



**HILT: High-Level Thesaurus Project Phase II**

**A Terminologies Server for the JISC Information Environment**

**Final report to JISC**

**Main report**

**Dennis Nicholson ▪ Ali Shiri ▪ Emma McCulloch**

**Additional Work: Rachel Heery (M2M appendix) ▪ Leonard Will (Evaluation)  
▪ Alan Dawson (Technical details on pilot) ▪ Simon Jennings  
Advisory input: Anu Joseph ▪ Gordon Dunsire ▪ Alan Gilchrist**

**Glasgow : Cente for Digital Library Research, 2004**

**Main Participants:**

- The Centre for Digital Library Research (CDLR) at Strathclyde University
- JISC representative
- mda (formerly the Museums Documentation Association);
- National Council on Archives (NCA);
- National Grid for Learning (NGfL) Scotland;
- Online Computer Library Center (OCLC);
- RDN representative
- FE Representative (Regional Centre)
- Scottish Library and Information Council (SLIC);
- Scottish University for Industry (SufI);
- UK Office for Library and Information Networking (UKOLN).
- Terminology experts, Alan Gilchrist and Leonard Will (external evaluator)

There was also involvement from, NLS, BL, and Wordmap.

**HILT Steering Group Members**

Louise Craven	Public Records Office
Gordon Dunsire	SLA/SLIC/CIGS
Graeme Forbes	National Library of Scotland
Rachel Heery	UKOLN
Helen Hockx	JISC/DNER
Kathryn Hughes	National Library of Wales
Simon Jennings	DNER/RDN
Ray Lester	Natural History Museum
Vanessa Marshall (corresponding member only)	National Preservation Office
Anne Matheson	Chairperson
Paul Miller	UK Interoperability Focus
Chris Rusbridge	Information Services, University of Glasgow
Diane Vizine-Goetz	OCLC

**HILT Management Group Members**

Fionnuala Cassidy	FE Representative
Elaine Fulton	SLIC
Alan Gilchrist	Advisor - The Cura Consortium
Stuart Holm	Museums Consultant acting for mda
Nick Kingsley	National Council on Archives
Joan Mitchell	OCLC
Leonard Will	External Evaluator - Willpower Information

## HILT Phase II Final Report

### Contents

Section	Title	Page
<b>0</b>	<b>Executive summary and Recommendations</b>	<b>4</b>
<b>1</b>	<b>Selected Key Points of Note, Including Illustrative Use Cases</b>	<b>10</b>
<b>2</b>	<b>Aims, Processes, Methodologies, Literature, User and Staff Surveys</b>	<b>14</b>
<b>3</b>	<b>Developing an Interim Specification</b>	<b>20</b>
<b>4</b>	<b>Building and Assessing the Pilot</b>	<b>24</b>
<b>5</b>	<b>Developing an Operational Server – Additional Requirements</b>	<b>32</b>
<b>6</b>	<b>Cost-benefit Analysis</b>	<b>36</b>
<b>7</b>	<b>Conclusions and Recommendations</b>	<b>41</b>
<b>Appendices</b>	<i>Published separately</i>	
<b>A</b>	<b>Methodologies</b>	
<b>B.1</b>	<b>Literature Survey</b>	
<b>B.2</b>	<b>Mapping Issues Literature Survey</b>	
<b>B.3</b>	<b>Mapping Exercises</b>	
<b>C.1</b>	<b>Service Survey Questionnaire</b>	
<b>C.2</b>	<b>Service Survey Results, Including Subject Schemes Used in JISC</b>	
<b>C.3</b>	<b>User Survey Overview</b>	
<b>C.4</b>	<b>User Survey Results</b>	
<b>D.1</b>	<b>User Workshop Overview</b>	
<b>D.2</b>	<b>User Workshop Questionnaire</b>	
<b>D.3</b>	<b>User Workshop Results</b>	
<b>E</b>	<b>RDN Subject Issues</b>	
<b>F</b>	<b>Notes on Possible Clustering-Based Enhancements to User Tool Set</b>	
<b>G</b>	<b>HILT and the IESR Shared Service</b>	
<b>H</b>	<b>Cost-Benefit Analysis Report</b>	
<b>I.1</b>	<b>Initial and Interim Service Specifications</b>	
<b>I.2</b>	<b>Operational Terminologies Server Specification; Pilot Description</b>	
<b>J</b>	<b>Delivering HILT as a JISC IE shared service (M2M report)</b>	
<b>K</b>	<b>Evaluator's Report</b>	
	<b>Glossary</b>	

## 0. HILT Phase II - Executive Summary

### 'Designing in' Consensus and Cooperation

This Final report is addressed to JISC, the funders of HILT Phase II, but may also be of interest to other organizations facing the problem of achieving and maintaining interoperability in the subject description and classification of distributed information resources. The project was funded to set up a pilot terminologies service for the JISC Information Environment, aiming to:

- a. Provide a practical experimental focus within which to investigate and establish subject terminology service requirements for the JISC I.E
- b. Make recommendations as regards a possible future service

This has been its main focus. From the first, however, it has been recognized that the successful resolution of the interoperability issue requires a constructive working relationship between JISC and other interested parties. This recognition is reflected in the project recommendations which propose that JISC begin a dialogue with key national and international players (see below for a possible list). It is also reflected in the proposed design itself, which assumes, amongst other things:

- Mapping between schemes, rather than preference for a single scheme
- The need for a facility to allow others to include their own (self-provisioned) mappings
- The existence of other terminology servers that will interact with the proposed JISC server to produce a range of terminology services

### Mapping Between Terminologies

It was assumed from the outset that the terminologies server would be the basis of a community process that would develop, maintain, and gradually improve interoperability of subject descriptions by mapping between terminology sets and that the aim of the project was to determine specific design requirements based on this approach. Not only was the focus on mapping in line with the community consensus in HILT I<sup>1</sup> which strongly favoured it over the adoption of a single scheme and other options, it also recognized that mapping schemes together was probably the approach most likely to be compatible internationally. Even if JISC were to adopt a single subject or class scheme across all services, it is unlikely that the same scheme would generally adopted elsewhere. Even if it were, mapping would still be necessary to deal with language variations across the world. It would also be necessary to deal with a key requirement of any terminologies server – the need to map subject terms used by information seekers to those used by staff working on the subject description of resources.

### Developing the Requirement

The specific design requirements of an operational server and the proposed approach to implementing it were drawn out over the lifetime of the project as the project team:

- Conducted literature reviews, a survey of services, user interviews, and terminology mapping exercises;
- Investigated, considered and discussed issues with colleagues, and with the project groups and other stakeholders, including the two terminology experts;
- Constructed an illustrative working pilot and assessed it via a user workshop and other means (the pilot is available at <http://hiltipilot.cdlr.strath.ac.uk/pilot/top.php>);
- Conducted a cost-benefit analysis of functionality levels and instantiation methods.

Full details of the approach taken are provided in the body of the report and summarised in the diagram on page 9 below.

### The Primary Purpose of the Server

At an early stage, a view was taken on the primary purpose of the server. It was agreed with the project management and steering groups that the function of the proposed terminologies server should be *to*

<sup>1</sup> See <http://hilt.cdlr.strath.ac.uk/Reports/Documents/HILTfinalreport.doc>

*optimize the ability of users to carry out successful subject searches<sup>2</sup> by providing a process that would, in time, permit JISC and JISC services to:*

1. Achieve and maintain as a high a level of interoperability as possible between:
  - The different standard subject schemes and versions of standard schemes in use in different services, both within and outwith JISC;
  - Amendments, additions, and extensions made to standard schemes across the services;
  - Terms used by users when composing search strategies.
2. Optimize both the consistency with which staff across the various services apply schemes in the subject description of materials, and the ability of users to formulate successful search queries, through the provision of information on descriptive term usage, appropriate training, and helpful feedback mechanisms (e.g. a ‘disambiguation’ facility to help clarify the subject of a user search).

### **Additional Design Considerations**

This perspective, together with the project research work outlined above, informed the outline specification for the development of an operational server included within the conclusions and recommendations set out below. Key design considerations arising out of the project research work included the following:

**DDC Spine:** The proposal is to map terminologies to a DDC spine. This has a number of advantages, including the fact that DDC is already extensively mapped to LCSH, has been used in other mapping projects such as Renardus<sup>3</sup>, and is translated into over 30 languages. It is also the only evident way of providing the proposed collections finding facility described below.

**Scheme Coverage, Other Mappings, Other Services, Other Funders:** The core proposal assumes that it is – initially at least – sensible to focus on DDC, LCSH, and UNESCO as the core of the server but provide, both functionality to permit interaction with other terminology servers, and facilities to permit other (self-provisioned) groups to create mappings to other schemes. The possibility of adding MeSH to the core set if a potential funding partner thought it desirable is also suggested. Other schemes, such as AAT, could also be considered on the same basis.

**UK Oriented Scheme Modifications Registries:** A UK oriented scheme modification registry would allow the extensions and amendments that service staff make to standard schemes to be harmonised<sup>4</sup> across the UK and presented to users undertaking subject searches. This would improve ongoing interoperability in this area, assist users in identifying terms not in standard schemes, and (potentially) help alleviate interoperability problems in legacy metadata. Additional regional extensions would entail greater costs but might be attractive to potential funding partners such as RE: SOURCE<sup>5</sup> and SLIC<sup>6</sup>.

**Collections Finding Facility:** Since the project was asked to look at ‘collection level requirements’, and since the JISC IE comprises distributed services with overlapping subject coverage but (in many cases) different subject schemes and practices in place, an additional requirement is to map user subject queries to JISC collections and advise users on which JISC collections might answer their queries, and what terms in the subject schemes used by the collections are required for searching.

**User Interface Facilities and Further Research:** Further research is required into the interface needs of users and the possible role of technology based mechanisms<sup>7</sup> for improving interoperability between terms used by users and existing subject metadata. The project has proposed that this be conducted in parallel with the development of the baseline server and the associated terminology mappings, UK

---

<sup>2</sup> Note that the aim is neither to improve precision at the expense of recall, nor to improve recall at the expense of precision, but rather to provide users with the information they require to do either of these things depending on their needs at a given time

<sup>3</sup> See <http://www.renardus.org/>

<sup>4</sup> Note: A UK oriented scheme modification registry would record agreed departures from standard schemes in use in the UK. Some, but not all, of these terms, would be UK-specific terms.

<sup>5</sup> RE:SOURCE. The Council for Museums, Archives and Libraries. See <http://www.resource.gov.uk/>

<sup>6</sup> The Scottish Library and Information Council. See <http://www.slainte.org.uk/slic/index.htm>

<sup>7</sup> An example is the clustering approach pioneered by the CHESHIRE project – see Appendix F

oriented scheme modifications registry, and staff updating and quality control facilities required to halt and reverse the decline in interoperability caused by existing subject description practices (see data in Appendix C.2). Further information is provided in the body of the report.

**Machine to Machine (M2M) Facilities and Interactivity Issues:** The terminology server is a shared service within the JISC Information Environment. Shared services are assumed to interact with other shared services and portals rather than directly with users. HILT research suggests, however, that a terminologies server may be atypical in this regard, requiring a degree of interaction with users (including staff users) that may make the provision of some centrally located user interaction the best approach on both economic and user support grounds. M2M facilities will still be required, of course. (see UKOLN report in Appendix J).

**Limited granularity mapping:** The option of mapping between subject schemes, user terms, and DDC at less specific levels of granularity only has been ruled out. The HILT view is that limiting mapping in this way would make it impossible to deal with a significant proportion of user subject queries. These tend, if anything, to be more, rather than less, specific than the levels of granularity available in standard schemes (It should be noted that there is no necessary connection between more general levels of granularity in subject description and 'collection level requirements'. The user need will most often be to map a subject search at a very specific level of granularity up to a collection classified at a higher level and then down again, within the local scheme used, to a level of granularity appropriate to the original query. Limited granularity mapping would not permit this.).

**Information Environment Services Registry (IESR):** The need to identify collections appropriate to particular subject queries and determine which subject and class schemes are in use in these services requires interaction between the proposed JISC terminology server and another shared service, the IESR. It also requires that IESR store data needed by HILT for these purposes. These requirements are specified in Appendix G of the report. They have been passed on to the IESR pilot site.

## Recommendations

The project recommends:

1. That JISC fund a development project to build a terminologies service for the JISC Information Environment and base it, at minimum, on the functionality and research work encompassed within option C from the cost-benefit analysis (see Section 6 and Appendix H), as follows:

1	DDC spine and term sets
2	LCSH mapping
3	UNESCO mapping
4	UK oriented modifications registry terms set creation
5	UK oriented modifications registry terms mapping
6	RDN terminologies harmonisation study
7	RDN-based clustering tool study
8	Interface needs user study (enhanced pilot with clustering)
9	Term match facility
10	Staff amend maps facility
11	Staff training module
12	Online user training module
13	Ability to host and map other schemes
14	Ability to interact with other mapping services
15	Processes to cope with scheme updates
16	Disambiguation facility
17	DDC collection identifier
18	Any hits test/rank facility
19	User terms monitor

The software functions listed in the above are taken to include M2M capability. In respect of the latter, it is proposed that the additional recommendations specified in the UKOLN report on M2M functionality be followed. These are specified in Appendix J of the HILT Phase II Final Report.

