



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

I.2	Operational Terminologies Server Specification; Pilot Description
J	Delivering HILT as a JISC IE shared service (M2M report)
K	Evaluator's Report
	Glossary

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 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 *optimize the ability of users to carry out successful subject searches² by providing a process that would, in time, permit JISC and JISC services to:*

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

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 Renardus³, 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 harmonised⁴ 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*⁵ and SLIC⁶.

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 mechanisms⁷ 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 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.

² 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

³ See <http://www.renardus.org/>

⁴ 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.

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

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

⁷ An example is the clustering approach pioneered by the CHESHIRE project – see Appendix F

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.

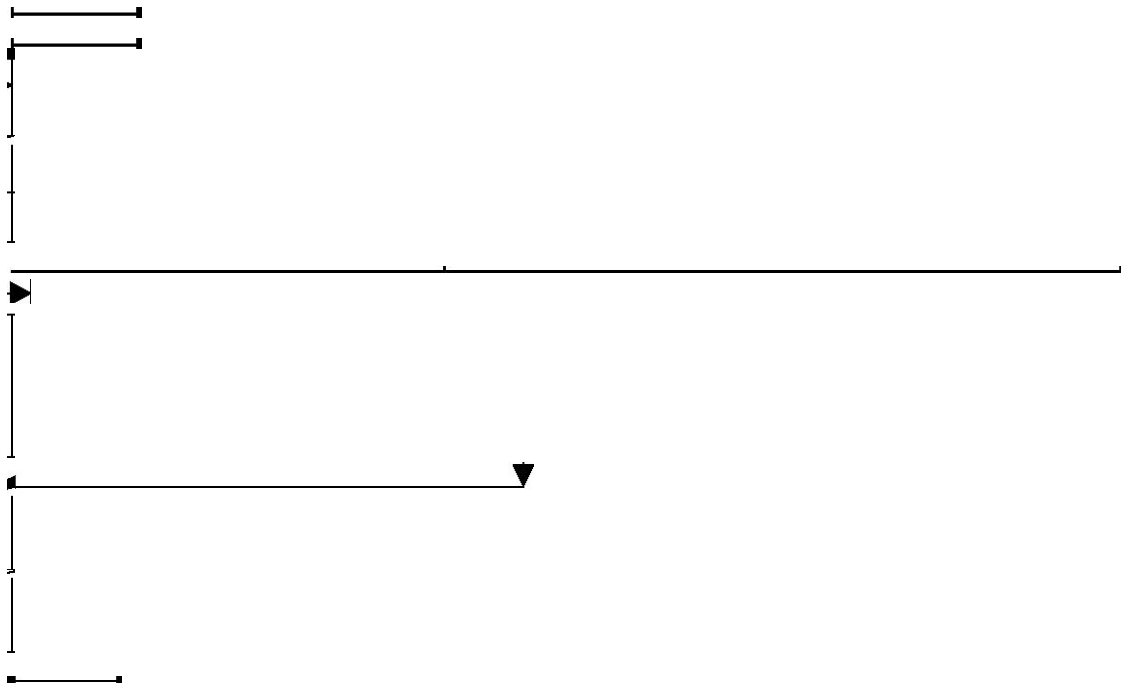
2. That it also consider whether there is value in adding UK regional scheme modification term sets and MeSH into the features list. The cost-benefit analysis figures suggest the additional cost of both will be £1,153,133 over a five-year period.
3. That it take a phased approach to the implementation, 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

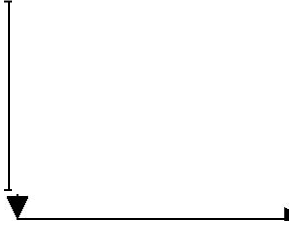
example)⁸ .

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

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

Overview of Project Processes, Including Dependencies Landscape





**HIL
T I
data
and
mod
el
and
HIL
T II
aims**

**Met
hodo
logie
s for
resea
rch
that
will
infor
m:**

- 1. Amendment and/or refinement of model**
- 2. A cost-benefit analysis of agreed design options**

**Research
on the
model
and its
cost-
benefit
analysis.
Includes
consultati
ons with
project
groups
via**

progress reports, demonstrations and related discussions. Feeds into pilot & specification and back into methodologies. Also includes cost-benefit analysis itself.

Initial model for terminology server

Continuously refined model of Terminologies server

Specification

**for
full
term
inol
ogie
s
serv
er**

**Con
clusi
ons,
Rec
om
men
dati
ons
Fina
l
Rep
ort**

**Contru
ction of
pilot
JISC
IE
termin
ologies
server**

**Project
outputs
are:
Pilot
termin
ologies
server;
Report
on
Cost-
benefit
analysis;
Specifi
cation**

for
operational
terminologies
server;
**Final
Report**
with
appendices on
methodologies
document,
surveys
,
workshop
etc.

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 parties⁹. 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 approach¹⁰ 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 model¹¹, 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 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.

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

¹⁰ See Appendix F

¹¹ See JISC Information Environment Architecture. Andy Powell & Liz Lyon, UKOLN, University of Bath <<http://www.ukoln.ac.uk/distributed-systems/dner/arch/>>

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 searches¹² by providing a process that will, in time, permit JISC and JISC services to:
 1. Achieve and maintain as 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

¹² 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.

¹³
These were 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.

¹³ Full details of the findings can be found at <http://hilt.cdli.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 I 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 software¹⁴ 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 project¹⁵ – 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 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).

¹⁴ See <http://www.wordmap.com/>

¹⁵ 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.

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 HASET 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 users¹⁶ 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.



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.cdli.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 .

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:

¹⁷

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

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 possibility¹⁸ 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.

¹⁸

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.cdlr.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 IESR¹⁹ 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 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.

¹⁹ Information Environment Services Registry – see project website at (add URL)

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 variations²⁰. 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.

²⁰

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://hilt.pilot.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

Address http://hilt.pilot.cdlr.strath.ac.uk/pilot/top.php

Search Sign In News Games Personal Mail My Yahoo! Yahoo!

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 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 identify 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 lists (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

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](#), [Landscape & area planning](#), [Music](#) ...
- [Computers, information & general reference](#) — [Associations, organizations & museums](#), [Bibliographies](#), [Computers, Internet & systems](#), [Encyclopedias](#), [Journalism, publishing & new media](#), [Library & information science](#) ...
- [History & geography](#) — [Biography & genealogy](#), [Geography & travel](#), [History](#), [History of Africa](#), [History of Asia](#), [History of](#) ...
- [Language](#) — [Classical & modern Greek languages](#), [English & Old English languages](#), [French & related languages](#), [German](#), [Italian](#), [Romanian & related languages](#), [Languages](#) ...
- [Literature](#) — [American literature in English](#), [Classical & modern Greek literatures](#), [English & Old English literatures](#), [Literatures](#), [German & related literatures](#), [Italian](#), [Romanian & related literatures](#) ...
- [Philosophy & psychology](#) — [Ancient, medieval & eastern philosophy](#), [Astrology, parapsychology & the occult](#), [Epistemology](#), [Ethics](#) ...
- [Religion](#) — [Christian denominations](#), [Christian pastoral practice & religious orders](#), [Christian practice & observance](#) ...

