Final Report Cover Sheet: HILT M2M Demonstrator Feasibility Study

Project Acronym	HILT
Project ID	N/A
Project Title	HILT M2M Demonstrator Feasibility Study
Start Date	Mid January 2005
End Date	End March 2005
Lead Institution	Strathclyde University: Centre for Digital Library Research
Project Director	Dennis Nicholson (d.m.nicholson@strath.ac.uk)
Project Manager & contact details	Emma McCulloch Centre for Digital Library Research Department of Computer and Information Sciences, Livingstone Tower, 26 Richmond Street, Glasgow, G1 1XH Tel: +44 (0)141 548 4752 Email: e.mcculloch@strath.ac.uk
Web URL	http://hilt.cdlr.strath.ac.uk/hiltm2mfs/ (M2M pages – main HILT site at http://hilt.cdlr.strath.ac.uk/).
Programme Name	Shared Services Programme
Programme Director	Leona Carpenter
Document Title	Final Report: HILT M2M Demonstrator Feasibility Study
Author	Dennis Nicholson
Date	March 31st 2005
Access	JISC and Internal

Document History	**Date**	**Comments**
Version 1.0	18.01.05	Initial draft, compiled by DN
Version 2.0	03.02.05	Interim report for JISC, compiled by DN
Version 2.1	04.02.05	New version of UKOLN report added
		Sent to JISC 7/2/5 as Interim report
Version 3.0	17.02.05	Sent to partners 11.02.05, to JISC 17.02.05
Version 3.1	31.03.05	Sent to JISC 31.03.05

HILT: High-Level Thesaurus Project M2M Feasibility Study

Final Report To JISC, 31st March 2005

Main author: Dennis Nicholson

CDLR ▪ EDINA ▪ BIOME ▪ UKOLN ▪ Willpower

HILT M2M Feasibility Study: Final Report

Contents

Main Participants:

The study entails collaboration between four partner organisations:

- **CDLR**, providing expertise in the design of the existing terminologies server pilot, terminologies mapping issues, report-writing, and project management.
- **EDINA**, providing expertise in M2M interface design, server-end programming requirements, and additional advice on client-end requirements.
- **BIOME**, providing expertise on client-end M2M requirements at BIOME, including the needs of BIOME users in respect of (transparent M2M) use of a terminologies service and RDN representation in the project.
- **Wordmap**, providing training in Wordmap APIs, expertise, and technical and software support to the project.

One of the terminology experts from earlier stages of HILT is also being consulted, as are relevant UKOLN personnel.

Consortium Agreement accepted by JISC partners CDLR, EDINA and BIOME and sent to JISC: 8th March 2005. Wordmap and Willpower were not included in the Consortium Agreement, since standard commercial financial arrangements were in place in each case.

Individual Participants

CDLR	Dennis Nicholson, Emma McCulloch, Alan Dawson, Anu Joseph, George Macgregor and Gordon Dunsire
EDINA	Christine Rees, David Medyckyj-Scott, Edward G Boyle, Ben Soares and Tim Stickland
BIOME	Bob Parkinson and Donald McKay
UKOLN	Rachel Heery and Andy Powell
Wordmap	Bill Hutchison and Dave Peacock
Terminology expert	Leonard Will (Willpower)

0. 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[1]-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[2])

The study was carried out within the allotted timescale, with this Final Report submitted to JISC on 31st March 2005 as scheduled. The detailed proposal for a follow-up project is currently under discussion and will be finalised – as agreed with JISC – by mid-April. It was concluded that an M2M pilot was feasible. A proposal for a follow-up M2M pilot project has been scoped, and is currently being costed.

Methodology and Outcomes

The project followed the methodology set out in Section 3.2 of this 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 this 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 5 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

[1] http://zthes.z3950.org/

[2] http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/

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[3] 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[4]. 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 seek to offer both SRW and Z39.50 services (perhaps through an SRW-Z39.50 gateway[5]) 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 (see Appendix D). With JISC’s agreement, two versions of this are being costed – a single server version and a distributed server

[3] 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

[4] 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/.

[5] 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.

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 I.E. 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. 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 is underway. It has been agreed with JISC that this can be provided shortly after the end of the study.

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 to be submitted by mid-April.

1. Background

Background: HILT I and II

Ensuring that FE and HE users of the JISC Information Environment[6] (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:

- 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 I[7]).
- 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).

The HILT Phase II pilot was based on commercial terminologies management software called Wordmap, which was adapted by the HILT team to meet the requirements of the terminologies server pilot.

Background: Outstanding Issues, including M2M Operations

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 machine-to-machine (M2M) interfaces suitable for being interrogated by other components in the JISC I.E. architecture. HILT Phase II has developed a range of facilities currently only available through a direct user interface. A HILT M2M interface would 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.

Background: M2M and HILT Phase II, including UKOLN[8] Recommendations

The HILT Phase II proposal indicated that it would be 'difficult in such a relatively small, relatively low-cost project to fully investigate M2M use of the pilot facility in an operational sense'. It therefore proposed to focus primarily on the use of the demonstrator service by end users and cover the M2M needs by 'examining the requirement for this on an ongoing basis at a mainly theoretical level'.

UKOLN undertook the examination of the M2M requirement and made the following 'concluding recommendations':

[6] http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/
[7] http://hilt.cdlr.strath.ac.uk/Reports/FinalReport.html
[8] http://www.ukoln.ac.uk/

- Provide M2M demonstrator services based on controlled vocabularies mapped within Wordmap. Develop SOAP[9] based interfaces between JISC IE components and Wordmap APIs (Application Programmers Interfaces). Use these services in the short term as an aid to firm up use cases, in the longer term as a basis for pilot service if this approach is still appropriate at that stage.
- Carry out investigative implementation of Zthes[10] based solution, whether data is exchanged using Z39.50 or OAI-PMH, with a view to taking advantage of standards based structured controlled vocabularies (particularly faceted vocabularies) as they become available from third party agencies.
- Track developments within the Semantic Web and eScience activities to ensure decisions made now concerning both syntax for structuring vocabularies and data exchange protocols take account of forward compatibility.

2. Aims and Objectives

Taking into account the possibility of a future Zthes-based solution using Z39.50 or OAI-PMH and syntax and data-exchange protocol implications of eScience and semantic-web developments, the project will:

- Investigate the feasibility of developing SOAP-based interfaces between JISC I.E. components and Wordmap APIs or a non-Wordmap alternative based on storing terminology mappings in an SQL compliant database.
- Determine the scope and cost of the provision of an actual demonstrator based on each of these approaches.
- Create an Interim Report by 7th February 2005 indicating early progress and submit to JISC.
- Conduct investigations on 1 and 2 above with the aim of presenting a draft final report to JISC by 18th March 2005.
- Finalise the recommendations and present the report to JISC, together with a project Completion Report, by 31st March 2005.

