HILT Phase III: Design Requirements of an SRW-compliant Terminologies Mapping Pilot.

Dennis Nicholson and Emma McCulloch

Aims

As Zeng and Chan (2004) note, Interoperability of knowledge organisation systems (KOS) is a key issue in today's networked environment. It is an issue likely to impact, in time, on the semantic web vision (Berners-Lee et al, 2001), but is more usually tackled at present in an information retrieval context[1]. Information services employ a plethora of different subject schemes to describe their resources. In some cases, they use recognised standards, in others 'in-house' or even uncontrolled schemes. Either way, the practice acts as a barrier to effective cross-searching by subject over distributed information services. The issue has attracted a good deal of interest in recent years. Potential solutions proposed include linking or switching between schemes, mapping, derivation/modelling (see for example Doerr, 2001; Chan and Zeng, 2002), and automatic or semi-automatic classification (see for example Koch and Vizine-Goetz, 1998; Godby et al, 1999; Ardo, 2004). CARMEN (2000), LIMBER (2000), Renardus (2002), and MACS (2005) are amongst a range of recent projects that have tackled the problem, and key international players such as OCLC (http://www.oclc.org/) have also done relevant work (see http://www.oclc.org/productworks/terminologiespilot.htm).

The HILT (HIgh-Level Thesaurus) project (http://hilt.cdlr.strath.ac.uk/), currently active in the area, is researching the problems of facilitating interoperability of subject descriptions in a distributed multi-scheme environment, aiming, ideally, to identify a generic solution. HILT Phase I (http://hilt.cdlr.strath.ac.uk/index1.html) found a UK community consensus in favour of improving interoperability via an inter-scheme mapping service. This idea was followed up in HILT Phase II (http://hilt.cdlr.strath.ac.uk/index2.html), which built a user accessible pilot terminologies mapping service based on a Dewey Decimal Classification (DDC) spine to investigate the approach. The subsequent Machine to Machine (M2M) Feasibility Study (http://hilt.cdlr.strath.ac.uk/hiltm2mfs/) then investigated, proposed, and costed a project to build an M2M version of the pilot and this led to the funding of HILT Phase III (http://hilt.cdlr.strath.ac.uk/index3.html). Further information on the earlier stages of HILT can be found in the final reports of the various phases, which are available online (HILT, 2002; 2003; 2005). A description of the Phase II pilot can be found in Nicholson et al, 2006.

HILT Phase III began in November 2005 and will run until January 2007. It consists of two overlapping stages. The aim in stage 1 – reported here – is to take a version of[2] the Phase II pilot service (http://hiltpilot.cdlr.strath.ac.uk/pilot/top.php)[3], extend its functionality in various ways, and create a centralised, single-server, M2M version of it built around SRW (http://www.loc.gov/z3950/agency/zing/srw/) and the SWAD-Europe (http://www.w3.org/2001/sw/Europe/) project's SKOS-Core (Miles et al, 2005). The aim in stage 2 – still at an early stage – is to look into the feasibility and design implications of adopting a distributed, multi-server, approach to service provision. Such an approach was identified in the M2M Feasibility Study as theoretically attractive - in that it might implement the kind of mapping-based solution HILT had envisaged to subject interoperability issues in a way that would spread the cost and effort over many organisations and a longer period of time. The research associated with it is scheduled for completion in time to impact on the final design of the pilot, but is insufficiently advanced for it to be covered further in this paper.

[1] Recent examples of work in the area are reported in Heery et al, 2001; Koch et al, 2001, Saeed and Chaudhury, 2002; and Vizine-Goetz et al 2005, but see Zeng and Chan, 2004 for a more comprehensive list.

[2] The initial pilot was based around Wordmap software (http://www.wordmap.com/) and this software may be used again in future. However, the plan for HILT Phase III is to use a PHP and SQL Server based version of the pilot (see http://hilt.cdlr.strath.ac.uk/hilt3/top.cfm).

[3] See also the worked examples in Appendix I of the HILT Phase II Final report (HILT, 2003) and at http://hiltpilot.cdlr.strath.ac.uk/pilot/examples/.

Methods (Design and Implementation of the Stage 1 Pilot)

The first stage in the process was the drawing up of a design requirements document, based on the following elements of work:

a) A re-examination of the analysis and development carried out in HILT Phase II on the design, implementation, and evaluation of the Phase II pilot
b) A re-examination of the analysis and development carried out in the HILT M2M Feasibility Study on the best approach to designing an M2M version of the pilot (e.g. using SRW and SKOS-Core) and on a number of use cases that provided a sketch of the scope of the functionality required in the M2M pilot.
c) Phase III analysis aimed at providing a base-line template for translating the facilities and functions of the Phase II pilot as determined under (a) to the M2M environment as scoped out under (b).
d) Phase III analysis aimed at providing the functionality required to meet the requirements of the five use cases drawn out by the M2M Feasibility Study.

