Project Information			
Project Acronym	HILT		
Project Title	HILT Phase IV and Embedding Project Extension		
Start Date	29/04/07	End Date	31/05/09
Lead Institution	University of Strathclyde (Centre for Digital Library Research)		
Project Director	Dennis Nicholson		
Project Manager & contact details	Emma McCulloch CDLR, 12.12 Livingstone Tower, 26 Richmond Street, Glasgow, G1 1XH Tel: 0141 548 5855; Email: e.mcculloch@strath.ac.uk		
Partner Institutions	University of Edinburgh (EDINA) (HILT Phase IV) University of Edinburgh (EDINA), Intute (Embedding Project)		
Project Web URL	http://hilt.cdlr.strath.ac.uk/hilt4/		
Programme Name (and number)	*Shared Infrastructure Services* *Repositories and Preservation Programme*		
Programme Manager	James Farnhill / Tom Franklin / Phil Vaughan		

Document Name			
Document Title	*Final report*		
Reporting Period	29/04/07-31/05/09		
Author(s) & project role	Dennis Nicholson, Emma McCulloch, and Anu Joseph		
Date	31/05/09	Filename	HILT IV Final ReportV1 7290509.doc
URL	*http://hilt.cdlr.strath.ac.uk/hilt4/documents/finalreport.pdf*		
Access	□ Project and JISC internal	□ General dissemination	

Document History		
Version	Date	Comments
1.0	30/10/08	DN first draft
1.1	14/04/09	DN first draft, including embedding dates and initial text
1.2	15/04/09	DN second draft, including Executive Summary
1.3	16/04/09	DN third draft
1.4	17/04/09	DN fourth draft
1.5	08/05/09	EM/AJ additions
1.6	27/05/09	Incorporation of feedback from Steering Group
1.7	29/05/09	Final editing by project team

JISC Final Report

HILT[1] IV and Embedding Extension

Dennis Nicholson

Emma McCulloch[2]

Anu Joseph

May 2009

[1] High Level Thesaurus

[2] Contact person – see cover sheet for email address

Table of Contents

Section	Section Title	Page
	Title Page	2
	Table of Contents	3
	Acknowledgements	4
0	Executive Summary	5
1	Background	7
2	Aims and Objectives	8
3	Methodology	10
4	Implementation	11
5	Outputs and Results	14
6	Outcomes	16
7	Conclusions	22
8	Implications	23
9	Recommendations	24
	References	25
Appendix	Appendix Title	
A	HILT Architecture and the Possibilities of Collaboration	27
B	HILT Demonstrators: Screenshots	30
C	HILT and other Projects	38
D	Work with HILT-Collaborators	39
E	Embedding Toolkit Requirements Document	48
F	Research Report	49
G	Evaluation Report	73
H	Terminologies Interoperability Centre Scoping Study Bid	83
I	Glossary	99

Acknowledgements

Acknowledgements are due to George Macgregor who contributed to HILT IV during its early stages, to JISC staff for the support they provided, to Kathleen Menzies and Gordon Dunsire for help with dissemination, to OCLC for ongoing support and advice, and to JISC for funding HILT.

The HILT team would also like to thank the partners involved in the HILT Embedding Project. Thanks to all at EDINA, Intute and CDLR for their contributions, particularly to Tim Stickland, Ben Soares, Ian Stuart, Phil Cross, Paul Meehan, Gordon Dunsire and Stephane Sechaud, who have undertaken the development work to embed HILT functionality within local services at each of the sites involved. Thanks are also due to the team at IESR, particularly to Leigh Morris and Jo Lambert, for their collaboration and interest throughout the project.

0. Executive Summary

Ensuring that Higher Education (HE) and Further Education (FE) users of the JISC IE can find appropriate learning, research and information resources by *subject search and browse* in an environment where most national and institutional service providers – usually for very good local reasons - use different subject schemes to describe their resources is a major challenge facing the JISC domain (and, indeed, other domains beyond JISC). Encouraging the use of standard terminologies in some services (institutional repositories, for example) is a related challenge. Under the auspices of the HILT project, JISC has been investigating mechanisms to assist the community with this problem through a JISC Shared Infrastructure Service that would help optimise the value obtained from expenditure on content and services by facilitating subject-search-based resource sharing to benefit users in the learning and research communities. The project has been through a number of phases, with work from earlier phases reported, both in published work elsewhere[3], and in project reports (see the project website).

HILT Phase IV had two elements – the core project, whose focus was 'to research, investigate and develop pilot solutions for problems pertaining to cross-searching multi-subject scheme information environments, as well as providing a variety of other terminological searching aids', and a short extension to encompass the pilot embedding of routines to interact with HILT M2M services in the user interfaces of various information services serving the JISC community. Both elements contributed to the developments summarised below.

HILT IV Results Summary

A range of pilot M2M terminology services based on SRU/W, SOAP, and SKOS, and a database of terminologies and pilot mappings were successfully built and tested by the project, as was an embryonic toolkit to help information services' technical staff to embed M2M interactions in user interfaces to improve subject retrieval, browse, and deposit services. Further details are provided in the body of the report, and on the project website and wiki. In addition, an evaluation was carried out and an associated report written. This is included below as Appendix G.

Various practical experiments to successfully embed terminology service interaction into JISC community services to create operational pilot subject browse and retrieve enhancements for service users were conducted - . The resulting demonstrations can be seen, along with other HILT demonstrators, on the HILT project demonstrators page at http://hilt.cdlr.strath.ac.uk/hilt4/demonstrators.html. In addition, illustrative screenshots are available in Appendix B.

A generic distributed subject interoperability and terminology services architecture was developed and shown (at a very basic level) to be a feasible proposition. Over time, this architecture will permit HILT-based services to grow and improve by incorporating other terminology services being developed elsewhere, both within JISC and the UK, and beyond[4]. As part of this work, it was determined that a terminology services registry is a key element of the architecture and that the core functionality required to build and run such a registry is already inherent in the HILT pilot services.

A report on research into various selected issues of relevance to the provision of an effective future entry-level service or its further refinement was compiled to help inform both the costs and requirements of an initial entry-level operational service and any future extension of this. The report is included in this document as Appendix F.

Against the above background, it was agreed with JISC that the best general option for a sustainable operational Shared Infrastructure Service based on project outcomes is a Terminologies Interoperability Centre offering a mix of standard 'plug and play' type M2M and toolkit facilities free at the point of use, including a training portal and an associated terminology services registry, more

[3] See http://hilt.cdlr.strath.ac.uk/dissemination.html for a list key HILT publications
[4] http://www.oclc.org/research/projects/termservices/resources/termservices-overview.pdf

flexible, charged-for, specially-scoped versions of these, tailored to the needs of individual services and institutions, and ongoing development via a mix of collaboration and externally funded R&D, as well as through JISC support.

It was further agreed that, prior to setting up such a Centre, a preliminary scoping study was required to provide a well-researched evidence base that would inform and guide a future 'soft launch' of a Terminologies Interoperability Centre (TIC) by:

- Putting in place service quality infrastructure to support the work of the Centre, including further development and testing of the components from HILT IV and work on a pilot terminology services registry. This will ensure that the standard services offered at the soft launch will be robust and usable in a range of JISC service and user environments.
- Determining service user and end user needs via iterative feedback from hands-on experience, utilising outcomes in TIC scoping and soft launch plans, creating mechanisms for an ongoing assessment of such needs, and identifying specific players to work with TIC during the soft launch period.
- Scoping in detail what free and charged-for services the Centre should offer and what they would cost.
- Producing a bid for TIC start-up costs, a programme of works, and a well-researched Sustainability Plan.

Finally, it was concluded that there is a need, both within the JISC community, and in the world at large, for a globally-scoped programme of collaborative research and development based on a common view of an inclusive architecture for subject interoperability service design. Accordingly, efforts were made to begin work towards agreeing a collaborative approach with other 'players' in the terminologies field. A paper on the idea was presented at an ontologies conference in Helsinki in November 2007, a paper on the topic published in the international version of the Signum journal in November 2008 (Nicholson, 2008), and, more recently (December 2008), steps have been taken to contact major European projects in the terminologies area to begin the process of talking about collaboration and about applying for FP7 funding to carry the work forward.

1. Background

Ensuring that Higher Education (HE) and Further Education (FE) users of the JISC IE can find appropriate learning, research and information resources by *subject search and browse* in an environment where most national and institutional service providers – usually for very good 'local' reasons - use different subject schemes to describe their resources is a major challenge facing the JISC domain (and, indeed, other domains beyond JISC). Encouraging the use of standard terminologies in some services (institutional repositories, for example) is a related challenge. Under the auspices of the HILT project, JISC has been investigating mechanisms to assist the community with this problem through a JISC Shared Infrastructure Service that would help optimise the value obtained from expenditure on content and services by facilitating subject-search-based resource sharing to benefit users in the learning and research communities. The project has been through a number of phases, with work from earlier phases reported, both in published work elsewhere[5], and in project reports (see the project website: http://hilt.cdlr.strath.ac.uk/hilt4/). Prior to the present phase (Phase IV), the project:

1. Established that the preferred approach of the various services in the JISC domain to resolving the issue is one based on mapping the various subject schemes together through a central shared service that provides users with the correct alternative terms to use in the various different schemes (HILT Phase I[6]).
2. Built an illustrative terminologies service pilot capable of taking a user-input subject term, identifying JISC collections relevant to the subject of the query and providing the user with the correct subject term to use for the subject scheme employed by any given identified collection (HILT Phase II).
3. Conducted a study that looked at the feasibility of turning this pilot into an M2M pilot service able to supply terminologies and mapping data for other services to use and scoped out an outline design for the pilot (HILT M2M Feasibility Study).
4. Built the M2M pilot and scoped out a design for the initial entry-level service described in Section 1 above (HILT Phase III).

The core HILT Phase IV project was funded to permit this initial entry-level service to be built, tested for user requirements and retrieval effectiveness, refined in line with the findings, and extended to permit the use of a range of distributed terminology services for interoperability - to research, investigate and develop pilot solutions for problems pertaining to cross-searching multi-subject scheme information environments, as well as providing a variety of other terminological searching aids. It was also to examine the level of need and interest amongst JISC services in respect of an operational service and, if appropriate, a scoping of the costs and requirements of a future operational phase of the service. The project was also charged with conducting a parallel programme of research into selected topics to help inform both the costs and requirements of an initial entry-level operational service and any future extension of this. In addition, a short extension was funded to explore the pilot embedding of routines to interact with HILT M2M services in the user interfaces of various information services serving the JISC community. This Final Report covers both elements of the project.

[5] See http://hilt.cdlr.strath.ac.uk/dissemination.html for a list key HILT publications
[6] HILT Phase I Final Report: http://hilt.cdlr.strath.ac.uk/Reports/FinalReport.html

2. Aims and Objectives

HILT Phase IV had two elements – the core project, whose focus was 'to research, investigate and develop pilot solutions for problems pertaining to cross-searching multi-subject scheme information environments, as well as providing a variety of other terminological searching aids', and a short extension to encompass the pilot embedding of routines to interact with HILT M2M services in the user interfaces of various information services serving the JISC community.

The core project

HILT Phase IV had the following aims and objectives:

1. The creation of an initial entry-level terminologies and subject interoperability service comprising
   (a) A freely available package consisting of an SRW client from the internet, instructions for making it interact with HILT, and illustrative user interface routines (which could be customised by local JISC information services) for using the client to exploit HILT facilities, terminologies, and terminology mappings.
   (b) A database comprising a range of subject schemes in use in the JISC IE, high-level mappings between these and (roughly) the top 1000 DDC sections[7], and a limited set of in-depth mappings in a subject area of interest to users of two representative JISC services.
   (c) A SOAP-based HILT requests and responses handler based around the eight search and retrieve functions identified in Phase III as meeting the needs of clients.
   (d) An SRW server to provide a standard interface to the SOAP requests and responses handler.
   (e) Client use of IESR and the HILT database of terminologies and mappings to identify collections appropriate to a user's subject request, determine the subject schemes they use, and provide subject interoperability by offering subject access via scheme hierarchies entered at a point appropriate to the user's subject interest.
   (f) Extend client functionality to allow (via IESR) the identification of terminology and interoperability services other than HILT[8]and their use to provide enhanced user services.
2. An examination of client user interface needs and retrieval effectiveness in respect of both the high level and in-depth mappings conducted using users, retrieval problems, and associated 'use cases' from the two services and the initial entry-level service described above.
3. The design and implementation of an integrated programme of project dissemination and survey activity.
4. Research into various selected issues of relevance to the provision of an effective future entry-level service or its further refinement – and a subsequent report on results.
5. The development and presentation of future development proposals, including an estimate of the costs of setting up and maintaining an operational service and funding ongoing research and development needs beyond an entry-level service based on information arising out of 1-4 above, together with discussions with JISC and the project Steering Group.

The HILT IV Embedding Extension

The short HILT IV Embedding Extension Project ran between January and May 2009 and had the aim of embedding the pilot use of HILT M2M terminology and interoperability facilities within the user interfaces of a number of information services in order to enhance functionality for the end-users of these services. The services in question are based at EDINA, Mimas (Intute), and CDLR. Specific objectives were to:

- Enhance Intute's user searches in various ways (e.g. by offering more specific terms when a user puts in a very broad one (like 'farming').

[7] Or the 1000 (or thereabouts) considered most useful by in-project experts.

[8] We have had discussions with OCLC and would expect, at minimum, to be able to incorporate pilot web service based terminology services developed by them as examples here, although we will also look more widely and will include any compatible terminology services funded under the last JISC capital programme if possible.

- Improve the Edina Depot[9] service by using HILT to guide user-provided subject tags via an M2M fed list of standard terms (e.g. from JACS) and extend this facility in case of a no match situation in JACS. For e.g. taking a user input search term and expanding it via HILT using DDC spine and which are then cross-related to find JACS equivalents.
- Enhance Scotland's Information[10] service by using HILT to offer a disambiguation stage with terms from DDC before finding the collections of interest.. The possibility of offering DDC terms in different languages would also be considered with German, French, and Welsh-Gaelic being possibilities for inclusion.
- Conduct generic work on optimizing the HILT spell check pilot mechanism for use in UK HE an FE.

.
A second extension project, funded in parallel at IESR in Manchester, also entailed HILT services embedding plans, and collaboration between the two in respect of HILT embedding is taking place. The relevant objective in this case was to:

- Use HILT M2M services to enhance the user browse interface to the IESR collection discovery service in a variety of ways by sending user terms to HILT, and enrich the interface by getting back broader, narrower, and related terms associated with various HILT schemes and using them to improve retrieval.

In addition to the above, the project also did some joint work with other projects. See Appendix C for brief details.

[9] http://depot.edina.ac.uk/
[10] http://www.scotlandsinformation.com/

3. Methodology

The overall methodology employed in the project was iterative development with ongoing refinement of outputs informed by live (albeit pilot) tests, background research, and internal discussions, with the following being the main areas of focus:

- Database structures;
- General and multi-lingual terminological and mapping work, the latter at both a detailed and a more general 'high' level;
- Development of central request and response functions;
- SKOS (and to a lesser extent MARC) mark-up issues;
- User interface issues and their relationships to request and response functions and to mark-up variations;
- Practical and theoretical architecture issues relating to a distributed approach;
- Work with users in services, including survey work, one-to-one work with a few selected services, and embedding work;
- Embedding toolkit design, development, and local and field testing;
- Internal and external evaluation work;
- Work on the probable best shape and form of a future service, including discussions with JISC informed by an initial attempt to develop a sustainability plan (helped identify the need for a scoping study)

In addition, there was an active programme of dissemination – aimed, appropriately for this stage of HILT, at other services and professionals in the field. There was also a programme of work focused on the ongoing development and presentation of interactive demonstrators of HILT capabilities. These were primarily web-based, although a video of the web-based toolkit demonstration was also made.

More detail on the approach taken and how it was organised is presented below in the implementation section. However, this is necessarily high level. As indicated above, the primary method of working was iterative development with ongoing refinement of outputs informed by live tests, background research, and internal discussions. This necessarily meant that much of the detail of the methodology emerged in the weekly and other meetings of the project team and their partners and advisors, so that reporting it in full would be a major task of minimal benefit outwith the project. That said, a good deal of the detail may be ascertained – and in a helpful form - from selected appendices, as follows:

Appendix D: Work with HILT-Collaborators
Appendix F: Research Report
Appendix G: Evaluation Report
Appendix H: Bid for Terminologies Interoperability Centre Scoping Study

and to a lesser extent from:

Appendix A: HILT Architecture and the Possibilities of Collaboration
Appendix C: HILT and other Projects
Appendix E: Embedding Toolkit Requirements Document.

4. Implementation

As indicated in the methodologies section, HILT phase IV employed a range of work strategies and methodologies, as appropriate to the range of tasks tackled, with the primary method of working being iterative development with ongoing refinement of outputs informed by live tests, background research, and internal discussions and the detail best represented by the various appendices listed above in Section 3. However, this took place within an overall structure with the following being key elements:

1. An in-depth examination of the user interfaces, subject schemes, and subject interoperability needs of the two JISC services chosen to be part of the project. This entailed detailed discussions with service administrators regarding current usage of local services and interface ergonomic issues, as well as desk-based research examining the schemes in use by the local services (e.g. coverage, semantic structure, exhaustivity, etc.) and how best to map these to the DDC spine to ensure optimum subject interoperability with other JISC services.
2. The subsequent compilation of a requirements document (see Appendix E) describing the user interface functionality development required, initial screen design needs and programming language issues, terminologies preparation and loading issues and associated database design questions, and HILT requests and responses functional requirements. The plan was to build on an extension of the facilities created in HILT phase III, rather than to redesign from scratch – meaning that no further methodologies for initial design and creation had to be considered. The plan for the requirements document itself was to follow a similar format and design as in HILT phase III[11] and for the Phase IV work in this area to essentially constitute an extension of the Phase III work. In practice – and as part of the aforementioned iterative process – some changes were applied, but this initial approach formed the basic structure for the iterative approach.
3. The programming and implementation of the SRW client, user interface routines, database, and requests handler elements of the *initial* version of the entry-level service for two JISC services that were to be the focus of some of the toolkit and related background functionality and terminological work (in the core project, these were Intute and CAIRNS, although this was extended by the embedding extension project, and some use was also made of Phase II embedding work with Go Geo!).
4. The creation of high-level mappings between the schemes used by the chosen JISC services and (roughly) the top 1000 DDC sections. These DDC numbers were to be identified through discussion with the terminology experts, OCLC and by HILT analysis of WorldCat. In the event, a better approach was found and they were identified via BUBL[12]. Mappings from the relevant JISC subject schemes to the numbers identified by the analysis were implemented in the HILT terminologies database. The methodologies for facilitating interoperability between satellite terminologies and the spine followed those used in HILT phase III, supplemented with guidance provided in Part 4 of the BS8723 (with possible refinement where indicated by advice from the terminology experts). Again, however, there was impact informed by ongoing research and discussion carried out within this basic approach.
5. The creation of in-depth mappings in a chosen subject area of each of the two JISC services. Again, the methodologies for facilitating interoperability between satellite terminologies and the spine followed those used in HILT phase III as described in 4 above, but with the same adjustments as detailed there.
6. The design and implementation of an evaluation programmes to test the initial entry-level service's 1) functionality 2) retrieval effectiveness and 3) interface in terms of effectiveness, helpfulness and ergonomics in each of the two clients and their associated service environments. See inset box below for an overview of how evaluations in general were conducted.
7. Processing of the results from the evaluations to inform an improved requirements specification for the entry-level service and its elements.

[11] HILT phase III requirements documents (V.4-10): http://hilt.cdlr.strath.ac.uk/hilt3web/requirements.html
[12] http://bubl.ac.uk /

8. The programming and implementation of the client, user interface, database, and requests handler elements of an improved *post-evaluation* version of the entry-level service.
9. An evaluation of the implications of allowing for the extension of the clients or the server to use other terminology and interoperability services[13] that might be discovered via IESR or similar services, either now or in future. See inset box below for an overview of how evaluations in general were conducted.
10. Determination of the likely impact of the need to deal with such terminology and interoperability services on collection and/or service level description requirements.
11. Various tests to demonstrate 'proof of concept' in respect of a distributed version of the HILT services (see inset last paragraph of the evaluation inset below).
12. The design and creation of an entry-level service dissemination programme to inform JISC service providers likely to benefit from an operational HILT service of the possibilities of the proposed service for their services and users (largely conducted via general dissemination and via the HILT-Collaborators interactive and survey work described in Appendix D).
13. The design and implementation of a survey to determine the impact of the dissemination programme on service providers and measure the likely demand for the proposed service (largely conducted via general dissemination and via the HILT-Collaborators interactive and survey work described in Appendix D).
14. An estimate of the costs of setting up and maintaining an operational service and funding ongoing research and development needs beyond an entry-level service. This was done by identifying the proposed nature of the services, specifying individual elements and estimating the likely costs associated with them, using the most reliable data available.
15. An associated proposal to JISC requesting funding to set up an operational service.
16. For the short Embedding Project Extension the following general steps were involved in overall planning and implementation:
    - An initial conference call was held with project partners to establish their key requirements and to formalise work outlined in the project bid in more detail.
    - A schedule was put in place to make contact with all partners on a weekly basis to chart progress. A wiki was also set up where partners were given a dedicated area to document issues and problems encountered whilst embedding HILT in their local services.
    - Guidance documentation was added to the wiki to provide support for partners' implementation of HILT.
    - Direct support was given by project staff when requested by partners, or when one of the weekly discussions highlighted a need for further input from the HILT team.

Evaluation programmes.

As indicated in the Phase IV bid and the Project Plan, it was unrealistic to predict the range of metrics required to enable a robust evaluation early in the project and a formal evaluation plan was not formulated until well into the second half of project development. There was, however, a planned and necessary focus on ongoing evaluation from very early in the project through the research, test, and discussion based iterative approach described earlier. In the event, it was decided that a good deal of the evaluation required to develop and refine effective working pilots would necessarily have to be of this kind and this was the main approach taken in respect of:

- Testing the functionality of the initial entry-level service;
- Testing the functionality as it develops;

[13] The services concerned are most likely to be pilot services available via our collaboration with OCLC

- Testing retrieval effectiveness in the initial entry-level service and its ongoing development in the first two thirds of the project;
- Comparing the high-level hierarchy-driven approach with the more in-depth mapping based approach;
- Testing the effectiveness, helpfulness, and ergonomics of the user interface in each of the two service-focused clients and their associated service environments (with assistance here from the staff of the services);
- User interface evaluation as the embedding work and, more particularly, development of demonstrators and the toolkit, progressed.

In addition to the above, a funded evaluation was conducted towards the end of the project (late 2008). This evaluation – based primarily on a user study involving both quantitative and qualitative elements and conducted independently of HILT by Ian Ruthven of the Computer and Information Sciences Department at the University of Strathclyde – looked at the overall HILT approach, concentrating on the utility of the HILT terminology server function. It did not attempt to evaluate the quality of the HILT mappings as such; however, it did consider the HILT functions and mappings within the context of stereotypical information searches. The full detail of the methodology applied within this external evaluation, together with the results and conclusions is provided in the external evaluator's report in Appendix G.

Both sets of evaluations, but particularly the internal set, were used to inform an improved requirements specification for the entry-level service and its elements (largely on an ongoing basis, as a more informed and developed understanding of issues and requirements was formed). Similarly, both sets, but particularly the external evaluation, will be used to inform future development strategies in respect of both technical and terminological developments.

The proposed evaluation of the implications of allowing for the extension of the clients or the server to use other terminology and interoperability services (e.g. the OCLC pilot terminology services) that might be discovered via IESR or similar registries, either now or in future, was undertaken internally. This was done through a largely theoretical, discussion based, approach, but also involved some practical testing of (a) interaction with the OCLC pilot services (b) general (as opposed to terminology) service discovery through an IESR simulation.

JISC also conducted its own study into HILT. During 2008, they funded an external institution to conduct an evaluation of the perceived value of HILT in the community. We saw the interim results of this and they showed a clear indication that the outcome would favour further development of the HILT services. Unaccountably, JISC elected not to publish the outcomes indicating that they felt that the recommendations were not supported by the research conducted. They provided no specifics in respect of this claim, despite a written request for such specifics sent to them in November 2008. We are continuing to ask for this information. Even if the claim about the recommendations is true (and we are far from convinced of this), there is no doubt that the interim outcomes seen by us indicated a clear response from those interviewed in favour of HILT and that outcome is, in our view, worth publishing. We are also unclear why (a) JISC agreed to the methodology used in the first place if they felt it inadequate (b) Why they didn't follow up with a second study with an acceptable methodology. We intend to continue to push for better information on this point. We did not do so in November 2008 because JISC started discussions in December 2008 in which they proposed the approach to the future of HILT described elsewhere in this report. However, we propose to do so now, since we feel the information should be in the public domain.

5. Outputs and Results

The outputs and results of HILT Phase IV fall under two headings: those achieved via the core project and those achieved via the embedding extension project.

Main HILT IV Project

The core project:

- Built and tested a range pilot M2M (SRU/W[14], SOAP[15], and SKOS[16]) based web services to deliver terminologies and terminology mappings to JISC and institutional information services, supporting the transparent enhancement of subject search facilities.
- Built and tested a database of terminologies (DDC[17], UNESCO[18], HASSET[19], IPSV[20], JITA[21], MeSH[22], CAB, GCMD[23], NMR[24], AAT[25], SCAS, JACS[26],), high level mappings to DDC (HASSET, IPSV, UNESCO, JACS), and limited sample 'deeper' mappings to MeSH and HASSET[27].
- Built and tested an embedding toolkit to offer information services the core software needed to begin building subject interoperability services for their users by interacting at M2M level with the above web services and database through routines embedded transparently in their service user interfaces - a programmer's toolkit to help build improved subject browse and retrieve facilities.
- Conducted various practical experiments to successfully embed terminology service interaction into JISC community services to create operational pilot subject browse, retrieve, and deposit enhancements for service users (a smaller piece of work preceding the embedding extension project mentioned above).
- Developed a generic terminology services architecture that will (over time) permit the HILT services to grow and improve by incorporating terminology services being developed elsewhere.
- Determined that a terminology services registry is a key part of this architecture and that the core functionality required to build and run such a registry is already inherent in HILT pilot services.
- Developed staff skill sets and experience associated with the problems of subject searching and subject interoperability within and across information services using different subject terminologies, with the best approaches to mapping new schemes into the database, and with an associated distributed architecture to permit the ready integration of new services into the JISC and global subject interoperability landscape.
- Determined in conjunction with JISC that the best general option for a sustainable operational Shared Infrastructure Service based on HILT project outcomes is a Terminologies Interoperability Centre offering a mix of standard 'plug and play' type M2M and toolkit facilities free at the point of use, including a training portal and an associated terminology services registry, more flexible, charged-for, specially-scoped versions of these, tailored to the needs of individual services and institutions, and ongoing development via a mix of collaboration and externally funded R&D, as well as through JISC support (for further details, see Appendix H, Section 1.1.4).

With these outputs and results in place, the project had developed to the point where it was in a position to tackle the more advanced embedding work required by the extension project (described below), and put forward to JISC proposals for future development towards the set up of an operational shared infrastructure service. It had in place a core of tested pilot facilities and terminology sets and an embryonic mappings set that could be the basis of a future service, an understanding of the

[14] Search/Retrieve Web Service (SRU/W): http://www.loc.gov/standards/sru/
[15] SOAP: http://www.w3.org/TR/soap/
[16] Simple Knowledge Organization System (SKOS) Core: http://www.w3.org/2004/02/skos/
[17] Dewey Decimal Classification (DDC): http://www.oclc.org/dewey/
[18] UNESCO Thesaurus: http://www2.ulcc.ac.uk/unesco/
[19] Humanities and Social Science Electronic Thesaurus (HASSET): http://www.data-archive.ac.uk/search/hassetSearch.asp
[20] Integrated Public Sector Vocabulary (IPSV): http://www.esd.org.uk/standards/ipsv/
[21] http://eprints.rclis.org/jita/
[22] Medical Subject Headings (MeSH): http://www.nlm.nih.gov/mesh/
[23] Global Change Master Directory (GCMD): http://gcmd.nasa.gov/Resources/valids/keyword_list.html
[24] National Monuments Record Thesauri (NMR): http://thesaurus.english-heritage.org.uk/
[25] Art & Architecture Thesaurus (AAT): http://www.getty.edu/research/conducting_research/vocabularies/aat/
[26] Joint Academic Coding System (JACS): http://www.ucas.ac.uk/figures/ucasdata/subject/
[27] Also included are LCSH mappings included with the DDC file provided by OCLC.

terminological, technical, architectural, and staffing requirements for future development, and a clear if outline vision of what was required to move towards a useful, robust, and sustainable service aligned with JISC's strategic aims[28].

Embedding Extension Project

The following partners embedded HILT functionalities within their services and demonstrators are listed below:

Intute: Suggests alternate terms corresponding to a user term input by retrieving terms from MeSH and HASSET, two schemes held by HILT, using the get_filtered_set API. The system also displays DDC suggestions for a user term input (using the get_ddc_records API). In the instance of a no match situation, the system offers spelling suggestions from HILT.

Demo: http://www.intute.ac.uk/search_hilt.html

EDINA: The Depot at EDINA looks for JACS terms corresponding to a user term (using get_filtered_set API). If there is no match, the system looks for any matches in DDC (using get_ddc_records API), which are then cross-correlated to find JACS equivalents. In the instance of a no match situation, alternate spelling suggestions are offered from HILT.

Demo: http://lucas.ucs.ed.ac.uk/cgi-bin/hilt-depot

CDLR: Scotland's Information Service offers DDC suggestions for a user term input before finding a collection or collections of potential relevance (using get_ddc_records). German and Welsh terms are also displayed (where available).

Demo: http://scone.strath.ac.uk/Service/SCONEServiceHilt/ddcsearchinput.cfm

Spelling suggestions from HILT also offered for a keyword search

Demo: http://scone.strath.ac.uk/Service/SCONEServiceHilt/IndexSpellCheck.cfm

Screen shots for these different demonstations are available in Appendix B

[28] http://www.jisc.ac.uk/aboutus/strategy/strategy0709/strategy_aims.aspx

6. Outcomes

Since the HILT work is ongoing, outcomes of HILT Phase IV and the associated embedding extension project are best described by indicating the current state of play in respect of various key areas of project development: *Service functions, database, and embedding toolkit*; *The HILT Architecture and its Implications*; *Project understanding of service and end-user needs*; and *The Requirements of a robust, useful, and sustainable service*.

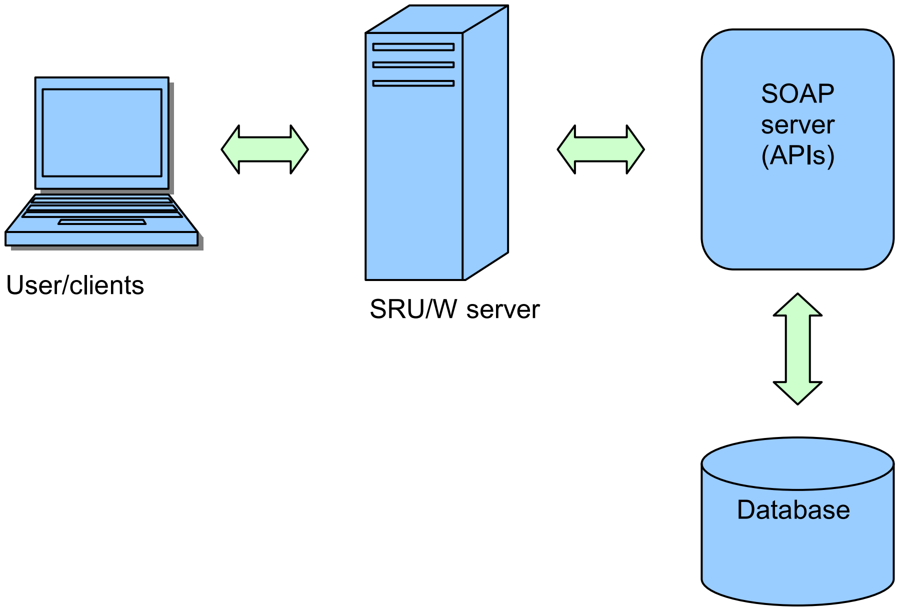
Service functions, database, and the embedding toolkit

A full description of these and how these function in practical circumstances requires greater detail than is possible here and is provided in McCulloch, 2008. The account below deals only with the technical details of the working facilities and with possible ways in which they can be used by information services through M2M interaction with HILT.

Technical Summary

Although the following description omits some of the detail, it gives a useful overview of the technical basis of the HILT services:

The system is designed in a modular fashion and the diagram below shows how each of the different modules interacts. This design facilitates the replacement or modification of the components if a need arises in the future. SQL Server 2005 is a highly scalable, programmable, secure database management system which is used for data storage. The SOAP server enables a distributed web based environment and provides the APIs to interact with the database. A SRW/U server is implemented to support M2M interaction and the APIs can be accessed through the SRW/U protocol.



Further information is provided in what follows.

SRU/W server: The SRU/W server facilitates M2M interaction between clients and the server. Index Data's SimpleServer (http://www.indexdata.dk/simpleserver/) is used to build the SRU/W server. This is a Perl module intended to develop Z39.50, SRU and SRW servers over any type of database. It is based on the popular YAZ toolkit and is robust, efficient, portable, and interoperable with all Z39.50, SRU and SRW clients. The server involves a SOAP envelop to wrap and transfer data. This implementation allowed us to have a distributed environment.

SOAP server: HILT uses a SOAP server to interact with the database and wrap the results in SKOS. REST is a recent alternative to SOAP; the project evaluated both options and decided on SOAP because it is more stable. This implementation is based on simple and easy-to-use NuSOAP (a group of PHP classes) which helps to create and consume SOAP web services based on SOAP 1.1, WSDL 1.1 and HTTP 1.0/1.1. It does not require any special PHP extensions. By using SOAP, it was possible to encapsulate database access through a few functions. It is also possible to replace SOAP with REST at a future date if the need arises. In this case, HILT functions would require to be implemented in REST.

