLabel>
          <skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/schemes/IPSV.xml"/>
</skos:concept>
<skos:concept rdf:about="#2633">
          <skos:prefLabel xml:lang="en">Tree planting</skos:prefLabel>
          <skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/schemes/IPSV.xml"/>
</skos:concept>
<skos:concept rdf:about="#1566">
          <skos:prefLabel xml:lang="en">Woodlands</skos:prefLabel>
          <skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/schemes/IPSV.xml"/>
</skos:concept>
<skos:concept rdf:about="#15">
          <skos:prefLabel xml:lang="en">Trees</skos:prefLabel>
          <skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/schemes/IPSV.xml"/>
</skos:concept>
</rdf:RDF>
</item>
</return>
</ns1:get_filtered_setResponse>
</SOAP-ENV:Body>
</SOAP-ENV:Envelope>
```

Figure 3: Example of M2M pilot terminology server response to Get_filtered_set in SKOS Core within SOAP envelope. Example illustrates filtered search for the IPSV term 'Arboriculture'. Note: example has been truncated for publication purposes.

The need for encoding the presence of mappings from multiple terminologies in response to client requests invokes the use of the SKOS Core MVS. In Figure 4 most of the mappings are deemed to be exact matches (i.e. `map:exactMatch`). This process enables the identification of a variety of terms from disparate terminologies associated with a particular DDC number

In: *Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007), 21-23 February 2007, Bangalore, India.* Documentation Research & Training Centre, Indian Statistical Institute, Bangalore. pp.109-120.

to be used to search relevant repositories or information services using the correct terminology to match local indexes. Such a function goes some way to ameliorating the current subject interoperability difficulties encountered within the JISC Information Environment. Note that there are several non-DDC terminologies represented in the example: the Global Change Master Directory (GCMD), Library of Congress Subject Headings (LCSH), IPSV, and the UNESCO Thesaurus. The SKOS Core MVS may prove useful for encoding terminology mappings within our model; however, the use of MVS within terminology services highlights particular issues which are further of further study. These will be discussed below in more detail.

5 Conclusion and further research

The M2M pilot implementation proposed in this paper offers terminology mapping as a principal function to enhance user access to disparately indexed heterogeneous digital objects held within multiple repositories, but it also offers terminology-based interactive query expansion functionality. We have demonstrated that SKOS Core can function effectively in a web services environment and that such an approach constitutes a flexible means of implementing various terminological functions for third party terminology services. Our experiments are currently being conducted within in a controlled environment (i.e. the JISC Information Environment); however, we consider the use of SKOS Core for a terminology mapping server (or terminology services generally) to be sufficiently flexible (and scalable, providing the necessary terminology sets exist) so as to permit similar approaches to be used in alternative contexts or global information environments. The wrapping of terminological data within SKOS Core provides a readable means of transporting data over networks (i.e. over HTTP using SOAP) and allows such data to be structured appropriately and modelled correctly, thus facilitating flexible re-use by clients. The authors therefore intend to continue this line of research, including a system and user evaluation of the M2M terminology mapping server as embedded within a client service. Results gleaned from this evaluative work are expected to be disseminated in a separate research paper.

A continuation of this research will demand that several areas undergo further study or investigation. In particular, our work has drawn to attention potential issues within the current draft of the SKOS Core MVS. The definitions of the MVS match types are based on the principles of set theory. Such an abstract paradigm can be useful as an arbitrary means of assessing equivalence in a variety of terminological scenarios; however, their abstractness can also cause uncertainty in practical application. For example, the use of majorMatch (`map:majorMatch`) and minorMatch (`map:minorMatch`) within our

```
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core#"
xmlns:map="http://www.w3.org/2004/02/skos/mapping#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:Concept rdf:about="#363.34">
        <skos:prefLabel xml:lang="zxx">363.34</skos:prefLabel>
        <skos:altLabel xml:lang="en">Disasters</skos:altLabel>
        <skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/schemes/DDC.xml"/>
        <map:narrowMatch>
                    <skos:Concept rdf:about="#sh 91000441"/>
        </map:narrowMatch>
        <map:exactMatch>
                <skos:Concept rdf:about="#2256"/>
        </map:exactMatch>
        <map:exactMatch>
                <skos:Concept rdf:about="#762"/>
        </map:exactMatch>
        <map:exactMatch>
                    <skos:Concept rdf:about="#143"/>
        </map:exactMatch>
</skos:Concept>
<skos:Concept rdf:about="#sh 91000441 ">
<skos:prefLabel xml:lang="en">Emergency management</skos:prefLabel>
<skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/schemes/LCSH.xml"/>
</skos:Concept>
<skos:Concept rdf:about="#2256">
<skos:prefLabel xml:lang="en">Natural disasters</skos:prefLabel>
<skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/schemes/UNESCO.xml"/>
</skos:Concept>
<skos:Concept rdf:about="#762">
<skos:prefLabel xml:lang="en">Natural hazards</skos:prefLabel>
<skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/schemes/GCMD.xml"/>
</skos:Concept>
<skos:Concept rdf:about="#143">
<skos:prefLabel xml:lang="en">Civil emergencies</skos:prefLabel>
<skos:inScheme
rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/schemes/IPSV.xml"/>
</skos:Concept>
</rdf:RDF>
```

Figure 4: Example of M2M pilot terminology server response to Get_non_DDC_records in SKOS Core with the Mapping Vocabulary Specification. Note: example has been truncated for publication purposes and therefore does not show full response or XML SOAP envelope.

