

HILT Phase III: M2M Pilot Demonstrator Project Proposal

1 Summary

This proposal is a follow-up to the HILT Machine to Machine (M2M) Feasibility Study¹, a ten-week project charged with determining:

- The feasibility of developing SOAP-based² interfaces between JISC Information Environment³ services and the HILT pilot demonstrator⁴ created under HILT Phase II⁵.
- The scope and cost of an actual M2M demonstrator.

In line with the outcomes of this study, an M2M demonstrator is proposed that will:

- Offer web-services access via the (SOAP-based) SRW protocol⁶, but be designed so that a possible extension offering other protocols (Z39.50⁷, or SRU⁸, for example) at a later date could be an option.
- Use SKOS-Core⁹ as the 'mark-up' for sending out terminology sets and classification data responses but be designed so that adding other formats such as MARC¹⁰ and Zthes¹¹ would be an option at a later date.

The development of the pilot M2M demonstrator requires the expertise of participants at CDLR¹² and EDINA¹³, together with some ongoing liaison with UKOLN¹⁴, the MIMAS¹⁵ IESR¹⁶ project, the HILT terminology advisors, and Wordmap¹⁷, a commercial partner providing liaison and a software license at no cost in order to keep this development path open.

With JISC's agreement, two versions of the demonstrator have been costed – a single server version (15 months) and a distributed server version (21 months), with the latter being the slightly more expensive option, but recommended as a better strategic approach and as cheaper in the long run. In both cases, the cost could be spread across two financial years should JISC feel this is helpful. In the case of the single server option, this would entail starting November 2005 and ending January 2007, with the total cost (£98, 263) breaking down as follows:

- Financial year ending July 2006 (9 project months) £58,958
- Financial year ending July 2007 (6 project months) £39,305

In the case of the dual-server option, this would entail starting mid-October 2005 and ending mid-July 2007, with the total cost (£132, 600) breaking down as follows:

- Financial year ending July 2006 (9.5 project months) £59,986
- Financial year ending July 2007 (11.5 project months) £72,614

The proposers would prefer that the start date be October or November 2005 depending on the project version chosen, but these dates are negotiable if JISC has other preferences. Note, however, that changing the start dates would also change how costs were spread over financial years.

¹ HILT Machine to Machine (M2M) Feasibility Study: <http://hilt.cdlr.strath.ac.uk/hiltm2mfs/>

² Simple Object Access Protocol (SOAP) 1.1: <http://www.w3.org/TR/soap/>

³ JISC Information Environment (IE): http://www.jisc.ac.uk/index.cfm?name=ie_home

⁴ HILT Pilot Demonstrator: <http://hilt.pilot.cdlr.strath.ac.uk/pilot/top.php>

⁵ HILT Phase II Project: <http://hilt.cdlr.strath.ac.uk/index2.html>

⁶ ZING Search / Retrieve Web Service (SRW): <http://www.loc.gov/z3950/agency/zing/srw/>

⁷ Z39.50: <http://www.loc.gov/z3950/agency/>

⁸ ZING Search & Retrieve URL Service: <http://www.loc.gov/z3950/agency/zing/srw/sru.html>

⁹ SKOS-Core: <http://www.w3.org/2004/02/skos/core/guide/>

¹⁰ MARC: <http://www.loc.gov/marc/>

¹¹ Zthes: <http://zthes.z3950.org/>

¹² Centre for Digital Library Research (CDLR): <http://cdlr.strath.ac.uk/>

¹³ EDINA: <http://edina.ac.uk/>

¹⁴ UKOLN: <http://www.ukoln.ac.uk/>

¹⁵ Manchester Information & Associated Services (MIMAS): <http://www.mimas.ac.uk/>

¹⁶ Information Environment Services Registry (IESR): <http://iesr.ac.uk/>

¹⁷ Wordmap: <http://www.wordmap.com/>

Since two of the five 'use cases' (see Appendix A) identified in the Feasibility Study apply to RDN¹⁸ needs, but RDN re-organisation makes their active involvement in the project impossible at this point, the project will examine these use cases in the context of locally available CDLR services such as the Glasgow Digital Library¹⁹ and Victorian Times²⁰ and provide the RDN with a report on issues that arise in relationship to RDN-specific needs. An RDN representative will also be invited to join the HILT e-mail discussion list.

As with previous phases of HILT, OCLC²¹ have again agreed to provide free access to the electronic files of DDC²² and of LCSH mappings to DDC²³.