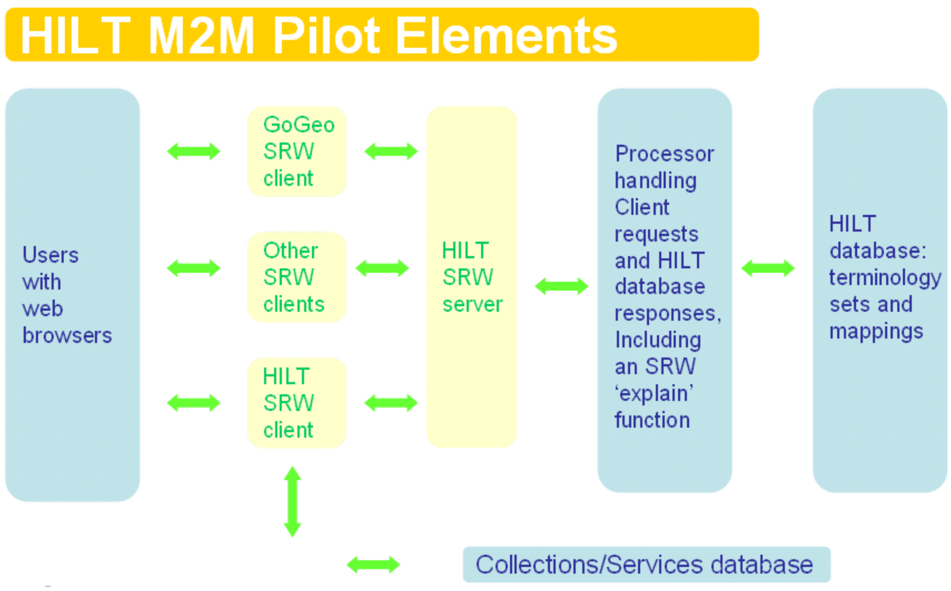
The requirements document specifies functionality required rather than the detail of how the functionality will be provided. It is a working document, subject to change as specific solutions are implemented and evaluated.

The second stage (begun in May 2006 and likely to be at least partially developed by September 2006) was the development and implementation of the Stage 1 pilot based on the requirements document.

Main Findings

The findings are presented in full in the detailed requirements document available at http://hilt.cdlr.strath.ac.uk/hilt3web/reports/h3requirements.pdf. The following is a brief summary of the key points:

The diagram below is an outline sketch of the proposed main elements of the M2M pilot and their functions and inter-relationships:


HILT M2M Pilot Elements
Users with web browsers
GoGeo SRW client
Other SRW clients
HILT SRW client
HILT SRW server
Processor handling Client requests and HILT database responses, Including an SRW 'explain' function
HILT database: terminology sets and mappings
Collections/Services database

Thus, in the working pilot:

○ Users will access the terminologies service via JISC services as shown.

- The JISC services will incorporate SRW clients which will have the functionality required to request, receive, and utilise the terminologies information from the HILT database required to service the various use cases identified by participating JISC services in the HILT M2M Feasibility Study:

Use case #1	Disambiguate the topic of a user's subject request; identify collections covering that topic and the subject schemes used; identify relevant terms from subject schemes used by collections identified; where possible, automatically retrieve relevant material from specific collections using the terms
Use case #2	User gets functions to: retrieve an enriched set of search terms based on a term originally input to a local service search; Map plural to singular terms; Map synonyms to main terms in thesauri; disambiguate terms such as COLD; Correct simple spelling/typographic errors. Having used these various functions, user selects one or more terms derived from the mapping process and these are used to search the requesting service database. Results are displayed in browser without substantial differences to the non-enhanced search.
Use case #3	User types a term into search box. The term is sent to HILT to generate a set of additional search terms that can be used to search the requesting service database. If any simple spelling or typographical errors are identified an intermediate screen offering an alternative spelling is presented along the lines of Google, "Did you mean?" After acquiring a correct spelling the term is sent back to HILT for further expansion. The original and derived terms are passed to the requesting service database, a search is run against it and a result set is returned. The user notices no substantial differences in the result set (apart from hopefully a larger number of results) between the non-enhanced query and a query enhanced first by via M2M interaction with HILT.
Use case #4	Browse of DDC offered in response to a 'no hits from HILT' situation after a service end request. Browse of another specified scheme in the database.
Use case #5	Provision of information on narrower and related terms from a given scheme.

- The SRW clients will do this via the SRW server which will translate SRW client requests into one of the HILT function calls listed below. They will also relay HILT responses back to the clients.

Function name	**Description**	**Use Case**
Get_explain	Get HILT service Explain file	All
Get_linked_records	Get records that include – or are directly or indirectly mapped to records that include – specified term or term phrase.	#1, probably not used in other contexts.
Get_linked_DDC_records	Get any DDC record that either includes the term specified, or that is mapped to by a record that includes the term specified.	#1, but might be used in other contexts.
Get_linked_nonDDC_records	Get any non-DDC record that includes a mapping to the DDC number sent.	#1, but might be used in other contexts.
Get_filtered_set	Get record and fields set that meets the specified parameters (UNESCO and LCSH only, say, or mappings or broader and narrower terms only, say).	Various sets of parameters needed for use cases #2 to #5.

- A (SOAP-based) HILT requests and responses handler will query the HILT database and pass back terminology-set responses wrapped in SKOS-Core.