Figure 2. Disambiguation page of the HILT Pilot Terminologies Service

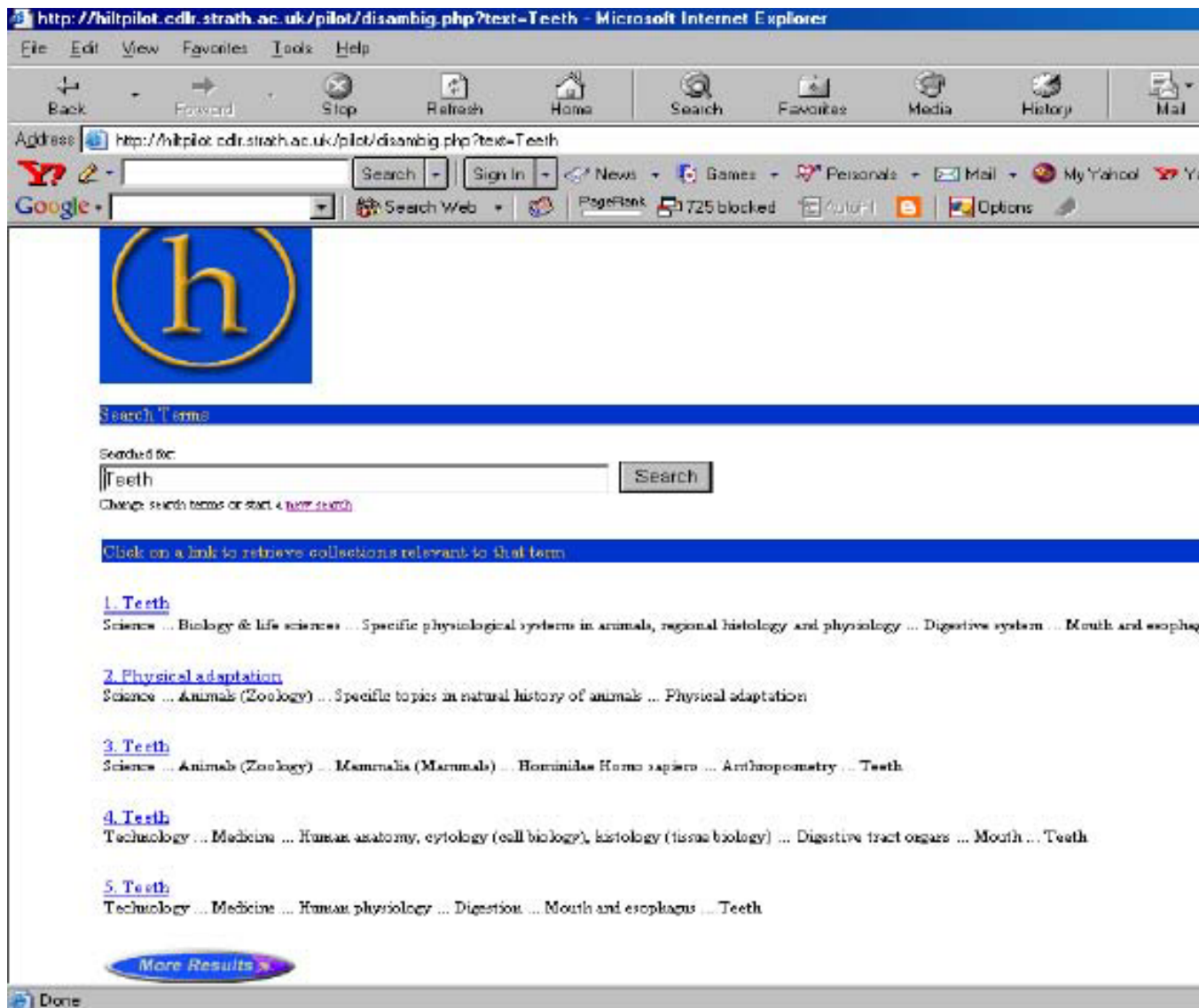


Figure 3. Collection selection page of the HILT Pilot Terminologies Service

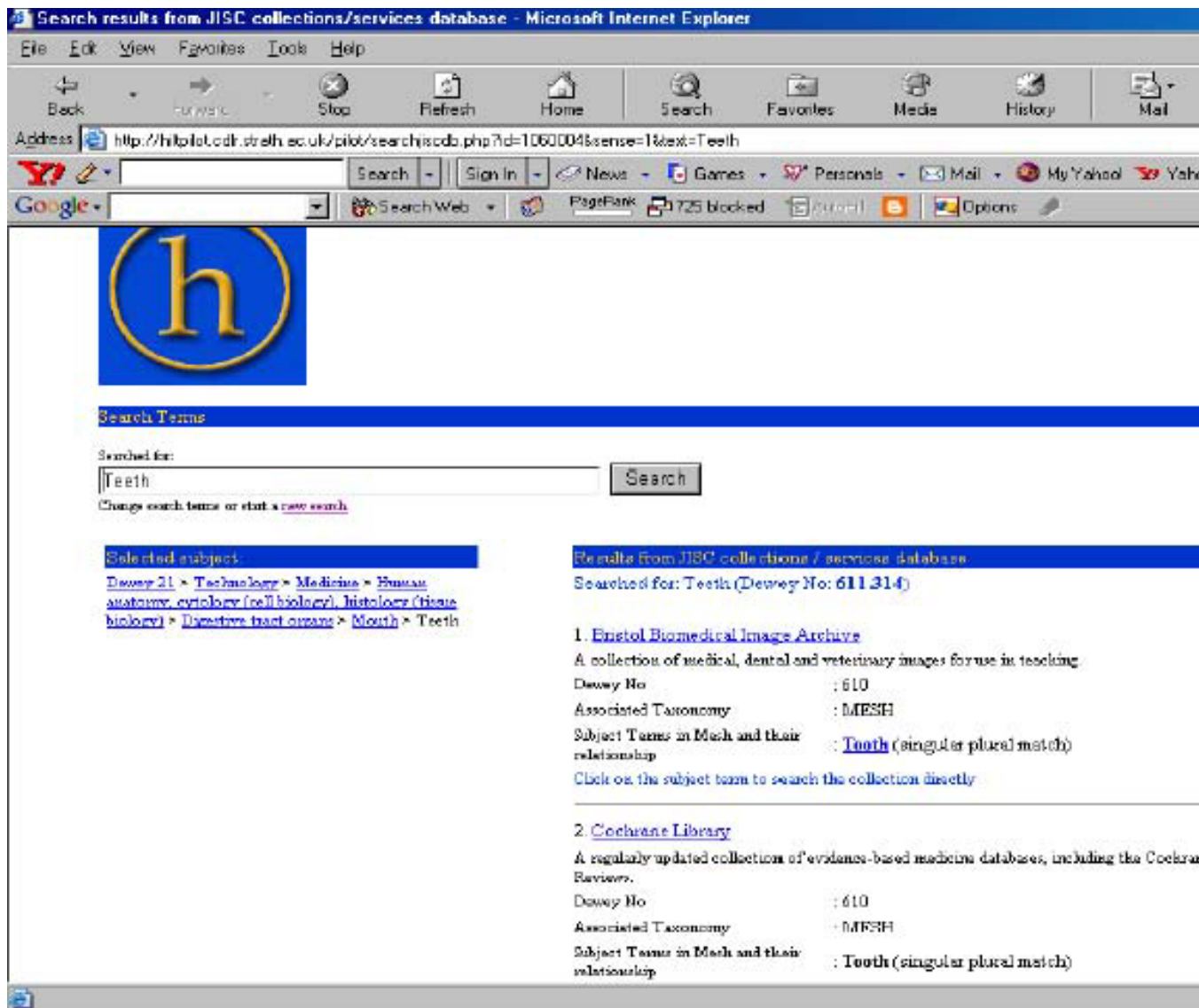


Figure 4. JISC collection found by the search term “Teeth”