2 The Problem Addressed

Background: HILT I and II and the M2M Feasibility Study

Ensuring that FE and HE users of the JISC IE can find appropriate learning, research and information resources by *subject search and browse* in an environment where most service providers use different subject schemes to describe their resources is a major challenge facing the JISC domain (and, indeed, other domains beyond JISC). Phases I and II of the HILT project:

1. Established that the preferred approach of the various services in the domain to resolving the issue was one based on mapping the various subject schemes together through a central shared service that would provide users with the correct alternative terms to use in the various different schemes (HILT Phase I²⁴).
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).

There are a range of issues that must be resolved before an operational JISC terminologies service can become a reality. Of these, one of the most important is the provision of its facilities via web-services protocols to enable Machine-to-Machine (M2M) interaction between terminology services and other components of the JISC IE architecture (services such as Go Geo!²⁵, the Glasgow Digital Library, or RDN hubs for example). HILT Phase II developed a range of facilities currently only available through a direct user interface. A HILT M2M interface will allow other machines to query the pilot server in the same way that end users can now, thereby permitting the various JISC services to provide terminology mapping services to their users in a transparent way.

The HILT M2M Feasibility Study Recommendations

The Feasibility Study concluded (see Appendix B)²⁶ that the development of an M2M pilot based on web-services protocols was feasible and that the preferred option was a pilot M2M terminology services demonstrator that would:

- Aim to create an M2M version of the current HILT pilot, but with facilities extended to take account of the five use cases drawn up under the HILT M2M Feasibility Study (see Appendix A).
- Use the SRW protocol only, but be designed so that a possible extension offering other protocols such as Z39.50 or SRU could be introduced at a later date.
- Use SKOS-Core as the 'mark-up' for sending out terminology and classification set responses, but be designed so that adding other formats such as MARC and Zthes would be a later option.

¹⁸ Resource Discovery Network (RDN): <http://rdn.ac.uk/>

¹⁹ Glasgow Digital Library (GDL): <http://gdl.cdlr.strath.ac.uk/>

²⁰ Victorian Times (VT): <http://www.victoriantimes.org/>

²¹ OCLC Online Computer Library Center: <http://www.oclc.org/>

²² Dewey Decimal Classification (DDC): <http://www.oclc.org/dewey/>

²³ LCSH to DDC mappings: <http://www.oclc.org/asiapacific/zchn/dewey/updates/numbers/default.htm>
<http://hilt.cdlr.strath.ac.uk/Reports/FinalReport.html>

²⁴ Go Geo!: <http://www.gogeo.ac.uk/>

²⁶ See also the Final Report at <http://hilt.cdlr.strath.ac.uk/hiltm2mfs/0HILTM2MFinalReportRepV3.1.doc>

It also noted that a two-server pilot which distributed the terminology mappings in HILT across two servers might be an attractive additional variation, noting that, whilst this was a more expensive option, and entailed undertaking more work and addressing additional technical issues, it also allowed a far more realistic pilot situation to be created, one that echoed the world of distributed terminology services envisaged in the JISC IE and the web services world generally. There is a case for saying that, if the future of terminology services is likely to be distributed (as appears to be true), then JISC needs to start investigating the issues sooner rather than later to ensure it has input to developing standards and positions in the area and can keep abreast of the needs of the JISC IE as it develops in this wider context. It is also likely that choosing the single server option will result in a need to fund a later two server project and that this later project will cost a good deal more than opting for a dual-server approach now. Not only will it be a different project with its own administrative overheads, but it will by that time have to take account of the implications of Full Economic Costing.

3 Project Aim, Pilot Software, Two Clients, Participants, Roles, Deliverables

Aim

The aim of HILT Phase III is to create an M2M version of the current HILT Pilot based on SRW and SKOS-Core, but with facilities extended to take account of the five use cases drawn up under the HILT M2M Feasibility Study (see Appendix A). With JISC's agreement, two versions of this have been costed – a single server version and a distributed server version. These are identical in all respects except one – that is, version two distributes the terminology service provided by the HILT pilot across two servers.