- A HILT database will encompass a range of subject schemes, together with illustrative mappings from parts of these schemes to the Dewey Decimal Classification Scheme.
- A collections database (emulating the JISC IESR service) will provide the pilot with information on collections covering particular services and the subject schemes they use.

Once developed, the pilot will be able to translate SRW requests for terminologies information and provide pilot data on a range of terminology sets (DDC, LCSH, MESH, UNESCO, JACS, IPSV, and AAT), on mappings of (selected) individual terms to DDC, and on collections covering particular subject areas and the subject schemes they use.

References

Ardo, A., Automatic Subject Classification and Topic Specific Search Engines - Research at KnowLib, Presented at *DELOS Regional Awareness Event: Between Knowledge Organization and Semantic Web: Semantic Approaches in Digital Libraries*, Lund, Sweden, 2004. Available online at: http://www.delos.info/eventlist/LUB1/Anders_Ardo/DELOS.PDF (accessed 20 January 2006).

Berners-Lee, T., Hendler, J. and Lassila, O., The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities, *Scientific American*, 2001, Available online at: http://www.sciam.com/article.cfm?articleID=00048144-10D2-1C70-84A9809EC588EF21 (accessed 20 January 2006).

Chan, L. and Zeng, M., Ensuring interoperability among subject vocabularies and knowledge organisation schemes: a methodological analysis. Presented at *68th IFLA Council and General Conference*, Glasgow, Scotland, 2002. Available online at: http://www.ifla.org/IV/ifla68/papers/008-122e.pdf (accessed 19th January, 2006).

CARMEN. Available online at: http://www.mathematik.uni-osnabrueck.de/projects/carmen/index.en.shtml (accessed 18 January 2006).

Doerr, M., Semantic problems of thesaurus mapping. *Journal of Digital Information*, 2001, **1**(8). Available online at: http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/ (accessed 20 January 2006).

Godby, C. J., Miller, E. J. and Reighart, R. R., Automatically Generated Topic Maps of World Wide Web Resources, *Annual Review of OCLC Research*, 1999. Available online at http://digitalarchive.oclc.org/da/ViewObject.jsp?fileid=0000002655:000000059193&reqid=9300 (accessed 20 January 2006).

Heery, R., Carpenter, L. and Day, M., Renardus Project Developments and the Wider Digital Library Context. *D-Lib Magazine*, 2001, **7**(4) Available online at: http://www.dlib.org/dlib/april01/heery/04heery.html (accessed 20 January 2006).

HILT, HILT Phase I Final Report, 2002. Available online at: http://hilt.cdlr.strath.ac.uk/reports/finalreport.html (accessed 20 January 2006).

HILT, HILT Phase II Final Report, 2003. Available online at: http://hilt.cdlr.strath.ac.uk/hilt2web/finalreport.htm (accessed 20 January 2006).

HILT, HILT M2M Feasibility Study II Final Report, 2005. Available online at: http://hilt.cdlr.strath.ac.uk/hiltm2mfs/0HILTM2MFinalReportRepV3.1.doc (accessed 20 January 2006).

Koch, T., Neuroth, H. and Day, M., Renardus: Cross-browsing European subject gateways via a common classification system (DDC). Presented at *IFLA Satellite Conference: Subject Retrieval in a Networked Environment*, Dublin, Ohio, 2001. Available online at: http://www.lub.lu.se/tk/renardus/preIFLA-demo.html (accessed 20 January 2006).

Koch, T. and Vizine-Goetz, D., Automatic Classification and Content Navigation Support for Web Services: DESIRE II Cooperates with OCLC *Annual Review of OCLC Research*, 1998. Available online at: http://www.lub.lu.se/tk/demos/class-ws/automatic.htm (accessed 20 January 2006).

LIMBER: Language Independent Metadata Browsing of European Resources. Available online at: http://www.limber.rl.ac.uk/ (accessed 18 January 2006).

MACS, Multilingual Access to Subjects. Available online at: https://ilmacs.uvt.nl/ (accessed 18 January 2006).

Miles, A., J., Rogers, N. and Beckett, D., SKOS-Core 1.0 Guide, An RDF schema for thesauri and related knowledge organisation systems, 2005.
Available online at: http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide (accessed 20 January 2006).

Nicholson, D., Dawson, A. and Shiri, A., HILT: A Pilot Terminology Mapping Service with a DDC Spine. *Cataloguing and Indexing Quarterly*, 2006 In press.

Renardus. Available online at: http://www.renardus.org/ (accessed 18 January 2006).

Saeed, H. and Chaudhury, A. S., Using Dewey Decimal Classification scheme (DDC) for building taxonomies for knowledge organisation. *Journal of Documentation*, 2002, **58**(5) 575-583.

Vizine-Goetz, D., Hickey, C., Houghton, A. And Thompson, R., Vocabulary Mapping for Terminology Services. *Journal of Digital Information*, 2004, **4**(4) Available online at: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/ (accessed 20 January 2006).

Zeng, M.L. and Chan, L.M., Trends and Issues in Establishing Interoperability Among Knowledge Organisation Systems. *Journal of the American Society for Information Science and Technology*, 2004, **55**(5).