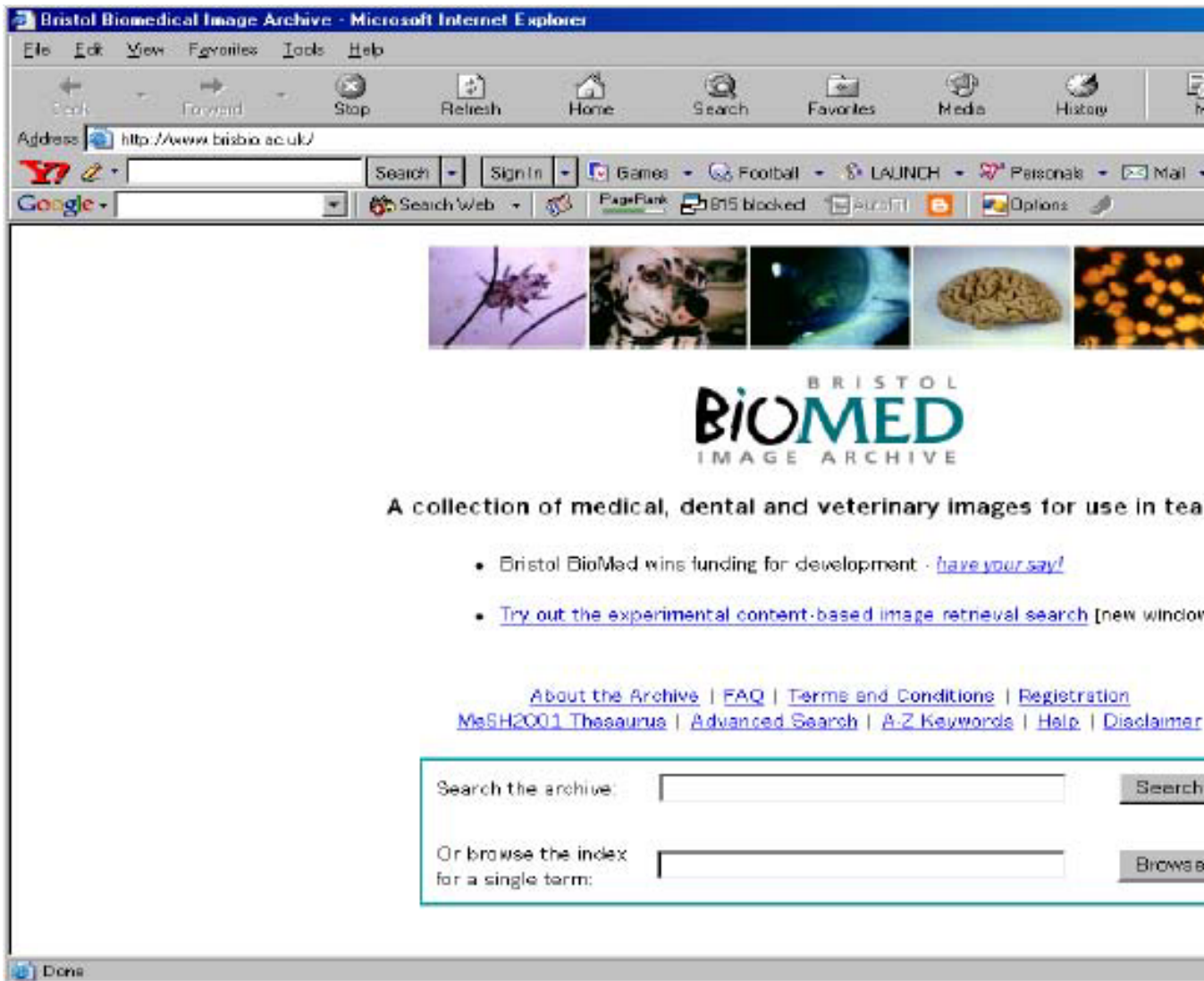


Figure 5. HILT pilot terminologies staff interface

Wordmap Taxonomy Management System: hilt

File Edit View Help

Master

- 📁 Dewey 21
 - 📁 Arts & recreation
 - 📁 Architecture
 - 📁 Arts
 - 📁 Dictionaries, encyclopedias, concordances of fine and decorative arts
 - 📁 Education, research, related topics of fine and decorative arts
 - 📁 Temporary and traveling collections and exhibits
 - 📁 Galleries, museums, private collections of fine and decorative arts
 - 📁 Historical, geographic, persons treatment of fine and decorative arts
 - 📁 Miscellany of fine and decorative arts
 - 📁 Organizations and management of fine and decorative arts
 - 📁 Serial publications of fine and decorative arts
 - 📁 Special topics in fine and decorative arts - History and description with reference
 - 📁 **Standard subdivisions of fine and decorative arts**
 - 📁 Appreciative aspects
 - 📁 Inherent features
 - 📁 Methodology
 - 📁 Special topics
 - 📁 The arts - Fine and decorative arts
 - 📁 Drawing & decorative arts
 - 📁 Graphic arts
 - 📁 Landscaping & area planning
 - 📁 Music
 - 📁 Painting
 - 📁 Photography
 - 📁 Sculpture, ceramics & metalwork
 - 📁 Sports, games & entertainment
 - 📁 Computers, information & general reference
 - 📁 History & geography
 - 📁 Language
 - 📁 Literature
 - 📁 Philosophy & psychology
 - 📁 Religion
 - 📁 Science
 - 📁 Social sciences
 - 📁 Table 1. Standard Subdivisions

Master

Saballites

Source Dewey 21-9UN (1 May 2003)

- 📁 Dewey 21
 - 📁 Arts & recreation
 - 📁 Computers, information & general reference
 - 📁 Associations, organizations & museums
 - 📁 Bibliographies
 - 📁 Computers, Internet & systems
 - 📁 Encyclopedias & books of facts
 - 📁 Journalism, publishing & news media
 - 📁 Library & information science
 - 📁 Administration of the physical plant
 - 📁 General libraries, archives, information centers
 - 📁 Library and information science
 - 📁 Education, research, related topics
 - 📁 Historical, geographic, persons
 - 📁 International organizations
 - 📁 Management
 - 📁 National, state, provincial, local
 - 📁 Permanent nongovernmental organizations
 - 📁 Operations of libraries, archives, information centers
 - 📁 Personnel administration
 - 📁 Reading and use of other information resources
 - 📁 Relationships of libraries, archives, information centers
 - 📁 Specific kinds of institutions
 - 📁 [Unassigned]
 - 📁 [Unassigned]
 - 📁 Magazines, journals & serials
 - 📁 Manuscripts & rare books
 - 📁 Quotations
 - 📁 [Unassigned]
 - 📁 History & geography
 - 📁 Language
 - 📁 Literature
 - 📁 Philosophy & psychology
 - 📁 Religion
 - 📁 Science
 - 📁 Social sciences

Ready

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 subjects²¹. 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 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.

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

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 SCONE²², 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 use²³.

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 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.

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

²³ 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](#)

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. This²⁴ applies whether or not there is to be a single central user interface available as a web service 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 retrieval²⁵ languages and skills, to carrying out effective subject searching
- 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.

²⁴

Advantages here are a common user interface for queries and avoidance of duplication of development effort

²⁵

There is a case for arguing that in the Information Age we need to consider retrieval languages and skills as the 4th '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 18th 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-benefit²⁶ ratio	Ranking
A	A	Do nothing option				
B	1	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

C	1+2	Option B plus direct and M2M disambiguation, collection finder, sample hits and collection ranking, user term monitoring, training	£926,096	742	801	1
D	1+3	Option B extended to AAT and MESH but without option C	£1,481,448	640	043	7
E	1+4	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
F	1+2+3	All 5 schemes, UK modifications terms set without regional variations, but with disambiguation, collection finder etc	£1,525,593	895	587	4
G	1+2+4	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
H	1+2+3+4	Everything: all 5 schemes; UK and regional term sets, disambiguation and related services	£1,664,738	1000	601	3

²⁶

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

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

Option	Mix	Description	Five Year Cost	Benefits Score	Cost-benefit ratio	Ranking
---------------	------------	--------------------	-----------------------	-----------------------	---------------------------	----------------

I	C+ MeSH	1+2+MeSH	£1,013,988	753	743	3
J	G+ MeSH	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 ²⁷.

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.

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

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 (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.