Pilot Software

The terminologies pilot built in HILT Phase II was based on an adaptation of the Wordmap software. In the period between HILT Phase II and the feasibility study an opportunity arose to build a pilot offering similar services based on a more generic solution (SQL Server was used). This alternative pilot²⁷ lacks the advanced terminology and mappings management facilities of Wordmap, but is purpose-built to provide the HILT pilot facilities themselves and is easier to work with on this front. In the Feasibility Study, this latter pilot was used to provide the simple SOAP demonstrator service²⁸ put in place to establish the baseline feasibility of an M2M pilot. The proposal as regards HILT Phase III is to use the SQL-Server based approach to build the M2M pilot but to continue to liaise with Wordmap as their product develops with a view to ensuring that this development avenue is not closed off as a possible future solution for an operational service. As indicated in HILT Phase II, the continued use of Wordmap could be advantageous in the long-term if, as HILT Phase II concluded, a multi-user interface for maintaining mappings in a distributed fashion is likely to be a need. Points against basing the M2M pilot on Wordmap are (1) that this interface is not needed for the M2M pilot (2) that whereas it was necessary to adapt the Wordmap database structure to provide the HILT Phase II pilot, the SQL Server version, having been specifically designed for HILT, is easier to work with in the experience of the CDLR HILT team. It does, however, remain true that the Wordmap staff management interface is a possible longer term need.

Two Clients

The proposal is that the project should develop M2M clients: one for users of Go Geo! at EDINA, and one for the HILT interface itself. Since the point of offering a SOAP-based SRW service is specifically platform independence, it has been agreed that taking a common approach and developing a single client would not be a sensible strategy – that, on the contrary, it would be better to develop (pilot versions of) two different clients. This will serve JISC and its users best by ensuring a more robust service that would be more likely to work with the various new clients that others in the community would need if they wanted to interface with the SOAP/SRW server. Accordingly, the proposal is that two clients be developed separately at Go Geo!, and HILT, but with help and advice from EDINA in the latter case. The needs of the RDN will be taken into

²⁷ <http://hilt.cdlr.strath.ac.uk/hilt3/top.cfm>

²⁸ <http://nevis.ed.ac.uk:8080/asp-misc/public/hilt.asp>

account by continuing to liaise with the service as their planned new hardware and software platform is identified and installed.

Participants and Roles

The proposed study requires collaboration between the following participants:

Participant	Role(s)
CDLR	Project management; Final and other reports; Dissemination; Web-site; Programming HILT – SRW, Program client for HILT; Overall co-ordination of M2M pilot design; HILT database redesign; Terminology mappings; Collections database adjustment; SKOS-Core work co-ordination; Co-ordinate testing; Evaluation; Analysis, programming re. Variant cases
EDINA	SRW server set up and support; Creation of client for Go Geo!; Advice and support for client programming work at HILT and BIOME
MIMAS	IESR and RDN advice
UKOLN	Advice on the JISC IE
Terminology experts (L. Will and A. Gilchrist)	Advice and views on terminology issues, classification issues, mapping issues, mark-up issues, the terminology services scene
Wordmap	Liaison on Wordmap software developments

Deliverables

The HILT Phase III deliverables will be:

- A working SRW/SKOS-Core based M2M pilot demonstrating M2M terminology services for the JISC IE based on the HILT Phase II pilot, illustrative extensions to cover the five use cases outlined in Appendix A below, clients to service the needs of two different service environments, and advice to RDN on issues that will affect their own needs as regards a client for accessing HILT.
- A final report on the project, together with details of future research and development requirements leading towards a future operational service.

4 Description of Work Proposed

Description of Work Proposed: Single server version

The ‘single server version’ of the project would last 15 months and would build a web-services version of the current HILT pilot with the following characteristics:

- It would use the SRW protocol only, but would be designed so that a possible extension offering other protocols (Z39.50, SRU) at a later date could be an option. This could have implications in areas such as how CQL²⁹ would be used to send queries, how terminology response sets were encoded, and for the implementation of the SRW ‘explain’ facility.
- It would use SKOS-Core as the ‘mark-up’ for sending out terminology and classification set responses but, again, would be designed so that adding other formats such as MARC and Zthes would be an option later on. SKOS-Core concept URIs would be used to identify concepts uniquely, so that a distributed version of the service could be a later option.
- It would have illustrative mappings needed to support the various use cases listed in Appendix A. This is likely to entail new (illustrative) mappings of LCSH³⁰, JACS³¹, UNESCO³², and

²⁹ Common Query Language (CQL): <http://www.loc.gov/z3950/agency/zing/cql/>

³⁰ Library of Congress Subject Headings (LCSH): <http://authorities.loc.gov/>

MeSH³³ terms to the DDC spine, together with mappings from specific terminology sets and mappings to cover areas highlighted as important in the use cases (synonyms, spelling and typing mistakes). These need to be sufficient to allow for realistic tests and evaluation under the various use cases but will only be illustrative.