3. Methodology, Including Research Plan, Standards, Evaluation

3.1 Overview

Overall Approach

The project will take the current pilot demonstrator and the subject schemes mapped within it (DDC, LCSH, UNESCO and MeSH[11]) as its starting point and be concerned only with the requirements and feasibility of building an M2M demonstrator of the current service[12] with particular reference to its use by a part of the BIOME RDN service. No significant new terminologies mapping work will be undertaken at this stage, although the possibility of looking at a small

[9] http://www.w3.org/TR/soap12-part1/

[10] See http://www.loc.gov/z3950/agency/profiles/zthes-04.html

[11] Note that MeSH in particular has only a few illustrative mappings in the current pilot, and UNESCO has only a few thousand. The whole of DDC21 is there, together with large numbers of mappings to LCSH provided with the OCLC DDC distribution.

[12] The service is currently provided only through the user interface.

number of additional MeSH mappings will be considered should this prove helpful in examining BIOME requirements (MeSH is one of the schemes in use at BIOME).

With this as the background, the following outline methodology is planned:

- Conduct in-depth preparatory consultations with partners, UKOLN, terminology experts
- Create initial draft Interim Report (internal) with outline research plan
- Conduct Wordmap API email consultations
- Develop and submit Interim Report to JISC (7th February)
- Create and develop outline Final Report based on Interim Report and detailed research plan for in-depth technical and costs stages of work (begins 3rd February)
- Conduct main technical investigations as detailed in the research plan
- Participate in Wordmap API related conference call and in follow-up consultations with partners, UKOLN, terminology partners
- Conduct cost assessments as detailed in the research plan
- Create and agree draft Final Report and send to JISC
- Finalise Final Report through partner discussions and consultations with JISC, UKOLN, terminology experts
- Submit Final Report
- Submit Completion Report
- Conduct dissemination via web-site, presentations, papers etc. (as and when appropriate during and after the study).

Note that the methodology will be examined on an ongoing basis to ensure flexibility in meeting project requirements as new data emerges from training sessions and other investigative work.

Standards

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 I.E. standards[13] will be adhered to where they are appropriate. The aim is to look at SOAP as the basis of the M2M functionality and to take cognisance of other standards such as Zthes in carrying out the study. 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 parts[14] and will consult on this as appropriate (Leonard Will is involved in the updating process and will act as an external consultant to the project).

Evaluation

This is a small ten-week project and evaluation will perforce be an ongoing process to ensure that conclusions reached are based on sound research and good practice.

[13] http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/standards/

[14] Called ‘Structured vocabularies for information retrieval’ - BS 8723.

Expert advice will be sought from project partners, cross-checked by other partners, and checked further through consultation with the external terminology expert and advisors at UKOLN and at JISC generally. The one deliverable is a Final Report which will express the findings of the project and will be evaluated by the various internal and external participants as it goes through successive stages: draft Interim Report, Interim Report, outline Final Report, draft Final Report, Final Report.

The factors evaluated will be:

- Compliance with appropriate standards, practices, and trends in relevant protocols (SOAP and Zthes in particular), terminology and thesauri construction, communication, use, and mapping, relevant communities (RDN, JISC I.E., semantic web, eScience)
- Validity of assumptions made and estimated unit costs applied in costing a pilot HILT M2M demonstrator project.

Quality Assurance

The procedures to ensure that project outputs comply with JISC technical standards and best practice are listed above. Evidence of compliance will be based on expert advice from the variety of sources, but particularly from UKOLN and the external terminology experts.

Dissemination

Dissemination of information will be via the HILT web-site, papers and news items in professional or academic journals, and presentations at seminars and conferences.

Sustainability

The question of an exit or sustainability plan is not relevant to this feasibility study. Assuming a positive result from the study, it is expected that there will be a follow-up application for funding to develop an M2M demonstrator, but this is seen as an additional process to be undertaken in discussion with JISC after the end of the study.

3.2 Research Plan

The research plan was agreed in the first few weeks of the project and followed in general outline. It was noted in the Interim Report that few adjustments to the methodology would probably be necessary as the work developed, and this proved to be the case. The changes needed were not major, involving in the main the order in which it proved practical to tackle some of the issues and the approach taken to reporting them.

Outline Research Plan

- Plot model of whole BIOME to HILT transaction set for a simple interaction involving all steps from subject query to retrieval from remote service using an actual subject search example likely to arise within BIOME.

- Design SOAP-based version of this model, identify syntax, data exchange protocol, and API interfacing issues that arise, and specify how the project will deal with them.
- Based on the resulting research, determine the feasibility of developing a SOAP-based version of the simple BIOME to HILT interaction. Conduct for both a Wordmap-based pilot and an SQL RDBMS-based pilot.
- Agree a representative set of use scenarios between BIOME and HILT[15], identify any new issues that arise from these, determine the feasibility of developing a SOAP-based version of BIOME to HILT interactions covering all use scenarios. Conduct for both a Wordmap-based pilot and an SQL RDBMS-based pilot.
- Examine the possible additional implications for delivering a SOAP-based version of BIOME to HILT interactions covering all use scenarios of a possible future need for a service also offering (1) a Zthes-based solution using SRW or OAI-PMH (2) a solution that takes account of the syntax and data-exchange protocol implications of eScience and semantic-web developments (see SWAD-Europe project). Determine whether changes to the design of the SOAP-based interface are required to ensure harmonisation with these possible future needs and whether such changes affect the feasibility of building a SOAP-based version of BIOME to HILT interactions covering all use scenarios. Conduct for both a Wordmap-based pilot and an SQL RDBMS-based pilot.
- Assuming that an adequate SOAP-based interface is feasible, using either Wordmap or an SQL system or both:

    - Agree the scope of a project for creating an operational BIOME to HILT M2M pilot based on the agreed use scenarios identified earlier.
    - Determine the cost of such a project using one or other or both Wordmap and SQL based solutions.
    - If both options are feasible, compare costs and benefits of each.
    - Make recommendations about a possible future project.