Database: HILT data is stored using SQL Server 2005 database software and contain 12 Vocabularies[29] (AAT, CAB, GCMD, HASSET, IPSV, JACS, JITA, LCSH, MeSH, NMR, SCAS and UNESCO) separate from DDC. Tables are designed in such a way to accommodate future versions of vocabularies and flat tables are also designed to improve performance. Interaction with the database is through stored procedures, which adds a layer of security and better performance. A full-text index facilitates the search and retrieval of data efficiently and SQL Server's ability to rank result records (based on the frequency of search terms), is useful in multi term queries. Database access is supported through the SOAP server.

APIs: As part of the HILT toolkit we have various APIs to offer, which are accessible using the SRW/U server. Results are wrapped in SKOS (Simple Knowledge Organisation System).

The APIs are:

1. get_collections

Returns collections classified under a specified DDC number or its stem, including subject scheme used. A process of truncation is used within this function in order to maximise retrieval potential and avoid a 'no hits' situation. For example, if a user searches for collections with DDC number 336.26, HILT returns collections matching 336.26, 336.2, 336, 330 and 300.

2. get_ddc_records

Returns DDC captions and numbers related to a subject term. The user can then choose the most appropriate to his/her interest. This function searches within DDC and any schemes that are mapped to DDC for an input term and returns only DDC instances.

3. get_non_ddc_records

Returns terms from schemes other than DDC by matching user terms to DDC notations, before identifying mappings to those particular numbers. This function also employs truncation to retrieve relevant mappigs. For an input 336.36, HILT returns mappings corresponding to 336.36, 336.3, 336, 330 and 300. It's up to the client to choose the mappings of their interest.

4. get_all_records

This function combines the output of get_ddc_records and get_non_ddc_records.

5. get_filtered_set

get_filtered_set enables a user to search a particular scheme or schemes directly for a specific term, together with its broader, narrower, related and non-preferred terms, if selected and where applicable. You can also search for the id of a term in a scheme (if known), to build a hierarchy of that scheme where available.

[29] Plus the DDC spine

6. get_sp_suggestions

HILT Spell Checker can suggest a list of words similar to the input word. This implementation is based on David Spencer's code using the n-gram method and the Levenshtein distance. An index (the dictionary) with all the possible words from the HILT database and a word list based on an English dictionary and on Intute keywords (Note: Intute keywords not yet added) is created and suggestions are based on this index. The index can be extended by including terms from other sources as well. A single term query returns multiple records whereas a multi term query returns a single record.

7. get_wordnet_suggestions

Returns the description about the input term from a large lexical database of English- WordNet. Implementation is based on JwordNet - Java interface to WordNet.

Toolkit: A toolkit has been designed to illustrate how these various APIs can be embedded within a service and provide intending implementers with a basic 'starter' package. Further details are available at: http://hilt4.cdlr.strath.ac.uk/toolkit/intro.cgi. A Perl version is available to download from: http://hilt4.cdlr.strath.ac.uk/toolkit.zip (Note: wiki account required for access; please contact project team) and a PHP version may be available at a later date.

Value: Possible Uses

As indicated earlier, a description of how the functions and the toolkit may be used in practice is provided in McCulloch, 2008. At present, these are pilot facilities, but the aim is to move gradually towards operational services as described below. Operational facilities will enable national, institutional, and other information service providers to access and use terminological and interoperability data to enhance their own services in a variety of ways, including, but not necessarily limited to, the following:

a. Improving recall in a subject search of one or more databases by enriching the set of terms known to a user by providing synonyms and related terms.
b. Providing the best terms for a subject search in a remote service that uses a subject scheme unfamiliar to 'home service' users (or in a cross-search of a group of such services).
c. Taking a user's subject term and using it to identify relevant information services via registries such as IESR (http://iesr.ac.uk/).
d. Generating an interactive browse structure where a scheme is arranged hierarchically.
e. Offering the ability to send a term from a chosen subject scheme and receive back data on broader terms, narrower terms, hierarchy information, preferred and non-preferred terms, and so on.
f. Providing cataloguing staff with information on subject schemes and inter-scheme mappings to assist in metadata creation.
g. Providing a spell-check mechanism to assist user searching.
h. Providing a service to assist user search formulation by providing information on search terms entered (e.g. what the term means, whether it has alternative meanings, whether there are synonyms that might be useful in a search and so on).

Uses of this kind can, of course, already be tested using the pilot facilities currently in place, and the embedding work carried on in both the core and the extension project has shown some of the practical applications of the HILT facilities.

The HILT Architecture and its Implications

A key element of the current state of play within the project is a developing understanding of the architecture required to support a rich and user adaptive set of relevant terminology and subject interoperability services, and the implications of this architecture for a distributed and collaborative approach to future service growth and associated research and development.

Arising out of JISC-funded work with UK user communities in the first phase of HILT, the project's original focus as regards solving subject interoperability problems was on mapping between subject schemes via a DDC spine. Mapping is an established and effective approach (see, for example, Mayr and Petras, 2008) and has attracted significant resourcing in other European countries (see, for example Agosti et al., 2007; Mayr and Petras, 2008). However, HILT now has an architectural approach that allows it (a) to adopt a range of interoperability strategies appropriate to specific use cases and cost-benefit levels (e.g. expensive deep mapping where the problem and user area justifies this, or high-level or browse-based retrieval where significant costs would be less justifiable), (b) to incorporate in the model - for the benefit of the JISC user communities – solutions offered and funded by other players, whether they be based on mapping to a spine, scheme to scheme, or some variety of automated approach.

The original assumption underpinning HILT was that it would be, in essence, a stand-alone service that would facilitate subject interoperability in the JISC Information Environment by mapping commonly used schemes to a DDC spine and providing the best terms to use in services using a particular subject scheme via these mappings. All through the progress of the project, it was very clear (a) that, even within the JISC communities this was likely to be a large and probably expensive task (b) that mapping to a DDC spine and, indeed, mapping in itself was only one of many possible approaches to the problem (c) that the variety of subject interoperability problems within JISC would probably be better tackled using a mix of these methods. However, the likely additional costs entailed in encompassing a variety of approaches in a HILT service meant that only the one approach – the aforesaid mapping via DDC spine - could be encompassed in a project with limited resources.

Fortunately, the move towards HILT offering M2M functionality has changed the picture somewhat. In the kind of web services environment now envisaged, it becomes possible to think in terms of a distributed and devolved approach to subject interoperability described in Appendix A (see Nicholson, 2008 for a more detailed description) – with information services, both within JISC and elsewhere, utilising M2M connections, not just to HILT, but to terminology services worldwide as they come onstream. These would be funded by a range of different players, and would utilise all kinds of different approaches to interoperability. Information services would find and interact with them via data obtained through infrastructural services such as IESR or dedicated terminology services registries. Other data from the same sources would allow the information services to elegantly handle the different kinds of data and different approaches to interoperability served up by these different terminology services.

With this as the backdrop, it becomes necessary to look at the question of research required for subject interoperability service development in the context of the distributed and devolved architecture it implies. The downside of this is that a *research for development* agenda that is more complex and wider in scope than might otherwise be the case (since it must handle not just interactions with one service of known functionality and scope, but (a) a large variety of different services of unknown functionality and scope, and (b) intermediate interactions with infrastructural services such as registries). The upside is that the distributed and devolved environment means (a) that JISC need only focus on those aspects of *research for development* needed by specific JISC communities at specific times (b) that other players in the field will likely seek to tackle elements of the *research for development* agenda required in their own domains, and (c) it is probable that some of the issues faced in the JISC communities will also be faced (and, in some cases, tackled) elsewhere.

There are a number of significant current or recent projects (CACAO[30], MACS[31], LIMBER, RENARDUS[32], KoMoHe[33], and STAR[34] are examples) and a good deal of ongoing and recent work (see, for example, Day, M. et al (2004); Landry, P. (2004); Tudhope, D., Koch, T., and Heery, R. (2006); Zeng, M. L., & Chan, L. M. (2006); van Gendt, M., et al. (2006); Vizine-Goetz, D. et al (2006); Macgregor G. & McCulloch E. (2006); Agosti et al. (2007); Mayr P. and Petras V. (2007); Binding C.

[30] http://www.cacaoproject.eu/
[31] MACS Project: https://macs.vub.ac.be/pub/
[32] RENARDUS Project : http://renardus.sub.uni-goettingen.de/
[33] KoMoHe Project http://www.gesis.org/en/research/information_technology/komohe.htm
[34] STAR Project http://hypermedia.research.glam.ac.uk/kos/star/

Et al (2008); Geser, G. (2008); Levergood, B et al (2008)) in the general area of terminologies, terminological interoperability in cross-searching and other scenarios, and terminology and terminology interoperability services. Much of it is of relevance, or potential relevance, to HILT, not just because the issues faced are similar, but because, in future, comprehensiveness, and the consequent need to collaborate and interoperate with other initiatives in the area (Nicholson, 2008), will be an issue.

This still emerging perspective on the likely collaborative technical and networked environment within which a future service will grow will impact on upcoming considerations relating to the shape and form of a sustainable service and on probable research and development paths pursued.

Project understanding of service and end-user needs

At present, HILT has only a limited understanding of the needs of the services likely to use HILT facilities and an even more limited understanding of the needs of the end users of such information services. Work within HILT III and HILT IV on embedding HILT facilities transparently into the user interfaces (or clones of those interfaces) of three services has been undertaken – the GoGeo! service, which used HILT terminological data to enrich user search term sets when searching distributed external services external to GoGeo!; the Intute service, which used HILT facilities in spell-check and in offering users a drop down list of terms related to the term input (e.g. *motherhood* and mother and *child relationships* as alternatives to *mothers*); and the CAIRNS distributed catalogue service which established a proof-of-concept demonstrator to show how HILT could be used to improve recall. These were small scale exercises, but nevertheless provided useful insights into the kinds of issues likely to be encountered as HILT aimed to offer its services operational to a range of disparate services with different aims and different terminological issues (it is safe to say that almost every information service operates in a unique subject terminologies environment making 'plug and play' solutions to all but a few core problems difficult to provide). They also helped inform the need for the embedding extension project and, to some extent, its shape and form. The embedding project consulted with three services on their individual needs. A pilot toolkit was then developed in line with these requirements, which was easily embedded into service demonstrators. Both of these processes led to the project gaining enhanced knowledge of user needs.

HILT IV also carried out some survey work with a number of individuals who joined a HILT email list (HILT-Collaborators@jiscmail.ac.uk) in response to a call for service staff who were potentially interested in using HILT pilot facilities. This work showed that there was a good deal of interest from the community in the kind of uses of HILT listed above. Twenty-nine people joined the list which was set up specifically to look at embedding work of this kind. A questionnaire asking what kinds of uses were of interest was answered by sixteen people. All indicated an interest in at least one of the above potential ways of using HILT, most indicated an interest in three or more ways. However, the work also showed that in many cases – via some practical tests of actual training and attempts to do simple embedding work - it would be difficult for the early stages of embedding work using HILT facilities to be supported remotely and that more work was needed to discover the best way of ensuring that embedding work at service sites was facilitated and well supported by HILT.

This work notwithstanding, however, it is safe to say that further in-depth work with both staff of services interested in using HILT's M2M services to support subject searching and interoperability functions within their service interfaces, and, more particularly, with the end-users of such information services, is needed to inform the process of transforming pilot facilities into a robust, useful, and sustainable service.

The requirements of a robust, useful, and sustainable service.

As noted earlier in this report, one of the outputs of HILT Phase IV was the conclusion that the best general option for a sustainable operational Shared Infrastructure Service based on HILT project outcomes was a Terminologies Interoperability Centre. More precisely, it was determined - based on an understanding of issues in the area of subject interoperability as developed over a number of phases of HILT, and a discussion with JISC itself on how to best fit future developments into JISC's strategic requirements and also provide a useful, reliable, and sustainable service for JISC-related communities - that the best route forward was to work towards the 'soft launch' of a Terminologies Interoperability Centre. Once launched, this would offer the community a mix of free and charged-for services and would be supported by a mix of JISC funding, externally earned income and R&D funding, and collaboration, and would provide:

- M2M and user-level access to terminology sets, the detail of those terminology sets, and data to facilitate interoperability between them.
- Open source software toolkits that would enable M2M interaction with HILT web services to be transparently embedded in the user interfaces of local, national and project information services.
- A basic architecture for terminology and interoperability services in the JISC Information Environment (and potentially beyond).
- A way of mounting and developing new terminologies and terminologies interoperability data required by the community, including JISC-specified work to facilitate improvements in subject access in and between the various JISC user communities and their external partners based on ongoing assessments of user and service needs.
- Advisory and M2M support services for projects, services, and other initiatives in JISC or JISC institutions where there is a subject description, subject retrieval, or subject interoperability facet.
- A JISC funded free advisory and training service on using the above facilities in local or national services and projects in the percentage of cases where this was relatively straightforward (plug 'n' play, but after a bit of advice and training).
- The development and hosting of a JISC-focused terminologies services registry.
- A charged-for consultancy service where the work and advice required by local and national services, projects, and organisations (both within and outwith JISC) was less straightforward or more sophisticated (because of unique client and client service circumstances and terminology sets)
- A portal for tools and training in the areas described above.
- A focus for wider work in the terminologies area, funded through a variety of sources, including non-JISC sources (for example though successful bids for European funding).
- Ongoing work to facilitate JISC involvement and leadership in a strategically important area with significance for both subject-based retrieval needs in research, learning, teaching, and elsewhere and semantic web developments.

The initial plan in respect of this was to move directly from HILT Phase IV to the proposed soft launch. However, initial consideration of this proposal quickly proved that an interim stage – a scoping study for the proposed Centre and its services – was a necessary and sensible first step. If funded, this scoping study will provide a well-researched evidence base that will inform and guide a future 'soft launch' of a Terminologies Interoperability Centre by:

- Putting in place service quality infrastructure to support the work of the Centre, including further development and testing of the components from HILT IV and work on a pilot terminology services registry. This will ensure that the standard services offered at the soft launch will be robust and usable in a range of JISC service and user environments.
- Determining service user and end user needs via iterative feedback from hands-on experience, utilising outcomes in TIC scoping and soft launch plans, creating mechanisms for an ongoing assessment of such needs, and identifying specific players to work with TIC during the soft launch period.
- Scoping in detail what free and charged-for services the Centre should offer and what they would cost.
- Producing a bid for TIC start-up costs, a programme of works, and a well-researched Sustainability Plan.

A bid to fund the scoping study was submitted to JISC in March 2009 and the outcome is awaited. The bid document (a Phase IV deliverable) is included in this report as Appendix H. The proposal was that this study would take place between June 2009 and November 2010, with an actual soft launch of the Centre taking place in December 2010. However, these dates no longer look feasible.

7. Conclusions

The Phase IV project has taken HILT to the point where the launch of an operational support service in the area of subject interoperability is a feasible option and where both investigation of specific needs in this area and practical collaborative work are sensible and feasible next steps. Moving forward requires detailed work, not only on terminology interoperability and associated service delivery issues, but also on service and end user needs and engagement, service sustainability issues, and the practicalities of interworking with other terminology services and projects in UK, European, and Global contexts.

In respect of the first of these, a Scoping Study is required to provide a well-researched evidence base that will inform and guide a future 'soft launch' of a Terminologies Interoperability Centre by scoping out the detail of a sustainable mix of JISC-funded and charged for services designed to help meet JISC strategic aims in respect of serving its stakeholders. The basic outline of what is required in respect of addressing subject interoperability issues is known. What is needed now is a clear, evidence-based, indication of what is required in detail to ensure a robust, sustainable service sensitive to community needs and the strategic aims of the JISC. Amongst other things, this requires intensive work in which specific JISC-related user communities (including both service and end users) are given hands on experience of the terminological tools as they become available as part of an iterative and ongoing approach to understanding, and implementing solutions to, user needs in specific communities.

In respect of the practicalities of interworking with other terminology services and projects in UK, European, and Global contexts, HILT sees a need, both within the JISC community, and in the world at large, for a globally-scoped programme of collaborative research and development based on a common view of an inclusive architecture for subject interoperability service design. Although HILT progress in this area is currently limited, efforts have nevertheless been made to begin work towards agreeing a collaborative approach with other 'players' in the terminologies field. A paper on the idea was presented at an ontologies conference in Helsinki in November 2007, a paper on the topic published in the international version of the Signum journal in November 2008, and, more recently (December 2008), steps have been taken to contact major European projects in the terminologies area to begin the process of talking about collaboration and about applying for FP7 funding to carry the work forward.

Discussions are at a very early stage, but if a proposal were put forward, it would be based on the idea that, despite the plethora of different means of tackling inter-KOS and inter-lingual interoperability currently in play and the size and complexity of the problem itself, it is both desirable and possible to facilitate gradual cumulative progress towards optimizing subject interoperability across the networked world through geographically distributed co-ordinated collaborative effort - that it can be done by agreeing a model set of requirements for interoperability service design and collectively pursuing a common research and development agenda based on it.

It is by no means certain that such a proposal will materialise in the event. However, a number of key players have at least indicated a clear interest in the idea.

8. Implications

The implications of the work carried out in HILT Phase IV and its extension project are set out in the Outcomes section above which covers implications for future work in the area, for other professionals, for the future of HILT development and research, for JISC itself, and for users.

9. Recommendations

The primary recommendations of the project are:

1. That JISC support the continuance of the HILT work by funding the proposed scoping study and the subsequent set up of a Terminologies Interoperability Centre.
2. That the HILT team explore the possibilities of European funding for an integrative collaborative subject interoperability project based on the idea of joint research and development across major European terminology projects based on the HILT distributed architecture or something like it agreed amongst partners.

References

Agosti et al. (2007). Roadmap for MultiLingual Information Access in the European Library. In *Research and Advanced Technology for Digital Libraries*, Springer Berlin/Heidelberg, Volume 4675/2007. Abstract online at: http://www.springerlink.com/content/g8126155w7518333/

Ardo, A. (2004). 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.

Binding C., Tudhope D., May K. 2008. Semantic Interoperability in Archaeological Datasets: Data Mapping and Extraction via the CIDOC CRM. Proceedings (ECDL 2008) 12th European Conference on Research and Advanced Technology for Digital Libraries, Aarhus, 280–290. Lecture Notes in Computer Science, 5173, Berlin: Springer. final preprint presentation

Day, M., Koch, T., & Neuroth, H. (2004). Searching and browsing multiple subject gateways in the Renardus Service. In: Dijkum, C. van, Blasius, J., Kleijer, H., & Hilten, B. van (eds.) *Recent developments and applications in social science methodology: proceedings of the Sixth International Conference on Logic and Methodology, August 17-20, 2004, Amsterdam, The Netherlands*. Amsterdam: SISWO Instituut voor Maatschappijwetenschappen. (CD-ROM). Online at: http://www.ukoln.ac.uk/metadata/publications/rc33-2004/renardus-paper.pdf

Doerr, M. (2001). Semantic problems of thesaurus mapping . Journal of Digital information, vol. 1, issue 8, 2001-03-26. Available at: http://jodi.tamu.edu/Articles/v01/i08/Doerr/ (accessed 24 July 2008).

Friesen, N. (2002). Semantic Interoperability and Communities of Practice. Available at: http://www.cancore.ca/documents/semantic.html

Geser, G. (2008). *STERNA Technology Watch Report. Full Report*. Ref: STERNA Del.6.5, 10 December 2008. Salzburg Research. Online at: http://www.sterna-net.eu/index.php/en/downloads

Godby, C. J. et al.(1999). 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 (

Koch, T. and Vizine-Goetz, D. (1998), Automatic Classification and Content Navigation Support for Web Services: DESIRE II Cooperates with OCLC Annual Review of OCLC Research, 1998. Available online at: http://digitalarchive.oclc.org/da/ViewObject.jsp?objid=0000003489

Landry, P. (2004). Multilingual Subject Access The Linking Approach of MACS. In *Cataloging & Classification Quarterly*. Volume 37, Issue 3/4.

Levergood, B. et al (2008). The Specification of the Language of the Field and Interoperability: Cross-Language Access to Catalogues and Online Libraries (CACAO). International Conference on Dublin Core and Metadata Applications, Berlin, 2008. Available at: http://dcpapers.dublincore.org/ojs/pubs/article/viewFile/933/929

Macgregor, G. and McCulloch, E. (2006) Collaborative tagging as a knowledge organisation and resource discovery tool. Library Review, 55 (5). pp. 291-300. Pre-print available at http://strathprints.strath.ac.uk/2335/1/strathprints002335.pdf

Mayr, P. and Petras, V. (2007). Building a terminology network for search: the KoMoHe project. In *Proceedings of the International Conference on Dublin Core and Metadata Applications 2008*. Available at: http://arxiv.org/pdf/0808.0518

Mayr, P., and Petras, V. (2008). Cross-concordances: terminology mapping and its effectiveness for information retrieval. In *IFLA 2008 Conference Proceedings*. Available at: http://eprints.rclis.org/13828/

McCorry, Helen C. (1991). Professional Notes : Computers II : Bad Language in Ethnography Records. Museum Management and Curatorship 10:4, 1991

McCulloch, E. et al. (2005). Challenges and issues in terminology mapping: a digital library perspective. Electronic Library, 23 (6). pp. 671-677. Available at: http://www.emeraldinsight.com/10.1108/02640470510635755

McCulloch, E (2008). Developing a pilot toolkit to improve subject interoperability between collections and services within the JISC Information Environment. Catalogue and Index. In press.

Nicholson, D. (2008). A Common Research and Development Agenda for Subject Interoperability Services?, Signum, Issue 5, 2008.

Olson, Hope A. (1994). Universal Models: A History of the Organization of Knowledge. Advances in Knowledge Organization. Frankfurt : INDEKS Verlag, 1994.

Spiteri, Louise F. (1999). The Essential Elements of Faceted Thesauri. Cataloging & Classification Quarterly, Vol.28 (4), 1999.

Tudhope, D, Koch, T., Heery, R. (2006). Terminology Services and Technology JISC state of the art review. Available at: http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.pdf

van Gendt, M. et al (2006). Semantic Web Techniques for Multiple Views on Heterogeneous Collections: a Case Study. In: Proceedings of the 10th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2006), Julio Gonzalo, Constantino Thanos, M. Felisa Verdejo and Rafael C. Carrasco (eds.), Springer Verlag, LNCS vol. 4172, pp. 426-437, Alicante, Spain, September 17-22, 2006. Available at: http://www.cs.vu.nl/~mtvgendt/STITCH-ECDL06.pdf

Vizine-Goetz, Diane, Andrew Houghton, and Eric Childress. 2006. "Web Services for Controlled Vocabularies." Bulletin of the American Society for Information Science and Technology 35 no. 5 (June/July). Available at: http://www.asis.org/Bulletin/Jun-06/vizine-goetz_houghton_childress.html

Whitehead, Cathleen. (1990). Mapping LCSH into Thesauri: The AAT Model. Beyond the Book: Extending MARC for Subject Access, edited by Toni Petersen and Pat Molholt. Boston: G.K. Hall, 81-96.

Zeng M and Chan L. (2004). Trends and issues in establishing interoperability among knowledge organization systems. Journal of American Society for Information Science and Technology, 55(5): 377 – 395.

Zeng, M. L., & Chan, L. M. (2006). Metadata Interoperability and Standardization – A Study of Methodology Part II: Achieving Interoperability at the Record and Repository Levels. In *D-Lib Magazine*, Volume 12, Issue 6. Available at: http://www.dlib.org/dlib/june06/zeng/06zeng.html

Appendix A: HILT Architecture and the Possibilities of Collaboration

An examination of the subject interoperability landscape and its various facets as reflected in current and recent work in the area points up three things. First, that the issue is of increasing concern to a wide and growing range of organisations (as the membership of the various projects listed in the aforementioned Zeng and Chan, 2004 will affirm). Second, that the problem is one of significant size and complexity[35]. Third, that it is likely to raise overlapping but nonetheless varying sets of issues in different community contexts probably best identified and tackled in those contexts, but also entails many elements of common concern for which a collaborative approach is either the best approach, or the only feasible one.

Taken together, these suggest that a move towards a collaborative approach is not only a desirable goal to aim for, but one necessary to a successful outcome. The problem is large and difficult to tackle, has potential for collaboration in the growing number of organisations with a stake in its resolution, probably requires action in a variety of communities using methods developed and tested in those communities (see, for example, Friesen, N. 2002), and has many common elements where division of labour is both feasible and desirable.

This latter point is particularly applicable to HILT. Its intended role as a JISC shared infrastructure service aiming to support the work of end-user focused information services facing a myriad of different problem situations and their various facets means its aim has largely been to address the interoperability problem in general, rather than some specific aspect of it. Moreover, since its primary focus is serving researchers, teachers, and learners in universities, and since this community as a whole is likely to need to find and access subject information located almost anywhere in the world and to do so using almost any of the available knowledge organisation systems (KOS[36]), and, potentially, any language, it is impossible to envisage a future service being a stand-alone enterprise. The need for HILT, and for any similar general service, will be to provide, not just direct access to the KOS and KOS interoperability data it provides, but also to offer services using it facilities to permit discovery of, and integrated interaction with, other sources of such data, as well as means of intelligently and transparently handling this M2M data in user interfaces to best meet the needs of individual user groups tackling particular types of task.

Accordingly, the HILT architecture assumes that one of the ways a future service will grow in functionality and coverage will be via the elegant inclusion of geographically distributed and administratively devolved Non-JISC KOS and KOS interoperability services based and financed elsewhere (e.g. by OCLC), as well as via (for example) particular JISC services offering their own interoperability data. In addition, it assumes that there will be collaborative and mutually supportive research and development taking place based on an agreed architecture that will, if it is to be inclusive – and hence widely accepted – be something close to that adopted by HILT.

The diagram below shows how the HILT model would work in practice. It assumes:

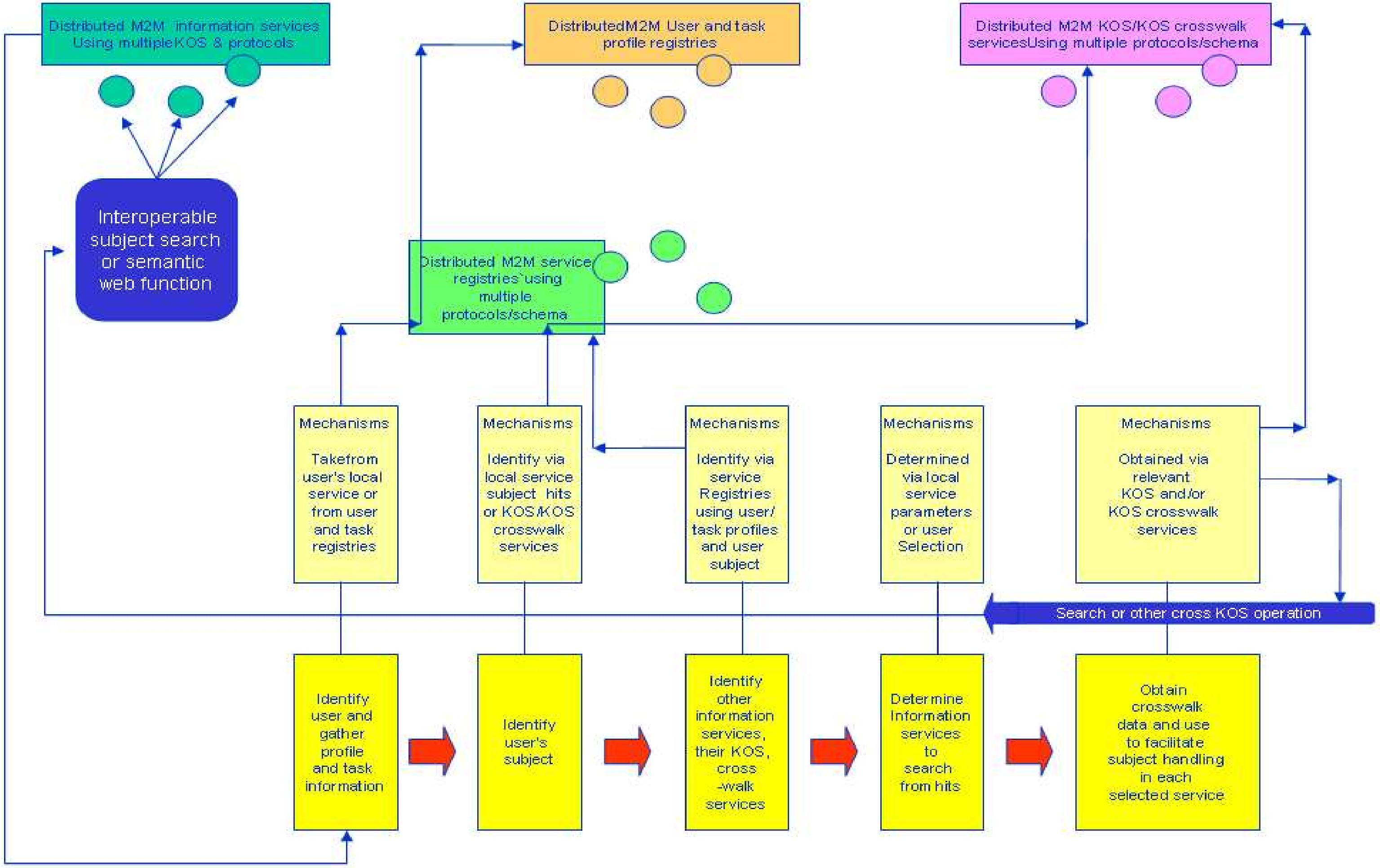
- Multiple service registries out there that 'know' about each other, so any given information service only needs to know about their own 'home' service registry

[35] A flavour of just how complex can be gleaned if we consider that simple subject retrieval raises a plethora of issues in its own right (see, for example, Spiteri, 1999; Olson, 1994; McCorry, 1991), even in the relatively simple circumstances of one user addressing one retrieval problem in one service. If it is to be effective, a common approach to the problem of subject interoperability must be adaptive to these and similar subject retrieval issues in both single and multiple scheme situations. Nor is it simply a matter of identifying a set of schemes to be made interoperable and using a single approach to solve the interoperability problem for these schemes. Many different approaches to this problem have been developed and applied. Zeng and Chan (2004) identify a number of methods for achieving and improving interoperability, including derivation/modelling, translation/adaptation, direct intellectual mapping, and mapping via a spine. Other potential solutions proposed include automatic or semi-automatic classification (see for example Koch and Vizine-Goetz, 1998; Godby et al, 1999; Ardo, 2004). All will have their own problems and will raise different issues for an effective common approach - see, for example, Doerr, 2001, McCulloch et al, 2005, and Whitehead, 1990 on issues related to the mapping approach.

[36] Taxonomies, classification schemes, thesauri, subject heading lists and similar – see Zeng and Chan, 2004 for a full description

- Multiple information services out there, recorded in and discoverable via one or more service registries, and classified according to their subject coverage
- Multiple user and task profile registries[37] out there, also recorded in and discoverable via one or more service registries
- Multiple KOS and KOS 'crosswalk' services out there, also recorded in and discoverable via one or more service registries (or possibly terminology registries)
- Many information services out there using many different KOS and different access protocols
- The user's start point is usually (but not necessarily) one of these information services
- The users home information service offers the user a facility to cross search other information services appropriate to her subject needs and uses a subject interoperability service (encompassing the whole of this diagram) to underpin this service.
- Most of the workings of this service are transparent to the user
- The assumption is that the user starts at one of the information services and needs to search one or more (possibly unknown) others
- The home service goes through these processess shown in the yellow boxes, using the services registries, user and task profile registries and KOS and KOS crosswalk services as necessary
- In a fully implemented system, the user's home information service would go through the following steps:
    - Use local information or information from user and task registries (possibly discovered via service registries) to gather relevant information on the context as indicated by the user's profile and her task
    - Identify the user's subject in relation to some standard scheme (e.g via user selected hits in a local database or by finding it in the preferred or non-preferred terms of a standard scheme through a KOS service identified via a service registry)
    - Identify (1) other information services relevant to the user's subject, user profile, and task, (2) Their KOS, (3) KOS crosswalk services appropriate to each individual KOS, either via local information, or service registries (NB These sets ((2) and (3)) would include HILT, but could also include other KOS crosswalk services, either using their own approach to mapping, or an alternative to mapping. An obvious example might be OCLC.) This is the key point about the inclusion of other KOS and KOS interoperability services in the georaphically distributed and administratively devolved architecture that HILT envisages.
    - Select the best information services to search either from user and task profile or user interaction
    - Obtain interoperability data for each relevant KOS via KOS and KOS crosswalk services
    - Use the data to facilitate user search of appropriate services, sometimes transparently, but often interactively (which, of course, is the point of the exercise)

[37] Note that the user and services registries are seen as future developments and are not essential to the running of a service; they will simply allow an enhanced service later.