- Additional programming to interface the HILT service with the SRW server (allowing inter-working between the SRW server and SQL Server APIs³⁴).
- Web clients designed to handle M2M interactions between Go Geo! and the HILT terminology server using SRW, CQL and SKOS-Core and a web-services based replacement front-end for the HILT service itself (this would be needed short term for testing purposes and ongoing research work, but it will also be a long-term requirement for a JISC terminology service).
- A local collections database as used in HILT Phase II. The present pilot does not use IESR but a simulated 'JISC collections' database for interaction between the terminology server and a collections database. It is proposed that, for the moment, this should continue to be the case, but that HILT and IESR liaise to ensure a harmonised approach through the inclusion of an IESR representative on the HILT management team e-mail discussion list.
- A database that extended the current SQL Server database structures to encompass the wider range of mappings and mapping types.
- Work to identify issues and solutions relating to the problems alluded to under the last section of Appendix A (headed 'Variant Cases').
- Ongoing liaison with Wordmap to keep this development path open.

Description of Work Proposed: Distributed server version

The 'distributed server version' of the project would last 21 months, build a web-services version of the current HILT pilot with the same characteristics as the single server version but would take one of the illustrative mappings listed above (JACS, say) out of the main server and set up a **second** terminology service. Mapping between JACS in server two and the DDC spine in server one would be achieved via the use of SKOS-Core concept URIs. For the purposes of the pilot, the assumption would be that the web clients for Go Geo! and HILT would already 'know' about the two servers and would 'talk' to one or the other depending on whether or not JACS was a factor (obviously, IESR would have to come into this longer-term, but it would not feature in the pilot at this stage only in discussions leading to future proposals). If (say) the Go Geo! client needed LCSH and JACS mappings, it would send a request to server 1 and get back (in the simplest case) the DDC caption appropriate to the subject sent, together with an LCSH mapping and the SKOS-Core concept URI for the concept. It would then send the SKOS-Core concept URI to server 2 and receive back the appropriate JACS term. The illustrative mappings in each server would have SKOS-Core concept URIs associated with them and these would be used to ensure intelligent linkings across the distributed service. Clearly, this is a very simple example of what would have to occur in reality in the long-term. What is suggested, however, is a pilot that will inform an investigation of the more complex issues and problems – a means of exploring and learning about issues rather than a solution for anything other than a few of them.

5 Associated Staffing Requirements and Other Cost Elements

The primary costs of the project will be the staffing costs of the various participants, comprising:

- Project management staff, terminology work research staff, and programming staff at CDLR
- Programming staff at EDINA
- Terminology experts consultancy work
- UKOLN advice and liaison
- Wordmap advice and liaison

³¹ Joint Academic Coding Scheme (JACS)

³² UNESCO Thesaurus: <http://www.ulcc.ac.uk/unesco/>

³³ Medical Subject Headings (MeSH): <http://www.nlm.nih.gov/mesh/meshhome.html>

³⁴ Application Profile Interface (API)

A breakdown of the tasks involved is shown in the table below.

There will be travel, accommodation and subsistence costs associated with the project as project participants are based in Glasgow, Edinburgh, Manchester, Bath and London and meetings are necessary for in-depth discussions on project issues. The project requires either two or three servers depending on the project version chosen, one SRW server and one or two terminology servers. No costs have been set against these by EDINA or CDLR. For the SRW server, EDINA will use a JISC funded server they already have as part of their Data centre provision: hence there will be no cost to the project. For the first terminology server, CDLR will use the SQL pilot server already in place as part of a CDLR-funded cloning of the JISC-funded HILT Phase II pilot. For the second terminology server, CDLR will donate space on an existing CDLR machine with no charge to JISC. The terminology servers and low-power machines and will not be suitable to support a future operational service.

Wordmap licensing costs are included but not itemised in the charges detailed by Wordmap.