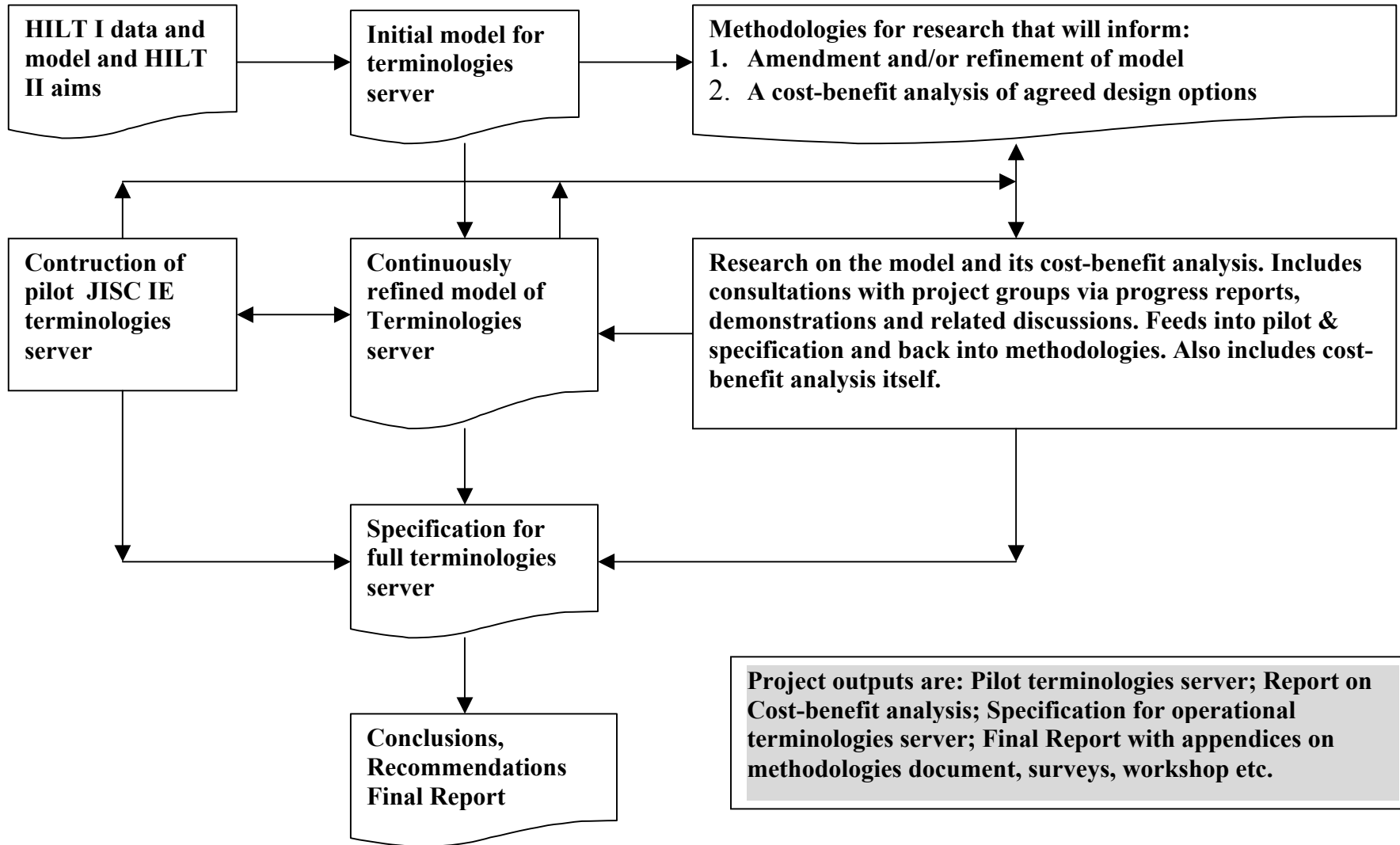
The cost-benefit analysis figures suggest the cost will be £926,096 over a five-year period, including project management, training, publicity, marketing, and redevelopment costs. However, costs may be revised in the light of detailed discussions with JISC should these recommendations be accepted.

2. That it also consider whether there is value in adding UK regional scheme modification term sets and MeSH into the features list. The cost-benefit analysis figures suggest the additional cost of both will be £1,153,133 over a five-year period.
3. That it take a phased approach to the implementation, spreading the cost of development, and of the additional research still required to inform aspects of service design, over 5 years in the first instance.
4. That it build in a regular review process that will permit, where necessary, the refocusing of aspects of the design to take account of changing circumstances, new research data, novel techniques and technologies, and other pertinent factors as they arise.
5. That the initial phase last two years and entail terminologies server development and other research specified in elements 1-15 in the table above, conducting 6-8 in conjunction with users and using the results to inform development beyond the initial two years (this implies further development of 16-19 as pilot elements in the first two years, followed by full development later).
6. That JISC build on the experience and relationships built up in HILT Phase II in any follow up project and involve the HILT team, the supplier of the Wordmap software, OCLC, and the various HILT stakeholders, but that they liaise with the team to determine how best to strengthen the approach taken by bringing in expertise from data mining and semantic web communities and professional expertise from other areas thought relevant (Input from internet search engine services like Google might be one example)<sup>8</sup>.
7. That JISC ensure that any follow up project takes account of the potential value of a mapping service of this kind to semantic web and semantic grid developments when considering the instantiation of design elements
8. That JISC work to begin a dialogue with key national and international players on how best to ensure cross-sectoral, cross-domain, multi-lingual, and international compatibility of the JISC terminologies server with other such developments – these to include OCLC and Library of Congress, other terminology scheme developers, RLN/RSLG, National Archives Network Consortium, mda, UK National Libraries, European and other National Libraries, UK players from other sectors (RE:SOURCE, SLIC, players from Museums and Archives), W3C, a representative from the RENARDUS project. It should also aim to include all communities working in or with JISC – HE and FE, e-learning and research, the semantic grid community, and so on.
9. That JISC consider funding an independent supporting study to explore, in conjunction with JISC itself, the best option for ensuring the long-term financial future of a terminology server and of other such shared services

---

<sup>8</sup> The main participants in HILT Phase II are listed on page 2 of this Final Report

## Overview of Project Processes, Including Dependencies Landscape



# **HILT Phase II**

## **Final Report**

## 1. Selected Key Points of Note, Including Illustrative Use Cases

### a. Involving Other Key Players

This Final report is addressed to JISC, the funders of HILT Phase II, but may also be of interest to other organisations facing the problem of achieving and maintaining interoperability in the subject description and classification of distributed information resources. The project was funded to advise on the design requirements of a terminologies server operating as a shared service in the JISC Information Environment and this has been its main focus. From the first, however, it has been recognised that the successful resolution of the interoperability issue requires a constructive working relationship between JISC and other interested parties<sup>9</sup>. This recognition is reflected in the project recommendations which propose that JISC begin a dialogue with key national and international players on how best to ensure cross-sectoral, cross-domain, multi-lingual and international compatibility of the approach to interoperability that underpins the proposed design of the server. It is also reflected in the proposed design itself, which assumes (for example)

- Mapping between schemes, rather than preference for a single scheme
- The need for a facility to allow others to include their own mappings
- The existence of other terminology servers that will interact with the proposed JISC server to produce a range of terminology services

### b. Limitations on the Work Carried Out

Project met its aims and produced the required deliverables, albeit requiring three additional months to do so (approximately half way through the 12 months of the project, the team asked JISC to allow three additional (unfunded) months for completion. JISC helpfully agreed to this, making the new project end-date September 2003. Even within this, the time and resources available to the project for carrying out the work fell far short of what would have been ideal – this despite the fact that the lead site donated an estimated 40K in additional staff time to the project, together with a server to run the pilot service (staff costs to JISC from lead site were 28K). HILT had originally pressed for a longer project and additional resources, but this had not been attractive to the funders and a compromise was reached. With hindsight, both JISC and HILT would have benefited if a two-year project and increased funds been agreed – a point worth noting for future reference.

The upshot of this circumstance was that HILT Phase II had limited resources and time available to research the complex array of issues associated with the provision of terminology services and to develop and ‘test drive’ a pilot service. Of necessity, therefore, the project focused on determining the functions that a terminologies server would be required to fulfill, on building and testing a limited functionality pilot based (as agreed) on the further development of the mapping-based approach specified at the end of HILT Phase I, and on gathering information likely to be of value when developing an operational server. This meant that areas of research work that might ideally have been carried out in HILT Phase II, but had not been part of the original bid (for example, a full practical examination of the value or otherwise of the Cheshire clustering approach<sup>10</sup> as a terminologies server tool) have had to be held over to any later phase of HILT.

These facts did not and do not undermine the validity of the conclusions reached by the project, but they did limit the extent of the work it was possible to do and should be borne in mind for future reference.

### c. Illustrative Use Cases: An Atypical Shared Service?

In the JISC Information Environment model<sup>11</sup>, it is assumed that the primary function of a shared service is to interact with other elements of the environment (portals, other shared services), rather than directly with users of the environment (in this case, both ‘end users’ and metadata professionals). In this regard, it

<sup>9</sup> A possible, but not necessarily complete, list is provided as recommendation 8 on page 8 above and in Section 7 below

<sup>10</sup> See Appendix F

<sup>11</sup> See JISC Information Environment Architecture. Andy Powell & Liz Lyon, UKOLN, University of Bath

<<http://www.ukoln.ac.uk/distributed-systems/dner/arch/>>

is worth noting that a terminologies server may be atypical. Machine to machine communication (M2M) will unquestionably be a key element of the server's function. A report on the M2M requirements of server design is a key project deliverable (see Appendix J below), and discussions have begun with JISC on funding an experimental M2M interface between the HILT Phase II pilot terminologies server and another shared service (not originally a project deliverable). There are, however, good reasons for also considering the inclusion of direct user interfaces (one end user, one staff) as one of the services offered by this particular shared service.

Use cases 1a, 1b, 2a and 2b below sketch out examples of roles the terminologies server will play in the I.E. and progressively spell out part of the case for the provision of direct user and staff interfaces.

#### **Use Case 1a: M2M from simple user query**

A user of portal A conducts a search of the portal database for documents on Railways. The portal has information on two other portals which probably have relevant material, including the fact that they each use different subject schemes from Portal A and from each other. Portal A queries the terminologies server on the best terms to use for 'railways' in these schemes. The server supplies the terms (e.g. railroads for Portal B which uses LCSH) and Portal A searches the other portals without the user being directly involved or even knowing of the existence of the terminologies server. One of the services returns no hits, but Portal A queries the terminologies server again for broader and narrower terms and repeats the search, and returns hits to the user via that route.

#### **Use Case 1b: Direct user query of the terminologies server**

[**Note:** This process is illustrated at the end of Section 4 below as a series of screen shots from the HILT Phase II pilot terminologies server]

A user conducts a search of the terminologies server using the term 'seal'. The server responds by asking the user to 'disambiguate' the term – it asks the user to specify whether she means seal, as in the sea animal or seal, as in stationery usage or seal, as in the packaging technology and so on. The user responds with one of these and the choice is mapped to a DDC number. Using successive truncations of the number, the terminologies server queries a database of JISC collections (IESR) classified by DDC, identifies collections likely to be relevant to the query, and obtains information on the subject schemes they use. It then uses its mappings of these schemes to DDC to identify the best term to use for the user's search in a particular collection, conducts searches of the collections, and shows the user terms and sample retrieval for each one. The user either uses these retrieved sets 'as is' or conducts more detailed searches of the most promising collections by accessing them directly and using local functionality.

#### **Note on points in favour of a direct user interface:**

Various parts of scenario 1b above could be handled by local portals, rather than a direct user interface for the terminologies server. For example, local portals could hold information on other collections likely to be of value to their users, and each could offer its own disambiguation facility based on an M2M interaction with the terminologies server. This approach would allow more flexibility at the portal end and may have some value in specific instances. As a general rule, however, a direct user interface to the terminologies server is likely to be more cost-effective and to offer users a more stable environment that will not change in its essential features from one portal to the next (although, of course, each portal could interpose its own look and feel on the central interface using style sheets.

**d. The Primary Function of a Terminologies Server – A Cautionary Note on Use Cases.**

The use cases presented above and below help illustrate some ways in which the terminologies server would be used in the IE. Two points should be noted, however:

- i. The four use cases presented are not exhaustive; they provide only a selective illustration of the roles played by a terminologies server.
- ii. They tend to disguise the primary function of the terminologies server. This is to optimise the ability of users to carry out successful subject searches<sup>12</sup> by providing a process that will, in time, permit JISC and JISC services to:
  1. Achieve and maintain as a high a level of interoperability as possible between:
    - The different standard subject schemes and versions of standard schemes in use in different services, both within and out with JISC
    - Amendments, additions, and extensions made to standard schemes across the services
    - Terms used by users when composing search strategies
  2. Optimise both the consistency with which staff across the various services apply schemes in the subject description of materials, and the ability of users to formulate successful search queries

**Use Case 2a: M2M from simple staff query**

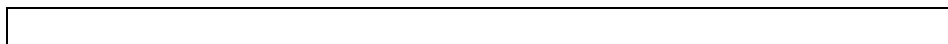
A member of cataloguing staff at Portal A is creating metadata for the first work on railways ever added to the database. He types 'railways' into the Portal A metadata form and clicks the 'get standard LCSH term' button. The portal queries the terminologies server, and receives back 'railroads' and associated terms. It automatically adds railroads in to the appropriate field in the metadata form but displays the associated terms also. These are not required, however. The staff member accepts 'railroads'.