Distributed M2M information services Using multipleKOS & protocols
DistributedM2M User and task profile registries
Distributed M2M KOS/KOS crosswalk servicesUsing multiple protocols/schema
Interoperable subject search or semantic web function
Distributed M2M service registries`using multiple protocols/schema
Mechanisms
Takefrom user's local service or from user and task registries
Mechanisms
Identify via local service subject hits or KOS/KOS crosswalk services
Mechanisms
Identify via service Registries using user/ task profiles and user subject
Mechanisms
Determined via local service parameters or user Selection
Mechanisms
Obtained via relevant KOS and/or KOS crosswalk services
Search or other cross KOS operation
Identify user and gather profile and task information
Identify user's subject
Identify other information services, their KOS, cross -walk services
Determine Information services to search from hits
Obtain crosswalk data and use to facilitate subject handling in each selected service

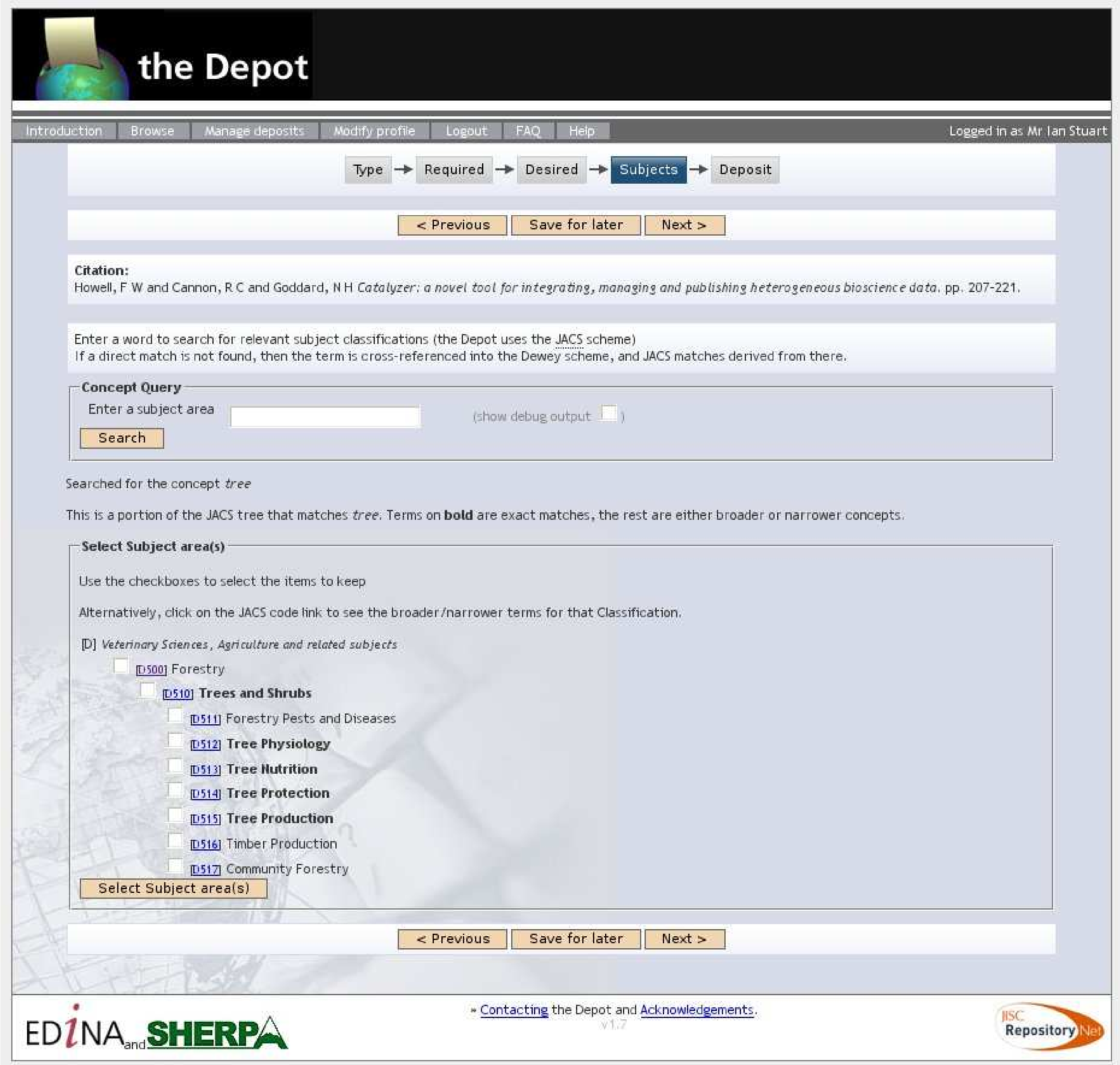
Appendix B: HILT Demonstrators: Screenshots

CDLR (Lead Site): HILT Toolkit: A demonstration of the HILT toolkit is available to view here http://hiltm2m.cdlr.strath.ac.uk/hilt4/demo_movies/toolkit.swf

Partner Sites: Embedding project partners have also created demonstrations of HILT functionality within their local services:

1) EDINA: The following screen shots demonstrate how embedding HILT terminology services can enhance the user experience in aspects of EDINA services. EDINA implemented the following three workpackages:

> WP1: Use of a user-tagging scheme (for Depot) - hence only allowing tagging from a particular subject classification scheme (JACS), so providing some consistency and the possibility of a more useful tag cloud/search of user generated tags (using hilt function get_filtered_set).



Citation:
Howell, F W and Cannon, R C and Goddard, N H *Catalyzer: a novel tool for integrating, managing and publishing heterogeneous bioscience data*. pp. 207-221.

Enter a word to search for relevant subject classifications (the Depot uses the JACS scheme)
If a direct match is not found, then the term is cross-referenced into the Dewey scheme, and JACS matches derived from there.

Concept Query
Enter a subject area ▭ (show debug output □)

Searched for the concept *tree*

This is a portion of the JACS tree that matches *tree*. Terms on **bold** are exact matches, the rest are either broader or narrower concepts.

Select Subject area(s)

Use the checkboxes to select the items to keep

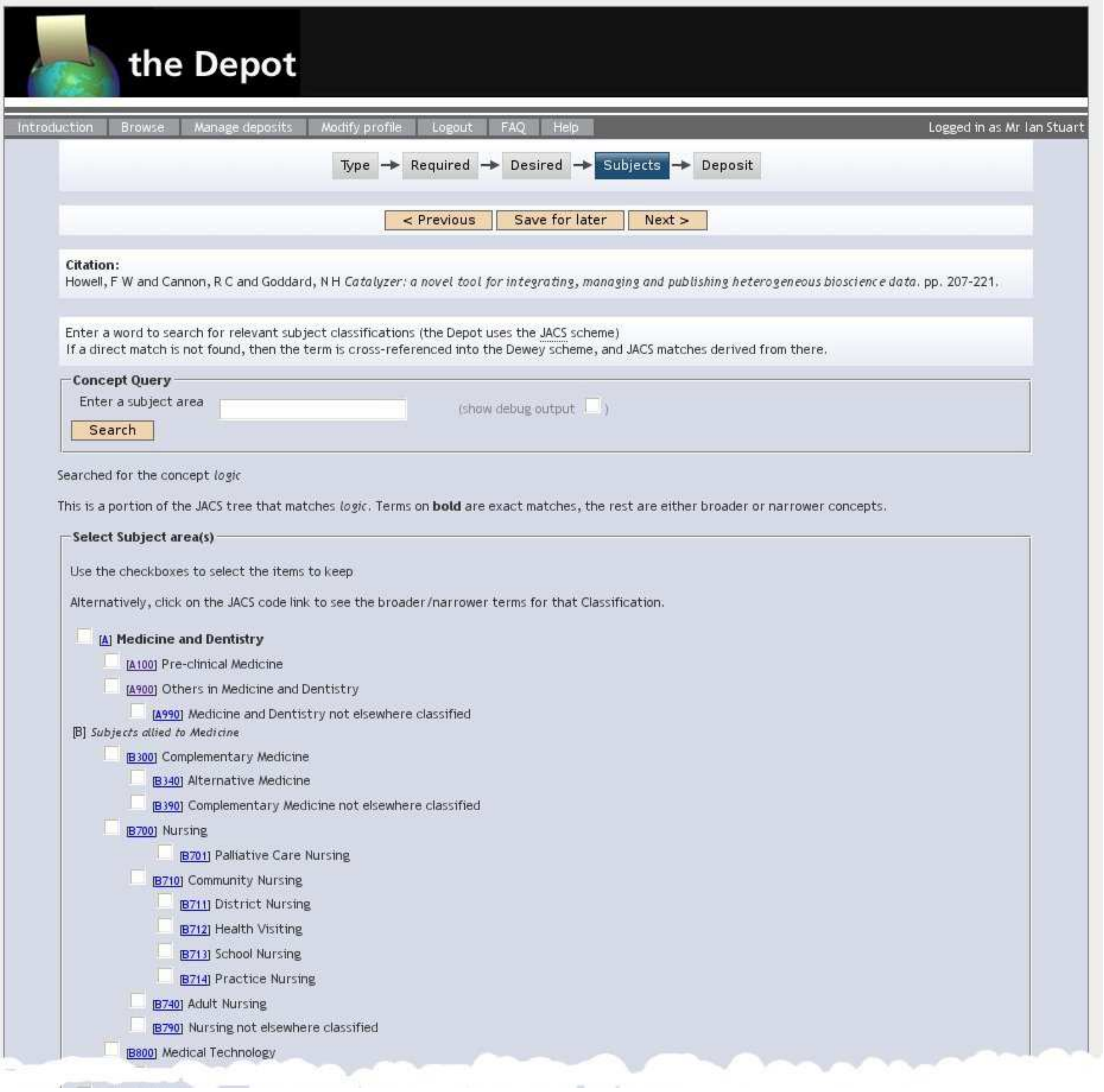
Alternatively, click on the JACS code link to see the broader/narrower terms for that Classification.

[D] *Veterinary Sciences, Agriculture and related subjects*
- □ [D500] Forestry
  - □ [D510] **Trees and Shrubs**
    - □ [D511] Forestry Pests and Diseases
    - □ [D512] **Tree Physiology**
    - □ [D513] **Tree Nutrition**
    - □ [D514] **Tree Protection**
    - □ [D515] **Tree Production**
    - □ [D516] Timber Production
    - □ [D517] Community Forestry

» Contacting the Depot and Acknowledgements.
v1.7

WP2: Demonstrate how mappings between JACS and DDC can be used to support (1) the classification of (learning, teaching, and research) materials against courses on deposit and (2) (via the remainder of the HILT database) enhancements to searching by end users of JACS and Non-JACS repositories within and out with UK HE, looking in particular at, the Depot and institutional repositories (jointly with the other two partners).

In the case of a no match situation in JACS, DDC will be searched for matches corresponding to the user term(s) input and JACS mappings for these DDC numbers will be returned (using HILT functions get_ddc_records and get_non_ddc_records).



Citation:
Howell, F W and Cannon, R C and Goddard, N H *Catalyzer: a novel tool for integrating, managing and publishing heterogeneous bioscience data.* pp. 207-221.

Enter a word to search for relevant subject classifications (the Depot uses the JACS scheme)
If a direct match is not found, then the term is cross-referenced into the Dewey scheme, and JACS matches derived from there.

Concept Query

Enter a subject area

(show debug output ☐)

Searched for the concept *logic*

This is a portion of the JACS tree that matches *logic*. Terms on **bold** are exact matches, the rest are either broader or narrower concepts.

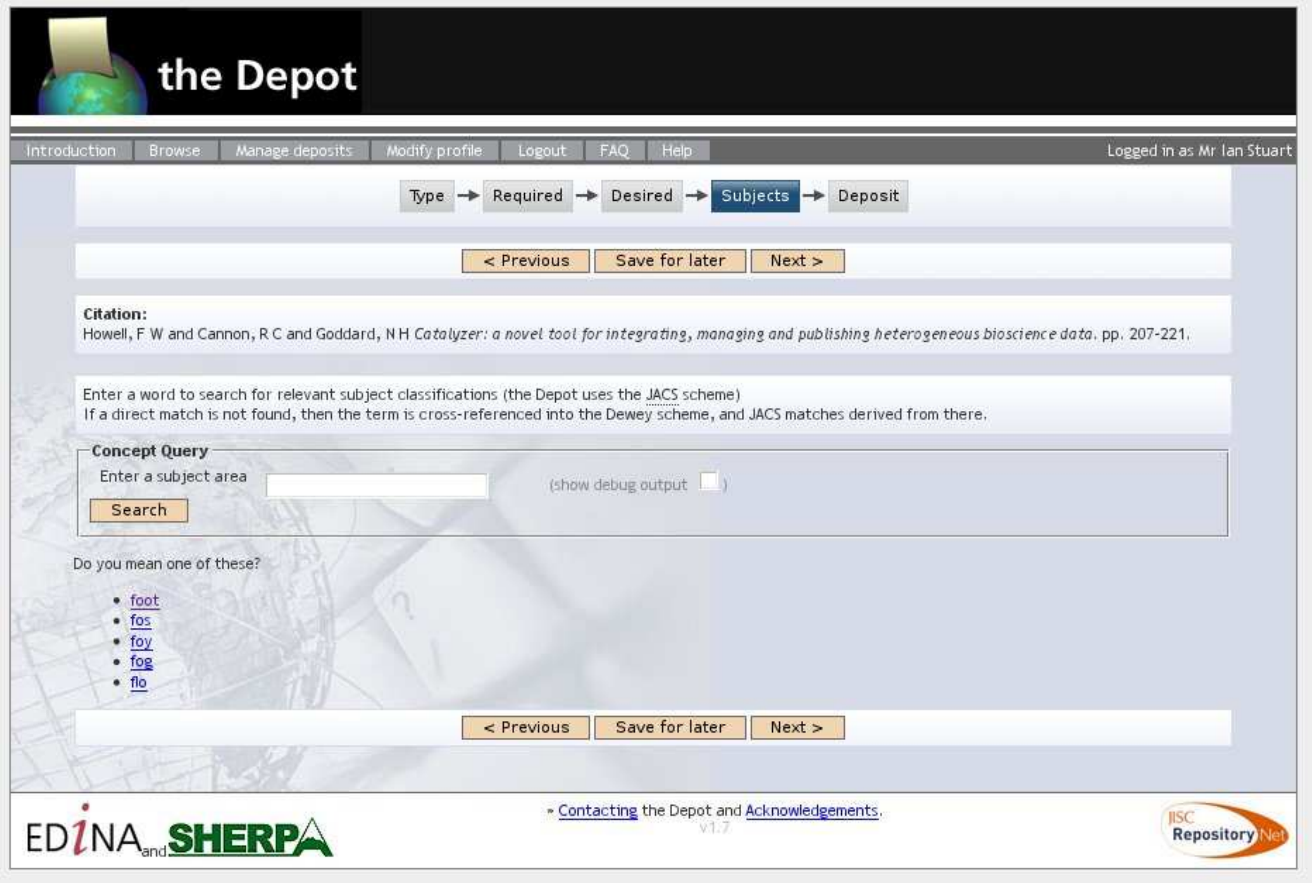
Select Subject area(s)

Use the checkboxes to select the items to keep

Alternatively, click on the JACS code link to see the broader/narrower terms for that Classification.

- ☐ [A] **Medicine and Dentistry**
  - ☐ [A100] Pre-clinical Medicine
  - ☐ [A900] Others in Medicine and Dentistry
    - ☐ [A990] Medicine and Dentistry not elsewhere classified
- [B] *Subjects allied to Medicine*
  - ☐ [B300] Complementary Medicine
    - ☐ [B340] Alternative Medicine
    - ☐ [B390] Complementary Medicine not elsewhere classified
  - ☐ [B700] Nursing
      - ☐ [B701] Palliative Care Nursing
    - ☐ [B710] Community Nursing
      - ☐ [B711] District Nursing
      - ☐ [B712] Health Visiting
      - ☐ [B713] School Nursing
      - ☐ [B714] Practice Nursing
    - ☐ [B740] Adult Nursing
    - ☐ [B790] Nursing not elsewhere classified
  - ☐ [B800] Medical Technology

WP3: If the user continues to receive no hits, spelling suggestions from HILT will be displayed.



the Depot
Introduction | Browse | Manage deposits | Modify profile | Logout | FAQ | Help
Logged in as Mr Ian Stuart
Type → Required → Desired → Subjects → Deposit
< Previous | Save for later | Next >

Citation:
Howell, F W and Cannon, R C and Goddard, N H *Catalyzer: a novel tool for integrating, managing and publishing heterogeneous bioscience data.* pp. 207-221.

Enter a word to search for relevant subject classifications (the Depot uses the JACS scheme)
If a direct match is not found, then the term is cross-referenced into the Dewey scheme, and JACS matches derived from there.

Concept Query
Enter a subject area
Search
(show debug output □)

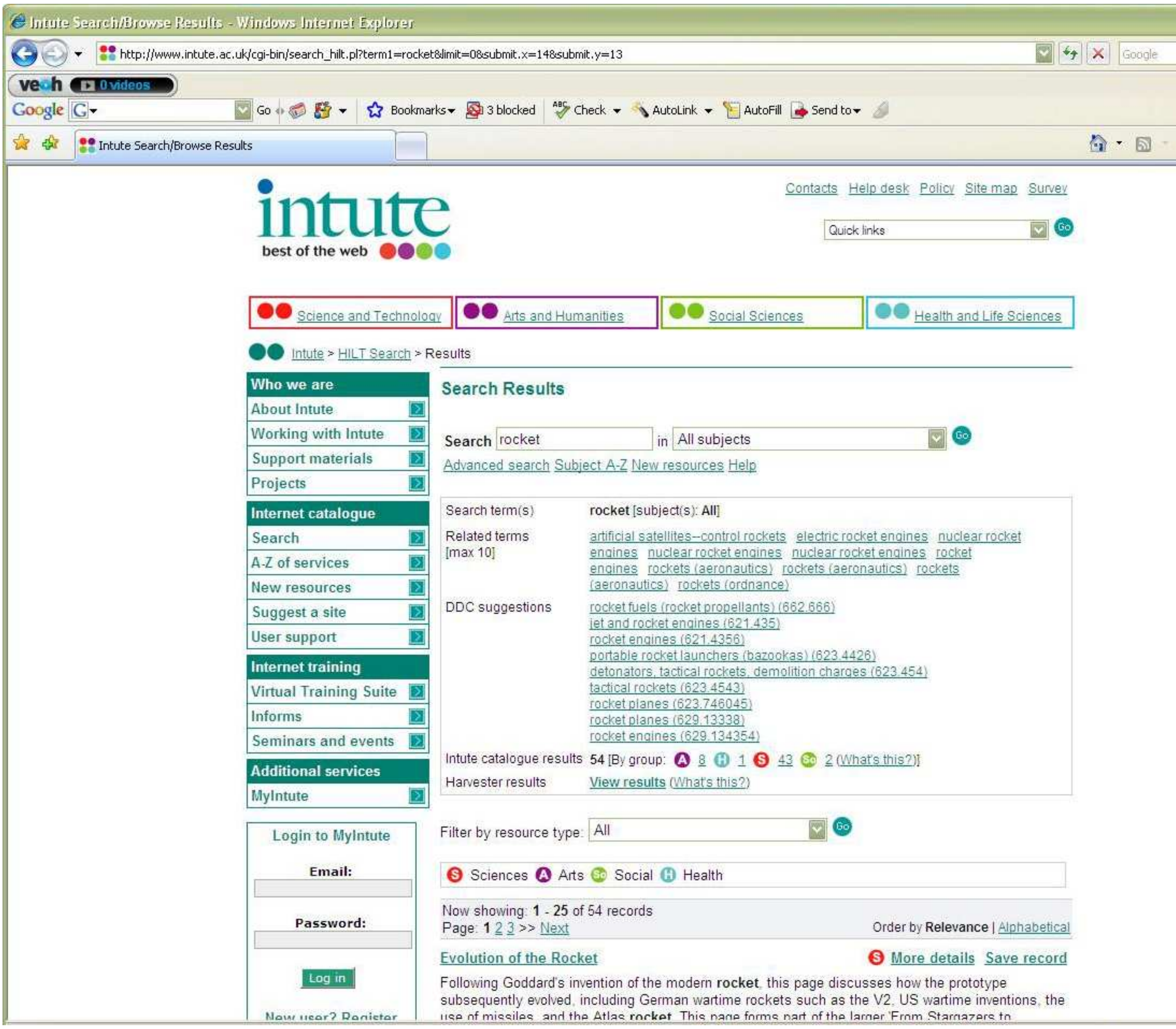
Do you mean one of these?

- foot
- fos
- foy
- fog
- flo

< Previous | Save for later | Next >
» Contacting the Depot and Acknowledgements.
v1.7
EDINA and SHERPA
JISC Repository Net

2) Intute: The following screen shots demonstrate how embedding HILT terminology services can enhance the user experience in aspects of the Intute service. Two workpackages were implemented:

> WP1: Demonstrate HILT's ability to support search functions by offering users narrower terms / broader terms / synonyms from all HILT vocabularies after a search has been carried out across the Intute service. 'Related terms' and 'DDC suggestions' from HILT are shown in the result screen below:



Search Results

Search rocket in All subjects Go

Advanced search Subject A-Z New resources Help

Search term(s)	**rocket** [subject(s): **All**]
Related terms [max 10]	artificial satellites--control rockets electric rocket engines nuclear rocket engines nuclear rocket engines nuclear rocket engines rocket engines rockets (aeronautics) rockets (aeronautics) rockets (aeronautics) rockets (ordnance)
DDC suggestions	rocket fuels (rocket propellants) (662.666) jet and rocket engines (621.435) rocket engines (621.4356) portable rocket launchers (bazookas) (623.4426) detonators, tactical rockets, demolition charges (623.454) tactical rockets (623.4543) rocket planes (623.746045) rocket planes (629.13338) rocket engines (629.134354)
Intute catalogue results	54 [By group: A 8 H 1 S 43 So 2 (What's this?)]
Harvester results	**View results** (What's this?)

Filter by resource type: All Go

S Sciences A Arts So Social H Health

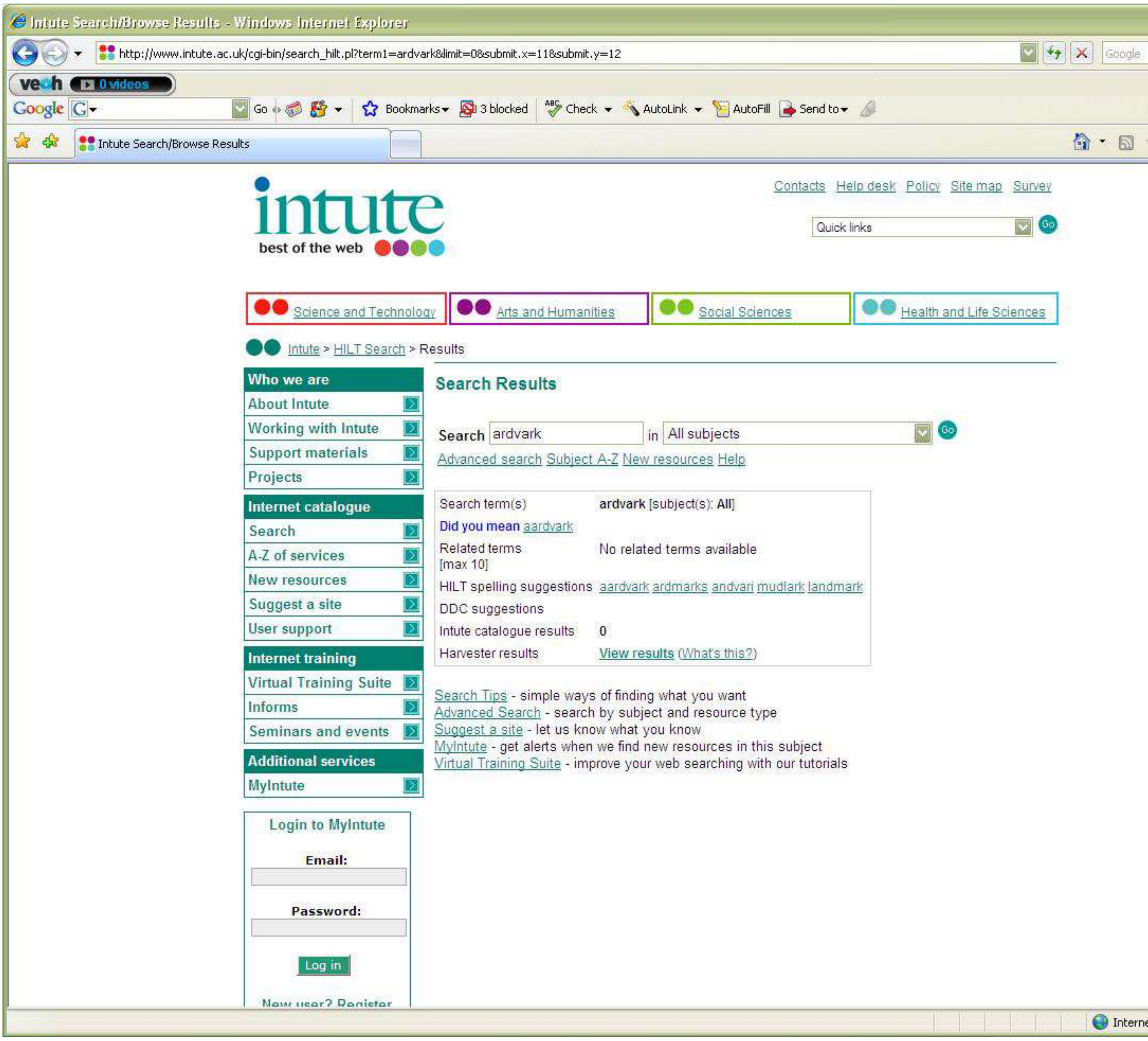
Now showing: **1 - 25** of 54 records
Page: **1** 2 3 >> Next

Order by **Relevance** | Alphabetical

Evolution of the Rocket S **More details** **Save record**

Following Goddard's invention of the modern **rocket**, this page discusses how the prototype subsequently evolved, including German wartime rockets such as the V2, US wartime inventions, the use of missiles, and the Atlas **rocket**. This page forms part of the larger 'From Stargazers to

WP2: Use of the spell-checker within Intute. This will be invoked in the event that a user receives no hits. So, for example, when no hits are returned for ‘ardvark’, the system will suggest searching for ‘aardvark’, along with other possible alternatives:


Intute Search/Browse Results - Windows Internet Explorer
http://www.intute.ac.uk/cgi-bin/search_hilt.pl?term1=ardvark&limit=0&submit.x=11&submit.y=12
Google
veoh 0 videos
Google G Go Bookmarks 3 blocked Check AutoLink AutoFill Send to
Intute Search/Browse Results
Internet

intute
best of the web

Contacts Help desk Policy Site map Survey

Quick links Go

Science and Technology | Arts and Humanities | Social Sciences | Health and Life Sciences

Intute > HILT Search > Results

Who we are
- About Intute
- Working with Intute
- Support materials
- Projects

Internet catalogue
- Search
- A-Z of services
- New resources
- Suggest a site
- User support

Internet training
- Virtual Training Suite
- Informs
- Seminars and events

Additional services
- MyIntute

Login to MyIntute

Email:

Password:

Log in

New user? Register

Search Results

Search ardvark in All subjects Go

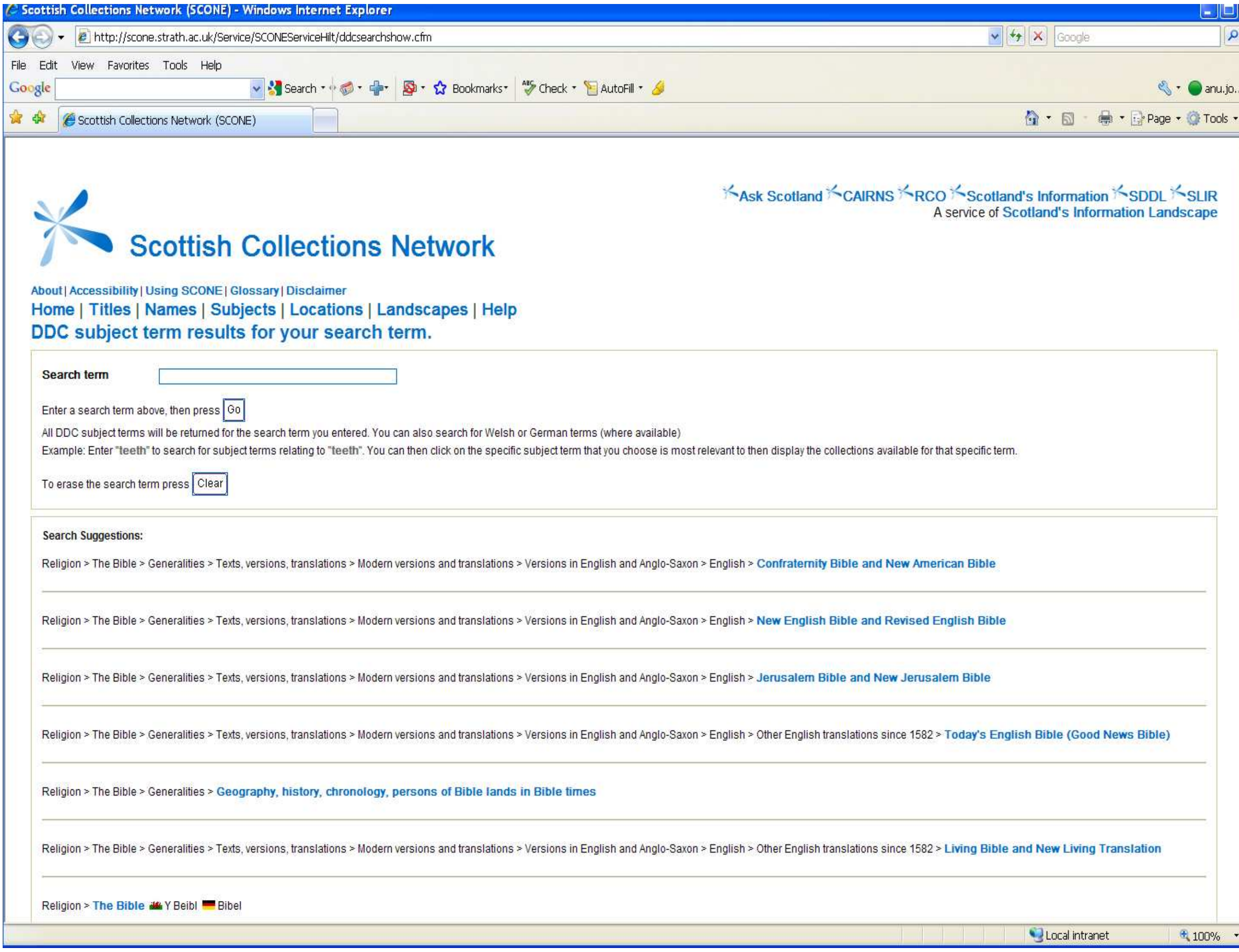
Advanced search Subject A-Z New resources Help

Search term(s)	**ardvark** [subject(s): **All**]
Did you mean aardvark	
Related terms [max 10]	No related terms available
HILT spelling suggestions	aardvark ardmarks andvari mudlark landmark
DDC suggestions	
Intute catalogue results	0
Harvester results	**View results** (What's this?)

Search Tips - simple ways of finding what you want
Advanced Search - search by subject and resource type
Suggest a site - let us know what you know
MyIntute - get alerts when we find new resources in this subject
Virtual Training Suite - improve your web searching with our tutorials

3) CAIRNS: The following screen shots demonstrate how embedding HILT terminology services can enhance the user experience in Scotland-wide services such as the CAIRNS (http://cairns.lib.strath.ac.uk/) distributed catalogue and the SCONE (http://scone.strath.ac.uk/Service/Index.cfm) collections database. Two workpackages were implemented:

WP1(a): Returning the DDC caption and number for a user term(s) input (using the HILT function get_ddc_records for term 'Bible'). You can also see Welsh and German captions corresponding to the English term (where available). See the last item in the screen shot.


Scottish Collections Network (SCONE) - Windows Internet Explorer
http://scone.strath.ac.uk/Service/SCONEServiceHilt/ddcsearchshow.cfm
File Edit View Favorites Tools Help
Google
Search
Bookmarks
Check
AutoFill
anu.jo.
Scottish Collections Network (SCONE)
Page
Tools
Ask Scotland CAIRNS RCO Scotland's Information SDDL SLIR
A service of Scotland's Information Landscape
Scottish Collections Network
About | Accessibility | Using SCONE | Glossary | Disclaimer
Home | Titles | Names | Subjects | Locations | Landscapes | Help
Local intranet
100%

DDC subject term results for your search term.

Search term

Enter a search term above, then press Go

All DDC subject terms will be returned for the search term you entered. You can also search for Welsh or German terms (where available)

Example: Enter "teeth" to search for subject terms relating to "teeth". You can then click on the specific subject term that you choose is most relevant to then display the collections available for that specific term.