4. Outputs and Results

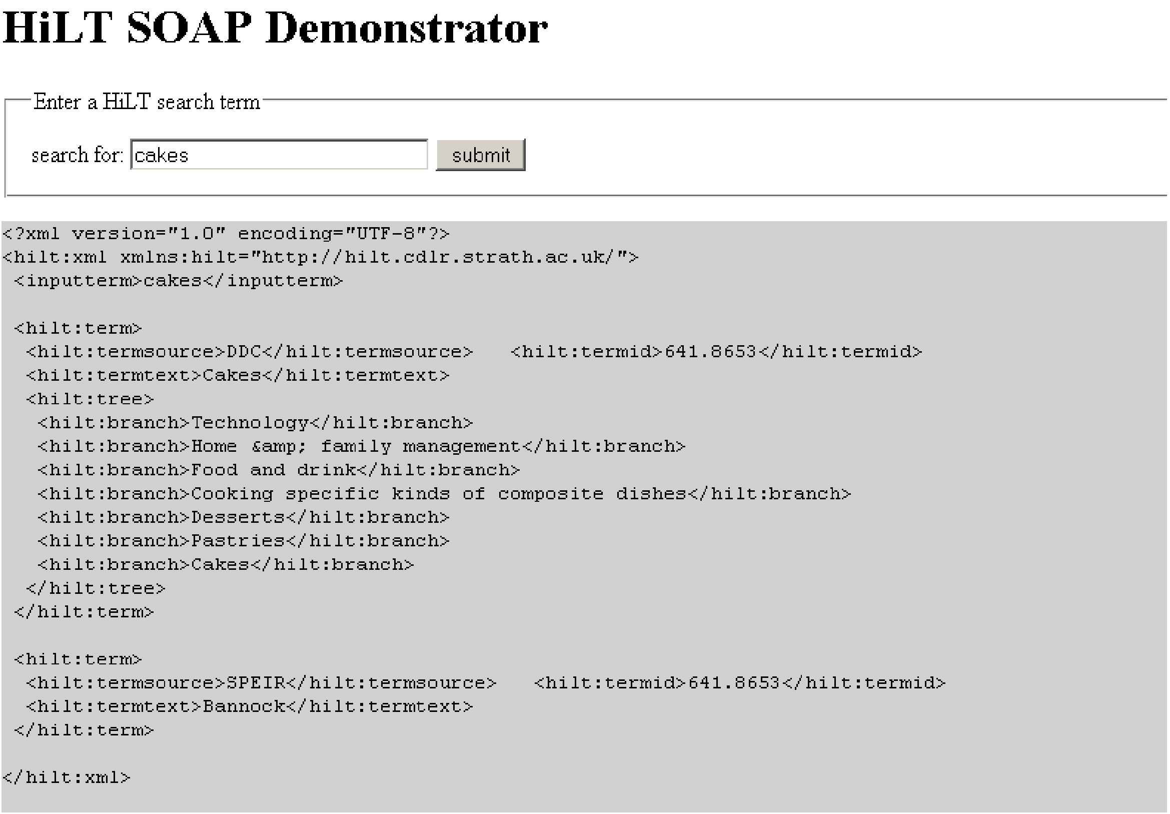
Outputs and results fall under the following headings:

SOAP Demonstrator

EDINA, working with CDLR, have put up a simple SOAP demonstrator at: http://nevis.ed.ac.uk:8080/asp-misc/public/hilt.asp

An illustration of a HILT server response via the SOAP demonstrator is shown on the next page. In this example shown, the term ‘cakes’ is input and a response is sent back in XML showing details of the appropriate DDC caption and (bottom of screen), a mapping of the Scots term ‘bannock’ from a terminology set used in the SPEIR project (see http://cdlr.strath.ac.uk/projects/speir.htm).

[15] http://www.w3.org/2001/sw/Europe/200311/thes/Use_cases_Thes_Service.html may be useful



HiLT SOAP Demonstrator

Enter a HiLT search term

search for: cakes

```
<?xml version="1.0" encoding="UTF-8"?>
<hilt:xml xmlns:hilt="http://hilt.cdlr.strath.ac.uk/">
 <inputterm>cakes</inputterm>

 <hilt:term>
  <hilt:termsource>DDC</hilt:termsource>   <hilt:termid>641.8653</hilt:termid>
  <hilt:termtext>Cakes</hilt:termtext>
  <hilt:tree>
   <hilt:branch>Technology</hilt:branch>
   <hilt:branch>Home &amp; family management</hilt:branch>
   <hilt:branch>Food and drink</hilt:branch>
   <hilt:branch>Cooking specific kinds of composite dishes</hilt:branch>
   <hilt:branch>Desserts</hilt:branch>
   <hilt:branch>Pastries</hilt:branch>
   <hilt:branch>Cakes</hilt:branch>
  </hilt:tree>
 </hilt:term>

 <hilt:term>
  <hilt:termsource>SPEIR</hilt:termsource>   <hilt:termid>641.8653</hilt:termid>
  <hilt:termtext>Bannock</hilt:termtext>
 </hilt:term>

</hilt:xml>
```

Note that this illustrates a simplified version of the real situation and is intended only to show the feasibility of the M2M interaction. Inputting ‘cakes’ as a term in Dewey for Windows returns five possible numbers with distinct captions: 641.815 Breads and bread-like foods; 641.8653 Cakes; 641.8659 Danish, French, related pastries; 664.7525 Pastries; 664.768 Formula feeds. The notes make it clear that some kinds of cakes occur at each of these places. This is a non-trivial problem and will introduce complications in the practical implementation, requiring human intervention. Also, the identical coding of each element as <hilt:branch> would not be appropriate in a full pilot which would have to encode the hierarchical relationships entailed.

Use Cases; Feasibility of M2M Pilot Based on Various Protocols and Mark-ups

This consists of the report included as Appendix D. Five use cases are described and the feasibility of building an M2M HILT demonstrator to deliver the services they imply using a range of protocols (SRW, OAI-PMH, and Z39.50) and mark-ups designed to handle thesauri and classification schemes (Zthes, SKOS-Core, MARC) is assessed. A summary of its content and conclusions is given in sections 5 and 6 below, which echo the Executive Summary in this Final Report.

Costing a Follow-up M2M Pilot

This work is currently in process. See under section 5 below.

5. Outcomes, Conclusions, Implications

Aims and Objectives

The project was asked to investigate the feasibility of developing SOAP-based interfaces between JISC I.E. 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-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[16])

The study was carried out within the allotted timescale, with this Final Report submitted to JISC on 31st March 2005 as scheduled. The detailed proposal for a follow-up project is currently under discussion and will be finalised – as agreed with JISC – by mid-April. It was concluded that an M2M pilot was feasible. A proposal for a follow-up M2M pilot project has been scoped, and is currently being costed.

Methodology and Outcomes

The project followed the methodology set out in Section 3.2 of this 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 this 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 5 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 all look adaptable to the task, although Zthes appears to be less suited to handling classification data than the other two are. MARC has at least one advantage in that some major thesauri are available in that format. SKOS-Core is more flexible and more suited to the Web Services perspective and the Semantic Web community.

[16] http://www.w3.org/2001/sw/Europe/reports/thes/1.0/guide/

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 seek to offer both SRW and Z39.50 services (perhaps through an SRW-Z39.50 gateway[17]) 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 (see Appendix D). With JISC's agreement, two versions of this are being 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

[17] 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.