**Use Case 2b: Metadata professional uses terminologies server directly**

A member of the cataloguing staff at Portal A is creating a metadata record for a work on 'Rail Services' using the Portal A metadata form. She clicks on the 'get standard LCSH terms' button and a new browser opens up offering direct access to the staff interface of the terminologies server but passing the term 'rail services' through to the server automatically using the appropriate M2M protocol. The server informs her that the standard LCSH term is railroads and also shows her sample retrieval from other collections using LCSH, enabling her to check that she is using the term correctly. It also informs her that if she wants to use an additional term more suited to UK users, the term that has been agreed for this across the UK and stored in the server's central mappings database is 'railways', allowing her to add an additional non-standard term without causing interoperability problems (note that an agreed list of UK oriented modifications to standard schemes requires central coordination of the kind made possible by a terminologies server).

**Note on the value of a direct staff interface:**

It would be entirely possible for local portals to all set up their own mechanism for searching other relevant portals to show their cataloguers whether or not they were applying a term correctly in particular instances, but doing this once through a central server would be more cost-effective than doing it many times in local portals.



**e. Cautionary Note on ‘Collection Level Requirements’**

The project was asked to focus in particular on the ‘collection level requirements’ of a terminologies server, and has, for the most part, done so. As is clear from use case 1b above, a key role of the server is assumed to be the identification of collections relevant to a particular subject search and the provision of advice on what subject scheme is used by the collection and what terms from that scheme are appropriate to the subject search in question.

It is worth noting, however, that the HILT team has not seen the focus on ‘collection level requirements’ as implying mappings between terms in one scheme to those in other schemes should only be carried out at less specific levels of granularity (shipbuilding as opposed to warships, for example). In this regard, the view taken has been that there is no necessary connection between more general levels of granularity in subject description and ‘collection level requirements’. The user need will most often be to map a subject search at a very specific level of granularity up to a collection classified at a higher level and then down again, within the local scheme used, to a level of granularity appropriate to the original query. Limited granularity mapping would not permit this (see also section 5 under *Limited Granularity Mapping*).

**Concluding Remarks**

Sections 2-7 describe the work carried out by the project, conclusions reached on terminologies server design, on the best way of progressing towards the development of an operational server, and on the thinking behind these conclusions. The results of an M2M requirements study carried out by UKOLN, and the report of the Project Evaluator are included as Appendices J and K respectively

---

<sup>12</sup> Note that the aim is neither to improve precision at the expense of recall, nor to improve recall at the expense of precision, but rather to provide users with the information they require to do either of these things depending on their needs at a given time

## 2. Aims, Processes, Methodologies, Literature, User and Staff Surveys

### HILT Phase II: Aims, Background, Approach

HILT Phase II was funded by JISC to conduct research into the problem of achieving and maintaining interoperability in the subject descriptions and classification of distributed information resources. More specifically, it was asked to set up a pilot terminologies service for the JISC Information Environment, aiming to:

- a. Provide a practical experimental focus within which to investigate and establish subject terminology service requirements for the JISC I.E., with particular reference to DNER, RDN, User, Collection Level, International Compatibility, and local, regional, national and UK-wide access considerations.
- b. Make recommendations as regards a possible future service, taking into account a range of factors, including the level and nature of user need, practicality, design requirements, effectiveness, functionality available in existing commercial software packages as against original development, and (above all) costs against benefits to FE and HE users of a full terminologies service focussed primarily on collection level needs.

As specified in the project plan, the findings of HILT Phase I provided the starting point for this work. These were<sup>13</sup> that:

- Many different subject schemes and practices are in use in UK services who believe that subject searching across their services is of value both to their users and their staff.
- There was a strong consensus across the Archives, Electronic Services, Library, and Museums communities in favour of a more practically focused follow-up pilot project that would develop a pilot service that would map subject schemes together, probably using a DDC spine.
- Further research was required into the effectiveness, level and nature of user need, practicality, design requirements, and costs against benefits of such an approach before a long term commitment to a possibly expensive service could be justified. This, it was determined, could best be done via a pilot project that would examine these and related issues.

The aim in HILT Phase II was thus to build a pilot terminologies server based on a mapping approach that would put in place a community process that would develop, maintain, and gradually improve interoperability of subject descriptions. Not only was this in line with the community consensus in HILT I (which strongly favoured the mapping approach over a range of other options, including the option of adopting a single scheme across the communities) it also recognised that mapping schemes together was probably the approach most likely to be compatible internationally. Even if JISC were to adopt a single subject or class scheme, it is unlikely that the same scheme would be adopted everywhere. Even if it were, mapping would still be necessary to deal with language variations across the world. It would also be necessary to deal with a key requirement of any terminologies server – the need to map subject terms used by information seekers to those used by staff working on the subject description of resources.

Project outputs were (1) a specification for an operational terminologies server, (2) a report on the associated cost-benefit analysis, (3) A report on the machine to machine (M2M) requirements of a terminologies server (compiled by UKOLN), (4) a final report on the project with recommendations with regard to the progression of terminologies server development and appendices on methodologies, surveys, the workshop, and other relevant areas of work, and (5) an illustrative pilot terminologies server (see Section 4 below). All were delivered.

---

<sup>13</sup> Full details of the findings can be found at <http://hilt.cdlr.strath.ac.uk/Reports/FinalReport.html> and on the HILT website generally

## Overview of Project Processes

The diagram at the end of the Executive Summary (see page 9 above) gives an overview of the approach taken to carrying out the work of HILT Phase II. In summary:

- An initial model for a JISC IE terminologies server was formulated from HILT 1 outcomes and HILT II aims and used to drive acclimatisation, training, and early adaptation work on the pilot server.
- Methodologies were developed to guide research that would inform the amendment and refinement of the model and the cost-benefit analysis process.
- A pilot service was developed as a research aid. This was based on Wordmap software<sup>14</sup> adapted to suit project requirements in various ways.
- Research was carried out on the model and on the cost-benefit analysis process. This included discussions with experts and stakeholders, literature searches, surveys, a user workshop, and other processes.
- A complex interaction between methodologies driven work, pilot design, other research, and the cost-benefit analysis process took place over the various stages of the project, leading to a continuously refined model, a specification for a full server, and final project conclusions. The cost-benefit analysis – based on a methodology developed by the JISC-funded INSIGHT project<sup>15</sup> – also informed recommendations regarding a follow up project that would begin to develop an operational service.

## Methodologies

Exploratory project research work was coordinated via a methodologies document. This developed over most of the lifetime of the project, moving through 7 major revisions. The last of these, included in this report as Appendix A, was completed in September 2003 just prior to the HILT Steering Group meeting which conducted the cost-benefit analysis process. It provides an outline account of the major areas of work carried out, together with an indication of its significance for the project. The summaries below give an overview of the document, specifying the main areas of research work carried out and their functions within the project. They also highlight outcomes of particular significance to the core thread of HILT Phase II (see paragraph below under ‘Core Thread...’). Full reports on the various pieces of work coordinated through the Methodologies Document, covering (where appropriate) further information on the detailed (as opposed to outline) methodologies used, are provided in the appendices noted in the summaries below.

### *Summary Details of HILT Methodologies Document*

Project and pilot evaluation and quality assurance and review methodology. (Section 0 of the document).

*Function:* To ensure that the members of the Professional Level Evaluation Group (PLEG), including the Project Evaluator, were consulted on the methodologies used, on conclusions reached, and on the quality of project deliverables.

*Significant Points to Note:* The project team made every effort to ensure ongoing consultation, although this was not always as easy to do in practice as it sounded in theory, partly due to practical considerations, partly due to the limitations of time and resources available to the project. The group might be consulted on a general proposed approach and their agreement obtained, for example, but changes to details might have to be made subsequently on which it was not practical to consult due to timescales. Or – as in the case of the cost-benefit analysis where PLEG was kept informed by email of developments but the major interaction was (as agreed) with the Steering Group – it might be more appropriate that the details of a methodology be agreed with a project group other than PLEG. In the last analysis, ultimate control in this area depends on the final

<sup>14</sup> See <http://www.wordmap.com/>

<sup>15</sup> See <http://www.mis.strath.ac.uk/predict/projects/insight/index.htm> and Nicol, David and Coen, Michael. [A model for evaluating the institutional costs and benefits of ICT initiatives in teaching and learning in higher education](#). *ALT-J - Association for Learning Technology Journal*. 11(2) 2003. p46-60.

element of project activity – the presentation of the team’s Final Report on activities, products, conclusions, and recommendations to the Project Evaluator and the Project Evaluator’s subsequent evaluation report (see Appendix K).

*Further Information:* Full reports on detailed work carried out, its outcomes, and the resulting conclusions are provided in this Final Report and its Appendices.



Literature Review (common thread covering issues from all areas of HILT work).

*Function:* To ensure that project progress was informed by issues and outcomes thrown up by relevant research reported in the literature.

*Significant Points to Note:* This was an extensive literature review, carried out to learn from prior research addressing mapping between and among terminologies and subject schemes in various subject areas and to shed light on the problems faced and issues addressed. It can be divided into three main parts.

The first part of the literature review investigated issues such as the integration of thesauri in a common subject area, subject switching, merging classification schemes with thesauri and the associated problems, and mapping between controlled vocabularies such as LCSH to thesauri. One of the studies, which has mapped Laborline thesaurus terms to LCSH, suggested 19 possible types of match between terms derived from the vocabularies, a point that informed elements of pilot server design and contributed to aspects of the specification for a full operational server.

The second part of the review looked into recent projects which aimed to use subject schemes for cross-searching and cross-browsing across electronic collections available on the web. The following projects were studied:

- CARMEN (2000)
- MACS (2000) (Multilingual Access to Subjects)
- LIMBER (2001) (Language Independent Metadata Browsing of European Resources)
- RENARDUS (2001)

Each of these projects investigated mapping issues with a different set of subject schemes. DDC was utilised in CARMEN, RENARDUS, and as reported in Saeed and Chaudhury (2002)). None of the projects tackled quite the same territory as HILT. However, all provided useful insights into the issues.

The last part of the literature review dealt with issues raised by the methodologies document. The questions covered were: subject schemes in use by JISC projects, collection strength testing methods, reasons for departure from standard schemes, user evaluation of terminologies: interfaces and usability, subject retrieval and subject queries, effectiveness of Cheshire clustering approach, approaches to solving subject interoperability, and the use of DDC for particular domains.

*Further Information:* Appendices B.1 and B.2 report in full on literature survey results.



Methodologies to ensure investigation examined representative services, subject schemes, and subjects within schemes as it developed views on the HILT model, mapping, functionality and interface features, cost-benefit analysis requirements, and so on (Section 1 of the Methodologies Document).

*Function:* To ensure that project developments were well-informed about subject description practices and issues across the range of JISC services

*Significant Points to Note:* The main product of this subsection of the Methodologies Document was a survey of JISC services and their staff. This informed the team in a general way as it dealt

with the range of issues listed above, but also helped inform the project in specific ways. The primary objective of the survey was to examine representative services, and the subject schemes they use, any implications arising from the need to use specialist thesauri, whether service staff modify standard terms to suit their needs and, if so, how and why.

An analysis of the data gathered through the survey indicated that the vocabularies such as DDC, LCSH, the UNESCO thesaurus, and the HASSET thesaurus (based on UNESCO) were the widely used subject schemes employed by JISC collections and services. There were also a number of services and collections who used in-house schemes.

The collections and services were also asked to provide reasons for the departure from standard schemes. The main reasons were:

- To accommodate new concepts or areas of knowledge
- To reflect user needs or demands
- Subject scheme is too broad

In addition a few of the services and collections noted other reasons such as geographic specificity, bilingualism, and cultural differences.

*Further Information:* A full report on the survey outcomes and the questionnaire used is provided in Appendix C.2



Methodologies to ensure investigation examined representative user types, tasks, and associated retrieval requirements and strategies as it developed views on the HILT model, mapping, functionality and interface features, cost-benefit analysis requirements, and so on (Section 2 of the Methodologies Document).

*Function:* To ensure that project developments were well-informed on user-associated issues.

*Significant Points to Note:* This subsection of the Methodologies Document had two main products. The first was a survey conducted by interviewing a range of users<sup>16</sup> on their views on, and approaches to, subject searching. The second was a user workshop which focused on the use of the pilot terminologies server and drew out information on a range of issues relevant to its design and on some of the assumptions that underpinned the design. Both informed the team in a general way as it dealt with the range of issues described above. In addition, the user interviews helped:

- To acclimatise project staff to problems and issues related to dealing with users and subject searching situations in a range of subject areas;
- Improve the design of a subsequent user workshop designed to evaluate aspects of the pilot terminologies server;
- Provide information on the nature of user subject searching requirements, on their willingness to consult a range of collections, on the level of specificity of the search terms they are likely to use, and the mix of search strategies likely to be employed.