Project facet	Roles	Additional Work for dual server version?
Project Management and set-up	CDLR	Yes, longer, more complex
Web-site and Project Plan	CDLR	No
Servers set up and maintenance	EDINA, CDLR	Yes, three servers interact
Set up SRW server, set up illustrative transaction between a requesting client, the SRW server, and a HILT response	EDINA, with some input from CDLR	No
Identify appropriate subject areas to cover in illustrative mappings	CDLR and terminology experts	No
Identify terminology set mapping requirements	CDLR and terminology experts	No
Design and set up extended HILT pilot database or databases	CDLR	Yes, two terminology servers
Add illustrative mappings to database or databases	CDLR	Yes, mapping across two terminology servers
Analyse mark-up requirements and associated needs as regards SKOS-Core mark-ups	CDLR	Yes, mapping across two terminology servers using SKOS-Core
Design and set up interface between HILT database or databases and SRW server – ‘code’ that will accept requests, ‘translate’ them into requests to SQL Server APIs, receive responses, wrap them in SKOS-Core, send them to the SRW server	EDINA and CDLR	Yes, two terminology servers
Adapt local collections database for new pilot requirements; liaise with IESR on their SRW service	CDLR	No
Specify, program and test Go Geo! SRW client; advise HILT on their client work	EDINA	Yes, client uses two terminology servers rather than one
Specify, program and test HILT SRW client	CDLR	Yes, client uses two terminology servers rather than one
Liaise with RDN, advise on their client needs	CDLR	
Set up SRW ‘explain’ facility	EDINA	No, just different
Launch and test pilot	CDLR	No, just different
Evaluate pilot under all five use cases	All involved	No, just different

	participants as appropriate	
Consider issues listed under 'Variant Cases'	CDLR and terminology experts	Yes, client uses two terminology servers rather than one
Re-work various aspects of pilot based on outcomes of tests and evaluations	CDLR and terminology experts	Yes, two terminology servers and client uses two rather than one
Draw conclusions, propose further R&D work, write Final Report	CDLR, EDINA and others	No, just different
Dissemination of progress, outcomes	CDLR, EDINA and others	Yes (longer period)

6 Start and Finish Dates, Project Plan, Scheduling

One server pilot: 15 month project starting November 2005 and ending January 2007

Two server pilot: 21 month project starting mid-October 2005 and ending mid-July 2007

In line with JISC practice, a detailed Project Plan will be written and submitted to JISC in the first three months of work. This will provide a more detailed workplan and schedule. At this stage, it looks likely that scheduling of tasks will be relatively simple, with all key tasks proceeding in parallel. However, this will be examined in detail when writing the Project Plan.

7 Project Management and Evaluation

Day to day management will be the responsibility of the project staff. This **Project Team** will report to a **Project Management Group (PMG)** consisting of the team and a representative from each participant. There will also be a **Project Steering Group (PSG)** comprising representatives from key stakeholders. Evaluation will be conducted within the project.

8 Risks

Risks	Probability	Severity	Score	Action to manage threat
Staffing	1/5	2/5	2	Use partners to fill any gaps, bring in new staff quickly
Organisational	1/5	1/5	1	Plan ahead, monitor daily, act early to fix
Technical	1/5	2/5	2	Adjust pilot as required; note for Final Report

9 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 standards³⁵ will be adhered to where they are appropriate. The specific standards that will impact on the project are SRW, SOAP, SKOS-Core, and (to the extent required to carry out the project aim) Z39.50, SRU, Zthes, and the MARC 21 Concise Format for Classification Data³⁶. The project is aware of the *British standard guide to establishment and development of monolingual thesauri* (BS5723:1987) (ISO2788-1986) and the *British standard guide to establishment and development of multilingual thesauri* (BS6723:1985) (ISO5964-1985) and of updating work going on to merge the two into one standard comprising both parts³⁷ and will consult on this as appropriate. It is also aware of current developments with respect to the Z39.19 'thesaurus standard'³⁸. Wherever appropriate, standard subject and classification schemes – such as LCSH and DDC - will be utilised in the pilot terminology services provided. Accessibility guidelines will be adhered to and the Technology for Disabilities Service (TechDis, <http://www.techdis.ac.uk>) used for guidance and advice.

³⁵ <http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/standards/>

³⁶ <http://www.loc.gov/marc/classification/eccdhome.html>

³⁷ Called 'Structured vocabularies for information retrieval' - BS 8723.

³⁸ <http://www.niso.org/standards/balloting.html>

10 IPR

Should the project be funded, the project partners will comply with the JISC requirements as regards project deliverables and IPR as agreed in the subsequent letter of award.