To erase the search term press Clear

Search Suggestions:

Religion > The Bible > Generalities > Texts, versions, translations > Modern versions and translations > Versions in English and Anglo-Saxon > English > Confraternity Bible and New American Bible

Religion > The Bible > Generalities > Texts, versions, translations > Modern versions and translations > Versions in English and Anglo-Saxon > English > New English Bible and Revised English Bible

Religion > The Bible > Generalities > Texts, versions, translations > Modern versions and translations > Versions in English and Anglo-Saxon > English > Jerusalem Bible and New Jerusalem Bible

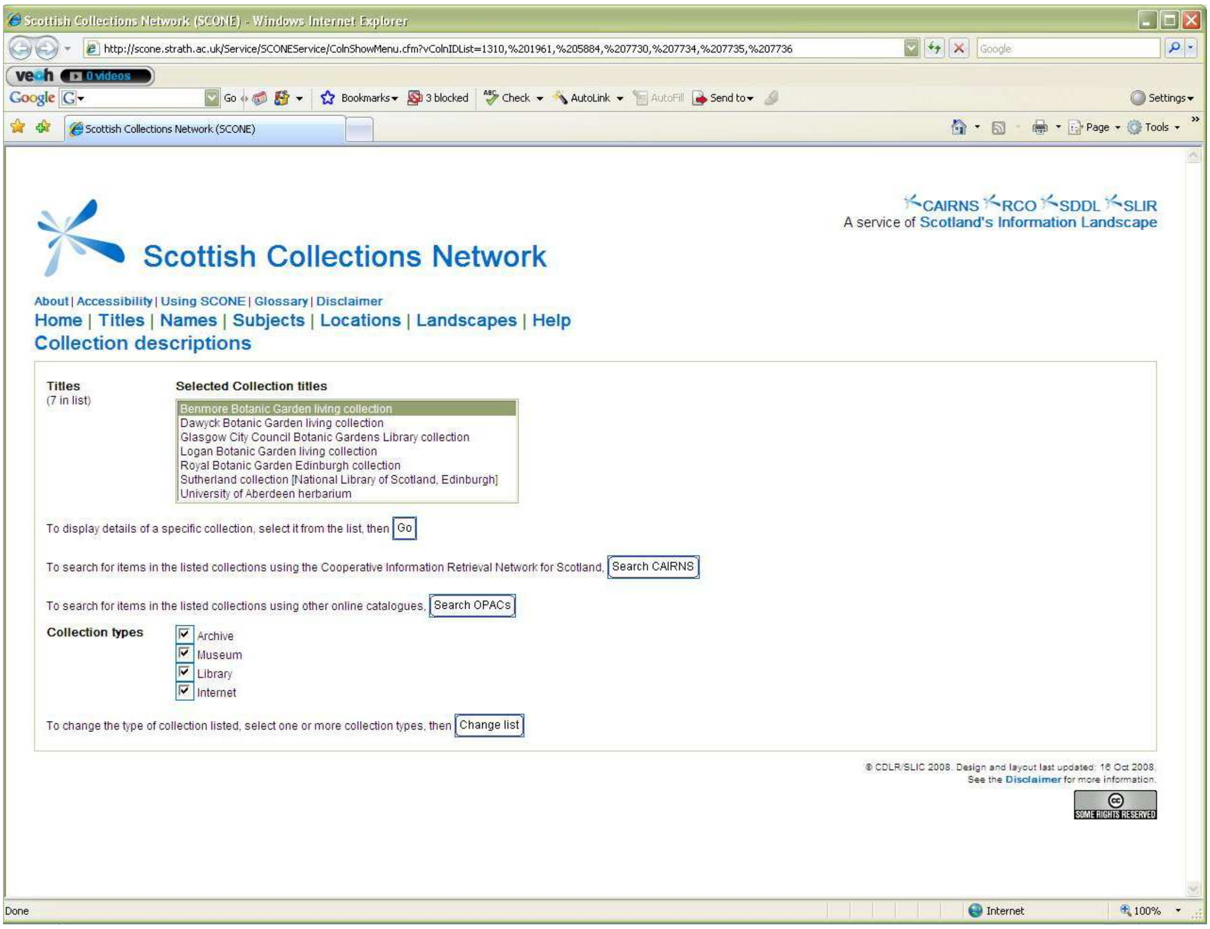
Religion > The Bible > Generalities > Texts, versions, translations > Modern versions and translations > Versions in English and Anglo-Saxon > English > Other English translations since 1582 > Today's English Bible (Good News Bible)

Religion > The Bible > Generalities > Geography, history, chronology, persons of Bible lands in Bible times

Religion > The Bible > Generalities > Texts, versions, translations > Modern versions and translations > Versions in English and Anglo-Saxon > English > Other English translations since 1582 > Living Bible and New Living Translation

Religion > The Bible Y Beibl Bibel

WP1(b): Retrieve SCONE collections using this DDC number:


Scottish Collections Network (SCONE) - Windows Internet Explorer
http://scone.strath.ac.uk/Service/SCONEService/ColnShowMenu.cfm?vColnIDList=1310,%201961,%205884,%207730,%207734,%207735,%207736
Google
Go
Bookmarks
3 blocked
Check
AutoLink
AutoFill
Send to
Settings
Scottish Collections Network (SCONE)
Page
Tools
Done
Internet
100%

Scottish Collections Network

CAIRNS RCO SDDL SLIR
A service of Scotland's Information Landscape

About | Accessibility | Using SCONE | Glossary | Disclaimer
Home | Titles | Names | Subjects | Locations | Landscapes | Help

Collection descriptions

Titles
(7 in list)

Selected Collection titles

Benmore Botanic Garden living collection
Dawyck Botanic Garden living collection
Glasgow City Council Botanic Gardens Library collection
Logan Botanic Garden living collection
Royal Botanic Garden Edinburgh collection
Sutherland collection [National Library of Scotland, Edinburgh]
University of Aberdeen herbarium

To display details of a specific collection, select it from the list, then Go

To search for items in the listed collections using the Cooperative Information Retrieval Network for Scotland, Search CAIRNS

To search for items in the listed collections using other online catalogues, Search OPACs

Collection types

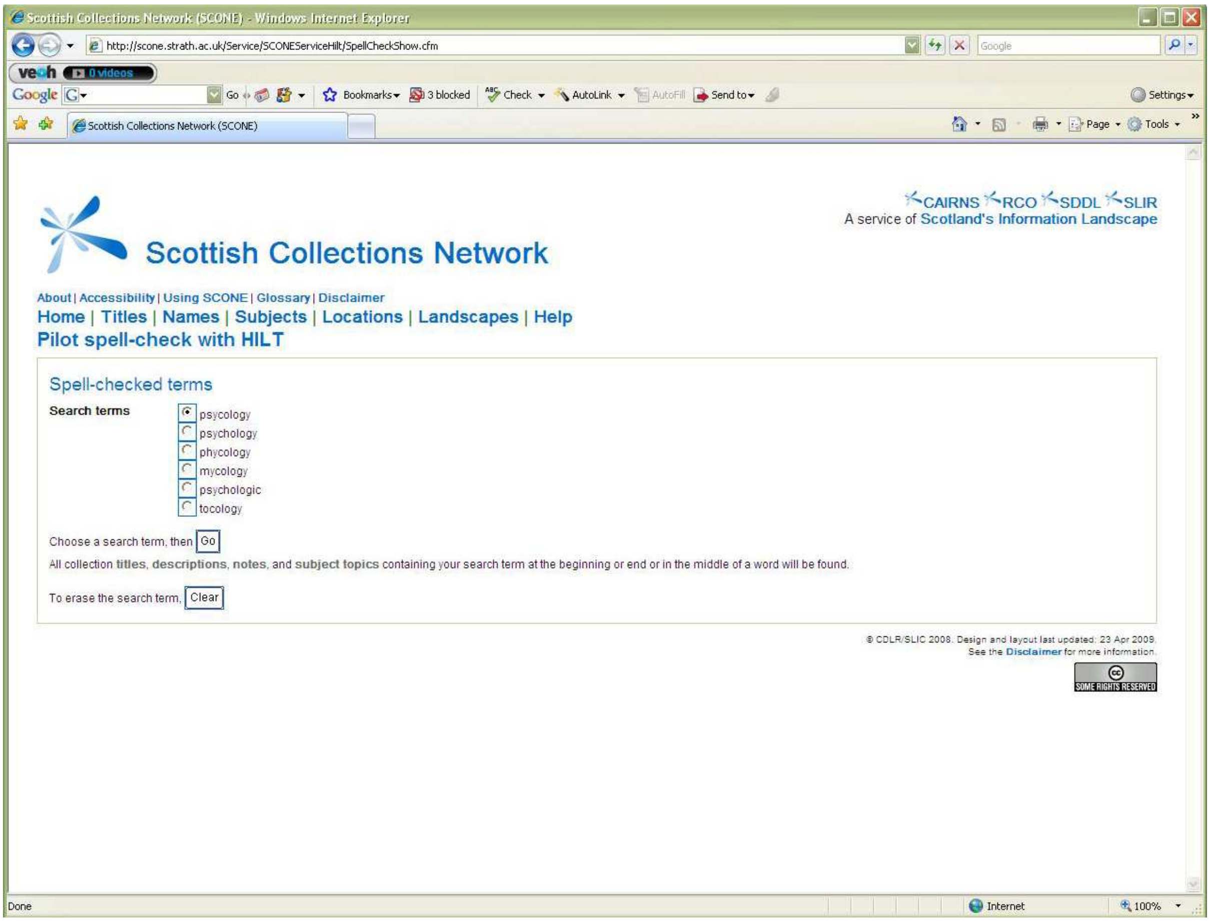
☑ Archive
☑ Museum
☑ Library
☑ Internet

To change the type of collection listed, select one or more collection types, then Change list

© CDLR/SLIC 2008. Design and layout last updated: 16 Oct 2008.
See the Disclaimer for more information.

SOME RIGHTS RESERVED

WP2: Retrieve suggested alternate spellings, similar to the user term(s) input:


Scottish Collections Network (SCONE) - Windows Internet Explorer
http://scone.strath.ac.uk/Service/SCONEServiceHilt/SpellCheckShow.cfm
Google
veoh 0 videos
Google G
Go
Bookmarks
3 blocked
Check
AutoLink
AutoFill
Send to
Settings
Scottish Collections Network (SCONE)
Page
Tools
CAIRNS RCO SDDL SLIR
A service of Scotland's Information Landscape
Scottish Collections Network
About | Accessibility | Using SCONE | Glossary | Disclaimer
Home | Titles | Names | Subjects | Locations | Landscapes | Help
Pilot spell-check with HILT
Go
Clear
SOME RIGHTS RESERVED
Done
Internet
100%

Spell-checked terms

Search terms

● psycology
○ psychology
○ phycology
○ mycology
○ psychologic
○ tocology

Choose a search term, then Go

All collection titles, descriptions, notes, and subject topics containing your search term at the beginning or end or in the middle of a word will be found.

To erase the search term, Clear

© CDLR/SLIC 2008. Design and layout last updated: 23 Apr 2009.
See the Disclaimer for more information.

Appendix C: HILT and other Projects

Gold Dust

HILT had a small involvement in the JISC funded Gold Dust project. Gold Dust's aims and objectives were:

- To explore the potential effectiveness of a recommender system based upon pervasively acquired user data;
- To explore the potential effectiveness of information obtained from RSS feeds for providing data to create PIPs;
- To examine the potential of various text-mining (and other) techniques to generate PIPs;
- To determine how well text-mining software performs the tasks of extracting terms to comprise the user's PIPs;
- In the light of user feedback, to examine the relevance and usefulness of items extracted from the Gold Dust database of current information.

  The project looked into the use of HILT's terminological services to assist the Gold Dust processes of providing go-lists, stop-lists and term-expansion for more effective PIP creation and matching. The results tended to suggest that the schemes in HILT were too high level to be useful in the particular context explored by Gold Dust.

  Further information on this work is provided in the Research Report in Appendix F below.

ERIS

The purpose of the ERIS project is to develop – in close partnership with researchers and their institutions' repository managers – a set of user-led and user-centric solutions that will motivate researchers to deposit their work in repositories, facilitate the integration of repositories in research and institutional processes and, as a result, develop the IRIScotland pilot into a trusted cross-repository resource discovery service, capable of providing access to a critical mass of Scottish research output. In order to achieve this overall aim, ERIS will pay particular attention to the requirements of research pooling – an innovative cross-institutional way of conducting research, which has been widely credited for having substantially contributed to Scotland's RAE 2008 successes.

One of the things the project will explore are the possibilities of *Subject access enhancements* – improving resource discovery and increasing the ability of institutions to showcase their subject strengths; this will involve machine-to-machine interaction with the HILT17 web services and the integration into DSpace, EPrints and Fedora of HILT-driven drop-down menus to assist in the subject cataloguing of resources, inter-subject scheme interoperability and enhanced subject retrieval through term expansion.

SLIC

The Scottish Library and Information Council (SLIC), an independent advisory body to the Scottish Executive, has been collaborating with the Centre for Digital Library Research (CDLR), based at the University of Strathclyde, since 2002 pioneering the development of a "Scottish Information Landscape" (SIL). This "landscape" encompasses a portfolio of online services useful to academic researchers, ordinary members of the public and those working within the library and information profession. These services are diverse but inter-linked, allowing libraries and others to share common data. SLIC recognises this group of services and embryonic services as including CAIRNS, SCONE, RCO, SDDL, SLAINTE, BUBL, SLIR, and, most recently, HILT.

Appendix D: Work with HILT-Collaborators

HILT IV carried out some survey and other work with a number of individuals who joined a HILT email list (HILT-Collaborators@jiscmail.ac.uk) in response to a call for service staff who were potentially interested in using HILT pilot facilities. This work showed that there was a good deal of interest from the community in the kind of uses of HILT listed above. Twenty-nine people joined the list which was set up specifically to look at embedding work of this kind (JISC staff indicated that 29 participants was a comparatively good response in comparison with similar exercises in other projects). The remainder of this appendix sketches out the kind of work done and the largely positive outcomes in respect of the interest the participants showed in using HILT and on the kinds of things they were interested in using it for..

Text Announcing the HILT-Collaborators List and the Engagement Exercise

The following was sent out to a variety of e-journals and e-lists asking for information services staff interested in using and testing HILT terminology services to get in touch with HILT staff if they were willing to work with us to look at some embedding practicalities.

~~~

Help us make HILT's terminology services useful in your information service

The JISC[38[1]]-funded HILT[39[2]] project is looking to make contact with staff in information services or projects interested in helping it test and refine its developing terminology services. The project is currently working to create pilot web services that will deliver machine-readable terminology and cross-terminology mappings data likely to be useful to information services wishing to extend or enhance the efficacy of their subject search or browse services. Based on SRW/U[40[3]], SOAP[41[4]], and SKOS[42[5]], the HILT facilities, when fully operational, will permit such services to improve their own subject search and browse mechanisms by using HILT data in a fashion transparent to their users. On request, HILT will serve up machine-processable data on individual subject schemes (broader terms, narrower terms, hierarchy information, preferred and non-preferred terms, and so on) and interoperability data (usually intellectual or automated mappings between schemes, but the architecture allows for the use of other methods) – data that can be used to enhance user services. The project is also developing an associated toolkit that will help service technical staff to embed HILT-related functionality into their services. The primary aim is to serve JISC funded information services or services at JISC institutions, but information services outside the JISC domain may also find the proposed services useful and wish to participate in the test and refine process.

Although the primary focus of the work is to improve interoperability during cross-search or browse by subject, the facilities offered can also be used for other purposes. Examples of possible uses include:

a Providing the best terms for a subject search in a remote service that uses a subject scheme unfamiliar to 'home service' users. HILT currently has the following KOS[43[6]] mounted and available: AAT, CAB, GCMD, HASSET, IPSV, LCSH, MeSH, NMR, SCAS, UNESCO, and DDC.
b Improving recall in a subject search of one or more databases by enriching the set of terms known to a user by providing synonyms and related terms.
c Generating an interactive browse structure where a scheme is arranged hierarchically.
d Taking a user's subject term and using it to identify available information services with subject coverage relevant to the query via collections and/or services databases such as IESR[44[7]] and SCONE[45[8]]

[38[1]] See http://www.jisc.ac.uk/
[39[2]] See http://hilt.cdlr.strath.ac.uk/ (HILT in general) and http://hilt.cdlr.strath.ac.uk/hilt4/index.html (HILT Phase IV, the current phase)
[40[3]] For an explanation, see http://www.oclc.org/research/projects/webservices/default.htm
[41[4]] http://www.w3.org/TR/soap12-part1/
[42[5]] http://www.w3.org/2004/02/skos/
[43[6]] Knowledge Organisation Systems – see, for example http://www.db.dk/bh/lifeboat_ko/CONCEPTS/knowledge_organization_systems.htm for further information.
[44[7]] http://iesr.ac.uk/
[45[8]] http://scone.strath.ac.uk/service/
~~~

The project is also looking to test other associated facilities it intends to offer for embedding in JISC or institutional information services – for example a spell-check mechanism and machine to machine delivery of Wordnet[46[9]] data.

The test and refine process is likely to begin towards the end of March 2008 and continue for at least six months beyond that. Individuals or services interested in participating, should begin by joining the HILT-Collaborators email list at http://www.jiscmail.ac.uk/cgi-bin/webadmin?SUBED1=hilt-collaborators&A=1

Note that, at this stage, both the facilities and the subject schemes are only being made available for testing purposes – to allow services to help us test and refine them (and, in time, evaluate their usefulness). They cannot and should not be built into operational services.

HILT Contacts

Emma McCulloch, Project Manager, e.mcculloch@strath.ac.uk
Anu Joseph, Programmer, anu.joseph@strath.ac.uk
Dennis Nicholson, Project Director, d.m.nicholson@strath.ac.uk

[46[9]] http://wordnet.princeton.edu/

Questionnaire and results

Questionnaire

Dear All

Thank you for joining the HILT-Collaborators list and for your patience in waiting for developments. This email is the first in a series you'll be receiving over the next few months – probably at a level of about one a week. Via these, we will aim to introduce you to some of the pilot web services HILT Phase IV is working on (and looking for your help in testing them). Initially, this will be done by demonstrating (a) functions available to retrieve useful terminological and interoperability data from HILT, (b) web-page based demos of some of the ways these functions can be used to build useful functionality within your own institutional or other services. Ultimately, though, we'd like to move beyond demonstration and get some or all of you experimenting with creating simple web pages of your own and using HILT functions within them to actually retrieve and display or otherwise use terminological and interoperability data using the pilot web services we've built. How far we can go with this last part of the process will depend to some extent on the skills each of you already has in this area, but we hope we can get all of you who wish to do this involved in at least a simple practical exercise of this kind.

Because our aim in doing this is to get you to help us test and refine both the pilot web services themselves and the ideas that underpin the approach, we will be asking for (and hoping to get) feedback from you on your experiences. This, we hope, will sometimes consist of informal reactions to what you experience, either in the form of comments and suggestions, or simply in the form of questions. However, we will also ask you to occasionally respond to a short questionnaire, the first of which you will find below.

The next email in this series will arrive in about a week and will begin the process of taking you through some of the HILT functions. The present email was just to tell you a bit about what will happen and to get you thinking about the issues we'll be looking at by responding to the short questionnaire below. Because the original call for volunteers contains background information you may find useful, I have attached it to this message for information. It may be helpful to read it before answering the questionnaire.

You can answer the questionnaire very quickly and simply. Just reply to this message and put either ALL, NONE, or the letters of the functions you think would be useful (e.g. if you think A and F would be useful but not the others, just put 'A and F' in your reply. You are also welcome to add any comments on any possible use you can think of we have not listed.

Questionnaire

HILT is building a pilot Shared Infrastructure Service (SIS) for the JISC Information Environment. The aim is to allow national, institutional, and other information service providers to access terminological and interoperability data at a machine to machine level that will allow them to enhance their own services in a variety of ways, including:

e Improving recall in a subject search of one or more databases by enriching the set of terms known to a user by providing synonyms and related terms.

f Providing the best terms for a subject search in a remote service that uses a subject scheme unfamiliar to 'home service' users (or in a cross-search of group of such services).

g Taking a user's subject term and using it to identify available information services with subject coverage relevant to the query via registries such as IESR (http://iesr.ac.uk/)

h Generating an interactive browse structure where a scheme is arranged hierarchically.

i The simple ability to send a term from a chosen subject scheme and receive back data on broader terms, narrower terms, hierarchy information, preferred and non-preferred terms, and so on

j Providing cataloguing staff with information on subject schemes and inter-scheme mappings to assist in metadata creation

k A spell check mechanism to assist user searching l A service to assist user search formulation by providing information on search terms entered (e.g. what the term means, whether it has alternative meanings, whether there are synonyms that might be useful in a search and so on)

Which of these do you think would be useful in services you are involved with or use?
That's all for now. Thanks again for participating in this collaboration.

Results

Sixteen out of the twenty-nine people on the list answered the questionnaire. All indicated an interest in at least one of the above potential ways of using HILT, most indicated an interest in three or more ways. The original emails are available if anyone wishes to consult them.

Additional Instructional Messages Sent Out

HILT Collaborators: Web services try out #1 (get_ddc_records SOAP Server Function)

As indicated in my previous message, what we plan to do initially is show you demonstrations of functions available to retrieve useful terminological and interoperability data from HILT. Ultimately, we hope to let you test drive these by embedding them in test web pages of your own and using them as true web-services. In the first instance, though, the plan is simply to let you see them in action and play with them via a web page based at HILT. You should use a recent version of Internet Explorer to do this. There are one or two little glitches in the demos we have not yet ironed out for the likes of Firefox and Opera.

The demo is here: http://hiltm2m.cdlr.strath.ac.uk/hilt4/hiltsoapclient.php. Click on the link, and try the following:

Click on the highlighted explain file link and scroll down to see a (currently very rudimentary) explain file. All responses to requests sent from this page appear in the lower half of the screen in black and grey. In some cases, it is not especially evident that the screen has changed, so make sure you know what was there before you sent the request. The explain file currently tells you what functions are available for requesting data from our pilot service and a little bit about what parameters they take. If you look carefully and compare the data in black and grey with the drop down options on the main screen, you'll see most of the functions are testable via the drop down lists, but two (get_parents and get_children) are not. These are available for use in machine to machine interaction between your home service and HILT but are not in the demonstration as yet.

Now click on the empty box next to Spelling Suggestions, type the string 'aerth' (without the quotes) into to the upper search box, choose get_ddc_records from the top drop down list, and click on the top search button to its right. Scroll down again to the lower half of the screen to see the result. HILT has no data to send you on 'aerth', but has suggested that you might have meant to type in 'earth'. Click on the highlighted earth and the lower half of the screen will change. You have retrieved Dewey Decimal Classification records from HILT indexed under 'earth'. You should get the same result if you go back to the search box at the top of the page, type in 'earth', choose get_ddc_records from the drop down list and hit the top search button. Try it!

Have a look at the records. You will see human-readable data such as a DDC number – e.g. 525 – and the DDC caption associated with the number: Earth (Astronomical geography). You will also see what looks like gobbledygook wrapped around the human readable stuff, like this - <skos:altLabel xml:lang="en">Earth (Astronomical geography)</skos:altLabel>. This is SKOS markup. It helps make HILT an M2M service by wrapping the human readable data in 'labels' that allows some computer code embedded in the remote service that receives the data from HILT to handle the human readable stuff intelligently. For example, it allows the computer code to distinguish between the DDC number 525 and the associated DDC caption Earth (Astronomical geography). There is no need to learn what it all is or means for the purposes of our tests. Whenever you need to know, we'll tell you.

Try playing around with the get_ddc_records function by typing in different words in the search box, selecting get_ddc_records function from the drop down list, and hitting search. Try it with and without spelling suggestions on and see what kinds of data you can retrieve. Please note that the data on DDC is the intellectual property of OCLC and is only made available for the purposes of testing the HILT pilot.

That is all for this week. Feel free to play with the other functions if you wish, but there are various things we will need to explain about these, so it may be better for you to wait till we move to the next function in the series in about a week from now.

Please let us know of any thoughts, comments, or questions you have once you have tried the above. Send these direct to us except where you would like others to see your thoughts or react to them.

HILT Collaborators: Web services try out #2 (get_filtered_set SOAP Server Function)

As indicated in my previous messages, what we plan to do initially is show you demonstrations of functions available to retrieve useful terminological and interoperability data from HILT. Ultimately, we hope to let you test drive these by embedding them in test web pages of your own and using them as true web-services. In the first instance, though, the plan is simply to let you see them in action and play with them via a web page based at HILT. You should use a recent version of Internet Explorer to do this. There are one or two little glitches in the demos we have not yet ironed out for the likes of Firefox and Opera.

The demo page for this is: http://hiltm2m.cdlr.strath.ac.uk/hilt4/hiltsoapclient.php (the same as for the last try out). Click on the link, and try the following:

Type the string 'asbestos' (without the quotes) into to the upper search box, choose get_filtered_set from the top drop down list, highlight IPSV in the list of schemes, and click on the top right search button. Scroll down to the lower half of the screen to see the result. Try the same with DDC, MeSH, LCSH and a combination of one or more of these (use CTRL click to highlight more than one scheme at a time).

Now try the same thing with other terms and other schemes or combinations of schemes.

Now try this. Type the string 'asbos' (without the quotes) into to the upper search box, choose get_filtered_set from the top drop down list, highlight IPSV in the list of schemes, and click on the top right search button. Scroll down to the lower half of the screen to see the result. You should be able to spot a term 'Antisocial behaviour and disorder' a few lines down. Now repeat the process but this time, in addition to doing all of the above, click OFF the checkbox that says 'Non Preferred Terms'. What happens? Now repeat the process again with 'Non Preferred Terms' clicked on but the other two boxes off. You should get the same list as with the first search of 'asbos'. This is because the preferred term:
<skos:prefLabel xml:lang="en">Antisocial behaviour and disorder</skos:prefLabel> is being retrieved via the non preferred term 'asbos'. If you look down the list under the preferred term, you'll spot:
<skos:altLabel xml:lang="en">ASBOs (antisocial behaviour orders)</skos:altLabel> You may also be able to work out how the SKOS 'wrapping' is distinguishing between preferred and non-preferred terms.

The get_filtered_set function is a powerful tool for requesting and retrieving data from HILT. If you try one of the asbos examples above and copy and paste the URL that results in to URL 'slot' at the top of your browser, you can begin to see how you might use it to query HILT via a web page in your own service. Try to get a feel for how different terms, schemes or scheme combinations, and combinations of preferred, non preferred and related term settings change the URL. For today's purposes, you can ignore everything to the left of the ? – this bit: http://hiltm2m.cdlr.strath.ac.uk/hilt4/hiltsoapclient.php and focus on this bit: term=vat&request=get_filtered_set&scheme%5B%5D=DDC&PT=true&NPT=true&RT=true. Look at the bits to the right of the sections highlighted in bold to work out how the function is used.

That is all for this week. Once again, feel free to play with the other functions if you wish, but there are various things we will need to explain about these, so it may be better for you to wait till we move to the next function in the series in about a week from now.

Please let us know when you've done this try out and let us know of any thoughts, comments, or questions you may have. Send these direct to me except where you would like others to see your thoughts or react to them.

I'm trying to strike a balance between scaring people off with too much gobbledygook and annoying the techies amongst you who may find my explanations a bit laboured. How am I doing?

HILT Collaborators: Web services try out #3 (get_collections function via SOAP and SRU)

As indicated in my previous messages, what we plan to do initially is show you demonstrations of functions available to retrieve useful terminological and interoperability data from HILT. Ultimately, we hope to let you test drive these by embedding them in test web pages of your own and using them as true web-services. In the first instance, though, the plan is simply to let you see them in action and play with them via a web page based at HILT. You should use a recent version of Internet Explorer to do this. There are one or two little glitches in the demos we have not yet ironed out for the likes of Firefox and Opera.

get_collections via SOAP

The demo page for this is: http://hiltm2m.cdlr.strath.ac.uk/hilt4/hiltsoapclient.php (the same as for the last two try outs). Click on the link, and try the following:

Scroll to the bottom of the page, go to the drop down list to the right hand side and select the get_collections function. In the search box to the left (still at the bottom of the page), and try typing in 510 and clicking on the search button. Scroll down to the grey area of the page. You will see that you have retrieved data on services whose subject coverage either *is* mathematics (510 is the Dewey decimal Classification system number for mathematics), or includes mathematics. Try it again with some other numbers: 910, 820, 120, 650, 330. Work out from the output what subjects are associated with these DDC numbers.

get_collections via SRU

Up till now, we have been looking at how to call the various HILT functions on our SOAP server. If you were actually calling HILT from within the web pages of your own local service, you'd probably be using a protocol called SRU and would send a get_collections request for 510 as follows:

http://bodach.ucs.ed.ac.uk:18113/hilt?version=1.1&operation=searchRetrieve&maximumRecords=100&query=hilt.get_collections+%3D+510

Click on the link and see what you get back. Try altering the URL to give you data on services whose subject coverage is described by DDC numbers 910, 820, 120, 650, and 330.

That's all for this time. *As ever, let us know how you got on.* The next email will be sent two weeks from today and will allow you to look more closely at the instructions you'd embed in your own service web pages if you were using the various HILT functions to retrieve terminological and mapping data from HILT. We'll then move on in subsequent weeks to focus on taking a specific SRU request for terminological data from HILT, looking at the machine readable output, showing how this might be displayed helpfully, and showing you an example of the 'parsing' script needed to take the M2M output and transform it into the human readable output. Once we've done that, you'll be in a position to:

1. Embed the SRU instruction and the parsing and display of the output in a webpage of your own (although the extent to which you will be able to do this will depend on the technical skills available to you at wherever you are based).
2. Change the instruction and the script to use and parse other SRU instructions (again, depending on the skills available to you locally)

Through all of this, we hope you will continue to give us feedback. One of the things we hope to be able to do at the end of this is to advise JISC on what level of support interested parties like yourselves will need from them to be able to successfully utilise HILT services to enhance services of your own.

HILT Collaborators: Web services try out #4 (embedding HILT interactions into your own service)

For a variety of reasons the focus of this trial will not be as advertised last time. Instead, we'll jump forward a little to: *taking a specific SRU request for terminological data from HILT, looking at the machine readable output, showing how this might be displayed helpfully, and showing you an example of the 'parsing' script needed to take the M2M output and transform it into the human readable output.*

The extent to which you will be able to complete this try out will depend on your local set-up and/or expertise. If you have access to a web server running PHP, you'll be able to do something close to an actual embedding of the script attached to this week's message (*get_filtered.php*) into your own local servcie and will know how to take it further if you wish to. If you don't have such access, or don't know whether or not you have, the best you'll be able to do is see something similar to what would have happened on your own web server by looking at a demonstration we've put on the HILT wiki. If you have access to a web server, but are using a scripting language other than PHP (Perl, say), you won't be able to go the whole way with what we've sent this week, but please get in touch and we'll try to accommodate your local needs.

If you do have access to a web server running PHP locally, put the file *get_filtered.php* that came with this week's email on the web server. Note the URL for the web page on your web server then point Internet Explorer browser at that URL. You should see something very similar to what you will see if you point your web browser at the HILT URL: http://hiltm2m.cdlr.strath.ac.uk/hilt4/toolkit/get_filtered.php. If you don't have access to a local web server running PHP, pointing your web browser at the HILT URL will show you what you would have got if you did. Either way, you can see how this result was achieved by saving *get_filtered.php* to your desktop and opening it with an editor such as Notepad or Wordpad. If you then search for the string *http*, you'll find the first of two SRU requests the script sends to HILT. A further search will find the second of these. These parts of the script interact with HILT to obtain terminological data regarding preferred, related, and non-preferred terms from the HASSET subject scheme on the word 'weeds' using the HILT get_filtered_set function:

http://bodach.ucs.ed.ac.uk:18113/hilt?version=1.1&operation=searchRetrieve&maximumRecords=100&query=hilt.scheme+%3D+HASSET%20and%20hilt.get_filtered_set+%3D+weeds%20and%20hilt.preferred+%3D+true%20and%20hilt.related+%3D+true%20and%20hilt.non_preferred+%3D+true%20and%20hilt.is_id+%3D+false

If you paste the above URL into your browser, you should see what HILT returns in response to this SRU request.

Roughly speaking, the remainder of the script takes the results you get when the SRU requests are sent to HILT and 'parses' the results to organize and display them as they are seen when you point your browser at either this URL: http://hiltm2m.cdlr.strath.ac.uk/hilt4/toolkit/get_filtered.php or the URL you get when the web page sent with this week's email is put on your own web server.

If you do have your own web server running PHP, you can try editing the URLs in the script we provided to call terms other than 'weeds' from HASSET. For example, try 'mother' or 'asbo'. Even if you don't have your own web server running PHP, you should be able to work out how you would go about doing this if you did. Either way, you should now know a little bit about how you might begin to use HILT to enhance local services.

In time – and probably with help from the toolkit HILT will eventually release - you could use methods like these to build enhanced capabilities into your own favourite information service. The demonstrator at: http://bubl.ac.uk/hilt4.htm shows one example. Type 'weeds' into the search box to get hits in the local database and other word choices. Then on the results page, click on 'Search beyond BUBL' and choose result 4 from the following page to find an Intute service, then click on the highlighted term 'weeds' under the Intute entry to carry out a search on Intute using the correct term and get appropriate results. Remember that only Internet Explorer gives the best results at the moment.

That's all for this time. *As ever, let us know how you got on.*

HILT Collaborators: Web services 'try out' #5: Tell us what you think

HILT has now completed the program of trials related to the HILT-Collaborators list. This is the final message in the programme we've been running over the last few weeks – and it's not really a try out this time, just another wee survey.

We'd be very grateful if you'd reply to this email giving answers to the following questions:

On a scale of 0 to 3, where 0 is not at all and 3 extremely

How interesting was this exercise for you?
How active were you in carring out the try-outs?
How active were you in providing feedback?
How important is it to you/your service that a future HILT service offers a high level of support and training for those wishing to enhance local functionality by interacting with HILT?

Please also feel free to send us any other comments that occur to you.

A couple of other points. First and foremost, thank you all very much for participating – HILT has found both the exercise itself and your feedback very useful. Second, whilst we have finsihed this programme as such, we'd be interested to hear from anyone who'd like to try to work with us further in this area on a one to one basis. If we get an avalanche of interest, we may find it difficult to resource this. My guess is, though, that we'll only get a few ;-)

We'd like to keep you all on the list for a few months and to send you occasional updates on the project – so you haven't heard the last of us I'm afraid...

Thanks again for your help

Responses to this last message

The responses, together with the responses and discussions arising from earlier efforts on the list, showed that in many cases – via some practical tests of actual training and attempts to do simple embedding work - it would be difficult for the early stages of embedding work using HILT facilities to be supported remotely and that more work was needed to discover the best way of ensuring that embedding work at service sites was facilitated and well supported by HILT.

Appendix E: Embedding Toolkit Requirements Document

The requirements document for the embedding toolkit is available on the HILT project website at http://hilt.cdlr.strath.ac.uk/

Appendix F: Research Report

HILT IV: Research Report

Dennis Nicholson, Anu Joseph and Emma McCulloch

Centre for Digital Library Research (CDLR), University of Strathclyde, Glasgow

May 2009

Contents

1.0 Introduction 51
2.0 HILT Overview 51
3.0 Research Issues and Findings 52
3.1 System Architecture 52
3.1.1 Database design 52
3.1.2 Data handling issues 52
3.1.3 Collections database 52
3.1.4 Distributed approach 54
3.2 Terminologies and Terminology Mapping 54
3.2.1 DDC 22 54
3.2.2 Terminological spine 54
3.2.3 Need for 3 digits 55
3.2.4 Satellite schemes 55
3.2.5 Mapping of schemes to DDC 22 55
3.2.6 Mapping methodology 56
3.2.7 Mapping types 58
3.2.8 Storing mappings 58
3.3 Web Services and Standards 59
3.3.1 SRU/W 59
3.3.2 SKOS 60
3.3.3 BS 8723 60
3.4 Toolkit and Functions 61
3.4.1 Provision – different programming languages – Perl and PHP 61
3.4.2 HILT APIs 61
3.4.3 Spell checker 61
3.4.4 Wordnet 61
3.4.5 Portlet 61
3.4.6 XCRI/WikiWord/Umbel 61
3.5 Responsibilities of Services 62
3.5.1 Interface design 62
3.5.2 Quality of cataloguing 62
3.5.3 Use of standard subject schemes 62
3.5.4 Currency of schemes 62
4.0 Associated Work 62
4.1 Gold Dust 62
5.0 Dissemination 63
6.0 Evaluation 64
7.0 Future Service Issues 64
Appendix A: HILT IV: High-Level Mapping Study – UNESCO to DDC 66

1.0 Introduction

Within the HILT Phase IV project, a collection of research issues relating to the provision of an effective future entry-level service, or its further refinement, were investigated. This report forms the HILT Phase IV deliverable documenting research into a) any possible alternative approaches to spine provision and their implications; b) the identification of preferred spines for specific query types where options exist; c) many-to-many mappings; d) guidelines for others wishing to produce HILT-compatible mappings themselves; e) searching with compound terms; f) mapping types required for effective user services at different service levels; g) mapping grading and coding; h) a list of terminology or related service types likely to enrich user experience if encompassed within the HILT architecture; and i) the possible value of providing a HILT portlet (based on the JSR168 or WSRP standards) as a way of providing services with a relatively easy way of incorporating useful core user interface features into local services.