The workshop also provided feedback that influenced the development of the terminologies server specification post-workshop, providing information on the usability of the pilot interface, the need for a UK oriented scheme modifications registry, the effects of training, and other matters.

*Further Information:* Further information on the workshop and its effects on project development is provided in sections 4 and 5 below. Full reports on both products, together with accompanying questionnaires, are provided in Appendices C.3 and C.4 and D.1 – D.3 respectively.




---

<sup>16</sup> These included students, intermediaries, lecturers and researchers

Methodologies to ensure (1) that the full functional requirement for an operational (as opposed to pilot) terminologies server was identified (2) That the extent to which it was implemented in the pilot was optimised (3) That the software used in the pilot was utilised in a way that faithfully reflected any specific requirements implemented and tested (Section 3 of the document).

*Function:* To guide the process of developing a specification for the pilot server, implementing it correctly, and using the pilot to inform the project as it developed the full specification for an operational server.

*Significant Points to Note:* See under further information below.

*Further Information:* The process of developing a specification for the pilot server, implementing it correctly, and using the pilot to inform the project as it developed the full specification for an operational server is described in main report sections 3 – 5 and developed further in sections 6 and 7.



Methodologies to ensure investigation examines terminologies server design options adequately and in a fashion useful to JISC (Section 4 of the Methodologies Document).

*Function:* To determine the options to be assessed in the cost-benefit analysis.

*Significant Points to Note:* The view of this issue developed over time, particularly in the context of meetings of the Steering Group (which was to conduct the cost-benefit analysis). The final position taken was that there were two levels at which cost-benefit analysis was appropriate. The first was functionality levels, the main determinant of costs and benefits. The second was instantiation methods. This assessment emerged as the team fine-tuned the methodology agreed for the cost-benefit analysis just prior to using it at a Steering Group meeting.

*Further Information:* The cost-benefit analysis and the options to which it was applied is covered in more detail in Section 6 below, and in Appendices H.1 and H.2.



*Element:* Methodologies to ensure the investigation conducts a fair and comprehensive approach to the cost-benefit analysis of the various options for terminologies server design agreed under Methodologies Document Section 4 (Section 5 of the Methodologies Document).

*Function:* To agree an appropriate approach to cost-benefit analysis, adapt it for HILT, and conduct it in an agreed fashion.

*Significant Points to Note:* It was agreed at both the Steering Group and the Project Management Group that the methodology developed by the JISC-funded INSIGHT project be adopted and adapted for HILT use, and that the cost-benefit analysis be carried out by the HILT Steering Group.

*Further Information:* The cost-benefit analysis was carried out as described in Section 6 below and in Appendices H. The outcome of the process is also described in these parts of the report.



### **General Influence of Methodologies-driven Work**

A significant point common to all of the above areas of project effort – and worth highlighting here - is that a primary influence of the various pieces of work was its contribution to forming and developing the team's view of the functions and specifications of an operational terminologies server and of the best approach to progressing its implementation through a follow up project. This influence could be straightforward – as in the case of the influence of the staff survey on the need for, and form of, a 'UK oriented modifications registry terms set' in an operational server. It could

also be less direct. For example, sometimes the act of carrying out a piece of research brought results unrelated to the primary focus of the research itself – stimulated thought, highlighted problems not previously considered, took the project down new avenues.

## **Core Thread of HILT Phase II**

The process of forming and developing a view on the functions and specification of an operational terminologies server and on the best approach to progressing its implementation is the focus of the remainder of this report. This process was the primary thread of HILT Phase II project work and is described in Sections 3-7 of the report. These follow the logical progression from initial specification to final report recommendations shown in the workflow diagram below. In reality, the process was more complex and less linear than is suggested by this logical progression – the process of developing the specification did not halt during the creation of the pilot, for example, and was influenced by the (ostensibly later) process of working on the detail of the cost-benefit analysis methodology. However, the order followed is an accurate reflection of how the core thread of the work progressed generally and it is sensible (and unavoidable in a report that is itself linear) that it be utilised to structure the remainder of the report.

### *Core Thread Diagram*

(Section 3)

An early outline specification, together with work on mapping issues and other relevant areas, was developed into an interim specification for a terminologies server.



(Section 4)

As far as possible within time, resources, and ease of adaptability of the software, the interim specification was implemented to create a pilot terminologies server. This was then assessed in various ways, including via a user workshop.



(Section 5)

The outcomes of the assessment process were utilised to identify the additional requirements of developing an operational server – these being taken to include both additional areas of functionality and areas where additional research would inform required functionality or its implementation.



(Section 6)

A detailed cost-benefit analysis process was developed based on an adaptation of a methodology developed by the JISC-funded INSIGHT project. This was used to perform a cost benefit analysis on the requirements identified in Sections 3 and 5, including those that relate to further research



(Section 7)

The outcomes of the cost-benefit analysis, together with project processes generally, then informed project conclusions and recommendations

### 3. Developing an Interim Specification

The process of forming and developing an interim view on the functions and specification of an operational terminologies server was managed through Sections 3.1, 3.3, and 3.4 of the Methodologies Document, as described under the headings tagged (3.1), (3.3) and (3.4) below. It culminated in the position expressed in the document *HILT Terminologies Server Pilot Specification (Version 3.0)* (see Appendix I.1(3)), a working discussion document to inform the programmers working to implement the pilot.

#### Agree initial service specification at outline level (3.1)

An 'initial service specification' was compiled and agreed with the Project Management Group and the Steering Group. This took the form of a description of a user's interaction with a Wordmap-based system and is included in Appendix I.1(2). The starting point for this early specification was a combination of the outcomes of HILT Phase I (see, in particular, I.1(1) and the aims of HILT Phase II.

Some elements of the specification are implied or assumed rather than spelt out:

##### *The Mapping Based Approach*

There was an assumption that the approach taken would be based on the HILT I finding that there was a strong stakeholder consensus in the UK that adoption of a single scheme by all communities, and even all services and institutions within a community, was not an approach likely to be widely favoured; that interoperability in respect of subject description should be based on the creation of an online service that would map between subject and class schemes and ensure consistency where schemes were amended or extended. (HILT I Final report, available on the HILT web site at: <http://hilt.cdlr.strath.ac.uk/Reports/Documents/HILTfinalreport.doc> ). This would permit stakeholders to use the scheme best suited to their needs but provide an ongoing online process that would, in time, ensure interoperability between services using different schemes<sup>17</sup>.

The aim in HILT II was to build on this outcome and determine the specific design requirements of a terminologies server based on the mapping approach.

##### *A Wordmap-based Approach*

It was stated in the bid for HILT II that the Wordmap software would be the basis of the pilot service and that, therefore, there would, in addition to the user interface, be a database of terminology mappings, and a staff interface to allow these to be viewed, used, amended and extended.

Taken together, these elements:

1. Facilitated the exploration of the possibilities of the Wordmap software as regards building a pilot server (Methodologies Document 3.2)
2. Provided the basis for drawing out an interim specification for an operational server

#### **Begin work on an interim specification by determining an initial list of end user, staff user, and schemes coverage requirements. (3.3)**

A consideration of the results of the user and staff surveys conducted by HILT II informed extensive discussions within the team and with the rest of the project groups. This process had two outcomes:

---

<sup>17</sup> The UK e-learning community has also indicated support for a mapping-based approach – See, for example, Duncan, C., Campbell, L, Graham, G. *(Not) an Idiot's Guide to Metadata*. Available at [http://www.estandard.no/docs/charles\\_duncan\\_april\\_2003/duncan-campbell-graham.doc](http://www.estandard.no/docs/charles_duncan_april_2003/duncan-campbell-graham.doc)

1. The development of a view on the general assumptions that should underpin server design and on the requirements implied by this view
2. The development of an interim position on subject schemes coverage.

Both were agreed with the Project Management Group and the Steering Group and are summarised below under the headings *Server Design: General Assumptions* and *Subject Schemes Coverage*

#### *Server Design: General Assumptions*

The view of these design requirements that emerged was developed within the project as the team:

- Conducted literature reviews, a survey of services, user interviews, terminology mapping exercises, and other preparatory work related to pilot construction
- Investigated, considered and discussed issues with colleagues, and with the project groups and other stakeholders, including the two terminology experts

The logic that underpins the position stems from a recognition that retrieval by subject in a distributed multi-scheme environment would be optimised if (1) standard subject schemes and class schemes were used 'as is' to describe resources (or, where schemes were modified, if the modifications were standardized across the UK through a central coordinating mechanism), (2) it was always clear to all assigning terms how the scheme and any modifications should be used for a given resource, and (3) all users seeking to retrieve resources had a complete knowledge of the scheme any agreed modifications and how it would be used to describe resources and applied that knowledge in retrieval attempts.

This, in turn, suggests that the requirement is for a terminologies server that is designed to:

- Improve accurate, consistent description by staff through training, feedback on items appropriately assigned particular terms, and the provision of a central coordination process to ensure country-wide consistency where changes and extensions to standard schemes are deemed necessary to help harmonise standard terminologies with terms used by users (a 'UK oriented modifications registry terms set')
- Improve accurate, informed searching by users by providing a coherent subject environment across services (partly through the first process above), information on standard and changed or extended terms used by staff, a user term disambiguation process, a find relevant collections facility, useful feedback mechanisms, training and acclimatisation modules, and processes to learn about user searching behaviours
- Map between terms used by user and terms used by staff utilising different standard schemes in different services
- Offer a process that will not only halt the deterioration in subject interoperability suggested by project research but also (possibly) provide a slow but sure means of dealing with subject interoperability problems in legacy metadata (there are reasonable grounds for holding that the creation of a 'UK oriented modifications registry terms set' and the coordination of its ongoing maintenance and development would do both in time). HILT surveyed staff at JISC services on whether they amended or extended standard subject schemes and found that they did. The main reasons given for doing this were: to accommodate new concepts or areas of knowledge (31%), subject scheme is too broad (27%), and to reflect user needs and demands (27%). Less common reasons given were: to facilitate geographic specificity (e.g. place names), to reflect bilingualism and cultural differences, subject scheme being too detailed, to reflect the services/collection sector (e.g. HE/FE) and to reflect the service/collection domain (e.g. libraries, museums, archives).

Since the project was asked to look at 'collection level requirements', and since the JISC IE comprises distributed services with overlapping subject coverage but (in many cases different subject schemes and practices in place), there is also a requirement to map user subject queries to JISC collections and advise users, either directly or via M2M functionality, of what service might answer their queries, and what terms in the subject schemes used are required for searching.

*Subject Schemes Coverage*

A survey of JISC collections and the many different subject schemes they use is included as Appendix C.2. The number of schemes listed is large compared with the few schemes used in the pilot and the few schemes considered for initial inclusion in an operational terminologies server. Moreover, the number of schemes used by services in the world at large, many of which are likely to describe resources of value to JISC users, is even greater. Clearly, it would never be practical for JISC to cover all of these schemes – and whilst there is perhaps a case for ultimately covering more schemes than can be encompassed within the initial phases of an operational service, it would neither be feasible, sensible, nor affordable to aim to cover all of these from the start. The approach proposed is therefore a gradual one that focuses on key schemes initially and assumes and encourages the involvement of other players in the creation of inter-terminology mappings, as follows:

**Term Sets Covered**

It is proposed that the initial focus be on mounting or creating term sets of class schemes and mapping between them, covering the following:

- A UK oriented modifications registry terms set, important because it will provide a means of mapping terms not in standard schemes but used by UK users to appropriate standard scheme terms and should also help resolve legacy metadata problems. Regional variations on a core UK non-standard terms set are also a potential requirement.
- DDC, important because it is well-used within JISC (see Appendix C.2) and internationally, and is the best approach to the provision of a spine, having a machine-processable hierarchical numbering system suitable for use in collection finding, and also being translated into more than 30 different languages.
- LCSH, important because it is well-used within JISC (see Appendix C.2) services and internationally, and because OCLC already have a mapping of a major portion of it to DDC and a programme for extending the mapping.
- UNESCO, a small term set well-used within JISC (see Appendix C.2) and particularly favoured in the archives community (as is LCSH)
- AAT, likely to be most popular in the Museums community, and important if working with that community is, or becomes, important to JISC.
- MESH, used at significant levels in the JISC community and internationally, and an example of a more subject-specific and detailed scheme that will provide a model for mapping other similarly subject-specific and detailed schemes

**Determine types of mapping problems likely to be encountered in building a terminologies server and specify mechanisms for solving these problems (3.4)**

This work informed the views of the team on server mappings database design and facilitated the process of creating illustrative mappings in the pilot server. Full reports are provided in Appendices B.2 and B.3 (although B.1 also has relevant material). One outcome was a recognition of the need for a field in the mappings database to specify relationship type (in recognition of the fact that there may be as many as 19 different types of relationship between terms in different schemes). Other points worth noting are summarised below.