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 I.E. 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. 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 is underway. It has been agreed with JISC that this can be provided shortly after the end of the study.

6. 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 to be submitted by mid-April.

Appendix A: Workpackage Outline, With Dates and Dependencies

	Task	**Responsibility**	**Start**	**End**	**Scheduled Outputs**	**Deliverables**	**Depends on steps**
1	In-depth preparatory consultations with partners, UKOLN, terminology experts	CDLR	17/1/5	28/1/5			**N/A**
2	Initial draft Interim Report with outline research plan (internal)	CDLR, All	17/1/5	28/1/5			**1**
3	Wordmap API email consultations	All	24/1/5	18/3/5			**1-2**
4	Develop, submit Interim Report to JISC	CDLR	31/1/5	7/2/5	**Interim Report**		**1-3**
5	Create and develop outline Final Report based on Interim Report and detailed research plan for in-depth technical and costs stages of work	All	3/2/5	25/2/5			**1-5**
6	Main technical investigations (detailed plan to be agreed at step 5)	All	14/2/5	18/3/5			**5**
7	Participate in Wordmap API related conference call and in follow-up consultations with partners, UKOLN, terminology partners	CDLR and others as appropriate	14/2/5	18/3/5			**1-6**
8	Cost assessments (detailed plan to be agreed at step 5)	CDLR, All	28/2/5	18/3/5			**7**
9	Draft Final Report created, agreed, and sent to JISC	CDLR, All	21/2/5	18/3/5	**Draft Final Report**		**1-8**
10	Finalise Final Report through partner discussions and consultations with JISC, UKOLN, terminology experts	CDLR, All	21/3/5	25/3/5			**1-9**
11	Submit Final Report	CDLR	28/3/5	31/3/5	**Final Report**	**Final Report**	**1-10**
12	Other elements: completion report	CDLR	28/3/5	31/3/5	**Completion Report**		**1-10**
13	Other elements: Dissemination	CDLR, All	17/1/5	31/9/5	**Dissemination**		**1-11**

Appendix B: Report on Conference Call with UKOLN 28 January 2005

Points Agreed in Conference Call with UKOLN 28 January 2005

Present: Rachel Heery (RH), Andy Powell (AP), Dennis Nicholson (DN)

Confirmation that primary concerns of HILT M2M study were:

- 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)

Notes:

- Choice of standards to be used depends on how HILT is to be placed on continuum from research project to production service. It might be sensible to look at 2 options for example, short term delivery of service using Wordmap, medium term based on standard mappings. Need to remember that Wordmap does not offer a standard structure for controlled vocabularies as a basis for interoperable query and data exchange. The Feasibility Study should consider the possibilities and should also discuss the options with JISC.

- It would also be sensible to take account of any availability of structured KOS (such as Dewey) from a third party (such as OCLC).

Process for dealing with the associated HILT M2M Feasibility Study issues:

1. Identify use cases in HILT based on BIOME examples (and perhaps have AP look at these to get a wider RDN view - note, project also agreed to take soundings from Go-Geo on this). Agreed that the number of use cases would have to be limited given the time and resource available.
2. Determine whether, given appropriate requirements for the use cases, the Zthes/Z39.50 or Zthes/SRW[18] request set can handle the required exchanges of terminology information between BIOME and HILT or whether extensions to the set would be required (hopefully not). Pose same question as regards OAI. Tabulate which use cases can and cannot be handled by the request sets for each protocol.
3. Based on the use cases, determine what is required in terms of query protocol and data exchange formats to handle the associated exchanges of terminology information between BIOME and HILT.
4. Determine whether the mark-up recommended for Zthes and the Zthes profile for SRW can be used or adapted to handle the mark-up required to deal with the exchanges of terminology information required for the various use cases. And whether Wordmap API can handle a Zthes structured query?
5. Determine whether the SKOS-Core mark-up can be used or adapted to handle the mark-up required to deal with the exchanges of terminology information required

[18] SRW is a SOAP implementation of Z

for the various use cases. And whether Wordmap API can handle a SKOS structured query?

6. Discover (by discussing with Diane Vizine-Goetz) how OCLC see the future in terms of standards for structuring and protocols for delivering DDC (and, if DV-G has information, LC and LCSH).

   DV-G's article at http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/ may have relevant information here.

7. Looking at the results from 3-6 above, and bearing in mind also that any solution for HILT should ideally be adoptable by any other future JISC I.E. terminology services, agree an approach that would be the basis of the HILT API for the proposed follow-up demonstrator project.

Appendix C: Executive Summaries from Interim Report and Draft Final Report

Summary of Progress as at 3rd February 2005 (Interim Report):

- The project began work as scheduled on 17th January and a Project Plan was submitted on 19th January. This identified successive versions of this report – from draft Interim Report to Final Report – as the main project management tool.
- A draft version of this Interim Report was prepared and the approach it described agreed. A second draft was prepared to meet the JISC deadline of 7th February and circulated to the team for comment.
- A management team email list (LIS-HILT-MGT@Jiscmail.ac.uk) was set up. This includes 5 CDLR, 5 EDINA, 2 BIOME, 2 UKOLN, 2 Wordmap and 2 OCLC staff members, plus two external terminology experts.
- A project web-site has also been set up (http://hilt.cdlr.strath.ac.uk/hiltm2mfs/).
- A meeting of technical and project management staff from BIOME, CDLR and EDINA was held in Glasgow and initial agreement reached on a number of matters. In particular, it was accepted that:
  1. The key elements of the draft Interim Report, concerned with the research plan and likely outputs and results from this (see Sections 3.2 and 4 below), were accepted as the way forward.
  2. A trip to Bath to undertake general training on Wordmap APIs was not the best use of the limited numbers of person hours available to the project. Accessing Wordmap expertise online, via email and by telephone, would be sufficient for the purposes of the study.
  3. The development of BIOME-based use cases would be the main basis for drawing out functional requirements for the M2M Feasibility Study.
  4. Discussion on use cases would take place on LIS-HILT-MGT and would be led initially by BIOME. Account would also be taken of Go-Geo requirements.
- A conference call between Andy Powell and Rachel Heery of UKOLN and Dennis Nicholson of CDLR took place on 28th January to discuss HILT M2M related protocol (SOAP, Z39.50, SRW, OAI) and mark-up (Zthes/SRW XML DTDs, SKOS-Core) issues and other project matters. A summary report is included below as Appendix B.
- As indicated in the research plan (see Section 3.2, first bullet point), a description of a simple use case has been mapped out by CDLR and a discussion on the issues related to a SOAP-based version of the related BIOME-HILT transactions begun. The initial draft of the use case is currently being discussed with EDINA to draw out issues related to designing a SOAP-based version that will include this and other transaction sets. The first draft of the outline use case and the first EDINA response is included as Appendix C below **[Note from DN – this Appendix C refers to Appendix C in the Interim Report, not Appendix C in this Final Report].**

- EDINA, working with CDLR, have put up a simple SOAP demonstrator at: http://nevis.ed.ac.uk:8080/asp-misc/public/hilt.asp
- Online discussion of a representative set of use cases - based on BIOME needs and supplemented with additional input from Go-Geo and UKOLN and with generic examples from HILT/CDLR is scheduled to begin soon.
- Work on a Consortium Agreement is well-advanced, but full agreement has not yet been reached. This is a matter of sorting out details rather than of fundamental disagreement.