This deliverable reports on the research topics on the above list, together with an indication of other work identified during the project as necessary for the development of an effective future entry-level service or its further refinement and, where appropriate, a report of such research carried out during Phase IV and the HILT Embedding Project Extension. Research into areas included in the list above are reported on under the headings below:

- System architecture
- Terminologies and terminology mapping
- Web services and standards
- Toolkit (main deliverable)
- Responsibilities of services
- Associated work
- Dissemination
- Future service issues

Before documenting project findings in relation to the above list, a brief overview of the HILT model will be presented to contextualise the research issues being addressed.

2.0 HILT Overview

HILT is developing a pilot toolkit to help facilitate cross searching and browsing across information services using different subject schemes.

Since few services within the JISC information environment use the same subject scheme to describe their resources, or even any standard subject scheme at all in many cases, there is a need to provide tools to enable users to cross-search and browse distributed collections offered by JISC services and projects. This is what HILT is working towards.

HILT has a number of terminologies stored within its database, including subject headings, thesauri and classification schemes. These subject schemes are reconciled by mapping them to a Dewey Decimal Classification Scheme (DDC) 22 spine.

HILT uses SRU/W, a web services protocol, meaning that functions can be invoked on a machine to machine (M2M) basis. HILT provides results marked up in SKOS (Simple Knowledge Organisation System). Once the SKOS output is received it is then the responsibility of individual services as to how they want to parse SKOS results for subsequent presentation to their users.

The model means that much of the HILT functionality called upon by remote services takes place in the background, without the end user having to be heavily involved in specifying detailed information needs. So where, for example, a user search term does not match an index term within their service of choice, HILT should be able to provide an appropriate synonym, broader or narrower, or related term with which a relevant collection or service can be searched, where permissible, using the OpenURL protocol. So the 'user' of HILT will often be an information service accessing HILT and taking results back to its own (human) end users.

The architecture supporting this model is distributed and designed to provide an extensible service.

3.0 Research Issues and Findings

3.1 System Architecture

3.1.1 Database design

SQL server 2005 is used for storing data and there are fifteen schemes including DDC 22 available in HILT. Though SQL server seems to support loading large XML files into the database, it didn't prove feasible for some schemes due to their size. Programmes were written (using PHP) to split these large files into smaller chunks before loading them into the database. Normalising the tables affected the performance, especially when these tables are searched using JOIN. Flat tables were created to improve performance along with normalised tables knowing that data will not change that often. These flat tables have to be generated every time data changes in the main table.

A problem was encountered with SQL server – it periodically hibernates and subsequently takes 30 seconds to respond when a query is issued. The reason may be due to DNS resolution delays (either at the client or server side) or encoding mismatch between query and indexes, leading to table scans. The query response time after a long break was high and this issue was tackled by scheduling a job activity (run a stored procedure) every 10 minutes.

3.1.2 Data handling issues

Note: see also 3.2.1 DDC 22 for specific data handling issues relating to this classification scheme and 3.2.4 Satellite schemes for examples of some of the issues encountered in handling other schemes.

3.1.2.1 Foreign characters

SQL server driver for PHP5 cannot handle UTF-8 data properly, and resulted in storing question marks instead of Arabic characters. The SOAP server that interacts with the database is also written in PHP and the server couldn't pass a query with foreign characters properly. We resolved the issue by storing terms with translated characters as well as English characters in the database. We hope to solve this issue in PHP6.

Foreign characters also appeared in the upload of a selection of Welsh language mappings to DDC 22, obtained in XML format. These terms were amended manually within the database before making the Welsh mappings available for use within HILT.

3.1.2.2 Unique identifiers

Unique identifiers to represent a term are important from a database design point of view as well as from a SKOS representational point of view. Many schemes came without these unique identifiers, which generated problems while building relationships - especially with duplicate terms in UNESCO. Though LCSH provides identifiers for its data, there are not unique either. Automatic identifiers generated by the database created an unforeseen identifier clash issue, across different schemes, especially when represented in one SKOS file. This has been solved by including the scheme name before the identifier.

3.1.3 Collections database

The get_collections API currently queries a local collections database held at CDLR, containing a list of JISC collections and services catalogued by DDC number. The get_collections API uses this database to identify information services of potential relevance to a user query.

The intention within HILT has always been to use IESR to fulfil this role within the HILT toolkit. Due to various limitations of IESR in the earlier stages of HILT, the local collections and services database was set up within CDLR to simulate the functionality of IESR. At the end of this phase IV, IESR is not yet able to provide the functionality that HILT requires. As a result the switch from the local collections and services database has not yet been made.

Changes are required to IESR in order to facilitate its integration with HILT functionality. If these changes cannot be accommodated, HILT will not be able to use IESR as its collections and services database and will have to consider other options to fulfil this function. This is not desirable since IESR has already been funded to provide this function and it is neither HILT's intention nor wish to duplicate work taking place elsewhere.

To integrate HILT with IESR, a number of functional requirements are necessary.

a) DDC uses ranges to denote coverage of certain concepts. In HILT, we are constructing scheme hierarchies using parent identifiers, which means that we have to include ranges as these parent identifiers, where appropriate. If a DDC caption is identified as useful, its parent may also be relevant to the user's query. Such ranges are in some cases expressed using greyed out zeros in DDC. For example, the 100 section shows:

DDC notation	DDC caption
100	Philosophy & psychology
100	Philosophy
100	Philosophy, parapsychology and occultism, psychology

Table 1: The 100 division of DDC [Source: WebDewey, OCLC]

The uppermost notation in the table above denotes coverage of 100-199. The middle notation in the table denotes coverage of 100-109. The third notation in the table does not denote a range, it indicates DDC 100 precisely. The representation of these three variations is difficult to accommodate within the HILT database since the DDC notation is used within the system as a concept's id. All ids within the database require to be unique.

As with the majority of ranges, and easily illustrated by 302-307 'Specific topics in sociology and anthropology' cataloguers are instructed to assign a single notation to a given resource, rather than assigning a range. For example, within 302-307 the instruction to "Class comprehensive works in 301" is given, along with the note stating that "Unless other instructions are given, class a subject with aspects in two or more subdivisions of 302-307 in the number coming last". The DDC rule of three[47] may also come into play, where resources covering three or more subjects equally are classed at the first higher number that includes all three. It may not be necessary to incorporate ranges within HILT to facilitate resource discovery therefore, since ranges should not be applied to individual resources. It is necessary however, to devise a way of handling the (seemingly) duplicate use of notations, in numerical terms, within HILT to enable the system to generate a browsable hierarchy of DDC using ids and parent ids and to display the broader hierarchy of any given search term (within the get_ddc_records function). Further research is required to fully examine the implications if incorporating DDC ranges into HILT.

b) HILT also requires that subject scheme information be available for collections and services listed in IESR. Without this, HILT will not be able to direct a user to an appropriate subject scheme for a given collection or service, meaning that potentially relevant mappings or browse hierarchies of the scheme in question will not be exposed to the user. HILT functions identify potentially relevant services and collections in response to a user query; the user should then be given details of the scheme used within that service/collection, and any terms mapped from that scheme to the DDC notation of relevance. This information is central to the HILT model to enable users to either enter the relevant mapped term within a service/collections search facility, or to search a service/collection remotely using the openURL protocol where available. Without the fundamental scheme information being made available via IESR, the HILT functions cannot be executed.

[47] http://www.oclc.org/dewey/resources/teachingsite/courses/choice_of_number_review.pdf

c) The OpenURL protocol enables HILT to query services and collections dynamically from within the HILT interface. The selection of a search term in relation to an Intute collection, for example, will invoke a new web browser window in which the relevant search will be performed without further action from the user. IESR does not currently store information on services' use of the OpenURL protocol.

d) Elimination of duplicate records and improved consistency of metadata. Progress has been made on this front but early in the project we were unable to delete duplicate records from IESR and many of the records varied hugely in their coverage.

Following detailed discussions with IESR staff Leigh Morris and Jo Lambert, as well as with Vic Lyte, we are now working with IESR to resolve current barriers to service integration between HILT and IESR.

3.1.4 Distributed approach

HILT successfully connected to an OCLC server and an IESR server using the SRU/W protocol. This and the fact that the OCLC service is theoretically discoverable through a services registry like IESR is taken to show that a distributed approach encompassing a wider, non-JISC, environment, is feasible.

3.2 Terminologies and Terminology Mapping

3.2.1 DDC 22

A new version of DDC was released during the HILT programme of work. Datasets therefore required to be updated from DDC 21 to DDC 22 at the outset of Phase IV, to ensure currency of data. This required the HILT team to obtain an XML file containing the new version of DDC from OCLC, complete with LCSH to DDC mappings. Additional time was spent loading the data into the HILT database and undertaking data cleaning to standardise notations, eliminate foreign characters and the like. Changes to the arrangement of some sections of DDC introduced new data handling challenges.

For example, within the revised version a number of structural changes are evident. There is wider use of ranges to provide guidance to the structure of DDC, which is not currently being handled effectively by HILT. Although WebDewey[48] presents various ranges, making use of greyed out zeros, as appropriate, this cannot be replicated within the HILT database since HILT requires all DDC numbers to have a minimum of three digits, in order to facilitate the truncation process, on which HILT relies for some of its functionality. Since DDC notations are used as ids within HILT, these notations also require to be unique. Within the HILT database, distinct notations are required for distinct captions in order to create a means of generating a browse hierarchy. Term identifiers (i.e. the DDC notations in the case of DDC) and parent identifiers are used to identify the structural arrangement of captions within the scheme.

The example shown in Table 1, which is echoed for each of the hundred divisions (i.e. those notations ending in 00) as well as all notations ending in a single zero, have so far been handled unsatisfactorily by HILT. The 'fullest' caption was retained and given the unique identifier of _00. The information in Table 1 was therefore collapsed, with 100 being represented by 'Philosophy, parapsychology and occultism, psychology' labelling this entire class, with no upper, or higher, levels beyond. This is clearly inadequate and does not represent DDC and its use of ranges accurately.

3.2.2 Terminological spine

HILT uses DDC as its terminological spine. DDC was decided upon as HILT's spine for a number of reasons. 1) The subject coverage of DDC is universal. 2) Its notational system means that it lends itself fairly well to truncation (and subsequent refinement of results sets), a key feature of HILT's disambiguation process where users are required to select the context of their search term from its

[48] http://www.oclc.org/dewey/versions/webdewey/

(possible) various instances throughout DDC. 3) DDC has been translated into over 30 languages[49], while new translations are being continually worked on.

Alternative spines have been, and will continue to be, considered. UDC is a possible alternative spine and pilot mappings are currently being created between UDC and DDC to enable HILT to assess its suitability, or otherwise, in relation to HILT's existing satellite schemes and in relation to different subject areas, services and collections.

Section 7.0 describes current thinking behind a possible sustainable future architecture. Such a model would allow many alternative spines to be used.

3.2.3 Need for 3 digits

The HILT system requires DDC notations to contain a minimum of three digits. DDC 22 has captions that overlap conceptually, and are denoted as ranges, shown using greyed out zeros. This arrangement is repeated for each of the hundred divisions (see Table 1 for example). Due to the need for unique identifiers this structural arrangement has not yet been satisfactorily represented within HILT. Further work is needed to devise a way of representing the notations that signify a range in such a way that is compatible with the HILT APIs and also with SKOS.

3.2.4 Satellite schemes

Fifteen schemes other than DDC have been uploaded to the HILT database. These are: AAT (Art and Architecture Thesaurus), GCMD (Global Change Master Directory), NMR (National Monuments Record), HASSET (Humanities and Social Science Electronic Thesaurus) JACS (Joint Academic Coding System), IPSV (Integrated Public Sector Vocabulary), UNESCO Thesaurus (United Nations Education, Scientific and Cultural Organisation Thesaurus), MeSH (Medical Subject Headings), CAB Thesaurus, JITA (subject scheme used within the ELIS[50] repository), LCSH (Library of Congress Subject Headings), RAE units of assessment, SCAS (Standard Classification of Academic Subjects - replaced by JACS in 2002), SPEIR (in-house scheme used by the CDLR SPEIR project), XCRI (eXchange of Course-Related Information). Selected portions from selected schemes are mapped to areas of DDC, as appropriate. Schemes were chosen based on their use within collections and services within the JISC Information Environment (IE). For example, an email-based survey revealed that both HASSET and MeSH were used within various areas of the Intute service.

3.2.4.1 Scheme variation

Typically, schemes are of different sizes, they exhibit different levels of granularity, they are structured differently, they cover different subject areas and so on. It follows that what works well whilst mapping one particular scheme to DDC, may not work well for any other scheme. As HILT incorporates fifteen different schemes, it is difficult to impose generic mapping rules to be followed in every case. Some of the issues encountered with specific schemes are documented elsewhere in this report.

3.2.5 Mapping of schemes to DDC 22

The decision(s) on which portions of chosen schemes to be mapped was based on the research premise that high and deep level mappings could be used to facilitate browsing and searching, respectively, within potentially relevant collections and services.

3.2.5.1 What to map

HILT maps concepts from several subject schemes to a DDC spine. This will facilitate a process of vocabulary switching to improve interoperability within, and across, services employing different schemes to describe their resources. A range of high and deep level mappings will be implemented in order to help the user to 1) identify hierarchical information including broader, narrower and related terms associated with a concept in a given scheme and 2) identify appropriate concepts in given schemes with which to search specific services employing those schemes.

This dual-approach of providing high and deep level mappings was taken to 1) offer the user search

[49] **http://staff.oclc.org/~dewey/dewey.htm**

[50] http://eprints.rclis.org/

and browse services, as mentioned above, and 2) to research the value of mappings at different levels of granularity.

The subject coverage of mappings to be included was decided on the basis of JISC collections and services with whom we were able to work during phase IV. What schemes to focus on and the extent of mapping implemented was largely dependent on this. HILT worked with CAIRNS (http://cairns.lib.strath.ac.uk/) and the Social Sciences section of Intute (http://intute.ac.uk/socialsciences/). An additional consideration was to ensure schemes selected covered the same subject area(s), to enable a direct comparison to be made between the effectiveness of search versus browse using the same set of mappings.

It followed that HASSET, UNESCO and IPSV would be mapped to the DDC spine at a high level. Deep level mapping would be focused within the subject areas of mental health and psychology within MeSH and HASSET.

Mapping work is also time consuming and costly. Whilst mapping portions of three different schemes – HASSET, MeSH and UNESCO - to DDC, HILT researchers calculated that the average mapping takes 7 minutes to create. Cost and resourcing issues may also influence decisions taking on what schemes, and what portions of schemes, should be mapped.

For the purposes of the HILT IV follow-up, the Embedding project, JACS has also been mapped to DDC in its entirety. To fit with the high and deep-level mapping model, all 1300 JACS codes/terms were mapped to the top 919 DDC notations, since this meets functional requirements of Edina and Intute-based services. Not all of these 919 notations have been mapped to, but the targets for mappings from JACS are restricted to these 919 notations. The project has also created mappings from the RAE subject headings to DDC, with a view to investigating the possibility of using HILT functionality to improve subject access in institutional repositories.

3.2.6 Mapping methodology

There is a fairly substantial list of reasons why, especially when we get down to individual examples, mappings are difficult to create (not least because of scheme variation, mentioned in 3.2.4.1.). Careful attention needs to be paid to the scope of concepts in individual schemes to ascertain the coverage of a term and the nature of its equivalence to a DDC notation.

In HILT, we have adopted a pragmatic approach, creating mappings that we consider useful to the user. So, where we can't attain an exact match to a particular DDC concept, we may opt to include a range of broader, narrower and related, including Boolean combinations, to try to give the user a range of potentially useful options.

As previously noted, HILT incorporates a range of high and deep level mappings.

3.2.6.1 High level mapping

The provision of high level mappings will enable users to enter an appropriate point of a browse hierarchy of a given scheme within the area relating to their subject interest. As such, HILT will undertake a mapping exercise from HASSET, IPSV and UNESCO to the top three levels of DDC. The top thousand DDC notations, and corresponding captions have therefore been identified. Equivalent concepts from HASSET, IPSV and UNESCO will then be identified and mapped to each DDC notation individually. The optimum mapping will be an exact match. If there is no exactly equivalent concept within a satellite scheme (in this case HASSET, IPSV and UNESCO), a combination of concepts that collectively constitute an exact match should be sought. Where combinations are adopted + and | symbols will be used to represent AND and OR respectively.
Where an exact match cannot be identified within a satellite scheme (either directly or via a Boolean combination), a narrower or broader match may be sought, with a view to pinpointing the next best match. Judgement should be used to decide which of narrower or broader is the closest match to the concept being mapped to. Where both are considered useful, both should be included. Where neither narrower nor broader concepts can be identified a related match may be identifiable.

3.2.6.2 Deep level mapping

It has been decided that the areas relating to psychology within HASSET and MeSH will be mapped to DDC, as this area has been identified as useful to intute.

HASSET has the following terms in this area:

```
PSYCHOLOGY
¦..APPLIED PSYCHOLOGY
¦  :..CLINICAL PSYCHOLOGY
¦     :..PSYCHOANALYSIS
¦     :..PSYCHOTHERAPY
¦        :..DRUG-PSYCHOTHERAPY COMBINATION TREATMENT
¦        :..HYPNOTHERAPY
¦  :..EDUCATIONAL PSYCHOLOGY
¦  :..OCCUPATIONAL PSYCHOLOGY
¦     :..MANAGERIAL CHARACTERISTICS
¦        :..LEADERSHIP
¦  :..SOCIAL PSYCHOLOGY
¦..DEVELOPMENTAL PSYCHOLOGY
¦  :..ADOLESCENT PSYCHOLOGY
¦  :..CHILD PSYCHOLOGY
¦  :..EMOTIONAL DEVELOPMENT
¦     :..EMOTIONAL IMMATURITY
¦     :..EMOTIONAL MATURITY
¦  :..INDIVIDUAL DEVELOPMENT
¦  :..MENTAL DEVELOPMENT
¦     :..LANGUAGE DEVELOPMENT
¦  :..PERSONALITY DEVELOPMENT
¦     :..PERSONALITY CHANGE
¦..PARAPSYCHOLOGY
¦  :..EXTRASENSORY PERCEPTION
```

Figure 1: Section of HASSET hierarchy [Source: http://www.data-archive.ac.uk/findingData/thesaurusInfo.asp?keyword=PSYCHOLOGY]

Within MeSH section F relates to psychology, as available at http://www.nlm.nih.gov/mesh/2008/MeSHtree.F.html

The methodology for establishing deep level mappings is the same as that for high level mappings, although reversed since we are unable to identify DDC notations/captions in advance. Terms to be mapped will be identified within each of the satellite schemes (HASSET and UNESCO), before actively seeking an appropriate match within DDC. The optimal outcome is to establish mappings in the following order of preference: exactMatch; narrowMatch/broadMatch; relatedMatch

3.2.6.3 Need for many-to-many mappings

At the uppermost levels of the DDC hierarchy, subjects often appear inter or multidisciplinary. In cases such as 000, for example, where the associated caption is ‘computer science, information & general works’, it is highly probable that many-to-one mappings will be required to capture equivalence for the caption as a whole. It is also probable that a ‘better’ match will be identified for corresponding UNESCO terms at a lower level of DDC, making many-to-many mappings necessary.

A decision has been taken to encode ‘computer science’ and ‘information’ as individual NTs of DDC ‘computer science, information & general works’. Likewise for ‘psychology’ and ‘philosophy’ in relation to DDC ‘psychology & philosophy’. This is because each of the terms form part of the subject being mapped to, but only partially. In other words the DDC term is broader than either of the mapped terms, individually.

Since this group of narrowMatches are sufficiently different from a narrowMatch dictated by hierarchical structure, another argument is that we should introduce a match type signifying ‘partial exact match’. It is arguable that each of the terms psychology and philosophy are more closely matched than other, perhaps more typical, NTs.

The need for Boolean operators (or SKOS classes) within queries is also relevant here for combined search, as is the issue of pre/post coordination of mapped terms. Such issues have been subject to discussion within the SKOS community and have not yet been resolved.

3.2.6.4 Boolean mapping

Boolean mappings are required to express equivalence relationships between satellite schemes and DDC. Since concepts rarely overlap in their entirety between schemes, it follows that AND and OR are useful for combining concepts, to determine a 'closer' match with the concept being mapped to/from.

3.2.6.5 Mapping study: UNESCO to DDC

A high-level mapping study was undertaken whilst mapping UNESCO to DDC. Findings of this exercise are documented as Appendix A. Note that this study took place early on in the project schedule; many of the issues have now been superseded by subsequent research. The study is included here, intended to be illustrative of the types of specific issue that may be encountered while undertaking terminology mapping.

3.2.7 Mapping types

To help prioritise the usefulness of mappings implemented SKOS mapping types[51] are being adopted within HILT to indicate the type of relationship between a particular term and a DDC notation.

A detailed research study was conducted in HILT III (McCulloch & Macgregor, 2008)[52] to determine the appropriate range of mapping types or equivalence relationships required, with which to categorise mappings from satellite schemes to the DDC spine. The outcome of this study was that five mapping types would be used within HILT, in line with SKOS standards. These are: exactMatch; broadMatch; narrowMatch; majorMatch and minorMatch. Subsequent to this study, major and minor match were deprecated, being replaced with overlappingMatch. Further discussion then concluded that relatedMatch should be used instead of overlapping, even although a significant number of contributors felt there was a clear need for both overlapping and related, which they viewed as significantly distinct (see also 3.3.2.1).

So the four mapping types used at the time of undertaking mapping work are exactMatch, broadMatch, narrowMatch and relatedMatch. Since the completion of this work a new SKOS mapping type has been introduced - closeMatch. This, and future changes, will require to be accommodated within HILT but the research team feel it would be better to wait for the standard to stabilise in the area of mapping before incorporating changes throughout the HILT APIs.

3.2.7.1 JACS to DDC: mapping type issue

Consultant terminology expert Leonard Will[53] observed, when creating high-level mappings (see 3.2.6.1) from JACS to the top three levels of DDC (as defined within the HILT project), very few cases of exact match were found. This was mainly because the limited range of DDC numbers used did not allow the exact JACS topics to be expressed. More precise numbers are available in many cases in DDC. Most of the JACS concepts are therefore shown as narrower than the DDC number to which they are mapped. This was deemed unreliable, though, because in many cases there was some overlap - the JACS concepts included things that are not in the DDC classes and vice versa. The lack of an 'overlap' match is a serious limitation, and guesses had to be made as to the direction in which most of the overlap occurred in deciding whether to show the match as broader or narrower.

3.2.8 Storing mappings

When undertaking terminology mapping in line with HILT's methodology, an Excel Spreadsheet should be established for each scheme being mapped, whether at high or deep level. DDC notations

[51] **http://www.w3.org/TR/2009/WD-skos-primer-20090317/#secmapping**

[52] McCulloch E. & Macgregor G. Analysis of equivalence mapping for terminology services, *Journal of Information Science* 2008 34(1) pp.70-92. Available at http://strathprints.strath.ac.uk/3173/

[53] http://www.willpowerinfo.co.uk/

should be listed in the leftmost column, with DDC captions in the column alongside. Mapped terms from a satellite scheme should be included in the third column, as single terms or combinations, as appropriate. The fourth column will show the type of mapping equivalence between the concepts being mapped. A notes field is also useful to accommodate any information that might justify the assignation of a particular mapping type, which may not be obvious otherwise. For example, if there is a USE FOR instruction to a term which, if preferred, would constitute an exact match.

DDC Class	DDC Caption	IPSV Term	Mapping Type	Notes
005	Computer programming, programs, data	Programming	exactMatch	
020	Library and information sciences	Library and information services	narrowMatch	
070	Documentary media, educational media, news media; journalism; publishing	Journalism + Newspapers + Communications industries	exactMatch	Communications industries: UF publishing

Table 3: Example mappings from IPSV to DDC (1)

Where different equivalence relationships are expressed for different mappings to the same DDC notation/caption, these should be stored on different rows of the Excel file, to facilitate parsing of results. For example, if a broadMatch and a narrowMatch are both identified as valid mappings to the same DDC notation/caption this data should be stored as follows:

DDC Class	DDC Caption	IPSV Term	Mapping Type	Notes
130	Parapsychology and occultism	Psychology	broadMatch	
130	Parapsychology and occultism	Occultism	narrowMatch	

Table 4: Example mappings from IPSV to DDC (2)

Where the match type is the same e.g. all narrowMatches, as many terms can be stored on a single row of the file, using Boolean operators as appropriate. For example, three distinct narrow matches should be expressed using the OR (|) operator.

3.3 Web Services and Standards

3.3.1 SRU/W

Implementation is based on Index Data’s SimpleServer – a simple Perl module intended to develop Z39.50, SRU and SRW servers. SimpleServer is based on popular YAZ toolkit which is robust, efficient, portable and inter-operates well with different Z39.50 and SRU/W servers. There is also a SOAP envelope involved, though it is transparent to the clients. See also 3.1.4.

3.3.1.1 Explain response

The EXPLAIN operation returns a ZeeRex (http://explain.z3950.org/) XML file that allows a client to find out the functional capabilities of an SRU/W server, and which indexes are available to use in CQL queries. It was hoped that more detail about the CQL structure, including controlled vocabulary values, could be given in the ZeeRex file. However, the ZeeRex format does not allow for this possibility, and so the considered recommendation is to create a Context Set reference document (a human readable document) explaining the CQL structure. The ZeeRex maintainers are considering adding a value to the index definition in ZeeRex to indicate whether the index contains controlled values. If this comes into operation then clients could use an SRU/W scan operation to find the controlled vocabularies.

3.3.1.2 Namespace declaration in individual records

SRU/W responses are XML documents containing an SRU/W specific wrapper or envelope. In the case of SRW there is a further SOAP envelope involved but this should be invisible to the application using the SRW client. In particular, the SRU/W searchRetrieve response may contain records that are transmitted in XML. These are part of the XML searchRetrieve response and so any XML namespaces used within these XML records must be declared so that the record data is within scope in the SRU/W XML document as a whole.

The most efficient, and conventional, place to declare these namespaces would be within the root element opening tag (along with other namespace declarations required for the SRU/W XML document). However, due to limitations of the software being used to front the SRU/W server (Indexdata's perl Net::Z3950::SimpleServer based on their popular yaz library) the namespace declarations for the records can only be inserted at the record data level. This means that for every record's namespaced tags to be in the scope of a declaration the namespace declaration must be inserted into the opening tag of every record. This is less efficient than desired, but it is not invalid XML. It makes no technical difference to the XML document, only to its size.

3.3.1.3 Caching

An SRU/W search request can specify the start record and the number of records to retrieve from the result set. This is so that records from a large result set can be retrieved a handful at a time for manageable browsing by the end-user. Unlike Z39.50, SRU/W does not give identifiers to the result sets, but rather uses the original CQL query issued in the request. Since this query won't change between requests for different pages of records it is suitable for caching the SOAP response (from the HILT SOAP server) in the SRU/W server ready for the next page request. In fact, the caching is now keyed on the SOAP method that the CQL query is translated into since the SRU/W client can request the hilt:matches or hilt:concepts part of the SOAP response by altering the CQL query. This provides a huge performance boost and greatly improves the response time for users subsequently requesting further results from an initial CQL query, or indeed the hilt:concepts.

3.3.2 SKOS

A working draft of the SKOS primer or, user guide, can be found at http://www.w3.org/TR/2009/WD-skos-primer-20090317/ SKOS is a developing standard and the accompanying reference document is located at http://www.w3.org/TR/skos-reference/. Both are published by the Semantic Web Deployment Working Group as part of the W3C Semantic Web Activity.

3.3.2.1 Limitations

SKOS hasn’t solved the issue with representing Boolean mappings, which force us to find a way to represent our data. HILT uses + sign to represent AND and | sign to represent OR in SKOS records. SKOS validation fails because of these signs in the output, but can easily adapt to any SKOS accepted solution in the future.

The change from SKOS mapping type overlappingMatch to relatedMarch (see also 3.2.7 and 3.2.7.1) is potentially limiting since, although thesaurus standards don’t provide a way to distinguish between these, it may be useful to have both since overlapping concepts, by definition, must be from the same facet of a scheme while related concepts can be taken from different facets of a scheme.

See also 3.1.2.2 Unique identifiers.

3.3.3 BS 8723

BS8723 Parts 1-4 were adhered to.

3.4 Toolkit and Functions

A toolkit has been developed to illustrate how different HILT functions can be usefully embedded within a service. The Perl based toolkit and documentation is available to download at http://hilt4.cdlr.strath.ac.uk/toolkit.zip and http://hilt4.cdlr.strath.ac.uk/toolkitDocumentation.doc respectively.

A demonstrator of the same is available at http://hilt4.cdlr.strath.ac.uk/toolkit/intro.cgi

3.4.1 Provision – different programming languages – Perl and PHP

3.4.2 HILT APIs

The HILT Toolkit uses the following APIs, accessible for testing at http://hilt4.cdlr.strath.ac.uk/hilt_SRU/W.cgi:

get_ddc_records

- Returns DDC captions and numbers related to a subject term. The user can then choose the most appropriate to his/her interest.

get_collections

- Returns collections classified under a specified DDC number or its stem, including subject scheme used. We will look at this function in more detail shortly.

get_non_ddc_records

- Returns terms from schemes other than DDC by matching user terms to DDC notations, before identifying mappings to those particular notations.

get_all_records

- Combines the functions of get_DDC_records and get_non_DDC_records.

get_filtered_set

- Allows specified fields from specific terminologies or combinations of terminologies to be searched.

3.4.3 Spell checker

A spell checker based on an index created from HILT's local database has provided added-value in HILT's search functionality (get_sp_suggestions). The Lucene spell checker is implemented in Java and the implementation is based on David Spencer's code using the n-gram method. In order to access Java classes in PHP (SOAP server is built using PHP), PHP-Java bridge is enabled in the server and the function is available as a web service. The existing class has also been extended to accept compound queries.

3.4.4 Wordnet

WordNet® is a large lexical database of English language terms, developed under the direction of George A. Miller. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. WordNet has also added value in HILT searches, using get_wordnet_suggestions, by helping users to choose their search term. WordNet suggestions are available as a web service and the implementation is Java based.

3.4.5 Portlet

After a brainstorming session with Edina staff, it was determined that the portlet (based on the JSR168 or WSRP standards) did not have the level of functionality that HILT required. In particular, it did not appear to be possible to pass parameters between two portlets on a service screen.

3.4.6 XCRI/WikiWord/Umbel

The possibility of integrating these projects with HILT was explored. Discussions with XCRI are ongoing; WikiWord data is not yet in a position to be used; Umbel can add values to HILT search functionality.

3.5 Responsibilities of Services

Optimum results from HILT are dependent on a range of external factors, some of which lie directly with the collections and services being consulted/searched by HILT. Responsibilities of individual collections and services that will have a direct effect on the value of HILT to users, as well as elements over which HILT had no control such as the way HILT is presented within, or incorporated into, a service include:

3.5.1 Interface design

The way in which HILT functionality is presented to users within services and collections, once embedded in their local websites, is largely the responsibility of the service itself. HILT offers a downloadable toolkit for integration but will not offer assistance in redesigning service websites to accommodate this.

As such, little work has been done in this phase of HILT relating to interface design since it is probable that individual services and collections will want to incorporate HILT in different ways, perhaps only using part of the toolkit in some cases or perhaps only using elements of the toolkit relating to a specific scheme or schemes, in line with how its resources have been classified.

See also 6.0 Evaluation.

3.5.2 Quality of cataloguing

The value of terms returned to HILT depends on how effectively they have been applied to individual records within services and collections. HILT can be used to identify exact matches, narrower/broader terms and so on, but the user will only retrieve useful resources from a service or collection if they have been effectively catalogued using appropriate terms.

3.5.3 Use of standard subject schemes

It is also crucial that services and collections apply subject schemes exactly according to the instructions/scope notes within a given scheme. Standard schemes should be used 'as is', that is, they should not be adapted to suit local needs if services and collections wish to be fully integrated with HILT. HILT maps standard schemes to a DDC spine. Unless a service or collection informs HILT directly, provides their adapted scheme for upload into the HILT database and requests that additional mappings be created from this locally adapted scheme to the DDC spine, that collection or services terminology will not be reconciled with those used in HILT.

3.5.4 Currency of schemes

In addition to employing a standard scheme (unless effort is to be made to incorporate local or in house schemes into the HILT infrastructure) it is also essential that the most recent version of any given scheme is used. HILT will update its terminological content in line with that of scheme providers so it is essential that services and collections do the same to maintain interoperability with HILT data to give users the best results.

4.0 Associated Work

See also 3.4.6 XCRI/WikiWord/Umbel.

4.1 Gold Dust

We analysed Personal Interest Profiles (PIPs) data collected as part of the Gold Dust project (http://www.hull.ac.uk/golddust/) and explored the possibility of matching these terms with any specialised subject area ontology in HILT. The outcomes are

Out of 5180 key phrases collected using two methods devised by Gold Dust, A and B, only 86 distinct HILT terms matched the terms identified using method A and 127 in method B[54]. There was overlap of

[54] The exercise here was not to compare methods A and B as defined by the GoldDust project, but to investigate whether HILT would be of any value in identifying terms that may feature in PIPs.

these terms across terminologies and detailed matching of terms in different HILT schemes are listed in the table below.

HILT scheme	HILT terms matched in Method A (out of 2730)	HILT terms matched in Method B (out of 2450)
CAB	55	69
Dewey	16	25
GCMD	7	6
HASSET	5	12
IPSV	2	7
JACS	3	0
LCSH	27	29
Mesh	24	28
NMR	1	54
SCAS	1	2
UNESCO	7	19

Table 5: Number of terms matched in individual schemes in HILT

5.0 Dissemination

An email list HILT-collaborators was established early on in the project to establish potential stakeholders' functional needs from a terminology service such as HILT. What do they want it to be able to do? What would they find useful? What level of technical expertise is available to them to embed HILT functionality in their local services? These, and other, questions were discussed and the HILT team received positive feedback on the APIs being developed. An archive of the HILT-collaborators mailing list is available at https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=HILT-COLLABORATORS

HILT IV work with the members of the HILT Collaborators email list showed that there was a good deal of interest from the community in these services. Twenty-nine people joined the list which was set up specifically to look at embedding work of this kind. A questionnaire asking what kinds of uses were of interest was answered by sixteen people. All indicated an interest in at least one potential way of using HILT, most indicated an interest in three or more ways.