*Terminology mapping Issues summary*

In order to inform the HILT project on practical issues and problems of terminology mapping, a series of testbed mapping exercises were carried out. The following provides a list of subject schemes used in the testbed mapping:

- UNESCO and MeSH: Health and medical section
- UNESCO and DDC: Health and medical section
- Wordmap Global Taxonomy and DDC: Health and medical section
- Mapping MeSH to DDC: Ethics section
- Mapping MeSH to DDC: Health services administration > Quality of health care

The aim of the testbed mapping was to investigate the extent of compatibility between different subject schemes - in particular between thesauri such as UNESCO and MeSH and DDC. The medical area was chosen as a) it was quite specific b) there were a number of JISC medical services and collections and c) the fact that DDC, UNESCO and LCSH have all medical sections.

The testbed mapping between UNESCO and MeSH showed that all the health related terms in the UNESCO thesaurus were covered in one way or another by the MeSH thesaurus. Most of the terms mapped were either exact match or cross-reference matches while a few of them were superordination and subordination matches (examples of the range of match types are presented in Appendix B.2, a small selection is shown in the table below).

The testbed mapping between UNESCO and DDC indicated that most of the mapped terms were either exact match or exact cross-reference match.

A mapping exercise between the Wordmap taxonomy and DDC demonstrated that DDC has a larger set of terms than Wordmap and half the terms mapped were either concept match or exact match. The remaining half included terms with one word in common.

Two separate testbed mapping exercises were conducted for MeSH and DDC to ensure the validity of the mapping as MeSH represents a specialist thesaurus while DDC is a general subject scheme. The testbed mapping between MeSH and DDC suggested the possibility<sup>18</sup> that the majority of MeSH terms could be mapped to DDC. The match types considered were exact match, cross reference match, concept match, subordination match, and super-ordination match. A few of these are illustrated in the table below. Further information is available in Appendix B.3.

MeSH	DDC	Match Type
Bioethics	174.957 Bioethics	Exact Match
Program evaluation	352.439 Management, performance, program audits	Concept Match
Principle-based ethics	170 Ethics	Super-ordination match

These mapping exercises provided useful insights into the practical issues and problems of terminology mapping in particular between specialist thesauri and general subject schemes such as DDC.

### **HILT Terminologies Server Pilot Specification (Version 3.0)**

The various processes above helped inform the creation of the document *HILT Terminologies Server Pilot Specification (Version 3.0)*, included in Appendix I.1(3) of this report. This is a statement of the team's interim position on server functionality requirements as construction on the pilot server began in earnest. It was a working discussion document used over several months to help coordinate the work on the pilot. No formal updates to it were produced.

---

<sup>18</sup> This was only a small sample

#### 4. Building and Assessing the Pilot

The process of developing and building the pilot server was managed through Sections 3.2, 3.5, 3.6, and 3.7 of the Methodologies Document, as described under the headings tagged as (3.2), (3.5), (3.6) and (3.7) below.

##### **Determine functionality available in pilot software (3.2)**

Utilising the initial (as opposed to Interim) service specification described under (3.1) in the last section of this report, the team investigated the possibilities of the Wordmap software as regards the instantiation of the functionality requirements indicated. This entailed attending Wordmap training, reading documentation, exploratory use of the software, and discussions with Wordmap technical support and training staff.

##### **Identify adequate mechanisms in the pilot software and in other areas of project work for implementing the requirements identified and implement these in a working pilot. (3.5)**

Utilising the general assumptions, subject coverage aspirations, and Specification Version 3.0 described in Section 3 of this report, the team investigated the extent to which the Wordmap software and other means available to the project (machine processing of files of subject terms sets, for example, or manual mapping) could be used to implement these in a working pilot. In the event, it was not possible to implement all aspects of the interim specification, although it was possible to implement most. Elements not implemented were not due to limitations in the Wordmap software, but to other factors (lack of programming or manual mapping time, failure of other processes (see UK oriented modifications registry terms set information below)).

The end result was the working pilot at <http://hiltipilot.cdli.strath.ac.uk/pilot/top.php> and that is illustrated to some extent in the screen shots provided later in this section of the report.

The following is an outline description of what it does and does not include:

##### *Inclusions: List of Features*

The pilot is based on a DDC spine and encompasses:

- Access to the whole of DDC 21, indexed on the DDC captions, standard sub-divisions, relative index, and the other schemes mentioned below
- Mappings of DDC to LCSH as provided by OCLC
- Illustrative mappings to UNESCO and MeSH
- An illustrative staff interface to the system based on standard Wordmap windows ‘drag and drop’ style interface but utilised in a HILT-specific way
- A user query interface
- A user query disambiguator (e.g. by lotus, do you mean the flower, the car, the software etc)
- A ‘find collections appropriate to disambiguated query’ function, based on a DDC truncation algorithm and (simulated) interaction with the proposed JISC IESR<sup>19</sup> shared service (HILT has fed its conclusions into the IESR shared service – see Appendix G)
- A ‘determine subject scheme used by retrieved collection’ function
- A ‘determine specific term from that scheme that maps to users query’ function
- A ‘find hits in retrieved collection’ function using this term
- Minimal on-screen user help

##### *Exclusions: UK oriented modifications registry terms set*

The project was not able to develop an illustrative ‘UK oriented modifications registry terms set’. It had been hoped that a machine-readable file mapping DDC numbers to captions created in a UK university library might provide a ‘first pass’ at this terms set. In the event, differences between the file and the terms in the DDC file provided by OCLC were not significant. This is not thought to invalidate the idea that such a set is needed. Data from the HILT staff survey shows that service

<sup>19</sup> Information Environment Services Registry – see project website at (add URL)

staff do amend and extend schemes for UK purposes and the User Workshop (see Appendix D.3) also provided some supportive evidence. Fortunately, the project does have information on the likely content of UK modifications terms set.

A 'UK oriented modifications registry terms set'. set is a set of terms not in standard schemes but likely to be used by UK users for retrieval purposes. It includes UK versions of terms in standard (often US-oriented) schemes (e.g. GP or General Practitioner for family doctor), regional variations of these, terms more specific than those in standard schemes, new terms, and other variations<sup>20</sup>. Because UK staff describing resources are aware of these variations between standard schemes and terms used by users and usually attempt to enhance standard descriptions with terms of this kind, it is likely that such a term set, once developed, will aid in the resolution of subject interoperability problems in legacy metadata. By creating a harmonised version of this term set, storing and maintaining it on a central service, and mapping it to standard schemes, we can remove interoperability problems created when staff use different UK variations for the same concept, aid users by showing them additional non-standard terms used by services, and map user searches to a range of standard schemes used in services more or less automatically.

#### *Exclusions: AAT*

No machine-readable mapping of AAT to DDC was available and project resources did not permit manual mapping before the beginning of the workshop described under (3.6) and (3.7) below.

#### *Illustrations*

The screen shots at the end of this section of the report illustrate the pilot and its use by users and staff.

### **Refine and extend the requirement and the pilot terminologies server (3.6)**

#### **Finalise requirement prior to cost-benefit analysis (3.7)**

The process of building on the interim specification to identify the additional requirements discussed in Section 5 below was informed by both the use of the operational pilot in various situations and the work entailed in building it. This process entailed three main elements:

#### *User Workshop*

A 'user' workshop at which a range of 'end users' and intermediaries (41 in total) were asked to use the pilot and respond to a set of questions. These were designed to elicit information that would permit HILT:

- To find out what students, lecturers, intermediaries think of the interface and its features and facilities (how could they be improved) [**primary aim**].
- To discover something about their subject retrieval behaviour and associated thought processes.
- To compare the terms they use with terms in the HILT database
- To compare terms used by students, lecturers, intermediaries to describe some documents by subject (URLs).
- To see whether there is any evidence in the results to suggest that learning or experience improves user performance in using the interface.
- To utilise the data we obtain to learn what we can about the efficacy of the general approach.

#### *Project Management Group Brainstorming Session*

An informal session was held with the Project Management Team looking at the pilot server in action and 'brainstorming' on functionality and other issues.

---

<sup>20</sup> That is, some of the terms are UK-specific, others are not

*Ongoing Discussion and Analysis by the HILT team*

The HILT team continued to discuss and analyse issues during both the pilot development phase and its operational phase before, during and after the user workshop and brainstorming sessions.

Outcomes from all three processes informed the specification of the additional requirements specified in 5 below and the associated Appendix I.2. Specific examples of this include:

- Helping to confirm the general approach to interface design
- Helping to refine the view stated below under 'limited granularity mapping'
- Providing confirmation of the need for a UK modifications terms set
- Helping to point up the need for a more complex disambiguation facility in the user interface
- Offering some support for the view that user performance in using the pilot terminology server may be influenced by training

These were influenced by the workshop in particular (for example, the attempt by a participant to search for GP (General Practitioner) and subsequent HILT follow-up work helped confirm the need for a UK modifications terms set). However, all three processes also contributed in a more general way to the team's overall perspective on the issue and helped finalise its view of requirements for an operational phase. The workshop also provided a wealth of information on user interface issues that will be of value in building an operational server (see Appendices D.1 – D.3 for a full workshop report).

**Screen Shots**

These begin on the next page.

Figure 1. Homepage of the HILT Pilot Terminologies Service

<http://hiltпилot.cdlr.strath.ac.uk/pilot/top.php> - Microsoft Internet Explorer


File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites Media History Mail Print Edit

Address <http://hiltпилot.cdlr.strath.ac.uk/pilot/top.php> Go

Y! Search Sign In News Games Personals Mail My Yahoo! Yahoo! Finance

Google Search Web PageRank 725 blocked AutoFill Options



The HILT terminologies server aims to identify JISC services and/or collections likely to have resources relevant to any subject query you may have. The process has three steps:

1. You enter your search term (or browse the subject hierarchies).
2. HILT looks for the best matches for your subject and asks you to clarify which is most appropriate.
3. You choose the most appropriate subject.
4. HILT tells you about possible services or collections that may interest you, tells you what subject schemes they use, and tells you the best term to use (if it can). It also allows you to connect through and do a search of the services and collections identified.

Enter your search term here...

Teeth Search

Use Double Quotes for a phrase search (Eg: "biology and life science") See [search tips](#) for details

...or browse by category

[Arts & recreation](#) → [Architecture](#), [Arts](#), [Drawing & decorative arts](#), [Graphic arts](#), [Landscaping & area planning](#), [Music](#) ...

[Computers, information & general reference](#) → [Associations, organizations & museums](#), [Bibliographies](#), [Computers, Internet & systems](#), [Encyclopedias & books of facts](#), [Journalism, publishing & news media](#), [Library & information science](#) ...

[History & geography](#) → [Biography & genealogy](#), [Geography & travel](#), [History](#), [History of Africa](#), [History of Asia](#), [History of Europe \(ca. 500 A.D.-\)](#) ...

[Language](#) → [Classical & modern Greek languages](#), [English & Old English languages](#), [French & related languages](#), [German & related languages](#), [Italian, Romanian & related languages](#), [Language](#) ...

[Literature](#) → [American literature in English](#), [Classical & modern Greek literatures](#), [English & Old English literatures](#), [French & related literatures](#), [German & related literatures](#), [Italian, Romanian & related literatures](#) ...

[Philosophy & psychology](#) → [Ancient, medieval & eastern philosophy](#), [Astrology, parapsychology & the occult](#), [Epistemology](#), [Ethics](#), [Logic](#), [Metaphysics](#) ...

[Religion](#) → [Christian denominations](#), [Christian pastoral practice & religious orders](#), [Christian practice & observance](#), [Christianity & Christian](#) ...