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

Glasgow : Cente for Digital Library Research, 2004

Appendix A: HILT Phase II: Methodologies Document

Document History	Date	Comments
Version 1.0	05.08.02	Early draft, compiled by DN
Version 2.0	07.08.02	Early draft, compiled by DN, first post-SG amendments
Version 3.0	21.08.02	Further rough detail added by DN
Version 4.0	04.09.02	Pre-User Workshop draft for discussion
Version 5.0	26.11.02	Surveys, interviews replace workshop 1; add Wordmap section(3)
Version 6.0	24.06.03	Upgrade 1-4 in line with recent thinking; merge 5-7; add chart
Version 7.0	01.09.03	Cost-benefit analysis Amendments; progress notes; pilot detail

Work still required:

None

Purpose of this Document:

- To set out the methodologies employed in HILT Phase II
- To set out, as part of this process, the project and pilot evaluation and quality assurance and review methodology employed by the project
- To show dependencies, order of progression

Project Steps and Associated Methodologies (Overview and Contents)

Section		Page:
0	Project and pilot evaluation and quality assurance and review methodology	2
	Literature Search	3
1	Methodologies to ensure investigation examines representative services, subject schemes, and subjects within schemes	4
2	Methodologies to ensure investigation examines representative user types, tasks, and associated retrieval requirements, and strategies	5

3	Methodologies to ensure (1) that the full functional requirement for an operational (as opposed to pilot) terminologies server is identified (2) That the extent to which this is implemented in the pilot is optimised (3) That the software used in the pilot is utilised in a way that faithfully reflects any specific requirements implemented and tested	6
4	Methodologies to ensure investigation examines terminologies server design options adequately and in a fashion useful to JISC	7
5	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 4	8
Annex A	Initial Service Specification Draft	10

Note: Research is carried out as specified in the Methodologies Document, but also influences the methodologies specified there (see Overview of project processes and dependencies landscape on page 9 of Final Report). In general, the content and form of the methodologies specified in any one section can be influenced by either the consideration or application of methodologies from any other section. In the main, however, the final shape and form of methodologies specified in a later section is more likely to be dependent on outcomes from an earlier section than vice versa.

Project Steps and Associated Methodologies

0 Project and pilot evaluation and quality assurance and review methodology

0.1 Task: Project and Pilot Evaluation methodology

Methodology:

Quality assurance of project products and processes and formative and summative evaluation at project level are ensured through the key roles played in the project by the Project Evaluator (PE), the Professional Level Evaluation Group (PLEG), and the Methodologies Document, as follows:

- The Project team, assisted by other project participants, will set out the methodologies to be employed in meeting project aims and objectives in a Methodologies Document, initially as a draft for discussion, ultimately as a final and agreed statement of intent that will guide project activity
- Successive drafts of this will be critically examined, refined, and ratified by the PE, the PLEG, and others
- Once agreed, the methodologies will be applied by the project team and others and the PE and PLEG will monitor the implementation of the agreed approach, the accuracy of the results recorded, and the validity of subsequent analyses, conclusions, and recommendations produced
- The PE will produce an Evaluator's Report that will be included in the Final Report and will influence final conclusions and recommendations.

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

Note: Literature search

A common thread running through many of the methodologies is the need to conduct searches of the literature on specific topics. To ensure an orderly approach to this, templates will be produced detailing specific questions to be researched, why the data is being sought, and providing 'prompts' for recording the who, what, why, where, when and how of any relevant report found in the literature, a judgement on the reliability of conclusions drawn by the authors of the report, and the implications for HILT Phase II aims, objectives, or outcomes. Questions to be researched include:

- What mapping projects are reported in the recent literature? Are any specific to JISC services?*
- What categories of mapping problem are reported? What specific examples of each type can be found? What are the implications for retrieval?*
- What schemes and practices are in use in JISC projects and initiatives?*
- What is there on methods of testing collection strength?*
- What is there on mapping specialist thesauri to general schemes?*
- What is there on indexing staff departures from standard schemes and the reasons for them?*
- Do particular subject areas in a universal scheme present particular problems?*
- What is there on user studies and terminologies and any of the following: retrieval, interfaces, choosing representative users, user study methodologies, monitoring software, how users express subject queries, user search strategies, the effect of training on retrieval effectiveness, user subject retrieval requirements*
- What is there on choosing queries for testing subject retrieval*
- What is there on different perspectives on what good retrieval is for given queries (different users, intermediaries, lecturers etc)*
- What is there on different approaches to solving the subject-based interoperability problem?*
- What is there on the effectiveness of the CHESHIRE clustering approach?*
- What is there on the benefits and deficiencies of different approaches to solving the subject query interoperability problem - methodologies for measuring these (particularly effectiveness of retrieval)?*
- What is there on the difficulties staff have assigning terms?*
- What is there on how dictionary compilers go about work to identify new words in the electronic age?*
- What is there on the problems of using DDC in particular domains?*
- What is there on costing methodologies in the terminology creation, compilation, mapping area?*
- What is there on expressing costs against benefits and ranking the results in this area?*

Note: We are looking at museums, archives, e-services, not just libraries in all of the above

Note: Information from this exercise informed the forward path of the project in a variety of ways. In particular:

- The ideas and perspectives brought by the project team to internal discussions and discussions with project groups
- The development of the model underlying the pilot and of the pilot itself
- The identification and handling of mapping issues

- The identification and handling of cost-benefit analysis issues

In the event, the idea of using templates to ‘ensure an orderly approach’ was found to be too difficult to implement in any formal sense, although the sketch of a design for these templates detailed above informed the approach taken when reading the literature and assimilating ideas into project perspectives

1 Methodologies to ensure investigation examines representative services, subject schemes, and subjects within schemes as it develops views on HILT model, mapping, functionality and interface features, cost-benefit analysis requirements and so on:

1.1 Task: Ensure examination of a good representative set of service types (IE landscape scope)

Methodology: Identify JISC collections using JISC web site as source but checking what future additions may be in the pipeline with appropriate JISC personnel. Take care to encompass cross-sectoral and cross-domain needs.

This task has been completed via the JISC services survey - see Appendix C.2.

1.2 Task: Ensure examination of representative subject schemes

Methodology: Identify JISC collections from JISC web site and survey them to find out about the schemes they use and their practices and staffing (e.g. trained cataloguers or not?). Use HILT 1 list to add any additional schemes the PLEG and PE think necessary or useful. Obtain any additional information available in the literature.

A large range of the subject schemes associated with JISC collections have been identified and listed. See Appendix C.2.

1.3 Task: Ensure examination covers subject strengths within general collections as well as special or subject collections in providing for users' collection level terminology needs in specific subject areas

Methodology: Identify representative general collections via 1.1 and examine TeRM and TeRM alternative effectiveness in at least two scenarios:

General service always offered to users with subject query
More subject specific approaches based on existing in-depth collection strength data (e.g. from SCONE CURL sites)

This was done with one service of general coverage in the pilot server. Simulated ‘collection strengths’ records were put in the ‘dummy’ collections database utilised to imitate the JISC IESR shared service. A brief account of collection strength assessment techniques has been provided in the literature review.

1.4 Task: Identify means of examining any implication arising from the need to use specialist thesauri such as MeSH

Methodology: Choose an example of a specialist thesauri and consider how best to integrate it with the other mappings. Log difficulties, problems, and possible solutions. Obtain any additional information available in the literature.

Some of the specialist thesauri such as HASSET, AAT, CAB used by JISC collections and services have been identified. A description of the mapping issues of specialist thesauri to classification schemes has been reported in the literature review. A mapping exercise has been carried out using MeSH (see Appendix B.3). No general conclusions have been drawn at this stage.

1.5 Task: Ensure research data is obtained on staff departures from standard terms in standard schemes and reasons for these departures (to help identify specific problem areas and range of problem areas - and, ultimately, to help determine mapping level requirements for TeRM)

Methodology: Use list of JISC collections, identify contact staff, do a mini-survey based on the questionnaire shown in Appendix C.1 (approach agreed with PLEG and others). Obtain any additional information available in the literature.

