HILT: A Pilot Terminology Mapping Service with a DDC Spine

D. Nicholson, A, Dawson, A. Shiri

D. Nicholson < d.m.nicholson@strath.ac.uk > is the Director of the Centre for Digital Library Research at Strathclyde University, A, Dawson < alan.dawson@strath.ac.uk > is a Senior Researcher and Programmer at the Centre for Digital Library Research at Strathclyde University, Dr A. Shiri < ashiri@ualberta.ca>is Assistant Professor at the School of Library and Information Studies, University of Alberta

Abstract

The role of DDC in the ongoing HILT (High-level Thesaurus) project is discussed. A phased initiative, funded by JISC in the UK, HILT addresses an issue of likely interest to anyone serving users wishing to cross-search or cross-browse groups of networked information services, whether at regional, national or international level – the problem of subject-based retrieval from multiple sources using different subject schemes for resource description. Although all three phases of HILT to date are covered, the primary concern is with the subject interoperability solution piloted in phase II, and with the use of DDC as a spine in that approach.

Keywords

HILT Project, Interoperability, Subject searching, DDC, Mapping terminologies

Introduction: HILT and DDC

The HILT project (1) began in September 2000 with HILT Phase I and ran for approximately 15 months (2). HILT Phase I was an investigation into the problems of cross-searching and browsing by subject in a distributed, multi-scheme environment and was charged with determining whether a consensus on the best solution to these problems could be reached in archives, libraries, museums, and electronic services in the UK. It was followed by HILT Phase II, which ran for a similar period between 2002 and late 2003 (3) and built an illustrative pilot terminologies server based on the consensus solution arrived at in the Phase I work (4). At the time of writing (January 2005), a short feasibility study is underway to determine whether the user facilities provided in the Phase II pilot can also be provided via a machine-to-machine (M2M) interface and to determine the likely cost of building such an interface. This is preliminary to possible Phase III work and is expected to lead to an attempt to build these M2M facilities into the pilot service in a full-scale HILT Phase III. The Dewey Decimal Classification (DDC) has featured in all three parts of the project to date – as a possible solution in its own right and then as a proposed spine for a mapping-based solution in Phase I, and as a spine for pilot terminology services in Phase II and in the Phase III feasibility study.

HILT Phase I: Overview

HILT Phase I was a collaborative investigation into the problems associated with cross-searching and browsing by subject in a cross-sectoral and cross-domain environment encompassing libraries, archives, museums, and electronic resource collections in the UK. Its principal aims were:

- To thoroughly research the problem, analyse and document its exact nature in detail, focusing on UK requirements across the various communities, services and initiatives, but setting the study firmly in the context of international requirements and standards.
- To analyse the data obtained, and discuss the results with the various communities, with an aim to reaching a consensus within the project on how best to apply the findings in relation to existing or new subject schemes and thesauri.
- To attempt to reach a similar consensus within the group of stakeholders generally, both at a stakeholder workshop and through other methods.

Reporting in early 2002 (2), the project determined that:

- Many different subject schemes and practices are in use in UK services.
- Subject searching across services is believed to be of value to users.
- There was a strong consensus across the archives, electronic services, library, and museums communities in favour of a more practically-focused follow-up project that would develop a pilot service providing mappings between subject schemes, probably using a DDC spine.
- Further research was required into the effectiveness, level and nature of user need, practicality, design requirements, costs and benefits of such an approach before a long term commitment to a (possibly expensive) service could be justified.

Further details of the project have been reported in detail elsewhere (5; 6; 4).

HILT Phase I and DDC

The key event in HILT Phase I was the stakeholder workshop. Here, well-informed stakeholders were presented with various options:

- Do nothing – on the basis that it was an unimportant problem, or that users could cope, or that solutions would be found by the artificial intelligence community or by commercial initiatives, or that the problem could not be solved.
- Set up a human process – such as a 'terminologies agency' – intended to lead to a solution in time.
- Adopt a base-level, gradual approach, with an eye on future developments – for example, apply a single scheme to collection-level descriptions of services, focus only on electronic services, or gradually create inter-community terminology 'cross-walks'.

- Adopt a single scheme under various circumstances – an existing scheme or an entirely new scheme, used in addition to a service's existing scheme or instead of it, with or without retro-conversion of legacy metadata.
- Set up a service that would provide mappings between subject schemes, or set up a pilot service of this kind so that further investigations could be conducted.

DDC was proposed as a strong candidate should a single scheme be the preferred outcome, as it offered the following advantages over other alternatives:

- It is owned by a major worldwide not-for-profit organisation with a clear commitment to continuing to maintain and develop it and a record of consulting key players.
- It is in use in over 200,000 libraries worldwide and in 135 countries.
- It is available in over 30 different languages, including languages with major world coverage such as English, Spanish, Arabic and Chinese.