Local intranet

Figure 2. Disambiguation page of the HILT Pilot Terminologies Service

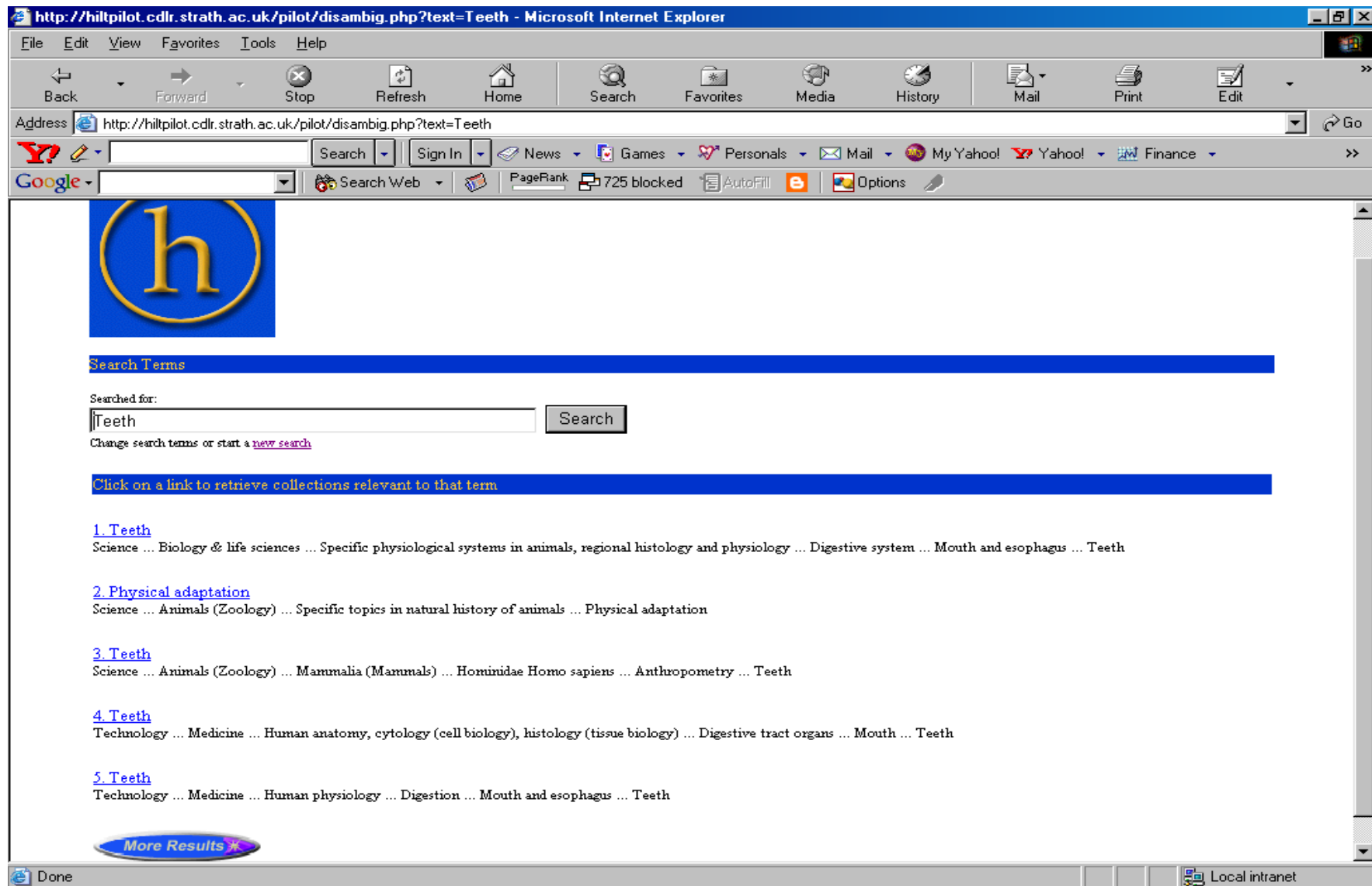


Figure 3. Collection selection page of the HILT Pilot Terminologies Service

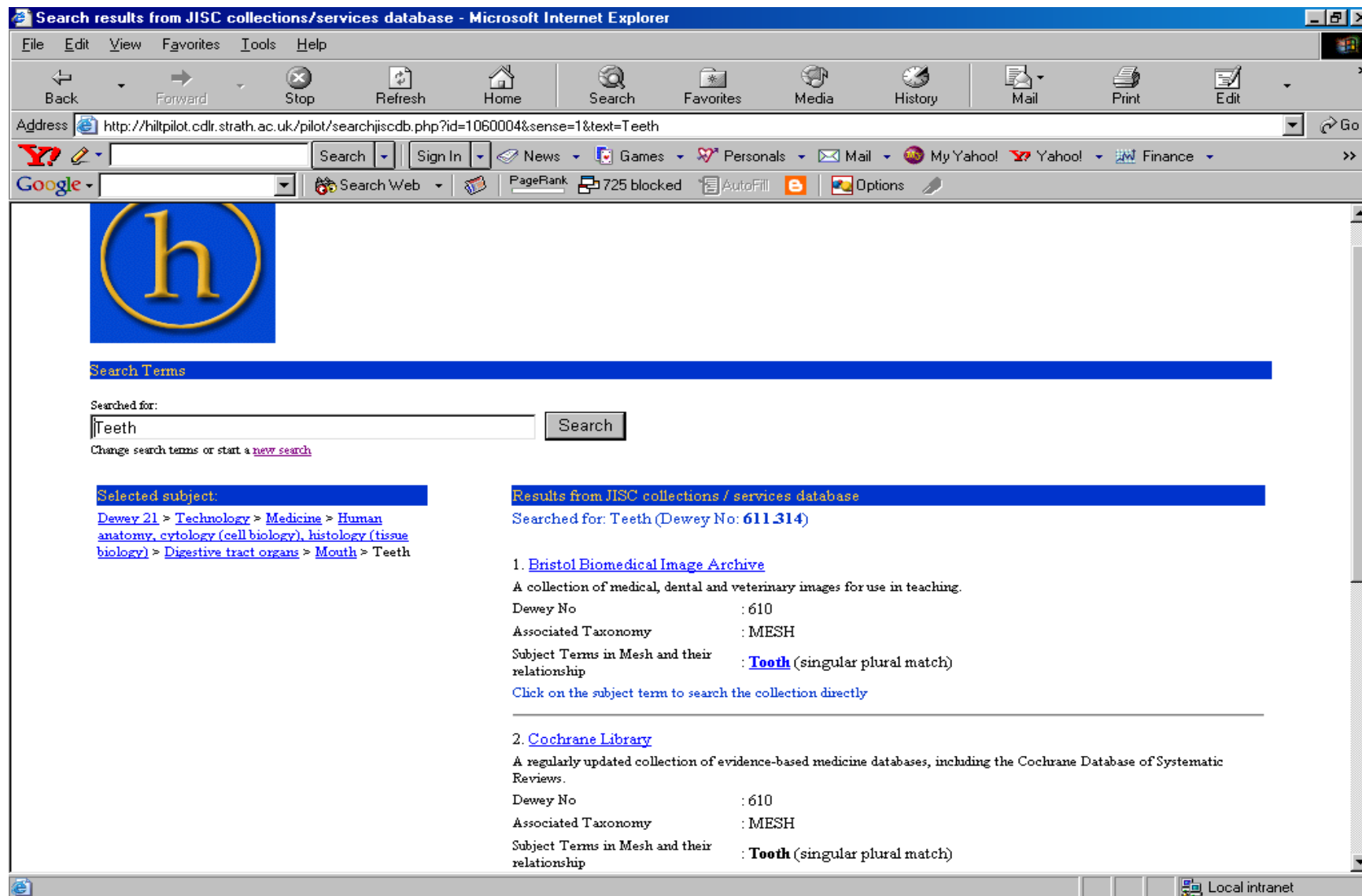


Figure 4. JISC collection found by the search term "Teeth"

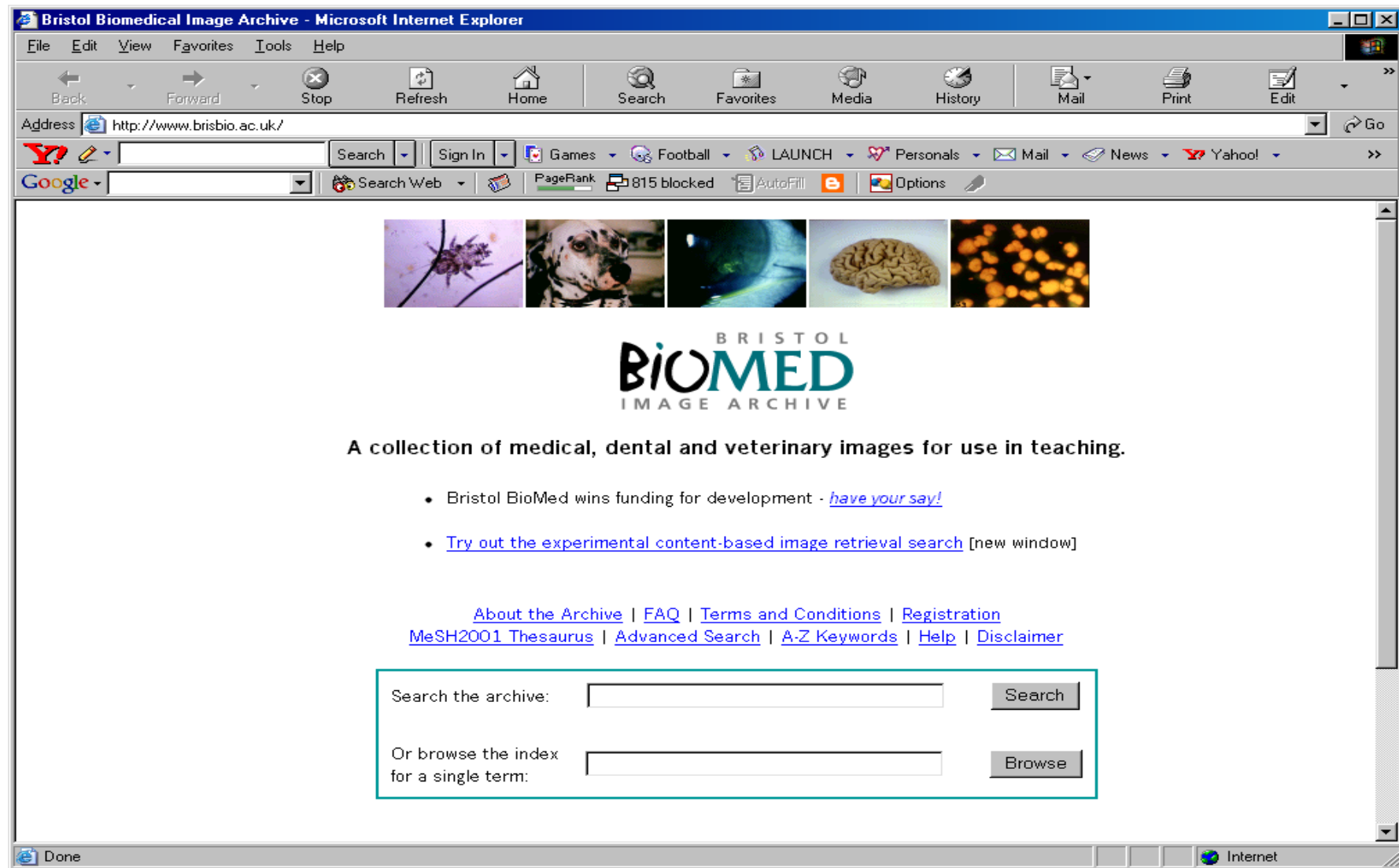
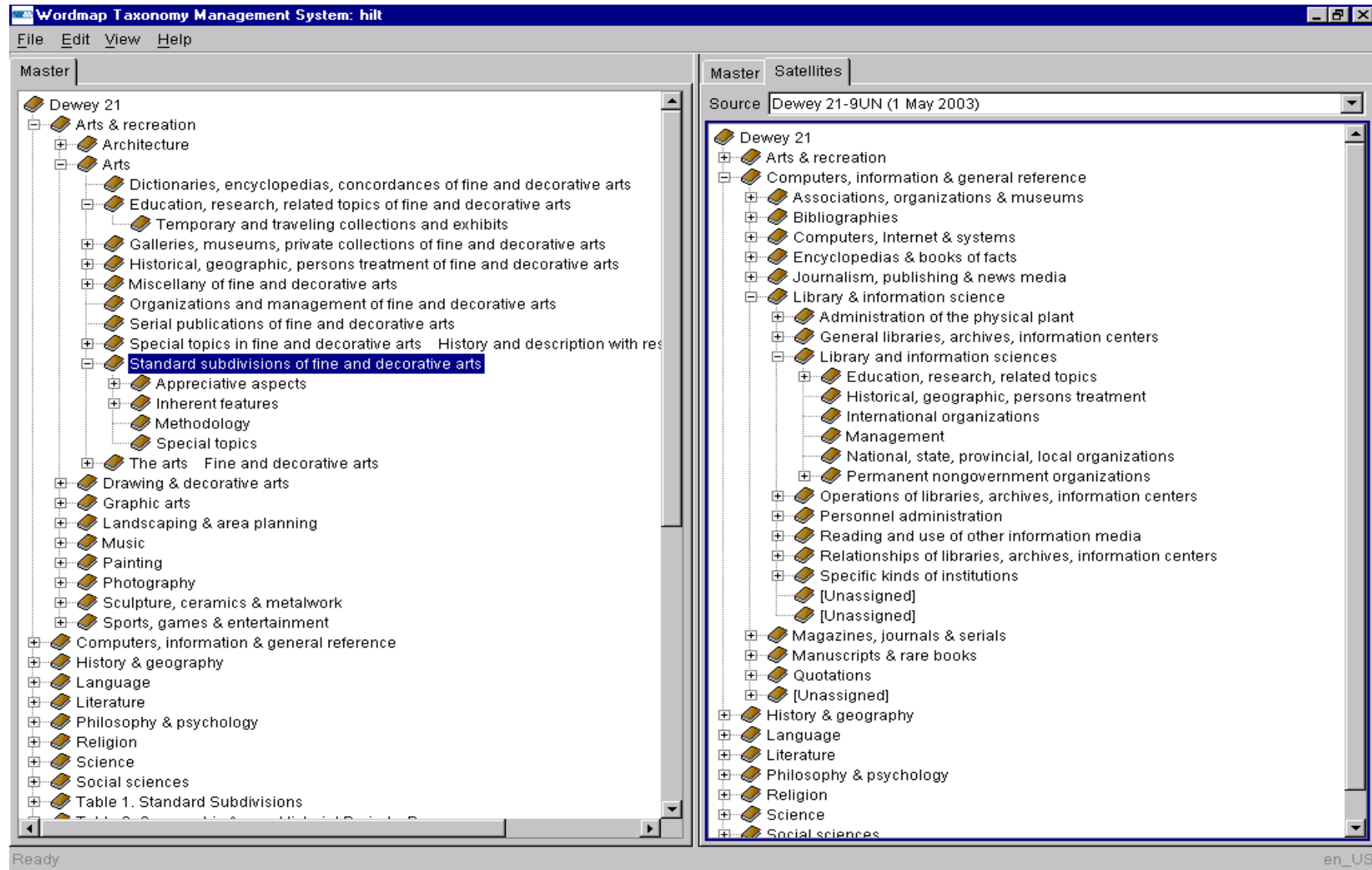


Figure 5. HILT pilot terminologies staff interface



## 5. Developing an Operational Server – Additional Requirements

Arising out of the assessments of the pilot server, and the added recognition that it did not implement the whole of the interim specification, a set of additional requirements for developing an operational server were identified, some relating to additional functional elements, some to exclusions, others to areas where research was felt to be required to illuminate the development process. Details of the range of additional requirements identified are summarised below. They are also combined, together with the remainder of the elements from the interim specification, in Appendix I.2, *The Development Requirement for an Operational Server*.