A wiki was established to facilitate exchange between the HILT project team and interested parties. The wiki was used to disseminate elements of the toolkit during development stages, for testing and feedback. Demonstrations of completed elements were also made available by this means. The wiki was later used to elicit discussion regarding technical embedding of HILT's APIs in EDINA-based, intute-based and CDLR-based services within the remit of the embedding extension project.

In addition, several talks, demonstrations and publications have been delivered throughout the project lifetime:

Presentations/demonstrations:

- HILT: Enhancing subject search through embedded web services, JISC Conference 2009, Edinburgh International Conference Centre, Edinburgh, 24th Mar 2009.
- Metadata and Scotland's information environment: potential benefits of Web 2.0, Metadata Issues and Web Services, CIGS Seminar, 30 Jan 2009.
- HILT IV Pilot Toolkit Demonstration, CIG 2008, University of Strathclyde, Glasgow, 3-5 Sep 2008.

Scheduled presentations/demonstrations:

- Nicholson, D. Signposting the crossroads: terminology web services and classification-based interoperability, Classification at a Crossroads – Multiple Directions to Usability, International UDC Seminar, The Hague, 29-30 Oct 2009.
- Nicholson, D. Looking at the Past and preparing for the Future, IFLA Satellite Meeting, Florence, August 2009.
- Embedding terminology web services to improve subject access, Where will it all end? – Emerging Technology in the Library, MmIT North West, Liverpool John Moores University, 17th Jun 2009.

Conference Papers:

- Nicholson, D. 'Optimising Interoperability in Multi-KOS Subject Searching: Framework for a Collaborative Approach? The Challenge of the Electronic Environment to the Organization of Knowledge - Second International Seminar on Subject Access to Information, Helsinki, Finland, 29-30 Nov 2007.
- Macgregor G. & McCulloch E. & Nicholson D. Terminology server for improved resource discovery: analysis of model and functions. Second International Conference on Metadata and Semantics Research. Corfu, Greece 11-12 Oct 2007. Available at: http://strathprints.strath.ac.uk/3435/
- Macgregor G. & Joseph A. & Nicholson D. A SKOS Core approach to implementing an M2M terminology mapping server International Conference on Semantic Web and Digital Libraries (ICSD-2007). Bangalore, India 21-23 Feb 2007. Available at: http://strathprints.strath.ac.uk/2970/

Journal Articles:

- Nicholson, D. A Common Research and Development Agenda for Subject Interoperability Services? Signum, Issue 5, 2008.
- McCulloch, E. & Macgregor, G. Analysis of equivalence mapping for terminology services, Journal of Information Science, 2008 34(1) pp.70-92. Available at: http://strathprints.strath.ac.uk/3173/

Submitted Journal Articles:

- Nicholson, D., McCulloch, E. & Joseph, A. HILT IV: Subject Interoperability through Building and Embedding Pilot Terminology Web Services. World Digital Libraries.

Practitioner Articles:

- Nicholson, D. & Menzies, K. BUBL, HILT, and the Scottish Information Environment: potentials of Web 2 and Web 3, WIDWISAWN, 7(1) Available at: http://widwisawn.cdlr.strath.ac.uk/issues/vol7/issue7_1_4.html

6.0 Evaluation

An evaluation of HILT was undertaken during Phase IV by Ian Ruthven of the Computer and Information Sciences Department at the University of Strathclyde. This study is included as Appendix G of the final report, which will be available at http://hilt.cdlr.strath.ac.uk/hilt4/documents.html.

7.0 Future Service Issues

Toward the end of HILT IV and the embedding project, it became evident that a sensible path to follow to sustain the future of an effective terminologies service to improve subject interoperability was to:

- Identify the likely elements and architecture of an inclusive subject interoperability service that could, in time, incorporate not only HILT and a mapping based approach built around a DDC spine but other terminology services across the UK and beyond applying either different approaches to mapping or different approaches to interoperability.
- Identify the research requirements associated with this architecture and its elements.
- Work towards an agreed multi-initiative approach to the associated research (and development) with partners elsewhere in the UK and in the world.

Although work in this area is still at an early stage, efforts have nevertheless been made to begin work towards agreeing a collaborative approach with other 'players' in the terminologies field. A paper on the architecture was presented at an ontologies conference in Helsinki in November 2007, a paper on the topic published in the international version of the Signum journal in November 2008, and, in December 2008, steps were taken to contact major European projects in the terminologies area to begin the process of talking about collaboration and about applying for FP7 funding to carry the work forward.

A range of issues would require further research and development should a collaborative approach prove favourable. These are fully documented in the project's final report, which will be made available at http://hilt.cdlr.strath.ac.uk/hilt4/documents.html following its acceptance by JISC.

Appendix A: HILT IV: High-Level Mapping Study – UNESCO to DDC

1. Introduction:

This document describes a HILT IV investigation into the mapping of UNESCO thesaurus to the top three levels of the Dewey Decimal Classification (DDC) Scheme hierarchy, or the top thousand DDC numbers (actually 919). Several satellite schemes are to be mapped to DDC at this level; the project has begun with UNESCO. In addition to these top level mappings, a deeper level mapping exercise is planned and is dependent on schemes in use within intute and across the JISC Information Environment.

Equivalence relationships identified between satellite scheme terms and DDC notations will be encoded using the SKOS Mapping Vocabulary Specification (MVS). It is of interest to consider Chaplan's mapping types in addition to the SKOS MVS. An initial attempt to reconcile the two approaches has been made, with a view to incorporating both approaches into the HILT mapping database. This may form part of an evaluative exercise undertaken later in the project. It will be of interest to consider whether the five SKOS mapping types are sufficient to characterise the nature of mappings between schemes or whether Chaplan's more detailed set is more valuable in the context of retrieval.

2. Aim:

Within the HILT project, terms from a number of subject schemes will be mapped to a central DDC spine. Mappings will then be used to provide subject interoperability by offering terms mapped to DDC, together with details of the type of mapping relationship evident. Subject access via scheme hierarchies entered at a point appropriate to the user's subject interest will be provided.

3. Methodology:

Notations corresponding to the top 3 levels of DDC (tens, hundreds, thousands) have been identified, together with captions. Exact matches are to be sought for each within the UNESCO thesaurus, taking into consideration the hierarchical context of each scheme and any associated scope notes and instructions. Where no exact match exists, the next 'best' match (likely to be narrower or broader) will be identified.

It is important to consider the hierarchical context of notations/terms since, for example, for DDC 192, the HILT database lists the term to map to as 'British Isles'. Only from the broader hierarchy or from detailed knowledge of DDC's structure do we know that this instance of 'British Isles' in fact relates to Modern western philosophy.

Mapped DDC numbers will then be input to the database, against mapped terms from each of the satellite schemes. In addition to the mapped number being entered, the type of relationship characterised by the mapping will be encoded in line with the SKOS Mapping Vocabulary Specification (MVS), as follows[55]:

exactMatch
broadMatch
narrowMatch
majorMatch[56]: Concept Match (CM) with significant overlap

[55] See section 5 (SKOS MVS developments) for details of recent development of the SKOS MVS and proposed resolutions to SKOS ISSUEs.

[56] Deprecated; proposal made to change to overlappingConcept

minorMatch[2]: CM with slight overlap. If not exact, narrower, broader or major, it's probably minor i.e. some type of relationship, but none of them quite fit.

Definitions and examples:

exactMatch: applies to semantically equivalent concepts with 100% overlap. Such terms need not be exact character-by-character matches. That is, they may exhibit spelling variation, intervening characters and so on.

broadMatch: applies to a match where a term from a satellite scheme is broader in scope than the equivalent DDC term. e.g. UNESCO: science; DDC: natural sciences & mathematics.

narrowMatch: applies to a match where a term from a satellite scheme is narrower in scope than the equivalent DDC term e.g. UNESCO: history; DDC: history & geography.

majorMatch[2]: conceptually equivalent terms with significant overlap (>50%), but not exactly equivalent. e.g. spheres – balls
Not all spheres are balls e.g. globes
Not all balls are spheres e.g. rugby ball
Since the majority of spheres are balls, but exceptions exist (in both directions), this example constitutes a MajorMatch.

minorMatch[2]: conceptually equivalent terms with some degree of overlap, although not significant (<50%). If terms are not deemed exactly matched or majorly matched, but there is some level of overlap between them. It is unlikely that terms exhibiting a minor match will exist within the same discipline/hierarchy. For example:
DDC

DDC number	DDC caption	UNESCO terms	Mapping type
005.8	Data security	CRIME	minorMatch

In the above example, Data security and CRIME are not exactly matched and one is neither narrower nor broader than the other. In some contexts however it is possible that there may be a degree of conceptual overlap. The extent of this overlap is likely to be encountered in limited circumstances so the terms may be deemed as a minorMatch. In practice, it is thought likely that minorMatch will be applied very infrequently.

4. Findings:

In undertaking the mapping of 919 UNESCO terms to the top thousand DDC captions the following issues were noted. One issue related to the nature of UNESCO itself is also documented here, after those relating more specifically to the mapping of UNESCO to DDC.

1. Need for many-to-many mappings

   At the uppermost levels of the DDC hierarchy, subjects often appear inter or multidisciplinary. In cases such as 000, for example, where the associated caption is 'computer science, information & general works', it is highly probable that many-to-one mappings will be required to capture equivalence for the caption as a whole. It is also probable that a 'better' match will be identified for corresponding UNESCO terms at a lower level of DDC, making many-to-many mappings necessary.

   A decision has been taken to encode 'computer science' and 'information' as individual NTs of DDC 'computer science, information & general works'. Likewise for 'psychology' and 'philosophy' in relation to DDC 'psychology & philosophy'. This is because each of the terms form part of the subject being mapped to, but only partially. In other words the DDC term is broader than either of the mapped terms, individually.

Since this group of narrowMatches are sufficiently different from a narrowMatch dictated by hierarchical structure, another argument is that we should introduce a match type signifying 'partial exact match'. It is arguable that each of the terms psychology and philosophy are more closely matched than other, perhaps more typical, NTs.
NB. The need for Boolean operators (or SKOS classes) within queries is also relevant here for combined search, as is the issue of pre/post coordination of mapped terms.

2. Indirect exact matches

   Where USE/UF relationships are evident within a scheme, mappings may be implemented as exact yet may not be immediately apparent. For example:

   UNESCO: Palaeontology
   UF Palaeobiology, Palaeobotany, Palaeozoology

   DDC: 560

DDC number	DDC caption
560	Paleontology Paleozoology

In the absence of the UF instruction UNESCO term Palaeontology would be considered a narrowMatch to DDC 560, since it constitutes a sub-area of the DDC caption. However, the UF instruction indicates that the term palaeontology would also be used for instances where palaeozoology might be adopted. As a result of this instruction, the UNESCO term is deemed to match DDC 560 exactly.

When it comes to DDC 561 however, this approach becomes problematic.

DDC number	DDC caption
561	Paleobotany; fossil microorganisms

The relevant UNESCO terms to map to DDC 561 are palaeontology (since it is UF palaeobotany) and fossils. Palaeontology is an exactMatch to paleobotany in DDC, however since 'fossil microorganisms' is part of the same heading HILT mapping methodology dictates that this is in fact a narrowMatch. Fossils is a narrowMatch in the sense that it is a subset of palaeontology, yet is broader than fossil microorganisms. There is therefore a degree of conflict here. The term microorganisms in UNESCO could also be mapped here. This term is located within UNESCO's biology hierarchy as follows:

Microbiology
    NT1 Bacteriology
        NT2 Microorganisms

and is already mapped to DDC 628.536 within the HILT database. The complete hierarchy for DDC 628.536 (below) indicates however, that this does not appear to be an appropriate match for the context of microorganisms denoted by DDC 561.

[DDC 628.536: Technology > Engineering and allied operations > Sanitary and municipal engineering Environmental protection engineering > Pollution control technology and industrial sanitation engineering > Microorganisms]

3. Equal 'narrowness'/'broadness'

   Where narrower matches are identified, comparable levels of granularity may not be evident, both within and across mapped schemes. Considering a mapping to DDC 000 computer science, information & general works, from a range of satellite schemes, including UNESCO we see the following:

AAT: computer programming
CAB: computers
UNESCO: computer science; information

Clearly, the three examples above are not equivalently narrower matches of computer science, information & general works. The UNESCO terms are immediately narrower, the CAB term is a narrower term of the AAT (2 levels narrower than DDC?) and the AAT term is a narrower term of the CAB term (i.e. 3 levels narrower than DDC?).

Within HILT, each of the above examples will be encoded as a narrowMatch, since each term is narrower than the DDC notation to which it is mapped. In retrieving a result set however, this aspect of the methodology does mean that the mapped terms presented in response to a query will not necessarily be equivalently narrower. This is really down to the nature of schemes themselves, their structures, levels of granularity and so on. Do we foresee problems with this approach?

4. Treatment of notes in DDC

To improve the consistency of mappings from satellite schemes to DDC, decisions have been taken regarding 'class here' and 'include' notes as they appear in the DDC schedules. Class here is to be treated as a concept match, encoded as either majorMatch/minorMatch[2], in line with the SKOS MVS, depending on the degree of conceptual overlap. Where notes under an area of DDC state 'include TERM A' TERM A from a satellite scheme will be deemed a narrowMatch of the DDC notation in question.

5. Consistent use of mapping types

In addition to the potential problem resulting from the assignation of the narrowMatch mapping type, further evidence has been uncovered that suggests the assignation of mapping types may be contradictory, depending on specific examples.

DDC 576:

DDC number	DDC caption	UNESCO terms	Mapping type
576	Genetics and evolution	GENETICS	narrowMatch
		EVOLUTION	narrowMatch

Each of the above terms is mapped as a narrowMatch since each form a subset of the complete DDC heading. However, both UNESCO terms 'Genetics' and 'Evolution' when considered independently, are clearly broader than DDC 576 (in contrast to the 'psychology & philosophy' case discussed in 1. above). This makes documentation of mapping guidelines problematic since many issues will have to be handled on a case-by-case basis.

For DDC 579:

DDC number	DDC caption	UNESCO terms	Mapping types
579	Microorganisms, fungi, algae	MICROORGANISMS	narrowMatch
		FUNGI	narrowMatch
		AQUATIC PLANTS	narrowMatch

Aquatic plants is also mapped as a narrowMatch since in UNESCO there is an instruction to USE aquatic plants for Algae. However, in the DDC schedules under 579 there is an instruction to class here microbiology and various other concepts. According to HILT mapping methodology, class here instructions (see 4. above) constitute a concept match (majorMatch/minorMatch[2]) between terms. This means that for the above example, microbiology would be mapped to microorganisms, fungi, algae as a majorMatch[2] or minorMatch[2]. Depending on how results are to be ranked within HILT therefore, a concept match may be treated as a 'better match' than a narrowMatch. Clearly when looking at the terms however, this may not necessarily be the case.

If there are seemingly more appropriate narrowMatches for a particular caption, should we leave it at that? Or, should all instructions (from both classification scheme and thesaurus in this case) be considered? Can class here and include notes be treated consistently across all cases? Should we exert more flexibility in our interpretation of these?

If such notes are to be treated consistently there is greater scope for a degree of automation being introduced to the mapping process; or at least the potential to speed up the intellectual mapping process in these cases. Will this lead to mapping errors or less value for the user, however?

6. Lack of qualifiers in DDC

   Not so much a mapping related issue but a potential problem for the disambiguation phase of HILT. Although, as already mentioned, the complete hierarchical information available will be considered in the mapping of schemes, and indeed in the presentation of initial matches to a user search term, until now we have largely assumed that duplicate terms in DDC i.e. at the end point of hierarchies are likely to be located in different disciplines or, if within the same discipline, for example, two instances of 'teeth' appear in the technology hierarchy, that the nature of the hierarchy itself will provide sufficient means to enable the user/client to differentiate between them.

   The process of mapping UNESCO to DDC however, has uncovered instances where identical terms belong to the same discipline, within very close proximity. For example, DDC 218 and DDC 233 both have the caption 'humankind' and both are located within the discipline of religion. The first instance relates to philosophy and theory of religion and the second to Christianity, Christian theology. How can HILT provide further information to the user in order to help the user choose between such instances? Is it a problem for HILT? What happens if a user wants to select two or more instances of a term? Will it be documented as a problem arising purely from DDC's structure?

7. Need for compound searching of mapped terms

   Even when mapping the highest levels of DDC we can see the need for pre-coordinated terms to be searched. For example, in the case of 'psychology & philosophy' as discussed earlier. This illustrates that UNESCO does not always have a suitable term to map to a DDC number that covers the full extent of the concept represented by that number. Often, terms mapped from UNESCO cover one aspect of the concept denoted within DDC. Looking at more granular subjects, for example:

   DDC 193:
   100
   Philosophy &
   psychology
   180-190
   Historical, geographic, persons treatment of
   philosophy
   190
   Modern western
   philosophy
   193
   *Germany and
   Austria
   Source: WebDewey

DDC number	DDC caption	UNESCO terms	Mapping type
193	Germany and Austria	PHILOSOPHY	broadMatch
		GERMANY	broadMatch
		AUSTRIA	broadMatch

In the above example, searching for either one of 'Germany' or 'Austria' or 'Philosophy' is unlikely to retrieve resources relevant to DDC 193. The SKOS classes – AND, NOT and OR[57] – may be applicable here.

8. Issues specific to UNESCO

UNESCO uses microthesauri. There are a total of seven microthesaurus headings within UNESCO as a whole. These have been included in the HILT database as top level terms. It has emerged however that some of these microthesaurus headings have duplicate preferred terms within lower levels of the microthesauri. If this was a consistent approach adopted by UNESCO it would be reasonable to delete the microthesaurus headings, using the duplicate preferred terms as terms to map to DDC as appropriate. It appears that some of the microthesaurus headings are unique however, creating uncertainty in how to handle them for HILT purposes. In addition, relationships exist beween microthesaurus headings and preferred terms. Such relationships would be lost if deleted and if not duplicated between preferred terms. Should we encode microthesaurus headings (MTs) as such? How would HILT handle these?

5. SKOS MVS Developments

During November 2007 the following developments have occurred:

1) majorMatch and minorMatch have been deprecated.
SKOS ISSUE-39 (http://www.w3.org/2006/07/SWD/track/issues/39) proposes to introduce skos:overlappingConcept

skos:overlappingConcept may be expanded to accommodate specific weightings if required. Whereas majorMatch and minorMatch indicated semantic overlap of <50% and >50%, further specification of the extent of overlap can be expressed using skos:overlappingConcept e.g. 0-30%.

2) SKOS ISSUE-39 also proposes to introduce skos:equivalentConcept in place of exactMatch.

The status of skos:related remains open.

3) Classes:
skosm: AND
skosm: OR
skosm: NOT

have been replaced by skos:Intersection; skos:Union; skos:Negation respectively.

4) ISSUE-39 states that instances like the use case proposed by HILT are probably out of SKOS core scope, that is "mapping links focused more on the conceptual mapping process than the essence of the conceptual mapping result".

5) Summary:

New set of mapping properties:
broader

[57] skos:Intersection; skos:Union; skos:Negation

narrower
related?[58]
equivalentConcept – replacing exactMatch?[9]
overlappingConcept – replacing major/minorMatch

New set of mapping classes:
skos:Intersection
skos:Union
skos:Negation

In light of the above developments, previous HILT mapping work will be revisited. Mappings previously denoted as major/minorMatch will be replaced by overlappingConcept.

6. Discussion

Issues documented in this paper will be tabled for discussion at the HILT IV Steering Group meeting scheduled for February 2008.

3/12/07

[58] Awaiting clarification from the SKOS mailing list on whether related is to be introduced and whether exact Match has been replaced

HILT IV Evaluation

Ian Ruthven
Department of Computer and Information Sciences
University of Strathclyde

Introduction

HILT serves an important function and one that will be of increasing use as our access to information resources becomes increasingly ‘anytime, anywhere’ and decreasingly mediated by informed information professionals. The HILT approach has the potential to significantly reduce the complexity of accessing online information resources but also opens other opportunities in helping online users to learn about the structure of organised resources. The challenges for the HILT service, of course, are many: the services that might use HILT are diverse and cannot be known in advance; the people organising these services have mixed technical ability; and the end-users of these services have different needs, varying expertise in online information access and highly variable motivations. As such the HILT approach of concentrating on the core technology and intellectual mapping of subject terms is sensible to allow flexibility to the end-user services.

To provide an overview of the HILT architecture, the model involves a central SRW/U server, a SOAP server (which interacts with databases) and SRW clients/web browser. Non-proprietary standards including the SRW/U have been adopted, enabling services to develop their own local user interfaces, capable of connecting to the HILT SRW/U server and employing HILT mappings within their local environment(s). Completing the model are two databases; one holding records of collections and services within the JISC (Joint Information Systems Committee) Information Environment and the other holding terminologies data including mappings from satellite schemes to the central DDC spine. The response to a user query is wrapped in SKOS (Simple Knowledge Organization System) by the SOAP server. The diagram below shows each of the different HILT modules and how they interact.

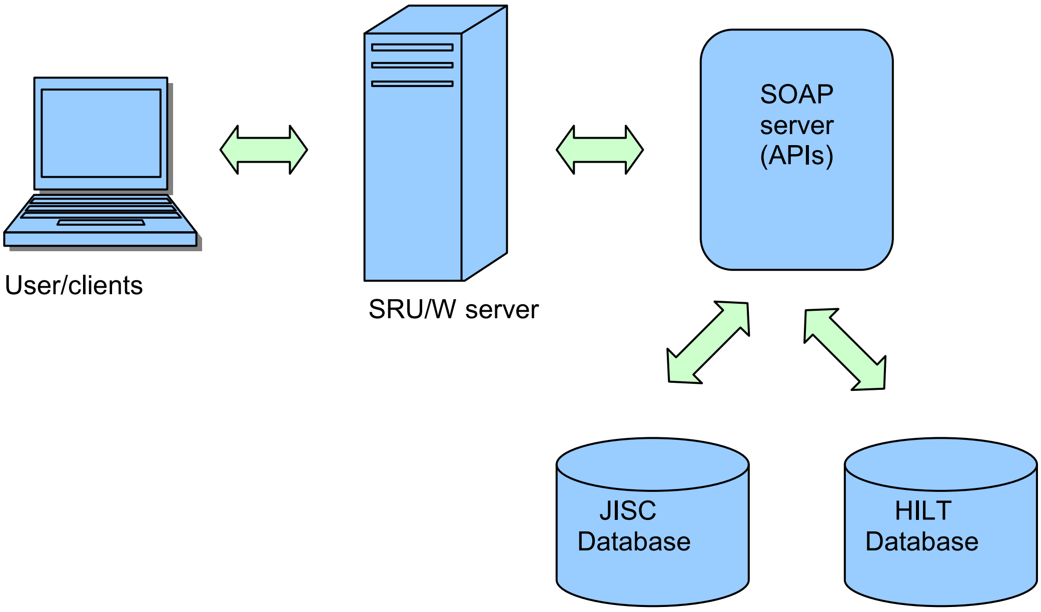


Diagram 1: Overview of HILT architecture

This evaluation looked at the overall HILT approach, concentrating on the utility of the HILT terminology server functions. It did not attempt to evaluate the quality of the HILT mappings as such; rather it considered the HILT functions and mappings within the context of stereotypical information searches. This report will outline the methodology used, its limitations, findings and summary recommendations.

Overall methodology

The evaluation was primarily informed by a user study in which data gathering was split into two sessions. The first session gathered quantitative data using selected topics from the HILT server. The second session gathered qualitative data, from interviews, data to help analyse the quantitative data from the first session.

The evaluation only used mappings available from the HILT server. In the absence of indexed or categorised documents there are certain measures which were difficult to assess directly, e.g. recall of documents or recall of mapped terms. However, the methodology did consider issues of precision in how representative users judge the mappings for different types of task.

The study complied with University regulations on ethics and participants were paid a small fee for participation to be paid from the consultancy fee.

Participants

24 participants were drawn from the MSc classes in Information Management and Information and Library Studies and BSc in Computer Science at the University of Strathclyde. All participants took part in the first session and 6 participants took part in the second session. This cohort meant the involvement of a medium sized group of literate adult searchers who are familiar with online searching and classification schemes. The participants were not informed about the specific purpose of the evaluation, rather each was asked to assess a set of HILT mappings that might be returned in response to a search on a local information resource such as a digital library. No documents, images or other information items were available for assessment under the HILT mappings returned: the participants were asked to judge how useful the mappings might be as a place to start investigating a collection of information objects. The participants were not informed of any of the indexing schemes being used in the study. Each participant was given £10 for participation in each session.

Tasks

The data gathering was centred on simulated work task situations [Bor03]. These are representative tasks that might be undertaken by the participant group. Each simulated work task situation asked the participants to judge HILT mappings with respect to a real-world task. The use of simulated work task situations have been shown to give more realistic assessments of utility of information as they contextualise the assessment of information according to realistic situations.

Each simulated work task situation has two parts – the task (e.g. collecting resources for an essay) and a topic (e.g. art deco).

Three simulated work task situations were developed:

- Finding images on a topic which was intended to simulate a well-formed precise information need in which the participant knew which topic was seen as important and this topic mapped exactly onto at least one HILT mapping
- Finding resources for a personal essay which was intended to simulate an ill-formed information need in that the task specifies particular information that is desired but the selected HILT mappings only partially match topic.
- Finding resources for a school project which was intended to simulate an exploratory information need, a situation in which there is a general need for information but where the participant would need to decide, and prioritise, useful places to start browsing/searching.

The topics of the task (the HILT mappings) were selected to investigate a range of topics from the Dewey classification scheme. Topics were chosen to balance a range of criteria:

1. to select topics that had sufficient mappings to be usable.
2. to select a mixture of topics of varying familiarity to the participant group. The participants were primarily arts graduates and computer science undergraduates. Topics such as fairy tales, learning, or digital libraries were felt to be fairly familiar topics to these groups, as confirmed by the output of the study; topics such as poisonous plants were predicted to be ones where the participants would have some general knowledge but lack specific knowledge and a topic such as modernism was predicted to be one in which participants would have low general knowledge.
3. to include topics that had a range of shallow and deep mappings
4. to include topics that covered as many of the indexing schemes as possible
5. to include some possibly ambiguous topics (e.g. memory which can be human or computer memory, or roses which can be the plant or type of wine)
6. to avoid topics that might cause distress to students. Some topics, particularly in the mental health area, were ideal on the above criteria but carried a risk of negative reactions from students suffering from the conditions. To minimise this we stuck to neutral topics such as anxiety, learning and memory.

The final chosen topics were (1) poisonous plants, (2) anxiety, (3) learning, (4) roses, (5) modernism, (6) memory, (7) art deco, (8) psychoanalysis, (9) new testament, (10) digital libraries, and (11) fairy tales.

Each participant was given two tasks of each work task type and each task had a different HILT topic (see section 5). Participants were grouped into two sets of 12 users, each set investigated 6 HILT topics: set 1 investigated topics 1, 2, 4 5, 6, and 7 and set 2 investigated topics 1, 3, 8, 9, 10, and 11. Topic 1 was included in both sets as a control to check for differences between the two sets of participants.

Each topic was run on the HILT server and all returned mappings extracted. Each set of mappings was analysed and 15 were selected to be assessed by the participants. For each topic a number of exact mappings and a number of partial mappings were chosen. The number of exact/partial mappings in the HILT output naturally varied depending on the topic chosen and the number of mappings available within the HILT server. For topics where there were a large number of mappings samples from within each class of mapping were chosen. For example the topic fairy tales returned a long list of fictional characters from which were chosen a small representative set. Within each set of mappings a small number of 'fictional' mappings were also included to test whether the participant could distinguish genuine from false mappings.

HILT lends itself to both browsing and querying approaches, although browsing interfaces are particularly well supported through the hierarchical relationships. One interesting question was how knowledge of the relative structure of the mappings influences which mappings are seen as useful, i.e. how can mappings be assessed by the mapping or does the participant need to know where in the hierarchy the mapping occurs? This was tested by two variants of the data collection form used. Mappings were either presented as presented by the HILT server (referred to as unstructured display) or presented with upper levels of the Dewey classification scheme (referred to as structured display). Figure 1 shows an example from the poisonous plants topic: the structured display shows the topics with the mapping being assessed shown in bold type and the upper levels of the Dewey classification scheme in non-bold. If a mapping was particularly deep then only sufficient upper levels were shown to contextualise the mapping being assessed. 'Sufficient' here was defined on a case-by-case basis, but in most cases this consisted of showing the upper two levels. In the non-structured information display only the bold mapping was shown.

botany: economic botany: poisonous plants
pharmacology: toxicology: poisons derived from plants and microorganisms
invertebrates: poisonous invertebrates
animal husbandry: poisonous food animals

Figure 1: Structured information display

Data collection – session 1

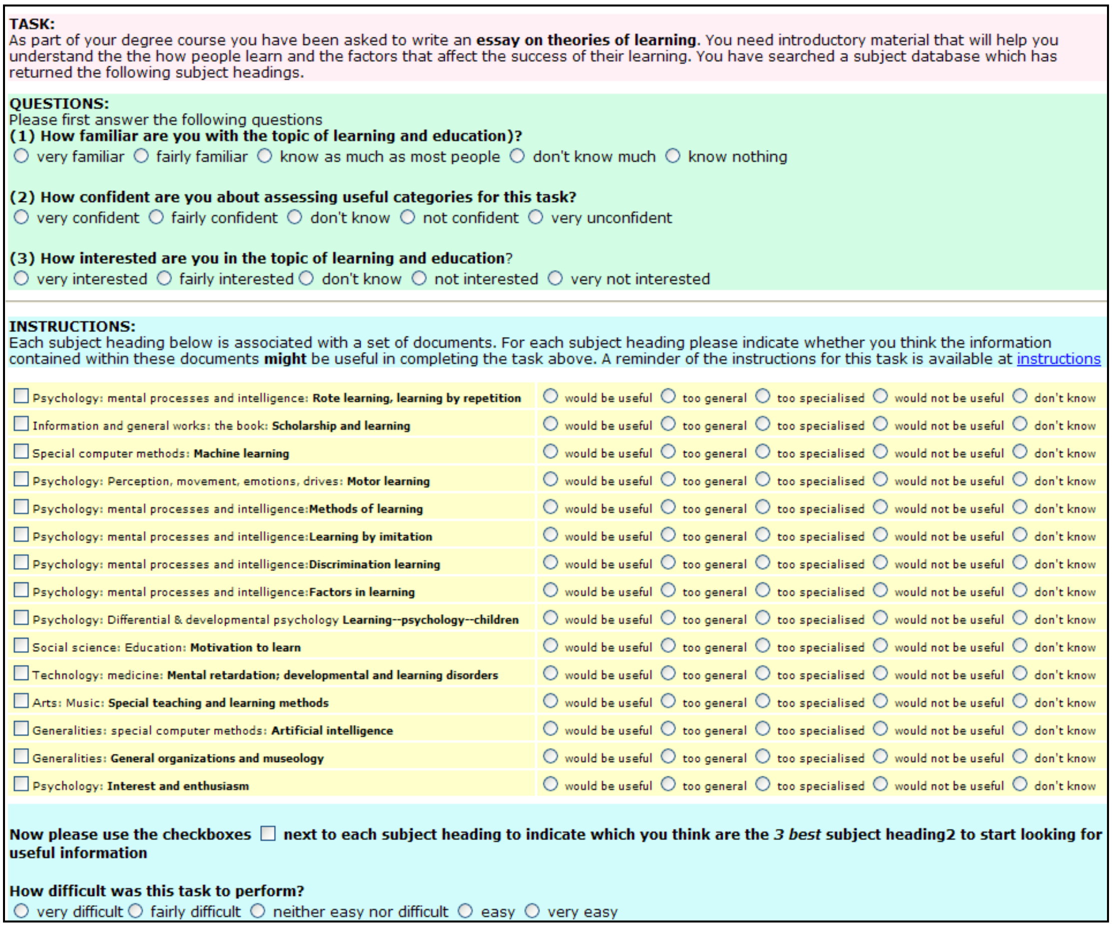
Quantitative data was collected through web forms. Each participant was given a set of instructions on the task, a practice task and guidelines on their rights as part of the ethics requirements of the University.

Session 1 took at most one hour in which the participants were asked to complete six web forms: three using structured information displays and three using unstructured displays. The order of HILT topics was rotated across participants within each set. The allocation of topic to experimental condition (type of simulated work task situation and structured/unstructured information display) was rotated across the participants so that each topic was used in an equal number of simulated situations and equal number of structured/unstructured displays. Each form was completed in turn and results emailed for analysis. The forms and questions were pilot tested with a sample group before being issued to the main participants.

Each form presented the participant with a simulated work task situation and asked the participants about their personal relationship with the topic: how familiar was the topic, how confident were they about their ability to assess the mappings and how interested they were in the topic. These were asked to help select participants for session 2.

Each participant was then offered 15 HILT mappings and asked to assess how useful the information associated with the mappings might be. The instructions asked them to rate the mapping as either leading to useful information, not leading to useful information, whether the mapping was to general, or too specific. A 'don't know' option was included in case the participant could not judge the mapping.

The final question on each form asked the participant to rate how difficult was the process of completing the form.



TASK:
As part of your degree course you have been asked to write an **essay on theories of learning**. You need introductory material that will help you understand the the how people learn and the factors that affect the success of their learning. You have searched a subject database which has returned the following subject headings.

QUESTIONS:
Please first answer the following questions
(1) How familiar are you with the topic of learning and education)?
○ very familiar ○ fairly familiar ○ know as much as most people ○ don't know much ○ know nothing

(2) How confident are you about assessing useful categories for this task?
○ very confident ○ fairly confident ○ don't know ○ not confident ○ very unconfident

(3) How interested are you in the topic of learning and education?
○ very interested ○ fairly interested ○ don't know ○ not interested ○ very not interested

INSTRUCTIONS:
Each subject heading below is associated with a set of documents. For each subject heading please indicate whether you think the information contained within these documents **might** be useful in completing the task above. A reminder of the instructions for this task is available at instructions

□ Psychology: mental processes and intelligence: **Rote learning, learning by repetition**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Information and general works: the book: **Scholarship and learning**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Special computer methods: **Machine learning**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Psychology: Perception, movement, emotions, drives: **Motor learning**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Psychology: mental processes and intelligence:**Methods of learning**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Psychology: mental processes and intelligence:**Learning by imitation**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Psychology: mental processes and intelligence:**Discrimination learning**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Psychology: mental processes and intelligence:**Factors in learning**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Psychology: Differential & developmental psychology **Learning--psychology--children**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Social science: Education: **Motivation to learn**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Technology: medicine: **Mental retardation; developmental and learning disorders**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Arts: Music: **Special teaching and learning methods**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Generalities: special computer methods: **Artificial intelligence**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Generalities: **General organizations and museology**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know
□ Psychology: **Interest and enthusiasm**	○ would be useful	○ too general	○ too specialised	○ would not be useful	○ don't know

Now please use the checkboxes □ next to each subject heading to indicate which you think are the *3 best* subject heading2 to start looking for useful information

How difficult was this task to perform?
○ very difficult ○ fairly difficult ○ neither easy nor difficult ○ easy ○ very easy

Figure 2: HILT evaluation form

5.1 Results – session 1

The participants declared a positive interest in about half the topics chosen, a positive familiarity with around half the topics and at least average confidence in assessing most of the topics. There were no significant differences in these variables between the conditions of structured/unstructured information display and type of simulated situation.