The survey of JISC collections and services provided information about: subject schemes in use by JISC collections, reasons for departure from standard schemes, examples of alterations and modified subject areas. See report in Appendix C.2.

1.6 Task: Ensure examination of representative subject areas within schemes

Methodology: Investigation via the literature survey, discussion at PLEG, general attempt to take a varied approach.

Both the user survey and user workshop sought to address a wide range of subject areas. This was achieved through assigning search tasks in different subject areas ranging from publishing and literature to architectural preservation, medicine, business, economics and technology.



2 Methodologies to ensure investigation examines requirements of representative user types, tasks, and associated retrieval requirements, and strategies as it develops views on HILT model, mapping, functionality and interface features, cost-benefit analysis requirements and so on:

2.1 Task: Conduct a survey by interviewing a representative group of FE and HE users, covering a range of subject areas, and including (as far as possible): students, PGs, researchers, teachers, supervisors, intermediaries using an approach agreed with PLEG.

Methodology: Compile a questionnaire (as shown in Appendix C.3) designed to investigate (1) what level of retrieval individuals such as students, lecturers, supervisors, librarians and other intermediaries think are required for particular subject related information tasks (2) what search terms and strategies these groups state they should and would use when searching for the information.

This exercise was carried out as described. The results are reported in Appendix C.4. The exercise helped acclimatise project staff to the problems and issues related to dealing with users and subject searching situations in a range of subject areas and also informed the design of a subsequent user workshop designed to evaluate aspects of the pilot terminologies server.

2.2 Task: Conduct a User Workshop a user to obtain feedback on the use of the pilot terminologies server in various real and simulated subject retrieval situations, on a range of issues relevant to its design, and on some of the assumptions that underpinned the design. Agree the approach with PLEG.

Methodology: Conduct workshop designed to give at least 25 users of various kinds online access to the pilot terminologies server and ask them to carry out a range of subject query exercises, some predetermined by HILT, others based on real exam or essay questions set in a range of subject areas. See Appendices D.1 to D.3 for full details of methodology employed and Workshop results and analysis.

The workshop was carried out and provided a wealth of information on user interface issues, subject retrieval behaviour, coverage of terms used by attendees in the HILT database, the effects of learning, and other issues. This information will be of value in building an operational server.



3 Methodologies to ensure:

That the full functional requirement for an operational (as opposed to pilot) terminologies server is identified

That the extent to which this is implemented in the pilot is optimised

That the software used in the pilot is utilised in a way that faithfully reflects any specific requirements implemented and tested

3.1 Task: Agree initial service specification at outline level

Methodology: Utilise HILT 1 recommendation and HILT 2 proposal to sketch out initial specification (attached as Figure A) and get agreement from PMG and SG

3.2 Task: Determine functionality available in pilot software

Methodology: Attend Wordmap training, read documentation, explore use of software

3.3 Task: Begin work on an interim specification by determining an initial list of end user and staff user requirements as regards functionality and on schemes coverage

Methodology: Determine initial list using survey and interviews described under sections 1 and 2 above

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

Methodology: Examine the literature and investigate particular subject areas; list types of problems encountered as regards relationship types between terms; specify possible HILT solutions

3.5 Task: Identify adequate mechanisms in the pilot software and mechanisms or processes from other areas of project work for implementing the requirements identified and implement these in a working pilot terminologies server.

Methodology: Compare results of 3.2 above with the results of 3.1, 3.3 and 3.4. Implement the requirement as far as possible utilising those mechanisms considered adequate, list any remainder as part of a full specification for a future operational server.

3.6 Task: Refine and extend the requirement and the pilot TeRM

Methodology: Hold workshop and conduct various end-user and staff user tests on the stage 2 pilot; extend requirement to encompass these; optimise pilot

3.7 Task: Finalise requirement prior to cost-benefit analysis

Methodology: Hold ‘brainstorming’ session with PMG – particularly the terminology experts – on the functionality required in the full-blown terminologies server and the mechanisms used in the software to implement it; agree final specification prior to cost-benefit analysis; make any appropriate adjustments to the pilot that are practical given project resources (fed into the process described in Section 5 below).

3.8 Task: Finalise requirement once results of cost-benefit analysis are known

Methodology: Adjust requirement to take account of results of cost-benefit analysis

3.9 Task: Determine M2M requirements

Methodology: UKOLN to determine likely M2M requirements of terminologies server in JISC IE



4 Methodologies to ensure investigation examines terminologies server design options adequately and in a fashion useful to JISC:

4.1 Task: Identify appropriate set of alternative approaches to assess and compare given project aims and associated resources

Methodology: Project team to outline proposed alternatives and discuss with SG, PMG, PLEG and JISC, aiming to finalise an agreed list shortly after the first workshop.

Options agreed as a result of 4.1 (various earlier versions existed, this is final one)

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

4.2 Task: Refine approach as understanding of requirement develops

Methodology: Discuss with SG and PMG in the context of early work on the cost-benefit analysis process¹ that will be used to compare design options

Subsequent to the Steering Group meeting that discussed the approach to be taken in the CBA, the HILT team realised that a two-stage process was necessary:

- INSIGHT CBA of different levels of terminology server functionality, each associated with different cost and benefit levels (does it have a DDC to LCSH mapping, does it have a disambiguation function, and so on)
- Subsequent analysis of different approaches to the instantiation of the terminology server (as listed in 4.1 above)

This approach was accepted by the Steering Group and applied as described in Appendix H and in Section 6 of the Final Report



5 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 4:

5.1 Task: Agree on cost-benefit analysis method suitable for the purposes of HILT

Methodology: Investigate possibilities through the literature, personal contacts, and discussion with project groups

Outcome: Agreed that the use of the INSIGHT methodology be investigated

5.2 Task: Examine the agreed cost-benefit analysis method in-depth to determine how best to apply it for HILT purposes

Methodology: Prepare a paper on the problems of using INSIGHT method in HILT. Discuss problems with SG and PMG and agree an approach

¹

As guided by draft paper issued for 01.07.03 and 03.07.03 meetings

5.3 Task: Conduct cost-benefit analysis

Methodology: Using agreed approach, conduct and report on analysis for discussion and criticism at SG and PMG. Record final outcomes and, if appropriate, possible alternative outcomes under a range of

parameter revisions

Steps 5.1 to 5.3 were carried out as described in Appendix H and in Section 6 of the HILT Final Report.

Annex A: Initial Service Specification Draft

	Process	Notes
	User enters system at TeRM server or at some other site that uses TeRM data	TeRM is built around an SQL compliant database using the ORACLE RDBMS
	▼	
Task not subject related, so go elsewhere	◀ User specifies task from drop down list	The Copac/clumps project, cc-interop will begin to identify tasks
	▼	
	Task is subject related	Obvious example is user says she wants to do a search by subject, but there may be other tasks that imply a subject search
	▼	
	User is prompted for subject terms	
	▼	
	TeRM interaction disambiguates subject and maps to DDC	In Wordmap user is given optional meanings of terms and asked to pick which she means (e.g. lotus, the flower, the software, the car, or...)
	▼	
	Task and DDC used to search CLD and CS collections database (a clone of SCONE for the pilot) that also contains service details	Implies a SCONE-like database, but with (1) DDC based collection strengths added, where they exist, for both special collections and general collections, with the latter mapped for strength below the general level where data exists, and (2)
	▼	
	Collections service identifies sub-group of collections relevant to task and to subject using DDC number, truncating it to find collections described at higher level of granularity	(2) Task relevance specified. As in SCONE, the top level of any collection hierarchy will, if it is an online service, have connection details attached. Services returned will be ranked according to the granularity level of ▼ their subject strength coding, with those with lower granularity levels ranked highest
	▼	
	User 'trims' list of collections to suit her purpose, deselecting those she decides not to use initially	User may spot services not relevant to her or simply wish to search less than the full list. Option to test results from highest ranked service may be available
	▼	
	Collections service details include subject scheme(s) used in each collection, and connection details	Specifying subject schemes necessary so that only appropriate terms are sent to any specific service, thereby avoiding the false drops that will occur if all terms from all schemes are sent to all

services



Further interaction with TeRM based on 'trimmed' list of services provides terminologies set needed to search for the user's subject in each chosen service, including standard terms from scheme(s) and (possibly) common UK alternative terms or service specific alternative terms

User clicks button which says something like 'get terms', TeRM sends back appropriate terms for each service. Different levels of service are possible, beginning with terms from standard unchanged scheme used by service, to this plus common UK alternatives, with or without standard terms, to service specific alternatives. The addition and maintenance of UK alternatives will raise costs, and the addition of service specific terms will raise costs even further. Costs against benefits of each approach will be measured by HILT and compared against alternative approaches such as clustering



User gets option of switching off all but one of the various schemes used by a given service and the user's own terms before searching

User will have the option to use her own terms, plus the standard term, plus UK or service alternatives for each service, or to switch any of these alternatives off.