Despite being presented as the HILT team's second preference, the option of a single subject scheme (based on DDC or anything else) was almost unanimously rejected by the workshop breakout groups.

However, DDC was identified as having a potential role in the option ultimately chosen by workshop stakeholders as the preferred route to consensus – the pilot mapping service option. One possible design for such a service entailed the use of a central spine and DDC was identified as a possible candidate for this role, primarily in view of the advantages presented above.

HILT Phase II: Overview

HILT Phase II was funded to set up a pilot terminologies server based on the mapping approach identified in HILT Phase I, an approach to interoperability used in a range of other projects (7; 8; 9; 10). The primary aim in this second stage of HILT was to provide a practical experimental focus within which to investigate and establish subject terminology service requirements for the JISC Information Environment (11) and make recommendations as regards a possible future service. There was also a requirement to consider issues such as user needs, collection level requirements, international compatibility, and costs against benefits. The question of whether or not the pilot service should have a central spine (as opposed to, say, mapping directly between user terms and individual schemes or between the schemes themselves) was left open at the outset, although DDC was identified as a strong candidate as a spine should that approach be adopted.

In the event, a spine-based approach was selected, with DDC as the preferred spine, in common with other recent initiatives (7; 12; 13; 14). The most important reasons for choosing a DDC spine were:

- A spine-based approach was likely to involve less labour-intensive (and, hence, less expensive) manual mapping than mapping between user terms and individual schemes or between the schemes themselves.
- DDC is already extensively mapped to LCSH and has been a favoured option by a wide range of other projects.
- Since multi-lingual mapping was a likely future requirement, a scheme available in more than 30 languages was seen as a leading candidate for the spine.
- The use of DDC was the only evident way of providing the proposed collections finding facility described below.

With this general approach agreed, the project then moved to deal with four primary concerns:

- Designing and building the pilot service.
- Conducting a user evaluation of the resulting service.
- Carrying out a cost-benefit analysis of alternative approaches.
- Making recommendations as to the creation of a possible future operational service.

Much of the remainder of this paper is concerned with a description of the pilot terminology server, with particular reference to the role played by DDC. Further information is available elsewhere on the user evaluation (15), on early work with users (16), and on the project outcomes as a whole (3; 17).

The HILT Pilot: A User's View

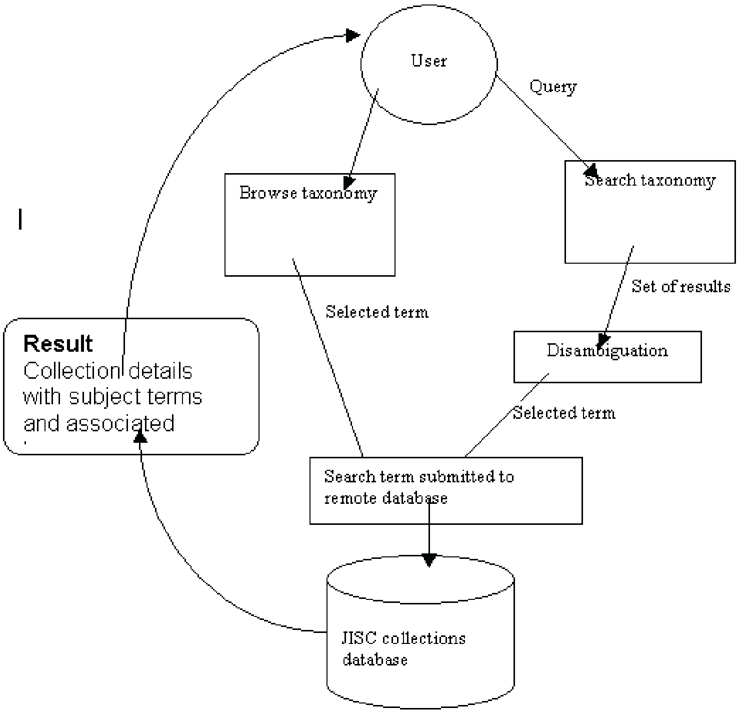
The HILT pilot is available at http://hiltpilot.cdlr.strath.ac.uk/pilot/top.php. Illustrative examples can be found at http://hiltpilot.cdlr.strath.ac.uk/pilot/examples/ and in (15). The following is a brief description of the steps involved:

1. The user enters a subject term.
2. The term is matched to the terminology set held by the pilot server (a set that includes terms in the DDC captions, index, and standard subdivisions, and terms from other schemes mapped to DDC).