Table 1 gives the overall number of HILT mappings judged according to the 5 categories useful/not useful/too general/too specific/don't know and split between the two categories of presentation. For the overwhelming majority of mappings the participants were able to make a judgement on the potential utility of the mappings as indicated by the low use of the 'don't know' category. The actual numbers shown in Table 1 are not of immediate importance as they are based on a small sample of the HILT database and some mappings were deliberately chosen to be ambiguous or deliberately false mappings. The accuracy of judgments will be considered below in section 6. What we can assess here are relative differences between the two displays – structured information display and unstructured information display and type of simulated work task.

structured	useful	not useful	too general	too specific	don't know
presentation	438	324	90	144	84
non-structured	useful	not useful	too general	too specific	don't know
presentation	335	330	152	155	108

Table 1: Number of HILT mappings assessed under each response category

In the structured presentation case there were significantly more mappings classified as useful than as not useful (p=0.004[59]), than as too general (p=0.000), too specialised (p=0.00) or don't know (p=0.000). In non-structured presentation there were significantly more mappings assessed as useful than as being too specialised (p=0.001), don't know (p=0.000) or too general (p=0.001) but here no significant difference between useful and not useful (p=0.704).

When no information was given on the place of the mapping with the subject hierarchy it appears to be more difficult to judge which mappings might be useful and there was a strong tendency to reject (not useful, too general, too specific) mappings that were seen as useful in the structured display– this was something followed up in session 2. This was reinforced when comparing across the two forms of display as participants judged significantly more mappings as being useful in the structured presentation case than the unstructured one (p=0.018), and significantly fewer judged as too general (p=0.004).

When comparing the categories across the types of simulated work task situation, Table 2, there were no significant differences between the tasks although the significance tests indicate that task might be seen to have an effect with larger numbers of participants. In particular when the information need was vague or ill-formed (exploratory task) fewer mappings were seen as not useful and when the information need was better formed (essay task) it was easier to make definite judgements on the utility of mappings, i.e. higher use of 'useful' and 'not useful' categories.

	useful	not useful	too general	too specific	don't know
essay	265	240	76	72	67
exploratory	253	194	97	113	63

[59] Using the non-parametric Wilcoxon test for statistical significance. Values where p<0.05 (i.e. less than 5% chance that such a difference could be found in the same population) are typically taken to infer that there is a real difference caused by the variable, in this case the presentation of mappings.

images	247	219	73	116	65

Table 2: Number of HILT mappings assessed under each response category for each type of simulated work task situation

Data collection – session 2

Session 2 involved a sample of the participants from session 1 who were interviewed on their responses. The aim behind the interviews was to learn something about how the participants made judgements on the selected topics.

Each participant was taken through their web forms in order and asked how they arrived at a decision on each mapping, what additional information would be useful to make a decision and, in some cases, also shown the original HILT output, the alternate information display (structured/unstructured) or the results from searching the term on alternate indexing schemes. In this section I will summarise the main findings across the interviews.

6.1 Results – session 2

As shown in many assessment tasks in information retrieval and information science, prior knowledge was a key feature in the assessment of the HILT mappings. If a participant felt confident in their knowledge of a topic then the task was perceived to be easier. In tasks where the participants had less prior knowledge a common strategy was keyword matching – judging mappings as useful or not according to the presence of keywords from the simulated work task situation to judge the mappings, e.g. selecting mappings that contained the word modernism for the modernism tasks and rejecting the others. The presence of structural information allowed participants to be more confident in their decisions as they disambiguated possibly useful mappings, e.g. knowing that the mapping 'literature-history and criticism' was a sub-topic of modernism allowed participants to more easily judge the possible utility of the mapping.

Assessor openness was also an important variable. This is a general term taken to indicate an assessor's attitude to assessing information. Previous studies in the Information Seeking and Retrieval literature have identified this as a separate variable from assessor knowledge or interest in material being assessed, e.g. [RBE07]. Although the variable is yet not very well-understood in the literature it appears to be a useful one in understanding differences between assessor judgments. Briefly, people assessing information objects can be roughly categorised as liberal or conservative. Liberal assessors are more willing to take risks in assessment, opting to over-estimate the amount of relevant material or likelihood of relevant material. Liberal assessors will also tend to see possible utility in information objects not seen by conservative assessors. Conservative assessors are ones who take a narrow, precise view of assessment and will tend to under-estimate the amount of relevant information or likelihood of obtaining relevant information.

In the context of this evaluation, liberal participants were more open to judging as useful mappings that might be seen as incorrect by conservative participants. The 'roses' topic, for example, contained few obviously correct mappings but did include the mapping 'pageantry'. Liberal participants were willing to at least consider this mapping to be useful for the image task – especially in the absence of more useful mappings – on the grounds that roses may be used in pageantry, other participants rejected this mapping as unsuitable.

Anchoring to expectations. Assessments on the utility of the HILT mappings were often anchored to the participant's expectations of the result [Bla80]. That is, the participants' assessments of the mappings were relative to what they expected to be presented with. For example, on a topic such as anxiety the participants expected to be presented with general information on general anxiety or types of anxiety disorders, on a topic such as poisonous plants they expected to see general information on types of poisonous plants and specific groupings or examples of poisonous plants. If the mappings offered, as a group, met the participant's pre-conceptions of what mappings should be offered then the participants generally viewed the mappings more favourably than they did not. Similarly, if the mappings matched the participants' preconceptions then they found it an easier task to assess the mappings.

A second aspect of anchoring related to the level of abstraction reflected in the topic. Some topics were conceptual, such as modernism, anxiety, learning, whereas others, for example roses, poisonous plants, were more concrete in nature. In both cases participants wanted a mixture of general topics (partly to orientate themselves) but also, particularly in the case of concrete topics, precise mappings reflecting specific instances of topics, e.g. examples of poisonous plants, examples of art deco artists or modernist writers. There was a distinct preference for the topics that offered deeper mappings; not only did these topics offer a richer set of mappings related to the topic but also gave the participants more ideas on what areas were available and ideas for further investigation.

The structured display helped by offering some context against why some topics seemed unusual or ambiguous. Ambiguity itself was an interesting issue. For some topics, such as memory, the ambiguity between human memory and computer memory was easily resolved by the participants and because the ambiguity was easily resolved the ambiguity was not seen as negative – rather it informed the participants of alternative uses of the topic being mapped. For other topics resolving the ambiguity was more difficult if the mapping shown could conceivably be mapped to more than one distinct area of interest – the mapping 'plants' for example could mean botanical plants or industrial plants. Structural information often helped resolve this ambiguity.

Participants were asked to suggest mappings that might be have been useful to have been included. Their ability to do this part of the interview was constrained by their knowledge of the topic but they could describe what type of mappings they preferred.

- in topics where they had little prior knowledge of the topic they often could not give concrete examples of useful mappings but stated that what they preferred were mappings that helped them understand the structure of the classification space. Here they stated a preference for higher level mappings that used keywords shown in the simulated situation to help orientate themselves in the classification hierarchy but, in addition, more precise mappings to show how a concept is commonly understood. Topics such as learning or memory were seen as good examples of the kind of display preferred here.
- in topics where they had more prior knowledge they could suggest specific mappings that might be useful. In the case of the 'fairy tales' topic, a topic commonly understood by all participants, they suggested names of authors, popular characters in (Western) fairy tale traditions. The HILT partial mappings supported knowledge discovery here in that some mappings, e.g. 'Psychoanalytic systems', suggested new ideas to the participant that would not have been considered but other partial mappings such as 'Gods and goddesses--classical—literature' were viewed negatively on the grounds that they were too unrelated to the topic under consideration – or at least the participants felt that there were more obvious mappings that could have been shown instead.
- in topics containing deeper, precise mappings (such as anxiety, memory) the suggested new mappings were variants on those already shown. Noticeably the participants expressed a preference for those topics for which deeper mappings were available on the grounds that the mappings were more obviously related to the topics being investigated, gave a more intuitive set of mappings and supported the participants own understanding of the topic.

None of the participants commented on the 'false' mappings introduced. The participants were asked about the display of the mappings in particular the order of the mappings and presentation style. All participants who expressed a preference preferred the structured information display on the grounds that it allowed them additional information to judge the possible utility of the mappings and also that it allowed them to predict what type of information might be indexed under the mapping.

Participants were asked about tag clouds as a, currently, popular method of accessing information. The data collection method asked participants to imagine the mappings being shown as the output of a search on an information resource such as a digital library. For such an information resource, none of the participants thought tag clouds would be useful or desirable, on the grounds that what they required was not a relatively small set of tags but more detail on the information structure of the digital library, effectively what information was available and how to obtain the information. Rather the participants expressed a preference for what might be termed facetted browsing or approaches that allow for orienteering behaviour, i.e. that allow some form of navigation towards a goal that is informed by the information made available. This preference was particularly strong when participants had low knowledge of the topic being searched.

A common criticism of the information display was the unusual ranking of mappings. The mappings were shown in the order in which they were output from the HILT server. This prioritised exact mappings before partial mappings. However, it also meant that mappings from different topical areas, in the case of ambiguous mappings, were intermingled. This was more obvious in the structured information display as the upper–level classification labels made it clear that the mappings were from different parts of the Dewey classification scheme. The participants expressed a clear preference for semantically grouping the output rather than displaying by degree of match, particularly in the case of ambiguous mappings.

A second criticism of the ranking of mappings was that, for compound topics such as 'poisonous plants', some form of ordering could be applied to the mappings such that more specific mappings – as reflected by more specific parts of the topic – would be shown first.

The use of partial mappings was interesting. As noted above partial mappings were positive in that they supported some form of knowledge discovery by the participants offering novel suggestions and supporting the participants understanding of the topic. However, the partial mappings were often not seen as a good substitute for deeper mappings. The topics for which deeper mappings were available were almost universally seen as more useful as a set and as individual entry points to investigating a topic. Whether a high-level or deeper-level mapping was seen as the *best* entry point depended partly on the task given. The essay task encouraged searchers to look for specific pieces of information and so deeper mappings were often seen as most useful. The other tasks were more vague and so higher-level mappings were seen as useful starting points.

For a small number of topics, the ones where there were few exact matches, the partial mappings were compared against the filtered sets from the alternative indexing schemes, e.g. comparing 'roses' on LCSH and on the whole HILT mappings. Here the filtered mappings were seen as much more relevant and intuitive than the partial mappings.

Although no significant differences between task types were found in the quantitative analysis, the interviews reflected differences in what information might be required for each task and how this affected the assessment process. The essay task deliberately gave more specific details on what information was required, for example the roses task asked participants to find information on varieties and care of garden roses whereas the images task simply asked participants to find images of garden roses. The mapping 'Diseases of roses' was often judged relevant to the former task but not the latter – as it was felt that images of diseased roses would not be suitable.

Summary and recommendations

In the following sections I will outline some recommendations based on this user study. Before this it is appropriate to acknowledge some limitations from the study. This study involved 24 participants of university level education and so may not be representative of the whole population who might use information services supported by HILT. However, the participants do form a group of people who are within the target group of such services, that is they are people who regularly make use of a wide variety of information services that might reasonably be facilitated by the HILT approach. Most, although not all, had English as a first language. No participant reported difficulty with the task descriptions within the study.

The topics chosen were deliberately intended to replicate people searching for both familiar and unfamiliar topics. End-users don't always search on topics with which they are familiar and the same information services are intended to support people who are topic experts and topic novices. An alternate methodology would have been to personalise search topics to individuals within the study or only ask participants to search on real information needs. I opted for the current approach to allow for some statistical generalisation across the participants searching on the same topics.

A further limitation was only showing a subset of mappings for some topics. However, the aim of the evaluation was not to evaluate the HILT mappings as such but to understand the decisions made upon the mappings that might influence how the HILT service currently operates and considerations for future. Naturally the choice of topics and individual terms that were shown to the participants does affect the values shown in Table 1 and Table 2 so these should not be taken as absolute values on the quality of the HILT mappings.

Finally participants were not asked to interact with a real search interface, although this would have allowed the exploration of searcher interaction as opposed to simply assessments of mappings. The purpose of HILT is to support end-user information services, not to provide interfaces for such services. Although we can imagine what such an end-user interface might look like, interfaces are complex constructs and participant's behaviour is affected by the specific details of the interface provided. Thus the specifics of any one interface can influence decisions on the appropriateness of the mappings shown. Although some of the findings indicate what features it might be useful to have in an end-user interface, I felt it more appropriate to abstract from the interface itself and only ask participants to judge the potential utility of mappings against a specific information seeking task.

7.1 Level of mappings

The participants expressed a strong preference for the presence of deep mappings. This is not surprising in itself: end-users always display a preference for precise, high-quality output from any information service. The real question is of balance between this quality of output from HILT against the obvious additional cost in providing the desired quality. This is a difficult question to answer but what the study does indicate is that deeper mappings were seen as most useful and those who benefit most from the deeper, more precise mappings were those that knew least about the topics they were searching or who were searching on more vague topics. This is not a trivial benefit from the deeper mappings; as noted above, we do not always search on topics we understand or for precise information needs. In such cases, people accessing online services facilitated by HILT will still require support. A natural benefit of investing in HILT as a centralized service is that more of this support comes automatically, resulting in cost-savings by multiple end-user services. As indicated in the interviews an orienteering approach to browsing, essentially breaking down an investigation of a information resource into small steps facilitated by information presentation, is a useful way of learning about an information resource. This is an increasing common method of interaction, particularly in web environments [TAAK04]. The structured display helped this form of interaction but it was made more useful by the deeper mappings.

A second benefit of deeper mappings relates to the size of the information resource being indexed. As information resources grow the number of items indexed under each classification heading will also grow. This raises interesting questions about how the number of items classified under an index heading relates to how likely an end-user is to find particular information items. Research in web environments suggests that many end-users use quite superficial investigation techniques to analyze the results of a search request. Recent research in Information Retrieval systems, e.g. [AV08], also indicate that individual information objects may become 'lost' in a system because they are not retrieved in a sufficiently high position in the output list to be visible to end-users. At present we cannot tell whether such results will have an impact on the services facilitated by HILT – we cannot, for example, estimate what size of collections may use HILT – but it is worth bearing in mind that HILT will be providing services for end-users whose method of interaction will be influenced by information seeking behaviours learnt in other online environments. Such behaviours may be superficial and such end-users may reject services that don't offer 'quick' solutions.

It is recommended that HILT continue with the deeper level mappings. At least, as in the current approach, continue these mappings within certain areas to allow for experimentation. It would also be useful at some point to experiment with indexing existing collections to gain a feel for how size of collections affect the 'findability' of information objects.

7.2 Information display

As noted above the participants preferred high-quality output. This is not simply the index terms themselves but also the contextual information available to interpret the output. In section 5.1 it was indicated that the presence of additional information on the terms themselves, in this case where the term occurred within the subject hierarchy, led to more terms being assessed as useful. The terms themselves were the same, what changed was the perception of the terms. This was reinforced in the interviews where participants felt more confident in deciding where they would start browsing based on additional structural information.

HILT does not supply interfaces to end-users, although the existing demonstrators do indicate the potential of HILT. It is recommended that given the level of development of the existing HILT demonstrators it would be worth investing some time in the presentation of the demonstrators themselves. They are functionally sound but some simple style sheets or presentation tools would give a better indication of the potential of HILT to end-user services.

7.3 Compound search terms

In the interviews several participants commented negatively on the output of searches on multiple word searches. In the case of 'poisonous plants', for example, the output will return exact matches but also keyword matches on 'poisonous' and 'plants' (including non-botanical uses of 'plant'). The participants' views on this were mixed. Whilst it was seen as sensible to return such partial matches, the participants commented that the ordering of the results (effectively the same order as results came from the HILT server) were not optimal in that mappings from different areas were intermingled, e.g. mappings describing biological plants, poisonous (plant-based or not) and non-botanical senses of plants were intermingled whereas a more useful ordering would group the mappings semantically so that users could eliminate large groups of 'incorrect' mappings. This only seemed to apply when using the *get-all-records* option. There are two arguments here – one could argue that such display-issues are the responsibility of the people who use HILT and HILT should not force such a grouping on the output. Alternatively it might be possible for the HILT output to be processed simply in order that mappings are organized by upper DDC numbers.

Semantically grouping mappings is difficult but it is recommended that the HILT team consider if simple groupings based on DDC numbers might organize the output into a slightly more useful form.

7.3 Use of SOAP

As noted in the introduction, HILT employs SOAP to wrap data for transfer. Recently there has been a move in internet systems development to using REST (Representational State Transfer). The question arises of whether HILT should maintain the SOAP development or consider moving to REST. The relative advantages of SOAP and REST is a very open question and informed as much by philosophical issues as technical ones. Many new web services and established ones are moving to REST on the grounds of simplicity and ease of starting a web service. REST uses URLs to exchange data; SOAP requires XML mark-up of objects. The biggest increase in REST users are new web service developers – REST allows for quick development of new services. There are large providers (such as Yahoo's web services) which use REST and Amazon has both SOAP and REST services. Established services such as Google so far have mostly stuck with SOAP.

The major advantage to REST is the lightweight development. However there are disadvantages not present in SOAP. SOAP allows for type-checking, and the xml formatted output is easier for parsing (REST output can be complex for end-users to parse and the output size can cause problems for URLs). SOAP is stable for future development, at least the near future, and it is difficult to judge exactly how the situation may change. However, given the nature of HILT, in particular that the output is typed, (possibly) highly structured, and that the end-users may not be highly technical, it is recommended that HILT continue development with SOAP rather than moving to another transfer protocol.

References

[AV08] Azzopardi,L. Vinay,V. (2008). Document Accessibility: Evaluating the access afforded to a document by the retrieval system. Evaluation Workshop at the European Conference in Information Retrieval.
[Bla80] Blair, D. C. (1980). Searching biases in large interactive document retrieval systems. Journal of the American Society for Information Science, 31, 4, 217-277.
[Bor03] Borlund, P. (2003). The IIR Evaluation Model: a Framework for Evaluation of Interactive Information Retrieval Systems. *Information Research*, 8, 3, paper no. 152.
[RBE07] Ruthven, I., Baillie, M., and Elsweiler, D., (2007). The relative effects of knowledge, interest and confidence in assessing relevance. Journal of Documentation. 63 4. 2007. 482 - 504.
[TAAK04] Teevan, J., Alvarado, C., Ackerman, M. S., and Karger, D. R. (2004). The Perfect Search Engine Is Not Enough: A Study of Orienteering Behavior in Directed Search. In Proceedings of the ACM Conference on Human Factors in Computing Systems (CHI '04)

Appendix H: Bid for Terminologies Interoperability Centre Scoping Study

Cover Sheet for Proposals

(All sections must be completed)

JISC

Shared Infrastructure Services Programme

Name of Lead Institution: Centre for Digital Library Research, University of Strathclyde

Name of Proposed Project: Scoping Study: Terminologies Interoperability Centre

Project Director: Dennis Nicholson, CDLR (Email: d.m.nicholson@strath.ac.uk)

Project Partner(s): EDINA, University of Edinburgh

Full Contact Details for Primary Contact:
Name: Emma McCulloch
Position: Research and Development Co-ordinator
Email: e.mcculloch@strath.ac.uk
Address: CDLR, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow, G1 1XH.
Tel: 0141 548 4752/4753

Project Length: 18 months

Start Date: June 2009 **End Date:** November 2010

Total Funding Requested from JISC: £276036

Funding Broken Down over Financial Years (April - March):

June 2009 – Mar 2010	Apr 2010 – Nov 2010	April 2011 – Mar 2012
£159,392	£116,644	

Total Institutional Contributions: £69009

Outline Project Description

The study will investigate the user and service sustainability requirements of a Terminologies Interoperability Centre based on the outcomes of the HILT project. In the past two years, HILT has:

- Successfully built and tested a range of pilot M2M terminology services based on SRU/W, SOAP, and SKOS, and a database of terminologies and pilot mappings.
- Successfully built and tested an embryonic toolkit to help information services' technical staff to embed M2M interactions in user interfaces to improve subject retrieval, browse, and deposit services.
- Conducted various practical experiments to successfully embed terminology service interaction into JISC community services to create operational pilot subject browse and retrieve enhancements for service users.
- Developed a generic terminology services architecture that will (over time) permit the HILT services to grow and improve by incorporating other terminology services being developed elsewhere.
- Determined that a terminology services registry is a key element of this architecture and that the core functionality required to build and run such a registry is already inherent in the HILT pilot services.
- Developed a staff skills and experience set appropriate to the above.
- Determined, in conjunction with JISC, that the best general option for a sustainable operational Shared Infrastructure Service based on project outcomes is a Terminologies Interoperability Centre offering a mix of standard 'plug and play' type M2M and toolkit facilities free at the point of use, including a training portal and an associated terminology services registry, more flexible, charged-for, specially-scoped versions of these, tailored to the needs of individual services and institutions, and ongoing development via a mix of collaboration and externally funded R&D, as well as through JISC support.

This scoping study will provide a well-researched evidence base that will inform and guide a future 'soft launch' of a Terminologies Interoperability Centre (TIC) by:

- Putting in place service quality infrastructure to support the work of the Centre, including further development and testing of the components from HILT IV and work on a pilot terminology services registry. This will ensure that the standard services offered at the soft launch will be robust and usable in a range of JISC service and user environments.
- Determining service user and end user needs via iterative feedback from hands-on experience, utilising outcomes in TIC scoping and soft launch plans, creating mechanisms for an ongoing assessment of such needs, and identifying specific players to work with TIC during the soft launch period.

- Scoping in detail what free and charged-for services the Centre should offer and what they would cost.
- Producing a bid for TIC start-up costs, a programme of works, and a well-researched Sustainability Plan.

1. Appropriateness and Fit to Programme Objectives and Overall Value to the JISC Community

The proposed Centre will help JISC meet its Strategic Priorities[60] in these areas:

Key Deliverable	Contribution of project
Aim 1, Priority 3	The scoping study will determine how a future Centre might best meet community needs in respect of cross-searching between JISC institutions and the NHS where such a need exists and where different subject description terminologies are in use
Aim 2, Priorities 1,4	The scoping study will determine subject searching and subject description interoperability needs in respect of cross-working between institutions and national services in the area of e-learning
Aim 3, Priorities 1,2,3,4	The scoping study will determine terminological interoperability needs in the research community in respect of robust central services to facilitate high quality research, particularly in respect of interoperable and scalable services to support sharing and accessibility where there is a subject interoperability element – e.g. across institutional repositories
Aim 4, Priority 2(2)	The scoping study will work with institutions in the area of improving subject access to information across institutional services such as Library Opacs and institutional repositories thereby helping them to effectively utilise pervasive ICT
Aim 5, Priority 3	The study will determine how the Centre can best contribute to JISC aims in the area of Business and Community Engagement(BCE) by aiming to work with groups such as the Strategic Content Alliance (SCA)to make it relevant and accessible to a BCE audience

1.1 Background

1.1.1 Background: Overview of the Requirement

Ensuring that HE and FE users of the JISC IE can find appropriate learning, research and information resources by *subject search and browse* in an environment where most national and institutional service providers – usually for very good 'local' reasons - use different subject schemes to describe their resources is a major challenge facing the JISC domain (and, indeed, other domains beyond JISC). Under the auspices of the HILT[61] project, JISC has been investigating mechanisms to assist the community with this problem and thereby optimise the value obtained from expenditure on content and services by facilitating resource sharing to benefit users in the learning and research communities. The Terminologies Interoperability Centre (TIC) scoped in the study proposed in this bid would help institutions and national and regional services to better serve their users in the various ways described below under *Background: The Value of These Services to the JISC Communities*. In consultation with JISC, it has been determined that the best general option for a sustainable operational Shared Infrastructure Service based on HILT project outcomes is a Terminologies Interoperability Centre offering a mix of standard 'plug and play' type M2M and toolkit facilities free at the point of use, including a training portal and an associated terminology services registry, more flexible, charged-for, specially-scoped versions of these, tailored to the needs of individual services and institutions, and ongoing development via a mix of collaboration and externally funded R&D, as well as through JISC support. The aim of the present project is to provide a well-researched evidence base that will inform and guide a future 'soft launch' of a Terminologies Interoperability Centre by scoping out the detail of a sustainable mix of JISC-funded and charged for services designed to help meet JISC strategic aims in respect of serving its stakeholders. The basic outline of what is now needed in respect of addressing subject interoperability issues is known. What is needed now is a clear, evidence-based, indication of what is required in detail to ensure a robust, sustainable service sensitive to community needs and the strategic aims of the JISC. Amongst other things, this requires intensive work in which specific JISC-related user communities (including both service and end users) are given hands on experience of the terminological tools as they become available as part of an iterative and ongoing approach to understanding, and implementing solutions to, user needs in specific communities.

[60] http://www.jisc.ac.uk/aboutus/strategy/strategy0709/strategy_aims.aspx
[61] Http://hilt.cdlr.strath.ac.uk/

As with HILT Phase IV, the project will require the expertise of participants at CDLR[62] and EDINA[63], and of the HILT terminology advisors, together with some ongoing liaison with MIMAS[64] who run IESR[65] and Intute[66]. It will also require ongoing support from OCLC[67] in respect of allowing the use of the electronic files of DDC[68] and of LCSH mappings to DDC[69].

NB Arising out of JISC-funded work with UK user communities in the first phase of HILT, the project's original focus as regards solving subject interoperability problems was on mapping between subject schemes via a Dewey Decimal Classification spine. Mapping is an established and effective approach (see, for example, Mayr and Petras, 2008) and has attracted significant resourcing in other European countries (see, for example Agosti et al., 2007; Mayr and Petras, 2008). However, HILT now has an architectural approach that allows it (a) to adopt a range of interoperability strategies appropriate to specific use cases and cost-benefit levels (e.g. expensive deep mapping where the problem and user area justifies this, or high-level or browse-based retrieval where significant costs would be less justifiable), (b) to incorporate in the model - for the benefit of the JISC user communities – solutions offered and funded by other players, whether they be based on mapping to a spine, scheme to scheme, or some variety of automated approach. Full references listed in Appendix *D.*

1.1.2 Background: Developments to Date

In the past two years, HILT has:

- Built and tested a range pilot M2M (SRU/W[70], SOAP[71], and SKOS[72]) based web services to deliver terminologies and terminology mappings to JISC and institutional information services, supporting the transparent enhancement of subject search facilities.
- Built and tested a database of terminologies (DDC[73], UNESCO[74], HASSET[75], IPSV[76], LCSH[77], MeSH[78], CAB, GCMD[79], NMR[80], AAT[81], SCAS, JACS[82],) and high level mappings to DDC (HASSET, IPSV, UNESCO, JACS).
- Built and tested an embedding toolkit that will offer information services the core software needed to begin building subject interoperability services for their users by interacting at M2M level with the above web services and database through routines embedded transparently in their service user interfaces - a programmer's toolkit to help build improved subject browse and retrieve facilities.
- Conducted various practical experiments to successfully embed terminology service interaction into JISC community services to create operational pilot subject browse, retrieve, and deposit enhancements for service users.
- Developed a generic terminology services architecture that will (over time) permit the HILT services to grow and improve by incorporating terminology services being developed elsewhere.
- Determined that a terminology services registry is a key part of this architecture and that the core functionality required to build and run such a registry is already inherent in HILT pilot services.
- Developed staff skill sets and experience associated with the problems of subject searching and subject interoperability within and across information services using different subject terminologies, with the best approaches to mapping new schemes into the database, and with an associated distributed architecture to permit the ready integration of new services into the JISC and global subject interoperability landscape.
- Determined in conjunction with JISC (see Appendix A) that the best general option for a sustainable operational Shared Infrastructure Service based on project outcomes is the kind of Terminologies Interoperability Centre to be scoped in the present project.

[62] Centre for Digital Library Research (CDLR): http://cdlr.strath.ac.uk/
[63] EDINA: http://edina.ac.uk/
[64] Manchester Information & Associated Services (MIMAS): http://www.mimas.ac.uk/
[65] Information Environment Services Registry (IESR): http://iesr.ac.uk/
[66] intute: http://www.intute.ac.uk/
[67] OCLC Online Computer Library Center: http://www.oclc.org/
[68] Dewey Decimal Classification (DDC): http://www.oclc.org/dewey/
[69] LCSH to DDC mappings: http://www.oclc.org/asiapacific/zhcn/dewey/updates/numbers/default.htm
[70] Search/Retrieve Web Service (SRU/W): http://www.loc.gov/standards/sru/
[71] SOAP: http://www.w3.org/TR/soap/
[72] Simple Knowledge Organization System (SKOS) Core: http://www.w3.org/2004/02/skos/
[73] Dewey Decimal Classification (DDC): http://www.oclc.org/dewey/
[74] UNESCO Thesaurus: http://www2.ulcc.ac.uk/unesco/
[75] Humanities and Social Science Electronic Thesaurus (HASSET): http://www.data-archive.ac.uk/search/hassetSearch.asp
[76] Integrated Public Sector Vocabulary (IPSV): http://www.esd.org.uk/standards/ipsv/
[77] Library of Congress Subject Headings (LCSH): http://authorities.loc.gov/
[78] Medical Subject Headings (MeSH): http://www.nlm.nih.gov/mesh/
[79] Global Change Master Directory (GCMD): http://gcmd.nasa.gov/Resources/valids/keyword_list.html
[80] National Monuments Record Thesauri (NMR): http://thesaurus.english-heritage.org.uk/
[81] Art & Architecture Thesaurus (AAT): http://www.getty.edu/research/conducting_research/vocabularies/aat/
[82] Joint Academic Coding System (JACS): http://www.ucas.ac.uk/figures/ucasdata/subject/

1.1.3 Background: The Value of These Services to the JISC Communities

The pilot services that will provide the backbone of the proposed Terminologies Interoperability Centre will enable national, institutional, and other information service providers to access terminological and interoperability data at a machine to machine level that will allow them to enhance their own services in a variety of ways, including, but not necessarily limited to, the following:

- Improving recall in a subject search of one or more databases by enriching the set of terms known to a user by providing synonyms and related terms.
- Providing the best terms for a subject search in a remote service that uses a subject scheme unfamiliar to 'home service' users (or in a cross-search of a group of such services).
- Taking a user's subject term and using it to identify relevant information services via registries such as IESR (http://iesr.ac.uk/).
- Generating an interactive browse structure where a scheme is arranged hierarchically.
- The ability to send a term from a chosen subject scheme and receive back data on broader terms, narrower terms, hierarchy information, preferred and non-preferred terms, and so on.
- Providing cataloguing staff with information on subject schemes and inter-scheme mappings to assist in metadata creation.
- A spell-check mechanism to assist user searching.
- A service to assist user search formulation by providing information on search terms entered (e.g. what the term means, whether it has alternative meanings, whether there are synonyms that might be useful in a search and so on).

NB HILT IV work with the members of the HILT Collaborators email list showed that there was a good deal of interest from the community in these services. Twenty-nine people joined the list which was set up specifically to look at embedding work of this kind. A questionnaire asking what kinds of uses were of interest was answered by sixteen people. All indicated an interest in at least one of the above potential ways of using HILT, most indicated an interest in three or more ways. In addition, two services (CAIRNS distributed catalogue and Intute) did some pilot embedding work in HILT IV and other work of this nature is planned as part of a small extension to the project running till May 2009.

1.1.4 Background: Conclusions regarding the best route to a useful and sustainable service

Based on an understanding of issues in the area of subject interoperability as developed over a number of phases of HILT, and a discussion with the appropriate JISC programme manager (see Appendix A) on how to best fit future developments into JISC's strategic requirements and also provide a useful, reliable, and sustainable service for JISC-related communities, it was determined that the best route forward was to work towards the 'soft launch' of a Terminologies Interoperability Centre. Once launched, this would offer the community a mix of free and charged-for services and would be supported by a mix of JISC funding, externally earned income and R&D funding, and collaboration, and would provide:

- M2M and user-level access to terminology sets, the detail of those terminology sets, and data to facilitate interoperability between them.
- Open source software toolkits that would enable M2M interaction with HILT web services to be transparently embedded in the user interfaces of local, national and project information services.
- A basic architecture for terminology and interoperability services in the JISC Information Environment (and potentially beyond).
- A way of mounting and developing new terminologies and terminologies interoperability data required by the community, including JISC-specified work to facilitate improvements in subject access in and between the various JISC user communities and their external partners based on ongoing assessments of user and service needs.
- Advisory and machine to machine support services for projects, services, and other initiatives in JISC or JISC institutions where there is a subject description, subject retrieval, or subject interoperability facet.
- A JISC funded free advisory and training service on using the above facilities in local or national services and projects in the percentage of cases where this was relatively straightforward (plug 'n' play, but after a bit of advice and training).
- The development and hosting of a JISC-focused terminologies services registry.
- A charged-for consultancy service where the work and advice required by local and national services, projects, and organisations (both within and outwith JISC) was less straightforward or more sophisticated (because of client and client service circumstances and terminology sets)
- A portal for tools and training in the areas described above.
- A focus for wider work in the terminologies area, funded through a variety of sources, including non-JISC sources (for example though successful bids for European funding).

- Ongoing work to facilitate JISC involvement and leadership in a strategically important area with significance for both subject-based retrieval needs in research, learning, teaching, and elsewhere and semantic web developments.

The aim of the present project is to build on the progress made by scoping out the detail of what kind of Shared Infrastructure Service will best meet the aims of JISC and the needs of those it serves. In particular, it will seek to work towards the provision of reliable, sustainable services specifically tailored the specific needs of key JISC-related communities (Institutions as represented by librarians, the e-learning community, the NHS community, and JISC partners in the wider world such as SCA).