Interim Recommendation

The work should continue along the lines described in this document (which echoes and supplements the Project Plan). In particular, the research plan outlined in Section 3.2 should be worked through as the primary mechanism for delivering project results and outcomes.

Executive Summary as at: 18th March 2005 (Draft Final Report):

Assessment: Use Cases, Protocols and Mark-ups

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 5 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 all look adaptable to the task, although Zthes appears to be less suited to handling classification data than the other two are. MARC has at least one advantage in that some major thesauri are available in that format. SKOS-Core is more flexible and more suited to the Web Services perspective and the Semantic Web community.

The picture that is beginning to emerge in respect of the follow-up pilot is that there are two sensible options. 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 seek to offer both SRW and Z39.50 services (perhaps through an SRW-Z39.50 gateway[19]) and would offer a choice of Zthes, SKOS-Core, and MARC mark-ups.

Another possibility 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 there 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. This would not be any cheaper than setting up the kind of service initially envisaged

[19] 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.

by HILT, but it would spread the cost over a number of players and the effort over a longish period of time. Obviously, the devil would be in the detail.

Assessment: Other Issues Arising from Use Cases

Wordmap and SQL Server APIs

None of the use cases mapped by project partners entailed a requirement for either additional development in respect of Wordmap and SQL Server APIs or the use of additional APIs.

HILT Programming Issues

For some use cases, additional programming will be required in the HILT service between the SOAP or SRW or Z39.50 server and the Wordmap or SQL Server APIs.

HILT Mapping and Database Issues

For some use cases, additional illustrative term sets and mappings of these to the DDC spine will be necessary – for example, in dealing, as proposed by BIOME, GoGeo, and RDN generally, with spelling and singular/plural issues.

Feasibility Assessment

Either of the two projects outlined above look to be feasible using either the Wordmap or the SQL Server options. There are a number of issues regarding whether it is best to use Zthes, SKOS-Core, MARC, or offer an option of all three and also about how best to use them. It may be sensible to make final decisions on this in the early stages of a practical pilot. There are also questions about whether or not it is sensible to look at both SRW and Z39.50, given that SRW/U is intended in time to replace Z39.50. This, however, is very much a decision for JISC. SRW may be the future, but Z39.50 is still heavily used at the moment.

Cost of a Follow-up Pilot

At present, no information is available on the likely cost of a follow-up project based on either of the options described above. This will be investigated in the final weeks of the project, aiming to produce a bid to JISC by March 31st 2005.

Draft Recommendation

It is recommended that two versions of a possible follow up project be costed, based on the options mapped out above. This work will begin in the week ending 18th March 2005.

Appendix D[20]: Assessment: Use Cases, Protocols and Mark-ups

Introduction

This is a first attempt to assess protocol and mark-up issues relating to the 5 use cases we drew up recently. Each use case is handled separately below. Note, however, that in most cases, common problems arise across the use cases. As a result, most of the assessment applied under use case #1 is valid for the later use cases and it is sensible to then only tackle new problems thrown up by the other use cases in the later assessments.

Summary of Interim Conclusions

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 5 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 all look adaptable to the task, although Zthes appears to be less suited to handling classification data than the other two are. MARC has at least one advantage in that some major thesauri are available in that format. SKOS-Core is more flexible and more suited to the Web Services perspective and the Semantic Web community.

The picture that is beginning to emerge in respect of the follow-up pilot is that there are two sensible options. 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 seek to offer both SRW and Z39.50 services (perhaps through an SRW-Z39.50 gateway[21]) and would offer a choice of Zthes, SKOS-Core, and MARC mark-ups.

Another possibility 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 there 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. This 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 longish period of time. Obviously, the devil would be in the detail.

[20] Appendix D is a working document and its style is informal

[21] 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.

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 is **off**, and teeth is the term, the server applies 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 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 matching algorithm and make it available to the community. Disambiguation would be a service end function based on data sent back from HILT.

Overview

There are four elements to consider in this use case under each protocol: (1) Can searches be formulated to the level of complexity required? (2) Can the 'switch' be handled? (3) Can the disambiguation stage be handled if necessary? (4) Can the expected response be adequately and appropriately formatted by Zthes, SKOS-Core and MARC?

SRW (1)

In SRW, CQL will allow sufficient complexity of search formulation to cover anything that is envisaged at this stage (and probably beyond).

See, for example, http://zing.z3950.org/cql/intro.html

SRW (2)

The 'switch' can be dealt with in one of three ways:

a. Using the recordXPath parameter of 'query' in the SRW searchRetrieve operation

To quote one source (http://www.loc.gov/z3950/agency/zing/srw/introduction.html):

'This parameter lets the client request a specific section of a record, rather than the entire thing. For example, the client may only want to display the title of the record, so rather than throw the rest of it away, it asks the server to return only the element that it needs'

On the face of it, if 'teeth' were sent with/without the recordXPath parameter this could signal the on and off conditions of the switch to the SRW server and the 'HILT API' could translate this to either invoke only the search_for_wordsets API in Wordmap (if switch set to 'off') or both that and get_features (if switch set to 'on').

A possible drawback here is that the default would be that the switch would be 'on' and it has been argued that off might be the preferred default.

b. Use the extraRequestData parameter

Using the extraRequestData parameter, which can legitimately contain a service-defined XML fragment (such as "<switch>on</switch>") in the searchRetrieve request may be a better alternative. More information at http://www.loc.gov/z3950/agency/zing/srw/service.html and http://www.loc.gov/z3950/agency/zing/srw/extra-data.html.

c. Multiple recordSchemata, each having different content corresponding to having the switch on and off (http://srw.cheshire3.org/docs/introduction.html).

This is not a preferred approach. Either a or b should be used.

SRW (3)

The disambiguation stage can be handled by the SRW 'scan' operation, followed by a search on the chosen entry. SRW(1) and SRW(2) then apply.