3. Based on mappings from the terminology set to DDC numbers, the server returns a number of possible subjects from DDC and prompts the user to choose the right one – to disambiguate the options.
4. The DDC number matching the term chosen by the user is submitted to a separate database of collections classified by DDC, to identify collections appropriate to the user's query. If there are no collections matching the DDC number precisely, then the DDC number is truncated until one or more collections are identified that cover a broader subject area than the search term.
5. The collections database sends back recommended collections to search, plus information on the scheme used, the term or terms from the scheme appropriate to the user's search, and, where technically possible, sample retrieval from specific collections

As is clear from the diagram below, an alternative browse-based route is also available, which takes the user down the DDC hierarchies to a specific subject of interest.

Figure 1: User interaction with the HILT pilot interface



Methodologies

Software and Data Conversion

The pilot server is based on an adaptation of a software package called Wordmap, offered by a company of the same name. The package – described in detail at http://www.wordmap.com/ - has three distinct elements:

- A taxonomy database (Oracle) holding terminology mappings.
- A simple user interface that interacts with the database according to staff specifications and user input and feedback.
- A powerful multi-user interface to support sophisticated staff interaction with the database for creation and maintenance of taxonomies and mappings.

The primary focus of the HILT project was on programming the simple user interface to interact with a database comprising data provided by OCLC supplemented by manual mappings provided by HILT. This offered:

- Access to the whole of the DDC 21 schedules, indexed on the DDC captions, standard subdivisions and relative index.
- Mappings of DDC to LCSH as provided by OCLC.
- Illustrative mappings from DDC to UNESCO and MeSH subject schemes, created by HILT.

Further information on the adaptation of the Wordmap database structure for HILT purposes and on processing and importing the files provided by OCLC is available in Appendix I.2 of the HILT Final Report (3).

Mapping

A literature review was conducted to investigate the problems and issues in integrating and mapping thesauri and classification schemes and the different types of mapping. A list of 19 match types was identified in the review (18), some mapping exercises were carried out, and example mappings were then selected to build into the server. Examples of common match types are provided in the table below:

Match type	First scheme DDC	Second scheme MeSH
Type 1: Singular plural	Teeth	Tooth

Match type	First scheme DDC	Second scheme LCSH
Type 2: Exact match	Teeth	Teeth

Match type	First scheme DDC	Second scheme UNESCO
Type 3: Concept match	Persons in late adulthood	Elderly

Search Algorithm

The search process is complex and not intuitively obvious. During testing it became clear that no single algorithm could give the best results for all search terms. The resulting process appears to give useful results for most search terms tried, but is not guaranteed to give the best results for all possible search terms. The system goes through the following steps when it receives a query.

1. Look for an exact match against the query term.
2. If there are five or more matches, the results are displayed. If there are between one and five matches, the system will adopt a pattern matching approach, looking for the term with any characters before or after it. For example, a search for ‘science’ would find ‘natural sciences’ and ‘science and mathematics’. The additional results are appended to the exact matches.
3. If there are no results found for step 2, the system will look for the search term and any characters after it (but not before). For example if the search term is ‘compute’ the system will then retrieve ‘computer’, ‘computerization’ etc.
4. If there are some results, the system offers a ‘more results’ option. This results in stemming of the search term (removing plurals, ‘ing’, ‘ed’ etc.), using the Porter stemming algorithm (19), and pattern-matching search as in step 2.
5. If no results are found following step 3, the system will adopt a pattern matching approach again, as in step 2.
6. If there are still no results, the system will parse the query to identify any individual words, remove stop words such as ‘the’, ‘in’, ‘and’ etc, and then run a search on the individual words using the same steps outlined above. The results will then be merged, de-duplicated, ranked and returned to the user.

Multi-word Queries

Before displaying results from multi-word queries, the system removes duplicates and assigns weights to individual items. If the same item has been retrieved as a result of a search with different words, that item gets a higher weighting (ranking) in the display of search results. The merging and ranking of search results gives the same effect as an AND search followed by an OR search. If the user types in boolean terms such as AND or OR these will be stripped out as stop words and ignored. Each remaining word is then handled individually.

If a search term is entered as a phrase in quotation marks, it will be treated as a single word and no parsing takes place.