11 Dissemination Strategy

Dissemination of information would be via the HILT Phase III Web-site, postings to appropriate e-mail lists, papers and news items submitted to professional publications and presentations at seminars and conferences. Key progress reports would be sent to relevant organisations, including, but not limited to, MIMAS (for RDN and IESR) and UKOLN. An active and successful dissemination programme would be a major aim throughout the project.

12 Proposed Exit Strategy

The project will make recommendations about the possible nature and cost of a future service. The partners will maintain the demonstrator service for a reasonable period of time beyond the end of the project, the exact time to be agreed with the JISC.

13 Project Contact

Dennis Nicholson, Director, Centre for Digital Library Research, University of Strathclyde,
Livingstone Tower, 26 Richmond Street, Glasgow, G1 1XH
Tel: 0141 548 2102 Fax: 0141 548 4523 Email: d.m.nicholson@strath.ac.uk

Appendix A: Use Cases to be Addressed in Proposed Project

Use case #1

Single two-stage process with a 'switch' used to turn stage two on and off.

~~~

Client sends request to HILT server for data on a subject search term ('teeth', say).

~~~

If request stage two switch at **off**, and teeth is the term, the server applies the Wordmap (or equivalent) search_for_wordsets function with teeth as 'search_term' parameter and returns all senses of wordsets (wordset id and the tree) that have word phrases that match 'teeth'.

~~~

If request stage two switch at **on**, server **also** applies the Wordmap (or equivalent) get\_features function and returns, **in addition**, a record for each feature of the wordset. The features retrieved are Dewey number associated with a term, and the mappings available. For example, in the case of one possible result of a search for teeth, the Dewey number is 611.314, and mappings are held in the database for LCSH (statistical mapping) and the MeSH taxonomy (singular plural match).

~~~

Some services would do the above in a single call, others as two separate calls. The use of the DDC number to search for appropriate collections in IESR would be a service end function, although HILT would also provide that option as an additional call and would maintain the code for the DDC algorithm and make it available to the community. Disambiguation would be a service end function based on data sent back from HILT.

Use case #2

BIOME/Go Geo!/RDN #1.

~~~

User types a term into service-end search box. Term is sent to HILT to generate an additional set of search terms that can be queried against the sending service database.

~~~

Web form created listing the original term, and the initially expanded/ derived terms, and presented back to the user

~~~

User given feedback on origin of derived term.

~~~

User selects terms from Web form for further expansion via HILT The results of the expansion are then inserted into the Web form.

~~~

User gets functions to:

Map plural to singular terms; Map synonyms to main terms in thesauri; disambiguate terms such as COLD; Correct simple spelling/typographic errors.

~~~

Having used these various functions, user selects one or more terms derived from the mapping process and these are used to search the requesting service database. Results are displayed in browser without substantial differences to the non-enhanced search.

~~~

The use case should allow for two possibilities – one is that user interaction is all handled at requesting service end rather than HILT end, the other that HILT will handle the interaction. The question of which is the best/most practical/most economic approach is most likely to be examined in the context of the likely M2M demonstrator project.

### Use case #3

BIOME/Go Geo!/RDN #2.

~~~

User types a term into search box. The term is sent to HILT to generate a set of additional search terms that can be used to search the requesting service database.

~~~

If any simple spelling or typographical errors are identified an intermediate screen offering an alternative spelling is presented along the lines of Google, "Did you mean?"

~~~

After acquiring a correct spelling the term is sent back to HILT for further expansion.

~~~

The original and derived terms are passed to the requesting service database, a search is run against it and a result set is returned. The user notices no substantial differences in the result set (apart from hopefully a larger number of results) between the non-enhanced query and a query enhanced first by via M2M interaction with HILT.

~~~

The question of whether it is better/more practical/ more economic for HILT to provide the 'did you mean' interface (as opposed to just the data that drives it) is again one for the future M2M demonstrator project.

Use case #4

Browse-based use cases.

~~~

Four situations to consider have been identified under this heading:

~~~

(a) Browse offered by HILT in response to a 'no hits from HILT' situation in response to a service-end request.

~~~

(b) Browse of appropriate scheme offered by HILT when requested by user in response to a particular term provided by HILT from the scheme in question.

~~~

(c) Browse of (a) handled by requesting service rather than by HILT.

~~~

(d) Browse of (b) handled by requesting service rather than by HILT.

#### **Use case #5**

Improved precision based use cases.