See http://www.loc.gov/z3950/agency/zing/srw/scan.html

SRW (4)

Zthes Response Mark-up

Although Zthes (http://zthes.z3950.org/profile/current.html) is not specifically designed to handle classification schemes, there appears to be general agreement that it might be adapted for the purpose and the Zthes developer has indicated a willingness to adapting it for use with classification schemes. Within the project, however, it is felt that this approach is probably not ideal. One approach might be to formulate the response in XML and seek to feed requirements into a Zthes classification scheme enhancement programme. Another is to use either SKOS-Core or MARC, both of which OCLC have found to be more suitable for encoding DDC responses.

A problem identified in DDC (and probably other schemes) is that the use of the thesaurus ideas of BT/NT is not an accurate way of describing the relationships in the DDC hierarchies. This is complicated by the fact that there can be different kinds of relationship and that the change of type of relationship is not usual made explicit. It may be possible to deal with this in Zthes by defining a new kind of relationship – a kind of classification scheme hierarchical relation 'catch-all'.

Zthes allows you to formulate new relationships alongside PT, RT, LE etc (Search for 'relationType' in http://zthes.z3950.org/profile/current.html. These could be used to deal with mappings.

SKOS-Core Response Mark-up

OCLC have stated that they have found both SKOS-Core and MARC more suitable for encoding DDC responses than Zthes.

SKOS-Core documentation states that SKOS-Core has been designed with classifications schemes (as well as thesauri) in mind - see http://www.w3c.rl.ac.uk/SWAD/deliverables/8.1.html#1. In addition, work has been done (http://www.w3.org/2001/sw/Europe/reports/thes/8.5/) on its use for PACS (Physics and Astronomy Classification Scheme). On the face of it, therefore, it should be adequate for HILT M2M pilot purposes. Unlike in Zthes, where it would appear the DDC class number would have to be used as the 'termID' (making it difficult to deal with a situation where a concept is relocated within a scheme during a revision), there is at least one way of keeping the two distinct – by using a 'concept URI' (http://www.w3c.rl.ac.uk/SWAD/deliverables/8.1.html#2.1) as a number for the concept and the 'externalID' (http://www.w3c.rl.ac.uk/SWAD/deliverables/8.1.html#2.5) to refer to the class number in DDC or elsewhere. A decision will have to be made as to how best to deal with the DDC caption – as a 'definition' or a 'scopenote'.

The BT/NT difficulty in DDC described above under Zthes would probably exist for SKOS-Core as well as Zthes and the solution would be the same.

As with Zthes, SKOS allows (http://www.w3c.rl.ac.uk/SWAD/deliverables/8.1.html#4.3) you to formulate new relationships. These could be used to deal with mappings. Also, SKOS work on dealing with mappings has been done – see http://www.w3c.rl.ac.uk/SWAD/deliverables/8.4.html which describes an SKOS mapping schema and 'gives recommendations and examples for defining mappings between concepts from different thesauri, and using these mappings to enable inter-thesaurus interoperability'. A point of note is that the mapping schema allows for the use of Boolean combinations to assist in mappings between concepts in different schemes.

MARC Response Mark-up

As already noted, OCLC have stated that they have found both SKOS-Core and MARC more suitable for encoding DDC responses than Zthes.

Preliminary research seems to suggest that we would have to use the MARC 21 Concise Format for Authority Data (http://www.loc.gov/marc/authority/ecadhome.html) rather than the MARC 21

Concise Format for Classification Data (http://www.loc.gov/marc/classification/eccdhome.html). The latter does not seem to allow for mapping, the former can, at least, deal with some aspects of the DDC information our records hold and allows for mappings to other schemes in the 7XX fields. JoDI article by Diane Vizine-Goetz et al[22] indicates ability to encode name or code of mapped vocabulary, mapped term, control number or unique identifier for mapped term, identity of the mapping organization.

Other points to note

Clearly, there will be a need to employ the SRW 'explain' operation in the proposed follow-up pilot http://www.loc.gov/z3950/agency/zing/srw/explain.html).

OAI-PMH (1)

Initial thoughts on the use of OAI-PMH to support searches of the HILT database are that this looks difficult, at best (http://www.openarchives.org/OAI/openarchivesprotocol.html). Current thoughts on this front are these:

1. HILT depends on a term, input by a user, being found somewhere in the HILT database – in fact, anywhere in the database. The records that term is matched to are sent back to the user.
2. OAI-PMH is not a search protocol, it is a protocol used to harvest records. Selective harvesting is possible, but only using dates (which is of little use to HILT) and SET membership (see information on sets at: http://www.openarchives.org/OAI/openarchivesprotocol.html#Set).
3. As far as one can tell, SET membership is not a dynamic thing. One has to be able to tell the harvester what sets are held from the start in response to the listsets request and it will then selectively harvest records according to a command much like listrecords SetSpec = 'physics'.
4. This is not adequate for HILT. It relies on being able to 'select' records based on any term a user may come up with, so OAI-PMH will not work for HILT searches.

OAI-PMH (2)

The switch proposed for use case #1 can be dealt with in OAI-PMH but only by using option (c) under SRW (2) above – that is, the least preferred option utilising different metadata formats for the different responses required by the switch.

OAI-PMH (3)

This is seen as being driven by scan in SRW (SRW (3)). The verb ListIdentifiers (http://www.openarchives.org/OAI/openarchivesprotocol.html#ListIdentifiers) could be used in OAI-PMH but would have to employ selective harvesting by sets, at which point it would fall foul of the problems indicated under OAI-PMH (1) above.

[22] Vizine-Goetz, D. *et al.* (2004), Vocabulary Mapping for Terminology Services, ***Journal of Digital Information***, Vol.4 No.4. Available: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/

OAI-PMH (4)

Since OAI-PMH should be able to handle sending back any of the types of 'mark-up' considered in this study using different metadata formats, the points made under SRW (4) apply.

Z39.50 (1)

Although Z39.50 doesn't use CQL, it should, like SRW, allow sufficient complexity of search formulation to cover anything that is envisaged at this stage (and probably beyond). Documentation can be found at http://www.loc.gov/z3950/agency/document.html.

Z39.50 (2)

There are at least two mechanisms that might be used to implement the switch required in use case #1:

1. The origin may ask the target for brief records rather than full records (see, for example, page 167 of http://www.loc.gov/z3950/agency/Z39-50-2003.pdf);
2. The origin may ask the target for one record format rather than another.

Again, option 2 is the same as option (c) under SRW (2) and is the least preferred option.

The examination of the Z39.50 protocol has not been exhaustive. It is, therefore, possible that other ways of implementing the switch exist.

Z39.50 (3)

It would appear that the disambiguation stage can be handled by the Z39.50 'scan' operation (see under 3.2.8.1, page 55 at http://www.loc.gov/z3950/agency/Z39-50-2003.pdf), followed by a search on the chosen entry. Z39.50 (1) and Z39.50 (2) then apply.

Z39.50 (4)

Since Z39.50 should be able to handle sending back any of the types of 'mark-up' considered in this study, the points made under SRW (4) apply.