Identifying Collections

A key element in the operation of the pilot server is the identification of network-accessible collections or services appropriate to the user's subject query. The collections database is stored on a separate server and uses different database software to the HILT server. Once the user has selected a term (either by browsing or by search and disambiguation), the system identifies any collections relevant to that term by searching the collections database, then displays the collection name and description along with any subject terms relevant to that collection. Searching the collections database involves the following steps:

1. The system retrieves the DDC number of the user search term from the Wordmap database, along with the corresponding features (subject terms, taxonomy and match type).
2. Collections relevant to the DDC number are retrieved from the separate collections database. If the DDC number is a range (e.g. 371.12-18) rather than a single number, the system retrieves all the collections matching that range.
3. The system also retrieves some broader collections. For example, if the DDC number of the term is 371.2134, the system retrieves collections with DDC number 371.2134, 371.213, 371.21, 371.2, 371, and 370. If there are no results for all these searches the system adopts pattern searching to retrieve related collections, e.g. all collections with DDC number starting 371.

4. If the selected term is from DDC table 2 or table 6 (standard subdivisions) rather than the main DDC schedule, the table number is converted to a DDC number (table 2 maps to the DDC 900s and table 6 maps to the DDC 400s) and then treated as a DDC number when searching the collections database. For example, a search for 'Cairo' will retrieve T2-621.6 from table 2, which will retrieve any collections with DDC number 962.

   If the collection allows remote searching by appending a variable search term to a fixed partial URL (as in the OpenURL standard), and if it uses one of the recognised taxonomies, then the system offers the user the option of dynamically searching the remote collection using the appropriate term provided by the terminology server. In order for this function to operate, the collections database has to include the partial URL to which search terms can be appended and remotely submitted. At present, this option is possible with only a small numbers of collections, a factor outwith the control of the project itself.

Outstanding Issues

As the worked examples at http://hiltpilot.cdlr.strath.ac.uk/pilot/examples/ make clear, it is possible to show that the terminology server will work intelligently for the terms chosen, both in terms of the identification of collections appropriate to the subject search in question and in terms of actual item-level retrieval from a specific collection identified by this means. However, the pilot server is only illustrative – it is not, at this stage, possible to claim that the approach will work in practice for the significant number of different schemes used by networked services in the world at large, or for the large number of types of subject query that would have to be handled on a large scale.

The approach adopted in the HILT pilot remains a possible route to a solution, but a good deal of additional research is needed before it will be possible to determine whether its potential can be realised in practice and, if so, how. Particular areas requiring attention are:

- User subject searching needs and associated interface design issues.
- A detailed, large-scale, multi-scheme examination of mapping issues that arise from these user needs.
- Issues associated with the requirement for the terminologies server to interact with users through other services on the network via machine-to-machine (M2M) interfaces.

The HILT M2M Feasibility Study

The HILT Phase II proposal indicated that it would be 'difficult in such a relatively small, relatively low-cost project to fully investigate M2M use of the pilot facility in an operational sense'. It therefore proposed to focus primarily on the use of the pilot server by end users and to cover the M2M needs by 'examining the requirement for this on an ongoing basis at a mainly theoretical level'. An independent examination of the M2M requirement was undertaken by UKOLN, which recommended (20) that an M2M follow-up project should aim to:

- Provide M2M demonstrator services based on controlled vocabularies mapped within Wordmap.
- Develop SOAP-based interfaces (21) between components of the JISC information environment and Wordmap application programming interfaces, and to use these services in the short term as an aid to specify use cases, and in the longer term as a basis for pilot service if still appropriate at that stage.

- Carry out investigative implementation of a Zthes-based solution (22), with a view to taking advantage of standards-based structured controlled vocabularies (particularly faceted vocabularies) as they become available from third party agencies.
- Track developments within the semantic web and eScience communities, to ensure that decisions made concerning the syntax for structuring vocabularies and the data exchange protocols would take account of forward compatibility.

With this in mind, JISC funded a short study into the feasibility of a project along these lines. This is charged with:

1. Investigating the feasibility of developing a SOAP-based interface between one of the JISC services and the HILT pilot server, whilst also taking into account the possibility of a future Zthes-based solution, relevant implications of work in the eScience and semantic web communities (23), and developments in vocabulary mapping generally (24).
2. Determining the scope and cost of the provision of an actual M2M demonstrator based on SOAP.

The study is due to report at the end of March 2005 and the report will be available on the HILT website by May 2005.

Conclusion

The DDC and mapping-based HILT pilot described in this paper may provide a basis for resolving subject interoperability issues in a distributed, multi-scheme information environment. However, a good deal of additional research is required before it would be safe either to conclude that it will, or to invest heavily in what is likely to be an expensive (and ongoing) enterprise. A cautious approach to forward development is necessary and appears to be the approach favoured by JISC. At present, the feasibility of building an M2M interface to the HILT pilot is being examined, with a positive outcome expected to lead, later in 2005, to the creation of a pilot M2M facility.