### **Preliminary Note on Areas Where Further Research is Required**

These fall into two categories. The first relates to the investigation of techniques like the clustering process utilised by Cheshire (see Appendix F). The team felt there might be a case for conducting practical tests designed to determine whether the technique has a role to play as a terminologies server user interface tool. However, conducting such tests was impossible within existing project schedules and staffing resources and had not been envisaged in the original bid. The second relates to the area of user subject searching requirements. HILT was able to learn a great deal from the work it did with users but concluded, both that there was more to learn about user needs in this area generally, and that a more complex user interface would have to be developed, requiring further testing by users on that front. The team also takes the view that the ongoing involvement of users as both the interface and the service generally develops is essential.

### **Assumptions Relating to Specific Elements of Server Design: Exclusions**

One result of the further consideration of design issues by the teams has been to exclude certain possible design elements from the specification:

#### *Direct Mapping of User Terms to Individual Schemes*

It has sometimes been argued in HILT fora that once a collection and its local scheme have been identified by the terminologies service via a DDC spine, the best way of identifying the correct local scheme term to use for searching is through a direct mapping from user term to the local scheme rather than indirectly via the DDC spine. This approach was rejected, partly because the project had insufficient resources to research it adequately, but mainly because the additional mapping required would have increased costs significantly, particularly in relation to LCSH. It is also an approach likely to give rise to difficulties in granularity levels between the local scheme, DDC, and the user term, and to increase the complexity of the disambiguation process. A study of the issue in any subsequent phase of the project should clarify whether there are any advantages in terms of retrieval accuracy.

#### *Alternative 'Open Access' Spine*

The project considered the idea of an International Standard Concept Number or ISCN scheme designed to replace the DDC spine. This might have constructive features but does not look practical in the short term. It may, however, be worth looking at as a long-term option – especially if it could be available free to anyone wishing to map their own scheme or ontology to a core 'spine' involving no licensing costs.

#### *Limited granularity mapping*

The option of mapping between subject schemes, user terms, and DDC at less specific levels of granularity only has been ruled out. The HILT view is that limiting mapping in this way would make it impossible to deal with the vast majority of most user subject queries. These tend, if anything, to be more, rather than less, specific than the levels of granularity available in standard schemes. It should be noted that there is no necessary connection between more general levels of granularity in subject description and 'collection level requirements'. The user need will most often be to map a subject search at a very specific level of granularity up to a collection classified at a higher level and then down again, within the local scheme used, to a level of granularity appropriate to the original query. Limited granularity mapping would not permit this.

**N.B.** This is regarded as a significant outcome of the project, and suggests that the idea of a 'high level thesaurus' which gives HILT its name gives a misleading perspective on the problem tackled by the project. There is a thesaurus-like structure to the database of mappings at the core of the terminologies server envisaged by the project. It provides some level of access to broader and narrower terms in the various schemes via the DDC spine. However, it would be inaccurate to describe it as 'high level' (although it does provide access to high level terms from more specific terms, particularly for the purpose of finding collections classified at higher levels of granularity).

#### *DDC auto-classification*

JISC was keen that HILT consider the possible value of DDC auto-classification in the context of a terminologies server. The team has given this matter some consideration but has concluded that this matter is out of scope for the project which was asked to focus on collection level requirements. It is unlikely that it would be either necessary or helpful to utilise this approach to classifying collections at collection level in order to make them findable via the HILT DDC-based collections finder. The number of collections is small and the effect on costs would be low. Moreover, it is almost certainly the case that manual classification would give better results. Use of the method to help classify items in JISC collections is a more likely approach. However, it is not clear how a terminologies server might contribute to the process and so difficult to assign any benefits arising from it to the terminologies server. It is probably true that its use to (say) classify all RDN collections by DDC would have a beneficial effect on interoperability within JISC, and that the provision of DDC indices in the hubs could enhance the terminologies server find collections facility by providing an automated indicator of collection strength in particular subjects<sup>21</sup>. However, it would not resolve the main problems tackled by HILT unless there was also a terminologies server carrying out the functions described in this document.

#### **Subject Schemes Coverage**

##### *Add Term Sets Not Included in Pilot*

Specifically:

- A UK oriented modifications registry terms set, important because it will provide a means of mapping terms not in standard schemes but used by UK users to appropriate standard scheme terms and should also help resolve legacy metadata problems. Regional variations on a core UK non-standard terms set are also a potential requirement.
- AAT, likely to be most popular in the Museums community, and important if working with that community is, or becomes, important to JISC.

##### *Support for Adding Additional Schemes*

Recognising that JISC cannot support all mappings on its own, two additional elements of server design are proposed:

- A facility to allow interaction with other terminology services providing similar mappings
- A facility to allow self-provisioned groups working with or within JISC to add and map their own terminologies

Both facilities should also be used to integrate the approach with other standard term sets such as LC name authority files (<http://>), and the planned EDINA geoXwalk digital gazetteer shared service (<http://www.geoXwalk.ac.uk/about.htm>).

**N.B.** A clear implication here is that JISC should aim to work closely across communities and Sectoral, domain, and national boundaries with other key players in this area [See Section 7, recommendation 8 for a possible list]. The approach to interoperability proposed here will diminish in value and effectiveness to the extent that it is out of harmony with approaches taken by

<sup>21</sup> See SCONE Final Report. Appendix A.4 at <http://scone.strath.ac.uk/FinalReport/SCONEFPNXA4.pdf>

other key players. This should include the Semantic Web community as well as more 'mainstream' terminology players. The semantic web vision clearly requires mechanisms for mapping between term sets (and presumably relationship sets). HILT aims to provide a subject structure rich in both entry term vocabulary and term relationships through mapping of various terminologies to DDC. This tool will provide a basis for understanding users' needs, the terminology they use and the ways in which their terminology can be mapped to subject schemes. HILT will have the potential to operate in an environment such as the Semantic Grid where a wide range of users interested in various areas, applications and tools in science areas require a consistent subject access to allow them to interoperate efficiently and in an effective manner.

### **Other Issues: Identifying Collections, Clustering, RDN, User Interface**

#### *Identifying Relevant Collections*

At its simplest, the process proposed to map user queries to collections, maps a user term like 'teeth' to DDC, truncates DDC to find that a higher level number covers Dentistry, finds a collection classified in a collections database as covering dentistry, and then checks that the user's more specific topic finds hits. The process can be improved by providing collection strength data for services with more general subject coverage, a process that be controlled, as suggested by the RSLP project SCONE<sup>22</sup>, by the informed professional judgment of JISC services and collection development staff linked to peer review and knowledge of user needs.

This implies that any follow up project work with subject and subject description experts in the JISC community to ensure the best approach to designing this mechanism and the subject description metadata it relies on.

A mechanism to map JISC users subject queries to JISC collections would help ensure that users get full value from the collections that JISC buys for the community by optimising their use<sup>23</sup>.

#### *Possible Clustering-based Enhancements*

HILT and HILT groups have recognised that the clustering facilities developed by the Cheshire Project may be one tool that a terminologies server could provide to assist users in searching at item level in collections where the local scheme is not yet mapped by HILT or where there are significant legacy metadata problems. However, the project had insufficient time and resource to investigate clustering in a way that would permit us to give information on whether or not it was of value in these specific circumstances (note that the project was not funded to do this). The same was true of other such 'data mining' techniques and of approaches taken by services like Google, and initiatives like RedLightGreen, all of which might (or might not) provide useful tools that an operational server might offer users. In respect of these, the project asks that JISC consider providing any follow up to HILT II with sufficient funds to fully investigate the possibilities of this type of approach in parallel with the development of the core terminology server facilities required to halt and reverse the decline in interoperability due to existing subject description practices. Appendix F details the current (limited) state of HILT research and analysis as regards this area. HILT cannot make recommendations on the value or otherwise of such techniques without further research into their precise effects on appropriate retrieval in respect of the wide range of subject-related tasks, and mix of services, subject schemes, and descriptive practices, likely to be encountered in the JISC Information Environment (Appendices C.2, C.4, and D.3 for data on these).

#### *RDN Subject Interoperability Issues*

The RDN has a number of problems in the area of subject description interoperability and is seeking a means of resolving them (see appendix E). Any follow up project should work with the RDNC with a view to determining how these problems can best be resolved in the context of the development of a terminologies server for the JISC IE. Since there is a need to investigate the

<sup>22</sup> See SCONE Final Report at <http://scone.strath.ac.uk/FinalReport/fpindex.cfm>

<sup>23</sup> See CERLIM work showing limited usage of JISC collections by JISC users at [JISC IE Joint Programme Meeting - Formative Evaluation of 5/99: The EDNER Project](#)

potential of the Cheshire and similar approaches as one terminologies server tool, and the RDN databases appear to provide an excellent testbed for this, the possible value of the approach to the RDN and to the JISC terminologies server could be investigated at the same time.

### *User Interface Considerations*

HILT Phase II was able to carry out a small scale user survey [See Appendices C.3 and C.4] to inform development of the pilot. It also: (1) designed a 'first pass' user interface providing term input, disambiguation, collection identification, hits testing, and (minimal) help facility (2) Obtained (at a workshop of 41 users) useful user feedback on the merits and demerits of the interface, real queries faced by users, the effects of training, and user thought processes in subject searching situations – information that will help inform the development of the interface (see Appendix D.3).

It is the view of the project team, however, that further investigation of user subject searching requirements is needed to inform the ongoing development of the 'user interface' to the server. This applies whether or not there is to be a single central user interface available as a web service<sup>24</sup> or a series of portal based interfaces supported by M2M protocols. Unless the design of all such interfaces – and, hence, any M2M facilities that underpin them – is based on sound knowledge of user requirements, their value, and the value of the service itself, will be impaired. The pilot interface was relatively basic and a more sophisticated development will be required in the context of an operational service. An example is the disambiguation facility which, in the pilot, can only cope with one user choice when, in reality, a user query will often be more complex than that. Ongoing work with users is required to ensure that the interface – and other server features – develop in line with the needs of real users.

The detailed design of the proposed investigation requires further discussion in the lead up to any follow up project. However, it should cover at minimum:

- A wide variety of users (lecturers, researchers, students, intermediaries in a wide variety of representative institutions and with varying levels of experience)
- An examination and categorisation of the types of subject queries that arise
- The possible need for a task-oriented interface for subject and other queries
- Implications for user profiling and landscaping
- The possible value of front ends that aim to 'stimulate' user thought as regards term selection (needs example)
- The effects of training, a common subject searching environment, and a knowledge of retrieval languages and skills, to carrying out effective subject searching<sup>25</sup>
- The problems of faced by RDN users in respect of subject searching and the possible role of a terminologies server incorporating a clustering facility in resolving them

### **Note on M2M**

The project was not asked to investigate M2M issues in any practical sense and did not have the resources to do so. It was asked to produce a Machine to Machine requirements report, a task delegated to UKOLN. This report is included as Appendix J.

### **Concluding Remarks**

These additional requirements, including the remainder of the elements from the interim specification, are combined in Appendix I.2, *The Development Requirement for an Operational Terminologies Server*. This requirement was also fed into the cost-benefit analysis process described in the next section of this report.

---

<sup>24</sup> Advantages here are a common user interface for queries and avoidance of duplication of development effort

<sup>25</sup> There is a case for arguing that in the Information Age we need to consider retrieval languages and skills as the 4<sup>th</sup> 'R'

## 6. The Cost-Benefit Analysis

The process of planning and conducting the cost-benefit analysis was managed through Methodologies Document Sections 4 and 5 as described under headings (4.1) to (5.3) below. It had as its focus the set of development requirements specified in Appendix I.2. Its aim was to measure the costs and benefits of different functionality levels and methods of instantiation for an operational server and so, inform the conclusions and recommendations presented below in Section 7.