Other Points To Note

For differences between SRW and Z39.50, see http://www.loc.gov/z3950/agency/zing/srw/z3950.html.
Z39.50 can be used for the service but will be more difficult to implement in various areas (e.g. dealing with the fact that connections are 'stateful', or that the scan facility is more complex in Z39.50 than in SRW).

Use Case #2

BIOME/GoGeo/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.

Overview

There are two versions of this use case to consider: the ‘simple version’ and the ‘version with switches’ covered below.

Use case #2: Simple Version

At its simplest, this use case raises only one set of new questions. It is possible to regard the requirements for singular/plural terms, synonyms, spelling and typographical errors as, at most, new mappings (synonyms at least would usually be there already). If, therefore, we assume that all that is required is the ability to send a term and get back all of the mappings as a result, the only new questions that arise relate to whether or not these additional ‘mappings’ can be marked up adequately and appropriately under Zthes, SKOS-Core, and MARC:

Zthes

Zthes allows formulation of new relationships alongside PT, RT, LE etc. These could be used to deal with the new ‘mappings’ involved.
~~~

SKOS-Core

As with Zthes, SKOS allows formulation new relationships. Again, these could be used to deal with the new ‘mappings’ involved.

MARC

Preliminary research seems to suggest that use of the MARC 21 Concise Format for Authority Data would be required (http://www.loc.gov/marc/authority/ecadhome.html) rather than the MARC 21 Concise Format for Classification Data (http://www.loc.gov/marc/classification/eccdhome.html). The latter does not seem to allow for mapping. The former can, at least, deal with some aspects of the DDC information our records hold and allows for mappings to other schemes in the 7XX fields. As noted, Vizine-Goetz et al[23] indicate the ability to encode name or code of mapped vocabulary, mapped term, control number or unique identifier for mapped term, identity of the mapping organization. Essentially, the new ‘mappings’ involved could be handled in this format, using these encodings.

Use case #2: Version with Switches

Looking beyond this simplest case, it is possible that we may need switches to turn such things as synonym mapping (unlikely?) and, spelling and typographical error checks off (if only for use case #1). This, however, presents the same problems as the switch described under use case #1 – which is to say, that there is no new problem here to comment on. These additional switches can be handled/ not handled in the various protocol and mark-up combinations to the same extent as the switch in use case #1 can be handled/not handled in these combinations.

Other points to note

This use case does, of course, raise issues in other areas (e.g. extra ‘mappings’ in the database).

[23] Vizine-Goetz, D. *et al.* (2004), Vocabulary Mapping for Terminology Services, ***Journal of Digital Information***, Vol.4 No.4. Available: http://jodi.ecs.soton.ac.uk/Articles/v04/i04/Vizine-Goetz/

Use Case #3

BIOME/GoGeo/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.

Overview

This use case raises no new issues not covered by ‘use case #2 version with switches’.
~~~

Use Case #4

Browse-based use cases

~~~

Four situations to consider have been identified under this heading:

~~~

(a) Browse DDC 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

Overview

Issues that arise under the four situations identified in this use case are:

Situation (a)

There is nothing new here as far as protocol and mark-up issues are concerned, simply an additional piece of programming to be added to the current **HILT** interface that either presents a browse DDC option automatically in a 'no hits' situation or, more likely, offers the option to browse or enter a new search.

Situation (b)

There is nothing new here as far as protocol and mark-up issues are concerned, simply an additional piece of programming to be added to the current **HILT** interface that allows the user to specify a browse of a scheme other than DDC in various situations (e.g. when a DDC search or browse highlights a mapped term from that scheme).

Situation (c)

An additional piece of programming must be added to the current **requesting service** interface that either presents a browse DDC option automatically in a 'no hits' situation, or, more likely, offers the option to browse or enter a new search. Offering the new search does not entail any new problem in respect of protocol and mark-up issues. Handling the browse interaction entails no new issue not covered by situation (d) below.

Situation (d)