~~~

Two of these are covered by browse use cases, and by disambiguation in use-cases 1 and 2.

~~~

We should probably also consider requests to HILT for information on narrower and related terms and (possibly) cross-scheme variations on this.

#### **Variant Cases**

Consideration needs to be given to the effects of having a phrase as the search term and of the effects of terms with large mappings. Are there searches or circumstances for which the effects of having the second stage switched on are such that they hit response times and where result-sets are excessively large? Also, are there cases where services that are running the DDC IESR search need to make additional calls to the HILT server for supplementary information? Use cases also need to consider the situation where requesting services, or services identified through IESR, use more than one subject scheme.

## **Appendix B: HILT M2M Feasibility Study Final Report: Executive Summary and Recommendation**

### **Aims and Objectives**

The project was asked to investigate the feasibility of developing SOAP-based interfaces between JISC IE services and Wordmap APIs and non-Wordmap versions of the HILT pilot demonstrator created under HILT Phase II and to determine the scope and cost of the provision of an actual demonstrator based on each of these approaches. In doing so it was to take into account the possibility of a future Zthes<sup>39</sup>-based solution using Z39.50 or OAI-PMH and syntax and data-exchange protocol implications of eScience and semantic-web developments.

After discussions with the main project partners, and with UKOLN, it was agreed that the primary concerns of the study should be an assessment of the feasibility, scope, and cost of a follow-up M2M pilot that considered the best options in respect of:

- Query protocols (SOAP, Z39.50, SRW, OAI) and associated data profiles (e.g. Zthes for Z39.50 and for SRW)
- Standards for structuring thesauri and thesauri-type information (e.g. the Zthes XML DTD and SRW version of it and SKOS-Core<sup>40</sup>)

The study was carried out within the allotted timescale, with a Final Report submitted to JISC on 31<sup>st</sup> March 2005 as scheduled. It was concluded that an M2M pilot was feasible.

### **Methodology and Outcomes**

The project followed the methodology set out in Section 3.2 of the M2M Final Report. The main outcomes were:

- A simple SOAP M2M demonstrator (see <http://nevis.ed.ac.uk:8080/asp-misc/public/hilt.asp>).
- A report assessing use cases, protocols and mark-ups.
- A draft follow-up proposal for discussion.
- This Final Report

The report assessing use cases, protocols and mark-ups is included in the Final Report as Appendix D, the draft follow-up project proposals as Appendix E. Both are summarised below.

### **Use Cases, Protocols and Mark-ups Summary**

Because it is a protocol designed for harvesting metadata rather than searching, OAI-PMH does not look appropriate for the task of providing the services required of HILT by the five use cases. SRW and Z39.50 both appear able to handle the issues that arise, although implementing a Z39.50-based M2M pilot service may involve greater complexity than would be entailed in implementing an SRW-based pilot service. On mark-up for returned classification, thesaurus, and mappings data, Zthes, SKOS-Core, and MARC<sup>41</sup> all look adaptable to the task, although Zthes appears to be less suited to handling classification data than the other two. MARC has at least one advantage in that some major thesauri are available in that format<sup>42</sup>. SKOS-Core is more flexible and more suited to the Web Services perspective and the Semantic Web community.

With this as background, there appear to be two sensible options as regards a baseline follow-up M2M pilot project. The simplest one would implement SRW, probably with SKOS-Core (but a case could be made for MARC and even ZThes). A more complex (and inevitably more expensive) version would

---

<sup>39</sup> <http://zthes.z3950.org/>

<sup>40</sup> <http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/>

<sup>41</sup> Although, in the event, it is likely that MARCXML (<http://www.loc.gov/standards/marcxml/>) rather than 'standard' MARC would be the choice for a practical pilot

<sup>42</sup> See Diane Vizine-Goetz, Carol Hickey, Andrew Houghton and Roger Thompson. Vocabulary Mapping for Terminology Services, Journal of Digital Information, Volume 4 Issue 4, Article No. 272, 2004-03-11, available at <http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/>.

seek to offer both SRW and Z39.50 services (perhaps through an SRW-Z39.50 gateway<sup>43</sup>) and would offer a choice of Zthes, SKOS-Core, and MARC mark-ups. A sensible compromise would be to implement the simplest approach, but ensure that the pilot design provided for later developments encompassing the more complex version. This implies a follow-up pilot that would:

- Use the SRW protocol only, but be designed so that a possible extension offering other protocols such as Z39.50 could be introduced at a later date.
- Use SKOS-Core as the 'mark-up' for sending out terminology and classification set responses, but be designed so that adding other formats such as MARC and Zthes would be later option.