### **Identify appropriate set of alternative approaches to assess. (4.1)**

### **Refine approach as understanding of requirement develops (4.2)**

This was done in conjunction with the Steering Group, it having been agreed that the Steering Group would themselves conduct the cost-benefit analysis. The list went through various changes, of which only the last two are relevant here. At the Steering Group meeting prior to the one at which the cost-benefit analysis was to be held, the group determined that the options to be examined were:

- A version where there is no central terminologies server and every JISC collection instantiates the functionality locally
- 'Home grown' development of server by in-house programming, or a variation of this based on WORDMAP
- A full commercially developed alternative to this
- A version based on development by OCLC

During the process of refining the methodology used for the cost-benefit analysis that took place after this meeting, the HILT team determined that a two level process was required – the first based on functionality levels, variations in which had most effect on costs and benefits, the second based on the instantiation methods listed above. This approach was accepted by the Steering Group and carried out as described below and in Appendices H.1 and H.2.

### **Agree on cost-benefit analysis method (5.1)**

It was agreed with the Steering Group and the Project Management Group that the cost-benefit analysis methodology developed by the JISC-funded INSIGHT project should be adapted for HILT purposes. Details of this are provided in Appendix H.

### **Examine the agreed cost-benefit analysis method in-depth to determine how best to apply it for HILT purposes. (5.2)**

The team adapted the process in conjunction with the Steering Group. Appendix H shows the final approach agreed and also includes some documented detail of the discussions that led to the final approach.

### **Conduct the cost-benefit analysis process (5.3)**

The Steering group of 18.9.03 conducted the cost-benefit analysis using the methods described in Appendix H (see, in particular, the Framework and Notes document), except that it did not have time to conduct the secondary process described under 'Experiment 2'. This was subsequently conducted to a limited extent by the HILT team with a view to determining the effects of adding regional term sets and MeSH to the two highest rated functionality grouping options.

#### *HILT Team View of the Cost-Benefit Analysis of Instantiation Methods*

This view was in the cost-benefit analysis documents presented to the Steering Group, and was referred to by the team, but was not otherwise discussed. In essence it is this:

As indicated above, there are four options to consider:

- A version where there is no central terminologies server and every JISC collection instantiates the functionality locally
- 'Home grown' development of server by in-house programming, or a variation of this based on WORDMAP
- A full commercially developed alternative to this
- A version based on development by OCLC

Of these, the first is arguably ruled out at the start on two counts:

- The absence of a central mechanism to support an ongoing process that will ultimately lead to interoperability means that key – arguably essential – benefits are not available through this route. The creation of a single UK non-standard terms set with mechanisms to support ongoing co-ordination is not possible without a central process and mappings to standard schemes could not be standardised either
- Since the service development and mappings and training and other elements that contribute to the cost of the enterprise would be duplicated across many JISC services on this model, the cost must turn out to be much higher than any of the other instantiation options

In short, it is safe to say that this first option would cost more than any of the other three and would fail to provide benefits that are key to the interoperability issue.

Comparing the remaining options is difficult in the present circumstances and has not been attempted here for two reasons. In the view of the HILT team:

- Comparative costings not based on a real tendering or bidding process are likely to be highly dubious and to yield questionable results that might well be overturned in a real bidding or tendering process (especially since benefits in each case are likely to be largely similar)
- There are good grounds for supposing that the ideal approach to building a terminologies server for JISC would be one that combined the strengths of all three approaches – for example, one that involved the various parts of the HILT team, OCLC, and a commercial developer like Wordmap

These points were in the documents presented to the Steering group but were not specifically discussed by them. The conclusions stated have been assumed to be correct in the conclusions and recommendations presented in Section 7 below.

#### *Results of the Cost-Benefit Analysis of Functionality Levels*

A cost-benefit analysis of functionality levels based on the INSIGHT model was conducted at the steering group meeting of 18<sup>th</sup> of September, 2003. The following steps were taken to carry out the cost-benefit analysis:

1. Identification of costs
2. Identification of benefits and their relationship to strategic objectives
3. Evaluation of benefits of various functionality levels
4. Conducting INSIGHT cost-benefit analysis (calculation of cost-benefit ratios)

As table a below shows, option C emerged as the most favoured option in terms of the cost-benefit ratio. This option entails:

- The creation of the basic interoperability process; staff services to support creation of UK modifications terms set, mapping to DDC, LCSH, UNESCO; Direct and M2M user advice on terms in these schemes; staff and user training
- Direct and M2M disambiguation, collection finder, sample hits and collection ranking, user term monitoring, training

The cost of this option has been calculated at £926,096 over 5 years, which is only more expensive than one other option, option B (a less developed system).

Option G, ranked second in terms of cost-benefit ratio, adds the option of regional scheme modifications. This results in an additional £130,000 onto the cost making it the fourth most expensive option. However, the benefit score for this option is second highest which means that option G emerges favourably when the cost-benefit ratio is calculated.

Table a

Option	Mix	Description	Five Year Cost	Benefits Score	Cost-benefit ratio <sup>26</sup>	Ranking
<b>A</b>	<b>A</b>	Do nothing option				
<b>B</b>	<b>1</b>	Basic interoperability process created; staff services to support creation of UK modifications terms set, mapping to DDC, LCSH, UNESCO; Direct and M2M user advice on terms in these schemes; staff and user training	£881,951	487	552	6
<b>C</b>	<b>1+2</b>	Option B plus direct and M2M disambiguation, collection finder, sample hits and collection ranking, user term monitoring, training	£926,096	742	801	1
<b>D</b>	<b>1+3</b>	Option B extended to AAT and MESH but without option C	£1,481,448	640	043	7
<b>E</b>	<b>1+4</b>	Option B extended to regional variations to the UK modifications terms set, but without option C or AAT and MESH	£1,021,906	592	058	5
<b>F</b>	<b>1+2+3</b>	All 5 schemes, UK modifications terms set without regional variations, but with disambiguation, collection finder etc	£1,525,593	895	587	4
<b>G</b>	<b>1+2+4</b>	DDC, LCSH, UNESCO, UK modifications terms set with regional variations, plus disambiguation and related services, but no AAT or MESH	£1,065,241	847	795	2

<sup>26</sup> For the sake of simplicity, the ratios have been multiplied by 1,000,000 and rounded up or down as appropriate

<b>H</b>	<b>1+2+3+4</b>	Everything: all 5 schemes; UK and regional term sets, disambiguation and related services	£1,664,738	1000	601	3
----------	----------------	---	------------	------	-----	---

It was of interest to note how the addition of MeSH, a specialist thesaurus would affect the cost-benefit ratios of the first two highest ranked options. Thus, additional options I and J (see table b below).were considered by the HILT team by conducting a selective version of ‘Experiment 2’ (see Appendix H for details).

The addition of MeSH to C and G lowers their cost-benefit ratios, but still leaves the resulting options I and J ranked higher than all other options (other than C and G themselves), suggesting that the addition of MeSH to the equation may also be worth considering under certain conditions.

**Table b: Cost-benefit analysis ratios for options I and J**

<b>Option</b>	<b>Mix</b>	<b>Description</b>	<b>Five Year Cost</b>	<b>Benefits Score</b>	<b>Cost-benefit ratio</b>	<b>Ranking</b>
<b>I</b>	<b>C+ MeSH</b>	1+2+MeSH	£1,013,988	753	743	3
<b>J</b>	<b>G+ MeSH</b>	1+2+4+MeSH	£1,153,133	855	741	4

Having considered these final permutations the overall conclusion is that option C is the most highly ranked ratio and therefore the most favoured option. Option G is the second most favoured option as it is the next most highly ranked, suggesting that the addition of regional terms to the equation may be worth considering in certain conditions. The addition of MeSH to C and G lowers their cost-benefit ratios, but still leaves the resulting options I and J ranked higher than all other options (other than C and G themselves), suggesting that the addition of MeSH to the equation may also be worth considering under certain conditions.

#### *Note on M2M Costings*

Since M2M versions of functions are a requirement of shared services, and an understanding of direct user facilities requirements is needed to design M2M versions these two elements were considered as single cost elements in the cost-benefit analysis of functionality levels and instantiation methods carried out by the project.

#### *Concluding Remarks*

The results of the cost-benefit analysis of functionality levels suggest that option C would be the best basis for a future development project, although option G (C plus regional term set mappings) is a close second that might find favour with potential funding partners such as RE: SOURCE and SLIC. The addition of MeSH to options C and G lowered their scores but still left them well above other options. Adding MeSH may also attract additional funding partners and has the added attraction of bringing a specialist thesaurus from a specific subject area into the proposed operational server.

Although option B (the baseline mapping option) scored much lower than option C (which includes B), the HILT team regard it as the core of the interoperability process and there is a case for scoring it higher. However, was not rated in this way by the Steering Group.

Option A (the ‘do nothing’ option) was taken out of the process because it caused practical difficulties with the assessment procedures. Instead, it was agreed that the project should note that

‘doing nothing’ was a possible option for JISC to consider. The HILT team do not believe it is a sensible option, and it was an option strongly rejected by the HILT Phase I Stakeholder Workshop on the issue<sup>27</sup>.

A detailed report of the process utilised in the cost-benefit analysis, including the use of the methodology developed by the JISC-funded INSIGHT project, the mapping of benefits to relevant elements of the JISC Strategy, lists of benefits, benefit elements, and cost elements, the involvement of the members of the HILT Steering Group in the cost-benefit analysis process, and the results from the process, is provided as Appendix H below.

---

<sup>27</sup> See HILT Phase I User Workshop Report, Conclusions section, at <http://hilt.cdlr.strath.ac.uk/Dissemination/WorkshopNew.html#Conclusion>

## 7. Conclusions and Recommendations

Having taken all of the above considerations into account, the project recommends:

1. That JISC fund a development project to build a terminologies service for the JISC Information Environment and base it, at minimum, on the functionality and research work encompassed within option C from the cost-benefit analysis (see Section 6 and Appendix H):

1	DDC spine and term sets
2	LCSH mapping
3	UNESCO mapping
4	UK oriented modifications registry terms set creation
5	UK oriented modifications registry terms mapping
6	RDN terminologies harmonisation study
7	RDN-based clustering tool study
8	Interface needs user study (enhanced pilot with clustering)
9	Term match facility
10	Staff amend maps facility
11	Staff training module
12	Online user training module
13	Ability to host and map other schemes
14	Ability to interact with other mapping services
15	Processes to cope with scheme updates
16	Disambiguation facility
17	DDC collection identifier
18	Any hits test/rank facility
19	User terms monitor

The software functions listed in the above are taken to include M2M capability. In respect of the latter, it is proposed that the additional recommendations specified in the UKOLN report on M2M functionality be followed. These are specified in Appendix J of this Report.

The cost-benefit analysis figures suggest the cost will be £926,096 over a five-year period, including project management, training, publicity, marketing, and redevelopment costs. However, costs may be revised in the light of detailed discussions with JISC should these recommendations be accepted.

2. That it also consider whether there is value in adding UK regional scheme modification term sets and MeSH into the features list (option G and option C or G plus MeSH respectively). The cost-benefit analysis figures suggest the additional cost of both will be £1,153,133 over a five-year period.
3. That it take a phased approach to the implementation, spreading the cost of development, and of the additional research still required to inform aspects of service design, over 5 years in the first instance.
4. That it build in a regular review process that will permit, where necessary, the refocusing of aspects of the design to take account of changing circumstances, new research data, novel techniques and technologies, and other pertinent factors as they arise.
5. That the initial phase last two years and entail terminologies server development and other research specified in elements 1-15 in the table above, conducting 6-8 in conjunction with users and using the results to inform development beyond the initial two years (this implies further development of 16-19 as pilot elements in the first two years, followed by full development later).
6. That JISC build on the experience and relationships built up in HILT Phase II in any follow up project and involve the HILT team, the supplier of the Wordmap software, OCLC, and the various HILT stakeholders, but that they liaise with the team to determine how best to

strengthen the approach taken by bringing in expertise from data mining and semantic web communities and professional expertise from other areas thought relevant (Input from internet search engine services from Google might be one example).

The main participants in HILT Phase II were:

- The Centre for Digital Library Research (CDLR) at Strathclyde University
- JISC representative
- mda (formerly the Museums Documentation Association);
- National Council on Archives (NCA);
- National Grid for Learning (NGfL) Scotland;
- Online Computer Library Center (OCLC);
- RDN representative
- FE Representative (Regional Centre)
- Scottish Library and Information Council (SLIC);
- Scottish University for Industry (Sufl);
- UK Office for Library and Information Networking (UKOLN).
- Terminology experts, Alan Gilchrist and Leonard Will (external evaluator)

There was also involvement from, NLS, BL, and Wordmap.

7. That JISC ensure that any follow up project takes account of the potential value of a mapping service of this kind to semantic web and semantic grid developments when considering the instantiation of design elements.
8. That JISC work to begin a dialogue with key national and international players on how best to ensure cross-sectoral, cross-domain, multi-lingual, and international compatibility of the JISC terminologies server with other such developments – these to include OCLC and Library of Congress, other terminology scheme developers, RLN/RSLG, National Archives Network Consortium, mda, UK National Libraries, European and other National Libraries, UK players from other sectors (RE:SOURCE, SLIC, players from Museums and Archives), W3C, a representative from the RENARDUS project. It should also aim to include all communities working in or with JISC – HE and FE, e-learning and research, the semantic grid community, and so on.
9. That JISC consider funding an independent supporting study to explore, in conjunction with JISC itself, the best option for ensuring the long-term financial future of a terminology server and of other such shared services