1.2 Aims and Objectives

The aim of this scoping study is to provide a well-researched evidence base that will inform and guide a future 'soft launch' of a sustainable Terminologies Interoperability Centre (TIC) as described in general outline above. Specific objectives are to:

- Conduct various 'service hardening' tests and developments in respect of existing SRU/W web services, associated SOAP functionalities and database routines, the embedding toolkit code, and a pilot terminology services registry, to ensure that the standard services offered at the soft launch will be robust, suitable, and usable in a range of JISC service and user environments.
- Work with four key communities (Institutions as represented by librarians, the e-learning community, the NHS community, and JISC partners in the wider world as represented by SCA) to determine service and end user needs via iterative feedback from hands-on experience, utilise user study outcomes in TIC scoping and soft launch plans, create mechanisms for an ongoing assessment of service and end user needs, and identify specific service and end user 'players' to work with TIC during a future soft launch period.
- Scope in detail what free and charged-for services should be offered and what they should cost.
- Produce a costed programme of works for TIC start-up, and a well-researched Sustainability Plan.

1.3 Project Outcomes

The key project outcome will be a well-researched evidence base that will inform and guide the future 'soft launch' of the proposed Terminologies Interoperability Centre in respect of both operational technical and support requirements and an in-depth understanding of user needs in key JISC communities. An important subsidiary outcome will be engagement with representative user groups that can be maintained beyond the project end via a social networking environment and an associated 'terminological tools showcase' giving service and end-users ongoing influence on TIC development.

2. Quality of Proposal and Robustness of Workplan

2.1 Overall Approach

The overall approach will be to use the methods described below in the work-packages section to move iteratively towards a clear understanding of the technical, support, user, set-up, and sustainability requirements of a robust and useful Terminologies Interoperability Centre aligned with JISC aims.

2.2 Project Outputs

Key output: a well-researched evidence base that will inform the future 'soft launch' and ongoing sustainability of the proposed Terminologies Interoperability Centre. **Subsidiary outputs:**

- Report on the technical and support requirements and cost structures of a sustainable Centre.
- Engagement with key JISC user groups, and a social networking environment with an associated 'terminological tools showcase' giving them ongoing influence on TIC development.
- A costed proposal and detailed programme of works for a soft launch of the proposed Centre.
- A well-researched Sustainability Plan for a Centre jointly funded by JISC and external finance.

2.3 Workpackages

The work will be conducted via four work-packages, focusing in turn on:

2.3.1 WP1: Project Setup, Management, and Dissemination work [led by CDLR]

This will be co-ordinated by CDLR but will involve both partners through a project management team working together virtually and in meetings. It will be responsible for project start-up activity, the creation of a new website (required to support an increased user and service, as opposed to project, focus), dissemination and outreach activities (via a Dissemination Plan), Project Plan production and maintenance, cross work-package co-ordination, planning, risk-management, scheduling, and project reporting, including interim and Final Reports.

2.3.2 WP2: User Engagement, Needs, Support, and Evaluation [led by CDLR]

General approach. The methods employed will focus on working directly with both end users and service provider users – for example giving users hands-on experience of tools and terminologies and working iteratively towards solutions that meet their needs as the tools and user experience of them develop. Key elements will be (a) workshops with hands-on opportunities designed to inform, engage (e.g. discussions on user experiences with subject retrieval, focus groups), and obtain useful feedback on needs (b) ongoing consultation via a social networking environment that will also provide access to tools etc as they develop and will encompass the use of community specific advisory groups/expert panels (c) the development a TIC profiling document describing use cases and stakeholder concerns and needs for the user groups recognised as key to TIC aims (d) an outputs evaluation programme.

Specific focus. The intention is to work with four target groups:

- HE and FE libraries (subject librarians, cataloguers/repository managers, systems librarians/repository developers, end-users of various kinds, national service providers (Intute, copac, suncat; other Mimas and Edina services; CAIRNS; M25 Inform; Intute; e-science organisations; IRS; Jorum; CETIS; and so on), representatives of Research organisations (RIN; Universities UK and Universities Scotland etc))
- E-learning (teachers, learners, librarians, VLE managers, VLE developers, national service providers (JORUM and so forth)
- NHS elibrary (and their services)
- A cross-domain group entailing Museums, Archives, Libraries etc – not just in HE (e.g. public libraries as well as HE libraries and work with the Strategic Content Alliance)

Scope. Various methods would be used to maximise involvement in the groups, for example:

- Contact representatives of other interested organisations and initiatives: e.g. OCLC, Becta, BL, NLS, NLW, EU terminology projects (MACS, KOMOHE, CACAO), Institutions and groups representing institutions (RLUK; SCURL; MLA; DAS; SLIC and others), National Libraries (BL; NLW; NLS), others who have indicated an interest, such as: the National Archives Network; IntraLibrary; UMBEL; The National Rural Knowledge Exchange http://www.nationalrural.org
- Use HILT Collaborators/other e-lists, and Ariadne, conferences etc to help 'seed' involvement
- Contact services and organisations not on the above list but at http://www.jisc.ac.uk/services/

Key threads. A designated staff member would have user engagement and lessons learned from this as their primary focus encompassing around 80% of their time and would co-ordinate the various key elements of work with users listed below, aiming to build engagement and understanding and draw out data to inform the Business Plan. The work would move interactively towards a reliable understanding of user requirements via the following threads:

- Four one day workshops at four UK sites with hands-on work, focus groups, and other activities designed to engage and inform, and set up an ongoing iterative process to allow improved understanding of needs – each to have a mix of the target groups, and be shared with IESR.
- Workshop demonstrations of terminology service embedding possibilities and uses and discussions on approaches, problems and issues aimed at finding out a more about real needs
- A purpose built social networking site would be set up and populated prior to the workshops and tools showcase added to it – creating a 'virtual lab' for ongoing interaction with user groups.
- During and after the workshops the environment would be used to show possibilities and get feedback aimed at pinpointing work needed in a soft launch; we'd also use it during the soft launch
- Discussions with these groups would be mainly on their own experiences with subject searching and terminology issues but would also go beyond terminology issues to discuss issues like the availability of local expertise, and how we might best pitch (a) an 'off the shelf' (and 'free') service (b) how we might cost a paid for consultancy (something that would also be discussed with consultants in the field). There would also be discussions on protocols, mark-up languages (SKOS, MARC, Zthes), and specific mapping and tools functionality needs.
- Work with the groups during and after the workshops to show them illustrative functionality and terminological improvements based on ongoing interaction.
- Work on the detail of how the chosen groups (or individual services or organisations/institutions from these groups) might be involved in a direct practical sense in a soft launch and what level of funding would be required by them for their involvement

Note on collaboration with IESR

The project will work with IESR to ensure that the work of both with users is increased in scope and impact and that JISC accordingly gets better value for money. In particular, it should be possible to (a) cover common user groups in more depth and detail by working together, (b) increase our respective engagement levels by working together (c) have more impact in terms of visibility of SIS working together (e) make our engagement 'richer' in terms of highlighting, discussing, and getting feedback on the possibilities of a mix of SIS services by working together. It is hoped that the collaboration might create an embryonic shared services 'outreach and needs assessment' initiative.

2.3.3 WP3: Infrastructure to support the TIC services [led by EDINA]

General approach. There will be a number of phases of work in this work-package:

- Investigating the options for creating, supporting and maintaining service quality infrastructure components of those developed during the pilot. Most of these components are currently based at CDLR and should benefit from a move to the EDINA service environment (months 1-3);
- examination of the implications of different business models resulting from WP4 (months 4-6);
- a more detailed study of the service requirements of the preferred business model (months 7-9);
- an extension of the work on the preferred model which will take account of responses to WP2 developments on user engagement (months 10-12);
- Preparing for the soft launch. (months 11-18).

This outline schedule is designed to meet the needs of the work whilst also ensuring that a draft bid and Sustainability Plan are ready for the February 2010 meeting of the appropriate JISC committee and a final bid and plan for a June 2010 meeting.

Specific focus and scope. The assumption underlying this work is that it is neither possible nor necessary to aim for immediate high volume use of the various functions and services post-project. Instead, what is required is a service that is reliable and robust to support both a soft launch and can be scaled to support an increase in use which it is anticipated will be gradual and largely predictable.

Key threads. Specific threads of work are as follows:

- Considering options and implementing a service quality version of the current database and SOAP server at CDLR.
- Considering and implementing a service quality version of the associated SRU/W server
- Implementing and running a full set of tests for the associated toolkit of user interface embedding code we plan to make available to National and institutional information services wishing to use out terminology services.
- Implementing and running a full set of tests of the pilot terminologies registry database once it was available.
- Putting in place procedures for an acceptable and supported level of testing of the toolkit routines with regard to operation with recent versions of a range of web browsers in common use.
- Considering infrastructure requirements, support and maintenance issues for a number of different business models to help determine the preferred model.
- Take account of changes to the toolkit, the SOAP functions, the database, and mark-ups and other options in response to user needs in the proposed infrastructure.
- Work preparatory to a soft launch.

2.3.4 WP4: Business Plan and Centre set up proposals: [Jointly led by CDLR and EDINA]

General approach. Work on the Business Plan and Centre set up proposals will develop in parallel with other parts of the work. It will be informed by work on service quality from WP3 and by work on user requirements from WP2 and will, in its turn, inform these other aspects of the study. This implies, and will necessitate, a phased approach to developing both the preferred business model and the associated Sustainability Plan and Centre set-up bid. The initial phase (months 1-3) will be based on the outline description of a possible Centre described elsewhere in this document, will draw on data already available to the project from past experience with pilot systems and from early work in WP2 and WP3, and will utilise the set of headings listed in Appendix B. It will produce detailed structures for the discussion and construction of the Sustainability and Centre set-up plans and a set of options for closer scrutiny and appraisal. These will then inform ongoing work in WP2 and WP3, which will in turn feed back into producing better and more precisely tailored plans for set-up and sustainability resulting in (a) Interim versions of the Sustainability Plan and Centre set-up plan and bid for discussion with the JISC programme manager (by month 6), (b) a draft version to for discussion by the JISC Integrated Information Environment Committee (by month 9), (c) a final version for submission to JIIEC in support of an application for funding (by month 13), (d) Operational versions to guide actual set up and begin full instantiation the implementation of the Business Plan (by month 18)

Specific focus and scope. The assumption underlying this work is that the core idea of an approach to sustainability based on a mix of JISC funded free ‘off the shelf’ services, charged for consultancy services, and externally funded and collaborative R&D work, already discussed with the JISC Programme Manager (see Appendix A) is the correct basic approach and that what needs to be investigated in detail is how best to instantiate the approach to produce a robust, reliable, useful, and sustainable service equipped to support the strategic objectives of the JISC.

Key threads. Specific threads of work are as follows:

- Initial work to produce detailed structures for the discussion and construction of the Sustainability and Centre set-up plans and a set of options for closer scrutiny and appraisal - based on the outline description of a Terminologies Interoperability Centre presented on page 4 of this bid, the technical and terminological experience of the partners, the set of headings presented in Appendix B below, and initial discussion between project partners and members of the Advisory Group
- Instigate ongoing work on partnership and collaboration possibilities via the creation of a list of possible partners (e.g. OCLC, Becta, EC terminology projects (Cacao, MACS, KoHoMe etc)).
- Interaction with the WP2 and WP3 work, including user interaction during and prior to workshop 1 and the associated F1 feedback event, the aim being, both to inform that work in respect of pinpointing specific areas in which to (a) seek user and stakeholder feedback (b) direct some of the specifics of service hardening work, and to inform the creation of an early interim version of the Sustainability Plan and Centre setup proposal for discussion with the JISC programme manager.
- Development of a draft Sustainability Plan and Centre setup proposal, based on the discussions with the JISC programme manager and on subsequent WP2 and WP3 work, particularly in relation to workshops 2-3 and the associated feedback events (F2 and F3 on the schedule); submission of the drafts to the JISC SIS Committee at its February 2010 meeting to inform it of ongoing developments and allow it to influence the final versions of each.
- Development of Final versions of the Centre setup proposal and the Sustainability Plan for formal submission at the June 2010 meeting of the JISC Integrated Information Environment Committee.
- Assuming a positive outcome from the June meeting, use of the documents, together with feedback from workshop 4 and the subsequent feedback event (F4), to build towards a soft launch of the proposed Centre by creating operational versions of the Plan and the set up proposal to guide actual instantiation work; linking this to ongoing WP3 soft launch set up work and to the (by that time) reduced but ongoing level of WP2 work with the soft launch user groups.
- If the June meeting has a negative outcome, an examination of alternative exit strategies.

2.4 Schedule [June 1st 2009 to November 30th 2010]

TIC Scoping Study	**1 J**	**2 J**	**3 A**	**4 S**	**5 O**	**6 N**	**7 D**	**8 J**	**9 F**	**10 M**	**11 A**	**12 M**	**13 J**	**14 J**	**15 A**	**16 S**	**17 O**	**18 N**
Project Management	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
Set up/website	▒																	
Dissemination	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
Project Plan	▒	▒	▒															
Co-ordination	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
Project reports	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
User Studies & evaluation		▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▨	▨	▨	▨
Workplan		▒		▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒				
Evaluation			▒		▒	▒	▒	▒	▒	▒	▒	▒	▒	▒				
Soc. Netw. Site		Setup m2/populate m2-3/maintain and develop thereafter																
Engage users			▒	▒		▒		▒		▒		▒	▒	▨	▨	▨	▨	▨
Workshops		▒	▒	▒	#1	▒	#2	▒	#3	▒	#4							
Gather feedback			▒	▒	F1	F1	F2	F2	F3	F3	F4	F4	▒	▨	▨	▨	▨	▨
Launch prepare										▒	▒	▒	▒	▨	▨	▨	▨	▨
Service quality instantiation	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
Mirror and test	▒	▒	▒															
Options tests			▒	▒	▒	▒												
Lead option test						▒	▒	▒	▒									
User changes test									▒	▒	▒	▒	▒					
Soft launch setup														▒	▒	▒	▒	▒
Sustainability Plan; Bid	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒	▒
Outline; options	▒	▒	▒															
JISC discuss draft			▒	▒	▒	▒												
Feb. Cttee. draft						▒	▒	▒	▒									
Final: June Cttee.									▒	▒	▒	▒	▒					
Versions for launch														▒	▒	▒	▒	▒

Members of Project Team: See under **Previous Experience of the Project Team** below.

2.5 Project Partners

The proposed study requires collaboration between the following participants:

Participant	Role(s)
CDLR	Project management; Final and other reports; Dissemination; Website; social networking environment; User studies; tools showcase; liaising with evaluation team; initial terminology services registry pilot; primary work on TIC start-up requirements and sustainability plan; advice on service set up; ongoing maintenance and development of current terminologies data, tools, embedding toolkit, documentation, development system. Liaison with advisory group etc.
EDINA	Ensuring service quality components are available - to ensure that the standard services offered at the soft launch will be robust, suitable, and usable in a range of JISC service and user environments; Working with CDLR to produce a bid for TIC start-up costs, a programme of works, and a well-researched Sustainability Plan. Contributing to pilot work on a terminology services registry (with CDLR)
Project advisors	Primarily advice and views on levels and approaches to charged for services in a future TIC, but also ongoing in other areas: terminology issues, classification issues, mapping issues, standards, as in previous projects.

The project will continue to existing advisors and stakeholders (such as UKOLN[83], the BL[84], the NLS[85] and NLW[86]) via the project Steering/Advisory Group. The primary costs of the project will be the staffing costs of the various participants, comprising: project management staff, user studies and terminology work research staff, and programming and technical staff at CDLR; programming and technical staff at EDINA; terminology expert consultancy work; evaluation work.

2.6 Project Management

Day to day management will be the responsibility of staff at each partner site. This **Project Team** will work together through regular virtual and face to face meetings. As in previous phases, there will also be a **Project Steering Group** comprising representatives from key stakeholders.

2.7 Risk Analysis

Risk	Probability (1-5)	Severity (1-5)	Score (P x S)	Action to Prevent/Manage Risk
Staffing	1/5	2/5	2	Use partners to fill any gaps, bring in new staff quickly. Both partners have backup staff.
Organisational	1/5	1/5	1	Plan ahead, monitor daily, act early to fix.
Technical	1/5	2/5	2	Should be small since the main aim is to move already tested pilot services and test for a soft launch. If necessary, manage via robust alternatives
Legal	1/5	3/5	3	Not relevant to scoping study but discussions with OCLC needed to determine exact position with future use of DDC. Early discussions suggest this will not be a problem.

2.8 Standards and Accessibility

The project will adhere to appropriate standards where these exist and will be advised in this by other participants, by UKOLN and by JISC generally. The JISC IE standards[87] will be adhered to where they are appropriate and open standards will be preferred where possible. The specific standards that will impact on the project are SRW, SOAP, and SKOS Core (used for terminologies mark-up). The project will adhere to the *British Standard for Structured Vocabularies for Information Retrieval* (BS8723) Parts 1-4, which greatly influenced SKOS Core. Two contributors to the standards are involved in the project[88]. The project is also aware of current developments with respect to the Z39.19 'thesaurus standard'[89]. Accessibility guidelines will be adhered to and the Technology for Disabilities Service (TechDis[90]) will be used for guidance and advice. The project will ensure that Deliverables conform to the Disability Discrimination Act (DDA) and Human Rights

[83] http://www.ukoln.ac.uk/
[84] British Library (BL): http://www.bl.uk/
[85] National Library of Scotland (NLS): http://www.nls.uk/
[86] National Library of Wales (NLW): http://www.llgc.org.uk/
[87] http://standards.jisc.ac.uk
[88] Alan Gilchrist and Leonard Will.
[89] Z39.19-2005: Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies: http://www.niso.org/standards/resources/Z39-19-2005.pdf
[90] http://www.techdis.ac.uk

legislation regarding accessibility. Web interfaces will conform to W3C current guidelines[91]. In addition, TIC will keep track of technical and other relevant developments.

2.9 Intellectual Property Rights

Should the project be funded, the project partners will comply with the JISC requirements as regards to project deliverables and IPR as agreed in the subsequent letter of award. It is not expected that this scoping study will produce anything with significant IPR issues.

2.10 Exit and Sustainability Plans

Once launched, the Centre would offer the community a mix of free and charged-for services and would be supported by a mix of JISC funding, externally earned income and R&D funding, and collaboration. However, the project itself is a scoping study that will produce exit and sustainability plans for the proposed Centre. It does not, in itself, need exit and sustainability plans.

Project Outputs	Action for Take-up & Embedding	Action for Exit
TIC start-up bid	Obtain funding for TIC start-up	Submit to JISC
TIC Sustainability Plan	Obtain funding for TIC start-up	Submit to JISC with above bid
TIC social network site	Future support via funded TIC	Await outcome of future bid

3. Engagement with the Community

General engagement will be managed via the Dissemination Plan developed and implemented via WP1; work with specific stakeholders as outlined in WP2 and as scheduled on page 8 above.

3.1 Stakeholder Analysis

Stakeholder	Interest / stake	Importance
Institutions; e-learning groups; NHS; SCA etc as described in WP2	Find distributed resources by subject where different schemes used or have them found	High

3.2 Evaluation Plan

An external evaluation will be carried out focusing primarily on assessing the reliability of data extracted from user work in WP2. There will be an early evaluation of the proposed approach to user engagement that will be used to guide WP2 work then a second phase later to evaluate the reliability of the outcomes. More precise details will be developed for the Project Plan.

3.3 Dissemination Plan

Dissemination would be via the website, the proposed social networking environment, workshops, postings to appropriate e-mail lists, papers and news items submitted to professional publications and presentations at seminars and conferences. WP1 will develop a scheduled Dissemination Plan.

4. Budget

Budgetary information removed.

5. Previous Experience of the Project Team

The team have all been involved in earlier phases of HILT – they are: Dennis Nicholson; Emma McCulloch; Anu Joseph (all CDLR); Christine Rees, Tim Stickland, Ben Soares (all EDINA).

[91] http://www.w3.org/WAI/Resources/#gl

Appendix A: Notes on meeting on HILT and Sustainability: London 12.12.08

Present: James Farnhill (JISC); Dennis Nicholson (HILT)

After some initial and wide-ranging discussions, **the following points were agreed**:

1. Although there might in time be elements of HILT that were exceptions, the facilities and expertise offered by HILT were never likely to be a 'service' in the strict sense that JISC understood the term. What was on offer was similar to the kinds of facilities on offer from the NacTem text mining centre, except that HILT was operating in the field of terminologies or Knowledge Organisation Systems (KOS) and interoperability between them rather than text mining. More specifically, it could offer (through a mix of JISC funding, and funding and/or collaboration from other sources, including Research Councils for related research, and possibly other organisations such as the British Library and OCLC) something like the following:
   - M2M and user-level access to terminology sets, the detail of those terminology sets, and data to facilitate interoperability between them
   - Open source software toolkits that would enable M2M interaction with HILT web services to be transparently embedded in the user interfaces of local, national and project information services
   - A basic architecture for terminology and interoperability services in the JISC Information Environment (and potentially beyond)
   - A way of mounting and developing new KOS and KOS interoperability data required by the community
   - A JISC funded free advisory and training service on using the above facilities in local or national services and projects in the percentage of cases where this was relatively straightforward (plug 'n' play, but after a bit of advice and training)
   - A charged-for consultancy service where the work and advice required by local and national services, projects, and organisations (both within and outwith JISC) was less straightforward or more sophisticated (because of the circumstances of the clients and their services and terminology sets)
   - A portal for tools and training in the areas described above
   - A focus for wider work in the terminologies area (for example though successful bids for European funding)
2. This being so, the way forward for HILT should be to move over the next two years or so towards a 'soft launch' of a National Centre that would offer these types of assistance and facilities to both JISC and the wider community. The likely cost would be in the region of £200k to £300k over two years.
3. In order to move towards a position where JISC funding for this 'soft launch' was made available, JF and DN would focus on re-purposing the draft Sustainability/Business Plan to support an application.
4. It was noted that the functionality required to provide a terminologies registry could almost certainly be provided through a combination of IESR's service registry facilities and HILT's storage and provision of M2M access to terminology sets and the detail of those sets. Depending on the outcome of discussions on this topic this could conceivably become part of what the Centre had to offer.
5. In a (relevant) aside to the above, DN noted that IESR played a role in what HILT had to offer but that HILT currently had to simulate IESR via an IESR-like database held at CDLR. This was because of factors such as inconsistency of data held across IESR and the omission or poor coverage of data needed by HILT. IESR and HILT were discussing what could be done about this, but some of the issues could only be addressed through a change in JISC's view of how the database should be populated and maintained. DN would ensure that JF received details of the various issues (for HILT) in respect of IESR.
6. In discussing a future iteration of the HILT Business Plan, it was noted that detail on what HILT could do *for specific audiences* within the JISC community (e.g. Intute Social Sciences searchers) needed to be made clear. A related point made by JF was that it was important in bidding for future funding to put HILT in a context that that is meaningful to the JISC communities (how HILT's facilities would help specific audiences in specific circumstances and specific (specified) ways).
7. The aim should be to have a new version of the Business Plan ready for the next meeting of the relevant JISC committee in March 2009.
8. JF made the useful suggestion that code that would facilitate interaction between repositories and HILT might usefully be built into repository software packages such as e-prints, DSpace, and Fedora. DN would consider how this might be done/funded.
9. A future iteration of the Business Plan should specify alternative approaches to providing what HILT/the National Centre might provide. It should also allow for work towards 'hardening' of the web services and toolkit facilities – for example, load testing and so on. Edina would, amongst other things, provide some of the technical capabilities (such as a hosting environment and call logging). There would be a need for outreach funding and other things such as help desk software and to contrast 'where we are' with 'where we

need to be' as a basis for planning/costing.

10. It was agreed that specifying closely what HILT would offer and who would use it would feed more directly into detailing measurable benefits in the Business Plan.
11. A discussion between HILT and CETIS would help inform HILT of requirements in the areas cover by CETIS.
12. It was agreed that it would not be sensible for JISC and its communities to rely on OCLC for the kinds of advice and facilities offered by HILT

Appendix B: Headings for consideration in WP4

It is envisaged that a bid to outline and support a service would include the following boxed items below (both in terms of written proposal and a requirement for work during this scoping phase):

Vision

Business Case – how it fits various strategies (JISC, CDLR, EDINA, other stakeholders)

Benefits

Risks

Overall costs & timetable

Business Model

Software Platform(s)

Core Services/additional services - m2m; users; support

Communities services offered to – designated (ones have to support); extended (ones who could use services but would not be eligible for level of support offered to designated community)

Transition: getting from where we are to service envisaged

Summary of Work packages (to support transition activity)

Timetable

Current & Projected usage

Evaluation

CDLR and EDINA roles – capabilities of each, key roles, decision making, reporting

Stakeholder/User – analysis & communication plan

To arrive at a single business model, it is envisaged that an options appraisal would be required, considering options of the items in the above box (benefits, risks, uncertainties, costs) to come up with a number of models to analyse and help towards a decision on a preferred model. Following this, it would be necessary to more fully detail/cost the preferred model put to JISC for consideration.

Business Plan

Approach

Deliverables

Cores Services

Additional Service options (if any)

Workpackages (for transition to service; for core services; for additional options etc)

Technical platform/environment for delivery of service: inc software, hardware, scalability, resilience, usability, security, maintenance

Risk register

Staff roles

Full costings

Appendix C: Glossary

AAT: Art & Architecture Thesaurus

DDC: Dewey Decimal Classification

EDINA: A JISC-funded national datacentre based at Edinburgh University Library, offering the UK tertiary education and research community networked access to a library of data, information and research resources.

FE: Further Education

HE: Higher Education

GCMD: Global Change Master Directory

Go Geo!: A tool designed to help users find details about geo-spatial datasets and related resources within the UK tertiary education sector and beyond. A trial service is provided by EDINA.

HASSET: Humanities and Social Science Electronic Thesaurus

HILT: High-level Thesaurus

IESR: JISC Information Environment Service Registry

intute: intute is a free online service providing access to the very best web resources for education and research. Formerly the Resource Discovery Network (RDN).

IPSV: Integrated Public Sector Vocabulary

JACS: Joint Academic Coding System

JISC: Joint Information Systems Committee

JISC IE: Joint Information Systems Committee Information Environment

LCSH: Library of Congress Subject Headings

MeSH: Medical Subject Headings

M2M: Machine to machine interaction

NMR: National Monuments Records Thesauri

OCLC: Online Computer Library Center

SKOS Core: Simple Knowledge Organization System (SKOS) Core supports the Resource Description Framework (RDF) description of language-oriented knowledge organisation systems (KOS), such as thesauri, glossaries, controlled vocabularies, taxonomies and classification schemes.

SOAP: Originally the Simple Object Access Protocol, but now more simply referred to as SOAP. Used to exchange XML-based messages over computer networks, normally using HTTP.

SQL: Structured Query Language

SRW: Search/Retrieve Web Service – Z39.50 Next Generation

SRU: Search & Retrieve URL – Z39.50 Next Generation

UKOLN: A centre of expertise in digital information management, providing advice and services to the library, information, education and cultural heritage communities. Based at the University of Bath and formerly known as the UK Office for Library & Information Networking.

UNESCO Thesaurus: United Nations Educational, Scientific and Cultural Organization subject scheme.

Use Case: A Use Case represents a series of interactions between a user (human or machine) and the system, utilising (in the present case) an M2M link. Typically, the interaction starts with an enquiry and leads to a resource that should answer that enquiry.

XML: Extensible Mark-up Language

Z39.19: ANSI/NISO Standard - Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabularies.

Z39.50: An international standard specifying a client/server-based protocol for searching and retrieving information from remote databases.

Zthes: The Zthes profile is an abstract model for representing and searching thesauri and specifies how this model may be implemented using the Z39.50 and SRW protocols.

Appendix D: References

Agosti et al. (2007). Roadmap for MultiLingual Information Access in the European Library. In *Research and Advanced Technology for Digital Libraries*, Springer Berlin/Heidelberg, Volume 4675/2007. Abstract online at: http://www.springerlink.com/content/g8126155w7518333/

Mayr, P., and Petras, V. (2008). Cross-concordances: terminology mapping and its effectiveness for information retrieval. In *IFLA 2008 Conference Proceedings*. Online at: http://eprints.rclis.org/13828/

Other works consulted:

Clavel-Merrin, G. (2004). MACS (Multilingual Access to Subjects): A Virtual Authority File Across Languages. In *Cataloguing and Classification Quarterly*. Volume: 39 Issue: 1/2. Springer: Berlin.

Cross-language Access to Catalogues and Online Libraries partners. *CACAO Presentation.* Online at: http://www.cacaoproject.eu/fileadmin/media/presentations/CACAO_presentation.pdf

Day, M., Koch, T., & Neuroth, H. (2004). Searching and browsing multiple subject gateways in the Renardus Service. In: Dijkum, C. van, Blasius, J., Kleijer, H., & Hilten, B. van (eds.) *Recent developments and applications in social science methodology: proceedings of the Sixth International Conference on Logic and Methodology, August 17-20, 2004, Amsterdam, The Netherlands*. Amsterdam: SISWO Instituut voor Maatschappijwetenschappen. (CD-ROM). Online at: http://www.ukoln.ac.uk/metadata/publications/rc33-2004/renardus-paper.pdf

Faaborg, A.J. Leveraging Metadata to Improve Information Retrieval in Directory Interfaces: http://alumni.media.mit.edu/~faaborg/research/cornell/hci_informationretrieval_finalPaper.pdf

Geser, G. (2008). *STERNA Technology Watch Report. Full Report*. Ref: STERNA Del.6.5**,** 10 December 2008. Salzburg Research. Online at: http://www.sterna-net.eu/index.php/en/downloads

Landry, P. (2004). Multilingual Subject Access The Linking Approach of MACS. In *Cataloging & Classification Quarterly*. Volume 37, Issue 3/4.

Mayr, P., Mutschke, P., and Petras, V. (2008). Reducing semantic complexity in distributed digital libraries: Treatment of term vagueness and document re-ranking. In *Library Review*, Volume 57, Issue 3. Pages 213 – 224.

Philipp Mayr and Vivien Petras. (2007). Building a terminology network for search: the KoMoHe project. In *Proceedings of the International Conference on Dublin Core and Metadata Applications 2008*. Online at: http://arxiv.org/pdf/0808.0518.

Tudhope, D. (2004). Semantic interoperability issues from a case study in archaeology. Online at: http://hypermedia.research.glam.ac.uk/media/files/documents/2008-07-05/SIEDL08-Tudhope-v3.pdf

Tudhope, D., Koch, T., and Heery, R. (2006). *Terminology Services and Technology. JISC State of the Art Review*. Online at: http://www.jisc.ac.uk/media/documents/programmes/capital/terminology_services_and_technology_review_sep_06.doc.

van Gendt, M., et al. (2006). Semantic Web Techniques for Multiple Views on Heterogeneous Collections: A Case Study. In *Lecture Notes in Computer Science.* Volume 4172/2006. Springer Berlin/Heidelberg. Abstract online at: http://www.springerlink.com/content/b886412v44t61u1w/

Zeng, M. L., & Chan, L. M. (2006). Metadata Interoperability and Standardization – A Study of Methodology Part II: Achieving Interoperability at the Record and Repository Levels. In *D-Lib Magazine*, Volume 12, Issue 6. Online at: http://www.dlib.org/dlib/june06/zeng/06zeng.html

Appendix I: Glossary

AAT: Art & Architecture Thesaurus

API: Application Programming Interface

CAB: Commonwealth Agricultural Bureaux

CAIRNS: Co-operative Information Retrieval Network for Scotland

CDLR: Centre for Digital Library Research

DDC: Dewey Decimal Classification

Depot: a UK national open access repository for researchers not yet having an institutional repository in which to deposit their papers, articles, and book chapters (e-prints)

EDINA: A JISC-funded national data centre based at Edinburgh University Library, offering the UK tertiary education and research community networked access to a library of data, information and research resources.

FE: Further Education

GCMD: Global Change Master Directory

Go Geo!: A tool designed to help users find details about geo-spatial datasets and related resources within the UK tertiary education sector and beyond. A trial service is provided by EDINA.

HASSET: Humanities and Social Science Electronic Thesaurus

HE: Higher Education

HILT Project: High-level Thesaurus Project

IESR: JISC Information Environment Service Registry

intute: intute is a free online service providing access to the very best web resources for education and research. Formerly the Resource Discovery Network (RDN).

IPSV: Integrated Public Sector Vocabulary

JACS: Joint Academic Coding System

JISC: Joint Information Systems Committee

JISC IE: Joint Information Systems Committee Information Environment

LCSH: Library of Congress Subject Headings

M2M: Machine to machine interaction

MARC: Machine readable cataloguing

MeSH: Medical Subject Headings

NHS: National Health Service

NMR: National Monuments Records Thesauri

OCLC: Online Computer Library Center

RAE: Research Assessment Exercise

REST: Representational State Transfer

RIN: Research Information Network

SCAS: Standard Classification of Academic Subjects

SKOS: Simple Knowledge Organization System. SKOS Core supports the Resource Description Framework (RDF) description of language-oriented knowledge organisation systems (KOS), such as thesauri, glossaries, controlled vocabularies, taxonomies and classification schemes.

SOAP: Originally the Simple Object Access Protocol, but now more simply referred to as SOAP. Used to exchange XML-based messages over computer networks, normally using HTTP.

SQL: Structured Query Language

SRU/W: Search/Retrieve Web Service and Search & Retrieve URL – Z39.50 Next Generation

UKOLN: A centre of expertise in digital information management, providing advice and services to the library, information, education and cultural heritage communities. Based at the University of Bath and formerly known as the UK Office for Library & Information Networking.

UNESCO Thesaurus: United Nations Educational, Scientific and Cultural Organization subject scheme.

Use Case: A Use Case represents a series of interactions between a user (human or machine) and the system, utilising (in the present case) an M2M link. Typically, the interaction starts with an enquiry and leads to a resource that should answer that enquiry.

VLE: Virtual Learning Environment

WSDL: Web Services Description Language

XML: Extensible Mark-up Language

Z39.50: An international standard specifying a client/server-based protocol for searching and retrieving information from remote databases.

Zthes: The Zthes profile is an abstract model for representing and searching thesauri and specifies how this model may be implemented using the Z39.50 and SRW protocols.