The HILT project has shown that the precision with which DDC can identify concepts gives it the potential to act as a mechanism for mapping between terms in different subject schemes. However, it remains to be seen whether this potential can be realised in a cost-effective production service to assist information retrieval by users of networked information services.

References

1. HILT, "High-Level Thesaurus Project" (2005) http://hilt.cdlr.strath.ac.uk/ (March 31, 2005).
2. HILT, "HILT Phase I Final Report" (2002) http://hilt.cdlr.strath.ac.uk/reports/finalreport.html (March 31, 2005).
3. HILT, "HILT Phase II Final Report" (2003) http://hilt.cdlr.strath.ac.uk/hilt2web/finalreport.htm (March 31, 2005).
4. D. Nicholson, "Subject-Based Interoperability: Issues from the High Level Thesaurus (HILT) Project," *International Cataloguing and Bibliographic Control* 32 (1) (2003).
5. D. Nicholson and S. Wake. "Interoperability in Subject Terminologies: The HILT Project," *New Review of Information Networking* Volume 7 (2001).
6. D. Nicholson, S. Wake and S. Currier, "High-level Thesaurus Project: Investigating the Problem of Subject Cross-searching and Browsing between Communities," in *Global Digital Library Development in the New Millennium*, edited by Chin-chich Chen. (Beijing: Tsinghua University Press, 2001).

7. CARMEN, Content Analysis, Retrieval and MetaData: Effective Networking, http://www.mathematik.uni-osnabrueck.de/projects/carmen/index.en.shtml (March 31, 2005).
8. LIMBER, Language Independent Metadata Browsing of European Resources, http://www.limber.rl.ac.uk/ (March 31, 2005).
9. MACS, Multilingual access to subjects, http://laborix.kub.nl/prj/macs/ (March 31, 2005).
10. RENARDUS, Evaluation Reports (2002), http://www.renardus.org/about_us/deliverables/d5_2/D5_2summ.html (March 31, 2005).
11. JISC, JISC Information Environment Architecture, http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/ (March 31, 2005).
12. R. Heery, L. Carpenter, and M. Day, "Renardus Project Developments and the Wider Digital Library Context", *D-Lib Magazine*, 7 (4) (2001), http://www.dlib.org/dlib/april01/heery/04heery.html (March 31, 2005).
13. T. Koch, H. Neuroth, and M. Day, "Renardus: Cross-browsing European subject gateways via a common classification system (DDC)". Paper presented at IFLA satellite meeting: *Subject Retrieval in a Networked Environment*, OCLC, Dublin, Ohio, 14-16 August 2001 (2001).
14. H. Saeed, and A. S. Chaudhury, "Using Dewey decimal classification scheme (DDC) for building taxonomies for knowledge organisation," *Journal of Documentation* 58 (5) (2002): 575-583.
15. A. Shiri, D. Nicholson, and E. McCulloch, "User evaluation of a pilot terminologies server for a distributed multi-scheme environment," *Online Information Review*. 28 (4) (2004): 273-283.
16. E. McCulloch, A. Shiri, and D. Nicholson, "Subject searching requirements: the HILT II experience," *Library Review* 53 (8) (2004): 408-414.

17. E. McCulloch, A. Shiri, and D. Nicholson, “Challenges and issues in terminology mapping: a digital library perspective. *Electronic Library* (2005). in press.

18. M. A. Chaplan, “Mapping Laborline thesaurus terms to Library of Congress subject headings: Implications for vocabulary switching,” *Library Quarterly*, 65(1) (1995):39-61.

19. Porter, Porter stemming algorithm, http://www.tartarus.org/~martin/PorterStemmer/ (March 31, 2005).

20. UKOLN, “HILT Final Report Appendix J” (2003), http://www.ukoln.ac.uk/metadata/hilt/m2m-report/hilt-final-report.pdf (March 31, 2005).

21. SOAP, “SOAP Version 1.2 Part 1: Messaging Framework,” http://www.w3.org/TR/soap12-part1/ (March 31, 2005).

22. “Zthes, a Z39.50 Profile for Thesaurus Navigation,” http://www.loc.gov/z3950/agency/profiles/zthes-04.html (March 31, 2005).

23. A.J. Miles, N. Rogers, and D. Beckett, “SKOS-Core 1.0 Guide, An RDF Schema for Thesauri and Related Knowledge Organisation Systems,” http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide (March 31, 2005).

24. D. Vizine-Goetz, C. Hickey, A. Houghton, and R. Thompson, “Vocabulary Mapping for Terminology Services,” *Journal of Digital Information* 4(4)(2004), Article No.272, 2004-03-11, http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/ (March 31, 2005).