An additional piece of programming has to be added to the current **requesting service** interface to allow the user to specify a browse of a scheme other than DDC in various situations (e.g. when a DDC search or browse highlights a mapped term from that scheme). There is nothing new here as far as mark-up issues are concerned. On the
~~~

protocol side, the difficulties in using selective harvesting using SetSpec to simulate searching apply for OAI-PMH; in SRW and Z39.50, the need to distinguish between browse DDC and browse some other scheme can be handled by specifying different indices in the associated scan requests.

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.

Overview

There is nothing new here as far as protocol and mark-up issues are concerned. The browse-based cases are covered under use case #4 and the disambiguation-based case is covered by use cases 1 and 2. The remaining situations described require new programming in either the **HILT** or the **requesting service** interfaces, but that is all. A record sent back for a term will already include such things as narrower and related terms, clearly marked-up as such, so a request for information on narrower or related terms on a term already presented should be something the requesting service has the data to handle. Further requirements can be handled by new requests for responses on narrower or related terms but this should not raise new protocol or mark-up issues.

The cross-scheme situation is slightly more complex, but not much. If the data sent back by HILT in response to a search request is for a DDC caption, but includes a mapped term from another scheme and the code for the scheme in question, a follow up search for that term in that scheme would send back information on narrower and related terms and so on for that term. Again, the requesting service would have the data it required to provide the user with information on narrower or related terms.

DMN 01.03.05

Appendix E[24]: HILT M2M Pilot Project Costing

The draft proposal below is under discussion with partners. A full proposal will be submitted to JISC in early April.

Note: Planned amendment to approach in bid

At time of submission of this report, discussion on the initial approach to costing described below is ongoing. However, one alteration to the likely approach has been agreed. Since the point of offering a SOAP service is specifically platform independence, it has been agreed that taking a common approach to developing a client would probably not be the most sensible strategy – that, on the contrary, it would be a better strategy to agree on differing approaches. This would ensure that we created 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 server. The implication of this is that, whilst EDINA could develop their own client and provide BIOME and HILT with assistance and advice in developing clients, the idea that they could develop an 'embryonic' or 'generic' client that would be adapted for others as proposed above is no longer the recommended approach. Three clients will be developed with help and assistance from EDINA.

1. Introduction

The proposal is that HILT Phase III (M2M Pilot) will 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 have been 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 pilots. This is 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 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 I.E. as it develops in this wider context.

2. Wordmap or SQL Server

At this stage, versions of the pilot based on either Wordmap or SQL Server are costed, but a single choice will have to be made in for the final bid. As indicated in HILT II, the continued use of Wordmap could be advantageous in the long-term if, as HILT II

[24] Appendix E is a working document and its style is informal

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 appear to be (1) That this interface is not needed for the M2M pilot (2) That whereas we had to adapt the Wordmap database structure to provide the HILT II pilot, the SQL Server version, having been specifically designed for HILT, seems to be easier to work with as far as HILT is concerned. Also, there are likely to be costs associated with using Wordmap for the M2M pilot and this is likely to be an issue for JISC given that the staff updating interface is not an M2M pilot requirement. It does, however, remain true that this is a likely longer term need. Comments on this welcome from all (including Wordmap – obviously)

3. Description of Work Proposed: Single server version

The 'single service 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 CQL 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 (Ben has (I think) suggested we could use Zthes as a profile and SKOS-Core as the mark-up here – comments welcome). 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 below in Annexe B. This is likely to entail new (illustrative) mappings of LCSH, UNESCO, and MeSH terms to the DDC spine, together with mappings from RDN-specific terminology sets and mappings to cover areas highlighted as important in the use cases (synonyms, spelling mistakes and typos). These need to be sufficient to allow for realistic tests and evaluations under the various use cases.
- Additional programming to interface the HILT service with the SRW server (allowing inter-working between the SRW server and SQL Server or Wordmap APIs).
- An embryonic web client to interact with the terminology server using SRW, CQL and SKOS-Core. The aim would be to design this so that it could be used in three contexts: as a BIOME client for interfacing with HILT, as a GoGeo client for interfacing with HILT, and as replacement front-end for the HILT service itself (this would be need short term for testing purposes and ongoing research work, but it might also be a long-term requirement for a JISC terminology service).
- A local collections database as used in HILT 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.

- A database that extended the current Wordmap and 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 Annexe A below (headed 'Variant Cases').

None of the use cases mapped by project partners entailed a requirement for either additional development in respect of Wordmap and SQL Server APIs or the use of additional APIs.

4. Description of Work Proposed: Distributed server version

The 'distributed service 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 (UNESCO, say), out of the main server and set up a second terminology service. Mapping between UNESCO in server 2 and the DDC spine in server 1 would be achieved via the use of SKOS- Core concept URIs. For the purposes of the pilot, the assumption would be that the web client for BIOME and GoGeo and HILT would already 'know' about the two servers and would 'talk' to one or the other depending on whether or not UNESCO was a factor (obviously, IESR would have to come into this longer-term, but it would not feature in the pilot at this stage). If (say) the BIOME client needed LCSH and UNESCO 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 UNESCO 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. Roles and funding sought

Please give an initial but relatively accurate 'guesstimate' of roles and funding sought, remembering:

a. We'll have to deliver within the amount awarded.
b. We'll have to justify estimated costs against the table in Section 6 below (or something like it).

Participant	Role(s) (please adjust role if appropriate)	Funding Sought	Extra if distributed
CDLR	Project management; Final and other reports; Dissemination; Web-site etc Programming HILT – SRW, adjust 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, support and related (Explain) Main work on generic web client for BIOME, GoGeo, HILT		
BIOME	Advice on BIOME needs, BIOME client programming; advice on subject areas, terminologies		
UKOLN/RDN	Advice on RDN needs, terminologies IE generally		
L. Will	Advice and views on terminology issues, classification issues, mapping issues, mark-up issues, the terminology services scene		
Wordmap	If Wordmap used, licensing, support, advice		

6. Cost element grid

Project facet	Roles	Cost: Single Server	Additional Cost Distributed server
Project Management, including web-site and Project Plan			
Equipment			
Set up SRW server, set up illustrative transaction between a requesting client, the SRW server, and a HILT response			
Identify appropriate subject areas to cover in illustrative mappings			
Identify terminology set mapping requirements			
Design and set up extended HILT pilot database			
Add illustrative mappings to database			
Analyse mark-up requirements and associated needs as regards SKOS-Core mark-ups			
Design and set up interface between HILT database and			

SRW server – ‘code’ that will accept requests, ‘translate’ them into requests to Wordmap or SQL Server APIs, receive responses, wrap them in SKOS-Core, send them to the SRW server			
Adapt local collections database for new pilot requirements			
Detail client requirements for BIOME, GoGeo, HILT			
Program and test generic client			
Adapt client for BIOME, GoGeo and HILT			
Set up SRW ‘explain’ facility			
Launch and test pilot			
Evaluate pilot under all 5 use cases			
Consider issues listed under ‘Variant Cases’			
Re-work various aspects of pilot based on outcomes of tests and evaluations			
Draw conclusions, propose further R&D work, write Final Report			

Annexe 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/GoGeo/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/GoGeo/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 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.
~~~

Annexe B: Executive Summary; Recommendation: Draft Final Report

Assessment: Use Cases, Protocols and Mark-ups

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 5 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 all look adaptable to the task, although Zthes appears to be less suited to handling classification data than the other two are. MARC has at least one advantage in that some major thesauri are available in that format. SKOS-Core is more flexible and more suited to the Web Services perspective and the Semantic Web community.

The picture that is beginning to emerge in respect of the follow-up pilot is that there are two sensible options. 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 seek to offer both SRW and Z39.50 services (perhaps through an SRW-Z39.50 gateway[25]) and would offer a choice of Zthes, SKOS-Core, and MARC mark-ups.

Another possibility 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 there 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. This 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 longish period of time. Obviously, the devil would be in the detail.

Assessment: Other Issues Arising from Use Cases

Wordmap and SQL Server APIs

None of the use cases mapped by project partners entailed a requirement for either additional development in respect of Wordmap and SQL Server APIs or the use of additional APIs.

HILT Programming Issues

For some use cases, additional programming will be required in the HILT service between the SOAP or SRW or Z39.50 server and the Wordmap or SQL Server APIs.

[25] 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.

HILT Mapping and Database Issues

For some use cases, additional illustrative term sets and mappings of these to the DDC spine will be necessary – for example, in dealing, as proposed by BIOME, GoGeo, and RDN generally, with spelling and singular/plural issues.

Feasibility Assessment

Either of the two projects outlined above look to be feasible using either the Wordmap or the SQL Server options. There are a number of issues regarding whether it is best to use Zthes, SKOS-Core, MARC, or offer an option of all three and also about how best to use them. It may be sensible to make final decisions on this in the early stages of a practical pilot. There are also questions about whether or not it is sensible to look at both SRW and Z39.50, given that SRW/U is intended in time to replace Z39.50. This, however, is very much a decision for JISC. SRW may be the future, but Z39.50 is still heavily used at the moment.

Cost of a Follow-up Pilot

At present, no information is available on the likely cost of a follow-up project based on either of the options described above. This will be investigated in the final weeks of the project, aiming to produce a bid to JISC by March 31st 2005.

Draft Recommendation

It is recommended that two versions of a possible follow up project be costed, based on the options mapped out above. This work will begin in the week ending 18th March 2005.

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

NKOS: Networked Knowledge Organisation Systems

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

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.