model can be difficult to apply as defined and it is unclear how a third party terminology service would incorporate such properties since the current definitions might be interpreted as implying knowledge of database content and indexes. It is noteworthy that similar application difficulties have been encountered by other researchers within different contexts (Liang & Sini, 2006). Current analyses indicate that our approach will require additional match types to those specified by the MVS. Although a pilot service could be implemented using the MVS alone, the lack of detail afforded in the

Specification would, we hypothesise, impose unnecessary cognitive load on the user since only minimal match type feedback would be provided. Any use of match types to assist in the ranking of results (according to the degree of concordance with users' preferred terminology) would also be limited. Future research will therefore aim to test this hypothesis by comparing the relative benefits of each approach for the purposes of user disambiguation.

Future work will also aim to optimise the way in which terminological data is modelled in SKOS Core. Since our system is accommodating numerous terminologies from different KOS, a generic approach has been necessary in the treatment of terminologies and it has therefore not been possible to model the nuances of every particular scheme. For example, LCSH structured headings use a delimiter (--) to denote the use of structured headings (e.g. 'Beach erosion--Monitoring'). Within our current system such a heading is not considered to represent two concepts and is therefore mapped as if it were one concept. Future work will aim to investigate the use of the MVS classes (e.g. AND, NOT, and OR) to optimise the way in which some terminological data is represented and to better accommodate compound concept searching.

6 Acknowledgements

This research is funded by the UK Joint Information Systems Committee (JISC) Shared Services Programme.

7 References

[1] Binding, C. & Tudhope, D. (2004). KOS at your Service: Programmatic Access to Knowledge Organisation Systems, *Journal of Digital Information 4*(4). Retrieved September 19, 2006 from http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Binding/

[2] Chan, L., M. & Zeng, M., L. (2002). Ensuring interoperability among subject vocabularies and Knowledge Organization Schemes: a methodological analysis, *IFLA Journal 28*(5/6) 323-327.

[3] Doerr, M. (2001). Semantic problems of thesaurus mapping, *Journal of Digital Information 1*(8). Retrieved September 19, 2006 from http://jodi.tamu.edu/Articles/v01/i08/Doerr/

[4] Efthimiadis, E., N. (1996). Query expansion, *Annual Review of Information Systems and Technology (ARIST)*, 31, 121-187.

[5] JISC. (2003). *Strategic activities: Information Environment.* Retrieved September 19, 2006 from http://www.jisc.ac.uk/index.cfm?name=about_info_env

[6] Liang, A., C. & Sini, M. (2006). Mapping AGROVOC and the Chinese Agricultural Thesaurus: definitions, tools, procedures, *New Review of Hypermedia and Multimedia 12*(1), 51-62.

[7] McCulloch, E., Shiri, A. & Nicholson, D. (2005). Challenges and issues in terminology mapping: a digital library perspective, *Electronic Library 23*(6), 671-677.

[8] Miles, A. & Brickley, D. (Eds.). (2004). *SKOS Mapping Vocabulary Specification*. Retrieved September 19, 2006 from http://www.w3.org/2004/02/skos/mapping/spec/

[9] Miles, A. & Brickley, D. (Eds.). (2005). *SKOS Core Guide: W3C Working Draft 2 November*. Retrieved September 19, 2006 from http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/

[10] Miles, A., Matthews, B., Wilson, M. & Brickley, D. (2005). SKOS Core: simple knowledge organisation for the Web, *International Conference on Dublin Core and Metadata Applications (DC-2005): Vocabularies in Practice*, September 12-15, 2005, Madrid, Spain. Retrieved September 19, 2006 from http://epubs.cclrc.ac.uk/bitstream/675/dc2005skospapersubmission1.pdf

[11] Nicholson, D. (2006). *HILT M2M pilot requirements document*, v6.0. Retrieved September 19, 2006 from http://hilt.cdlr.strath.ac.uk/hilt3web/reports/h3requirementsv6.pdf

[12] Nicholson, D., Dawson, A. & Shiri, A. (2006). HILT: A pilot terminology mapping service with a DDC spine, *Cataloging & Classification Quarterly 42*(3/4), 187-200.

[13] Shiri, A., Nicholson, D. & McCulloch, E. (2004). User evaluation of a pilot terminologies server for a distributed multi-scheme environment, O*nline Information Review 28*(4) 273-283.

[14] Vizine-Goetz, D., Hickey, C., Houghton, A. & Thompson, R. (2004). Vocabulary mapping for terminology services, *Journal of Digital Information 4*(4). Retrieved September 19, 2006 from http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/

[15] W3C. (2006). *Semantic Web Best Practices and Deployment Working Group*. Retrieved September 19, 2006 from http://www.w3.org/2001/sw/BestPractices/

[16] Zeng, M. L. Chan, L: M. (2004). Trends and issues in establishing interoperability among Knowledge Organization Systems, *Journal of*

In: *Proceedings of the International Conference on Semantic Web and Digital Libraries (ICSD-2007), 21-23 February 2007, Bangalore, India.* Documentation Research & Training Centre, Indian Statistical Institute, Bangalore. pp.109-120.

the American Society for Information Sciences and Technology 55(5), 377-395.