User searches services, either singly or in groups, using preferred term sets

This is assumed to be a broadcast search using Z39.50 plus, where facilities exist, other alternative protocols.

Glasgow : Cente for Digital Library Research, 2004

Appendix C.1: HILT II Survey of practices in Services and Collections

Thank you for agreeing to participate in this survey.

HILT Phase II focuses on terminology and thesauri requirements at the collection level in the JISC Information Environment. As such, we are aiming to identify a representative set of service types, subject schemes in use within services and collections, implications arising from the need to use specialist thesauri, and how and why service staff modify standard terms to suit their needs. The following questionnaire is

designed to address these issues. Responses will help to determine mapping level requirements for the pilot server (TeRM; Terminologies Route Map), that HILT will develop in line with user, service, and expert evaluator outputs.

****We estimate that the questionnaire should not take more than 10-15 minutes to complete****

Background information

Name of institution:

Name of collection/service:

A1. What subject scheme(s)/taxonomy(ies) are in use at your service/collection? (If no schemes are in use please leave blank)

DDC (Dewey Decimal Classification) LCSH (Library of Congress Subject Headings) UNESCO thesaurus (United Nations Educational, Scientific, & Cultural Organisation) AAT (Art & Architecture Thesaurus) MeSH (Medical Subject Headings) UDC (Universal Decimal Classification) SHIC (Social History and Industrial Classification) In house scheme (a scheme developed specifically to suit the need of your service/collection) **Other, please specify below**

A2. How many individuals are involved in the subject description of items in your service/collection?

Of these, how many are:

Qualified or experienced information professionals

Qualified or experienced in the subject of this collection

Neither of the above, but given on-the-job training

A3. Do those involved in subject description sometimes have to make alterations to your chosen scheme(s)?

Yes

No (If 'No', please skip to question A9)

A4. When this is done, what are the reasons:

subject scheme too detailed?

subject scheme too broad?

to accommodate new concepts or areas of knowledge?

to facilitate geographic specificity? (e.g. place names, geographical features or geographical areas not included in a used scheme)

to reflect bilingualism?

cultural differences? (e.g. British/American/Canadian/Australian view of history)

to reflect user needs or demands?

to reflect your service/collection sector? (e.g. HE or FE)

to reflect your service/collection domain? (e.g. libraries, museum, archives)

don't know

If other, please provide details

A2. How many individuals are involved in the subject description of items in your service/collection?

Of these, how many are:

Qualified or experienced information professionals

Qualified or experienced in the subject of this collection

Neither of the above, but given on-the-job training

A3. Do those involved in subject description sometimes have to make alterations to your chosen scheme(s)?

Yes

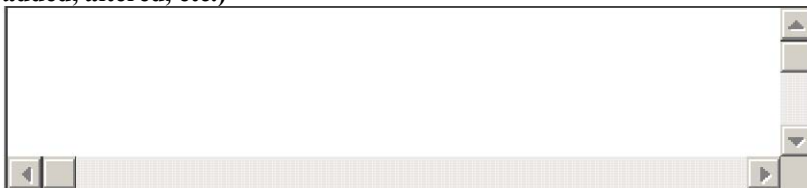
No (If 'No', please skip to question A9)

A4. When this is done, what are the reasons:

subject scheme too detailed? subject scheme too broad? to accommodate new concepts or areas of knowledge? to facilitate geographic specificity? (e.g. place names, geographical features or geographical areas not included in a used scheme) to reflect bilingualism? cultural differences? (e.g. British/American/Canadian/Australian view of history) to reflect user needs or demands? to reflect your service/collection sector? (e.g. HE or FE) to reflect your service/collection domain? (e.g. libraries, museum, archives) don't know **If other, please provide details**

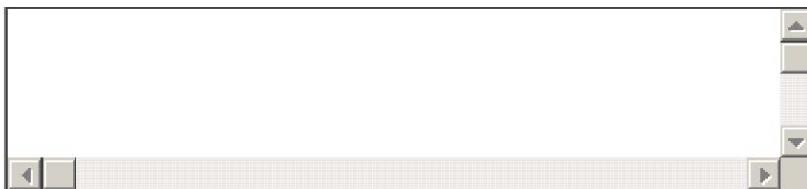


A5. Please give an example of the alteration(s) mentioned above in question A4 . (e.g. Terms you have added, altered, etc.)



A6. Are there any particular subject areas which require regular modification in this way? Please give details

(If you are referring to more than one scheme please be clear in your comments to which scheme you are referring)



A7. Approximately how many times in the past 6 months have you adopted the following techniques when describing material not well catered for by your chosen scheme(s)

Consulted other services/collections?

Consulted other schemes

Employed a 'best match' or partial match policy in your service/collection?

Asked colleagues? **If other, please give details**



A large rectangular text area with a light gray background and a thin border. It features a vertical scrollbar on the right side and horizontal scrollbars at the bottom, indicating it is a scrollable input field.

A8. Is there any mechanism in place whereby changes and modifications are recorded?

Yes

Please give details: 

No

Don't know

A9. Are there any particular subject areas that usually don't require you to make changes to the standard scheme? Please comment



A large rectangular text area with a light gray background and a thin border. It features a vertical scrollbar on the right side and horizontal scrollbars at the bottom, indicating it is a scrollable input field.

A10. Are you aware of any situations where users of your service/collection are attempting to search using terms not used in your standard scheme(s)?

Yes,

Please give examples:



No

Don't know

A11. Please use the comment box below to volunteer any further information you feel may be appropriate to disclose.



Thank you for participating with this research.

Appendix C.2: Results of Survey of Collections and Services

HILT Service/Collection Survey Results

Overview

Table 1 below shows the services who were sent the survey results. Services marked in bold replied to the survey.