A further possible variation is a two-server pilot, perhaps using SKOS-Core concept URIs as the basis for mapping between different schemes on the two servers. On the face of it, there is the basis in this for an approach that might ultimately lead to a matrix of servers being available with mappings between schemes being based on URIs and being built up slowly but surely over a long period of time. This might implement the kind of solution HILT had envisaged to subject interoperability issues in a way that would spread the cost and effort over many organisations and a longer period of time. Such an approach would not be any cheaper than setting up the kind of service initially envisaged by HILT, but it would spread the cost over a number of players and the effort over a longer period of time. If the one server pilot option were chosen, SKOS-Core concept URIs should be used to identify concepts uniquely, so that a distributed version of the service could be a later option.

### **Proposed Follow-up Project**

After discussion within the project, it was concluded that HILT Phase III (the proposed M2M Pilot) should aim to create an M2M version of the current HILT Pilot, but with facilities extended to take account of the five use cases drawn up under the HILT M2M Feasibility Study. With JISC's agreement, two versions of this are costed – a single server version and a distributed server version. These are identical in all respects except one – that is, version 2 distributes the terminology service provided by the HILT pilot across two servers. This is likely to be a more expensive option, and entails undertaking more work and addressing additional technical issues. However, it also allows a far more realistic pilot situation to be created, one that echoes the world of distributed terminology services envisaged in the JISC I.E. and the web services world generally. There is a case for building the single service version first, then treating the distributed version as a new project or a new project stage. However, there is also a case for arguing that building a single server version first may result in a set-up that could prove difficult to adapt to a distributed set up. It might also be suggested that, if the future of terminology services is likely to be distributed (as appears to be true), then JISC needs to start investigating the issues sooner rather than later to ensure it has input to developing standards and positions in the area and can keep abreast of the needs of the JISC IE as it develops in this wider context. This is largely a matter of strategy and of cost – and the project has left the matter in the hands of JISC (with the agreement of the relevant Programme Director). More detail on both options is provided in Appendix E of the Final Report. A position on whether the pilot should be based on Wordmap or on a more generic SQL-based solution will be taken in the context of the project costing exercise. There is a case (see Appendix E) for each of these options, and it is not impossible that this issue may also require JISC involvement in a decision.

### **Costs**

An exercise to cost a follow-up project based on either a single or distributed solution as described above has been undertaken.

### **Recommendation**

It is recommended that JISC fund one of the two versions of the follow up project outlined above, basing their decision on a formal and costed bid submitted mid-April.

---

<sup>43</sup> SRW-Z39.50 gateways are known to exist. It would be interesting to determine whether a Z39.50-SRW gateway also exists. This would allow an SRW-based service to be created with Z39.50-based requests also supported through the gateway.

## Glossary

**API:** Application Programmers Interface

**BIOME:** BIOME is a collection of gateways providing access to evaluated, quality Internet resources in the health and life sciences, aimed at students, researchers, academics and practitioners.

**DDC:** Dewey Decimal Classification

**DTD:** Document Type Definition

**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.

**e-Science:** Research Councils UK (<http://www.rcuk.ac.uk/escience/>) describe e-Science in the following terms 'In the future, e-Science will refer to the large scale science that will increasingly be carried out through distributed global collaborations enabled by the Internet. Typically, a feature of such collaborative scientific enterprises is that they will require access to very large data collections, very large scale computing resources and high performance visualisation back to the individual user scientists'.

**FE:** Further Education

**HE:** Higher Education

**Go Geo:** A tool designed to help users find details about geo-spatial datasets and related resources within Great Britain tertiary education and beyond. A trial service provided by EDINA.

**HILT:** High Level Thesaurus

**IESR:** JISC Information Environment Service Registry

**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

**OAI-PMH:** The Open Archives Initiative Protocol for Metadata Harvesting

**OCLC:** Online Computer Library Center

**RDN:** Resource Discovery Network

**Semantic Web:** A collaborative initiative led by the W3C, the Semantic Web provides a common framework that facilitates data sharing and reuse across application, enterprise, and community boundaries.

**SKOS-Core:** SKOS Core supports the RDF description of language-oriented knowledge organisation systems (KOS) such as thesauri, glossaries, controlled vocabularies, taxonomies and classification schemes.

**SOAP:** Simple Object Access Protocol

**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.

**Wordmap:** A commercially available taxonomy management software application that supports management of multiple controlled vocabularies.

**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.