Tables 2 and 3 show results of particular relevance to HILT Phase II, indicating:

1. That DDC, LCSH, UNESCO and MeSH are used at significant levels within JISC
2. That services often amend and extend standard schemes
3. Why services amend and extend standard schemes

Results

Table 1: Services/Collections surveyed*

Service/Collection	Response received (marked)
--------------------	-------------------------------

1. 1970 British Cohort Study
2. A2A
3. AMICO Library
4. Archives Hub
5. ArcHSearch
6. Art Abstracts
7. Association for Computing Machinery (ACM journals)
8. AVANCE
9. AXIS
10. Bartholomew Digital Map Data
11. BEI and ERIC (Education Literature Datasets)
12. BioOne
13. BIOSIS Previews
14. Bristol Biomedical Archive
15. British Crime Survey
16. British General Election Studies
17. British Household Panel Study
18. British Universities Newsreel Project Database
19. Bureau van Dijk Databases for FE and Bureau van Dijk Databases for HE

20. Cambridge Scientific Abstracts
21. Cambridge Structural Database
22. Census Knowledge Base
23. Central Postcode Directory (PostZon)
24. Charleston Advisor
25. Cochrane Library
26. Compendex - FE Subscriptions to Ei Databases
27. Compendex -HE Subscriptions to Compendex Databases
28. COPAC
29. CRC Press
30. CrossFire
31. Dental Images
32. Digimap
33. Early English Books Online (EEBO)
34. Economist.com
35. EEVL
36. EEVL (Aerospace and Defence section of Engineering)
37. Electronic Law Reports

38. Elsevier Science
39. Elsevier ScienceDirect
40. Embase
41. Emerald MCB
42. ESDU Engineering Validated Data
43. Euro-barometer Survey Series
44. Eurotext
45. Family Expenditure Survey
46. Family Resources Survey
47. Farm Business Survey
48. General Household Survey
49. GENUKI Genealogy Information Server
50. Grove Dictionaries of Art, Opera and Music
51. Health Survey for England
52. Historical Abstracts
53. History Data Service
54. HUMBUL
55. IMF Statistics
56. Index to The Times, 1790-1980

57. Info4education
58. Infotrac OneFile and Custom Newspapers
59. Inspec
60. International Bibliography of the Social Sciences
61. International Passenger Survey
62. Internet Archaeology
63. ISI web of Science
64. Joint Unemployment and Vacancies Operating System
Unemployment Statistics
65. JSTOR
66. KnowUK and KnowEurope
67. Labour Force Survey & Quarterly Labour Force Survey
68. Lexis Nexis Executive product set
69. Literature Online and LION for Colleges
70. Mossbauer Effect Reference Database
71. National Art Slide Library
72. National Child Development Survey
73. National Diet and Nutrition Surveys
74. National Food Surveys

75. OECD Main Economic Indicators Databank
76. ONS Databank
77. Organisation for Economic Co-operation and Development (OECD)
78. Oxford English Dictionary (OED) Online
79. Oxford Reference Online
80. Project Muse
81. PsycINFO
82. SALSER
83. SciFinder Scholar
84. SCRAN Resource Base
85. Social and Political History of Great Britain
86. Social Attitudes Survey
87. SOSIG
88. St Andrews University Library Image Collection
89. Statistical Accounts of Scotland
90. Television Index
91. UK Data Archive
92. UKBORDERS

93. UNIDO Industrial Statistics Databank
94. Update: The Farming and Countryside Index
95. Visual Arts Data Service online catalogue
96. Vital Statistics for Wards (1981-1991)
97. Wiley Electronic Reference Work
98. Wiley InterScience
99. xreferplus
100. Zetoc

Of the 100 collections and services surveyed a total of 49 responded.

*Note that the list of JISC services/collections as posted on the JISC website (<http://www.jisc.ac.uk/index.cfm?name=collbrowse>) has changed since this survey was undertaken.

Table 2: Schemes in use by JISC services/collections

Scheme	No. of services collections using scheme
DDC	24
LCSH	25
UNESCO	27
AAT	
MeSH	
UDC	1
SHIC	
In house	47
SSD	16
HASSET	18 (based on UNESCO)
IBSS	1
APA	1
CareData	1
LIR	1
ACM Computing Classification	1

A review of the websites of JISC collections and services also indicated that at least 4 services make use of

MeSH thesaurus to provide subject descriptions for their collections.

Table 3: Reasons for departures from standard schemes

Reason for departure	No. of services quoting reason
Subject scheme too detailed	1
Subject scheme too broad	21
To accommodate new concepts or areas of knowledge	24
To facilitate geographic specificity (eg place names)	3
To reflect bilingualism	2
Cultural differences	2
To reflect user needs or demands	21
To reflect your services/collection sector (eg HE/FE)	2
To reflect your service/collection domain (eg. libraries, museums, archives)	1

Appendix C3. HILT II Questionnaire and Interview

This exercise is designed to investigate what level of retrieval individuals such as students, lecturers, supervisors, librarians and other intermediaries think are required for particular subject related information tasks. It also aims to discover what search terms and strategies these groups would use when searching for the information.

TASK: Essay on Own Subject Area

You have been asked to find information for an essay relating to the current status of *your own subject area* in the UK. You are told that the library has paid so that you can have free access over the web to 6 services with different content that each have relevant information.

1. Which of the following do you think is the best strategy to adopt?

- A) Choose one at random or the one you are most familiar with and study only the material from that service
- B) Look at all of the services but study some material from a couple of them in depth
- C) Use all of them, identify all relevant resources and study all in depth
- D) The minimum required to ensure a reasonable grade
- E) Some other variation of the above
- F) Something else? Please specify

2. What would you actually do in practice?

3. Which search terms you would use or suggest using when searching for this topic?

4. When conducting your search would you enter a single term or would you combine terms in any way? (If so, how?)

5. Please go through questions 1. - 4. again, thinking of each of the tasks below in turn

- a) compiling a bibliography on publishing techniques
- b) finding a specific book about Robert Burns
- c) identifying key articles on the history of architectural conservation
- d) general study of journalism software
- e) preparing for a test on statistical methods/tests
- f) preparing for a discussion based tutorial on article writing
- g) planning a presentation to your tutorial group on poster design

Appendix C.4: Results of User Interviews

Overview

- 1. Context**
- 2. Participant Details**
- 3. Summary of Interview Processes**
- 4. Summary of Results**
- 5. Detailed Results: Questions 1 to 3**
- 6. Detailed Results: Question 4**

1. Context

The user interviews helped ensure that the project took account of different user types, and different subject-retrieval related 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. The exercise helped acclimatise project staff to the problems and issues related to dealing with users and subject searching situations in a range of subject areas and also informed the design of a subsequent user workshop that aimed to evaluate aspects of the pilot terminologies server.

2. Participant Details

Distribution of Users by Educational Level and Institution

Education level	No of users
PhD	4
MSc	7
BA	2
HND	4
Misc.	13
Total	30

Table1. Distribution of users by their educational level

Institution

N

Caledonian University	6
Napier University	5
Glasgow College of Building and Printing	4
Strathclyde University	
University of Sheffield	9
Scottish Library & Information Council	3
Glasgow College of Commerce	1
Public Records Office	1
	1

Table2. Distribution of users by institution

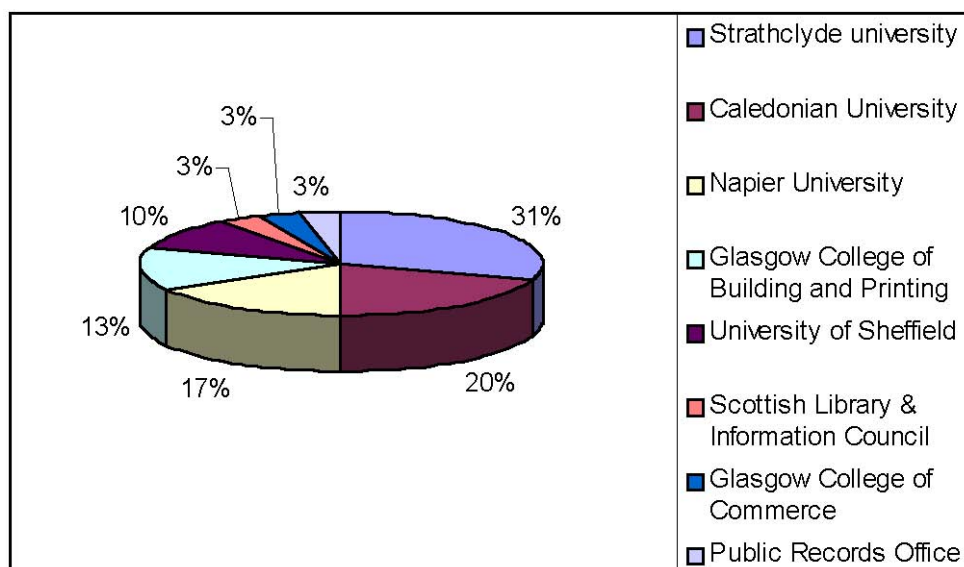


Figure1. Distribution of users by institution

Course and Subject Areas of Users Participating in the Interview

Course/subject area	N
Medical/biological/applied sciences	9
Library intermediary	5
Library and Information science/Archive	4
Finance	3
Construction and architecture	4
Psychology	1
Publishing	1
Business administration	1
English	1
Fashion marketing	1

Table3. Course and subject areas of users

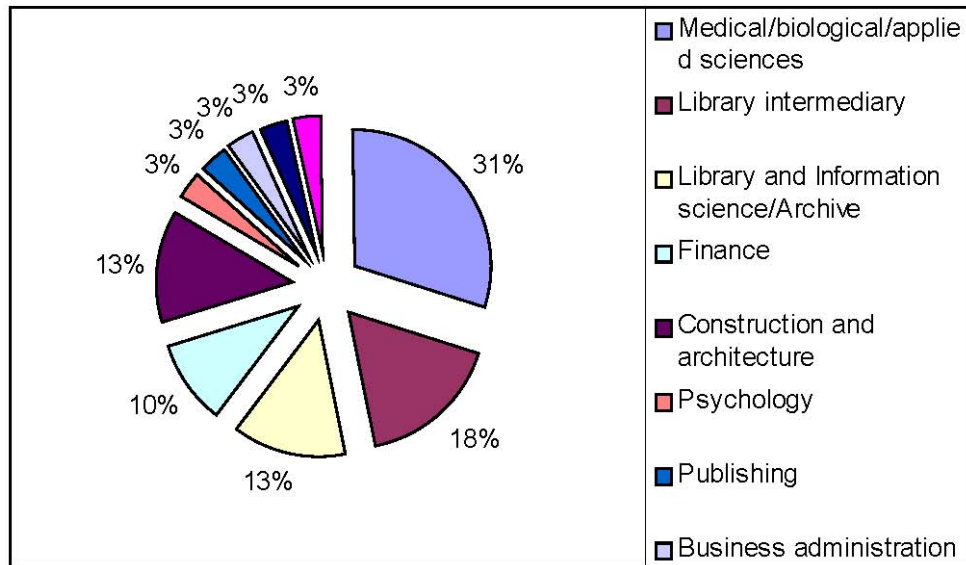


Figure2. Distribution of users by course/subject area

3. Summary of the Interview Process

The exercise was designed to investigate what level of retrieval individuals such as students, lecturers, supervisors, librarians and other intermediaries thought were required for particular subject related information tasks. It also aimed to discover what search terms and strategies these groups would use when searching for the information.

Participants were told that the library had paid to give them free access over the web to 6 services with different content that each had relevant information but different content. There were then asked to imagine they were asked to tackle each of these subject retrieval tasks in turn:

- Finding information for an essay relating to the current status of their own subject area
- Compiling a bibliography on publishing techniques
- Finding a specific book about Robert Burns
- Identifying key articles on the history of architectural conservation
- General study of journalism software
- Preparing for a test on statistical methods/tests
- Preparing for a discussion based tutorial on article writing
- Planning a presentation to your tutorial group on poster design

And to then give answers to the four questions below for each of the tasks.

1. Which of the following do you think is the best strategy to adopt?

- A) Choose one at random or the one you are most familiar with and study only the material from that service
- B) Look at all of the services but study some material from a couple of them in depth
- C) Use all of them, identify all relevant resources and study all in depth
- D) The minimum required to ensure a reasonable grade
- E) Some other variation of the above
- F) Something else? Please specify

5. What would you actually do in practice?
6. Which search terms you would use or suggest using when searching for this topic?
7. When conducting your search would you enter a single term or would you combine terms in any way? (If so, how?)

4. Summary of Results

This was an exploratory survey, conducted partly to retrieve information on how users of various kinds viewed different approaches to subject searching in different circumstances, partly to acclimatize project staff to the problems and issues of subject searching and to the users the terminologies server aimed to support, and partly to help inform the design of a user workshop planned for later in the project (see appendices D.1 to D.3).

Questions 1 and 2 were designed to work in tandem. It was hoped that if participants were asked, both what they felt they should do in the circumstances described, and what they thought they would actually do, they would be more likely to give an honest assessment of probable actual behaviour when answering question 2 (whilst at the same time indicating what they thought their ideal approach would be). In the event, whilst in every scenario asked about, the number of respondents who said they would conduct the relatively comprehensive searches covered under options B & C in practice was significantly lower than the number who saw these approaches as ideal, a surprising number stated that they would still be quite comprehensive in practice. For example, when tackling a paper in their own subject area, 24/30 respondents saw either option B or C as the ideal approach and 17/30 (just over half) saw them as approaches they would adopt in practice. Even amongst students, the figures were 11/17 and 7/17.

Leaving aside the ‘finding a specific book on Robert Burns’ example, where looking at all six collections was almost certainly unnecessary, the lowest number of all participants opting for either B or C was 13/30 and 8/30 for ideal and actual behaviour respectively (general study on journalism software and tutorial on article writing). The corresponding numbers for students were of a similar order.

Although the survey was fairly ‘rough and ready’, it is interesting to note that significant numbers of participants said they should, and claimed they would, use a number of different collections for the tasks proposed if such collections were readily available – a point in favour of the collections finding facility included in the pilot and proposed for an operational server.

The results appear to show that participants saw the different tasks as requiring different approaches. There is also some evidence that there was some logical basis for the perspectives held. Finding a single book on Robert Burns, for example, is not seen by many as requiring a search of all six collections, and more would use all six collections in compiling a bibliography or finding key articles on a topic than would do so for a tutorial on article writing. However, it is difficult to discern any overall pattern in the results for the different tasks or to draw any particular conclusions from them.

Question 3 aimed to discover something about the levels of specificity of search terms that users were likely to use when conducting subject searches, and also about their likely approaches to formulating searches.

In the event, the tasks set by the project generally predetermined the responses of the participants to this question and only the first topic (an essay on the participants’ own subject area) produced useful results. These seemed to show a variety of granularity levels, some very general like ‘analytical chemistry’, others fairly specific like ‘quartz exposure’ and ‘microfiphages’, suggesting that the server and the mappings it would be built around would have to cover a wide range of specificity levels, a view borne out by results

from the later user workshop (see appendix D.3).

Question 4 aimed to find out something about the complexity of the search strategies a terminologies server would have to cater for. The results suggest that a range of strategies – free text (6/177), single term (31/177), multiple terms (97/177), and Boolean combination (43/177) – might have to be dealt with (see table 33 in Section 6 below). An analysis of the search strategies implied by answers to question 3 showed a similar pattern (see table 34 in Section 6 below).

More detail on the results of the four questions and eight scenarios are presented below. Questions 1-3 are handled together and in a great deal of detail. Question 4 is handled separately at the end of the report on questions 1-3 and the detailed results are presented in summary tables only.

5. Detailed Results: Questions 1-3

TASK: Essay on Own Subject Area

You have been asked to find information for an essay relating to the current status of *your own subject area* in the UK. You are told that the library has paid so that you can have free access over the web to 6 services with different content that each have relevant information.

1. Which of the following do you think is the best strategy to adopt?

- 1 both of which entailed looking at all six of the services said to be available
- A) Choose one at random or the one you are most familiar with and study only the material from that service
 - B) Look at all of the services but study some material from a couple of them in depth
 - C) Use all of them, identify all relevant resources and study all in depth
 - D) The minimum required to ensure a reasonable grade
 - E) Some other variation of the above
 - F) Something else? Please specify

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
A	3				3
B	7	3	1	4	15
C	4	2	1	2	9
D	1				1
E	1				1
F	1				1
Total	17	5	2	6	30

Table 1: Ideal strategies of users by group

2. What would you actually do in practice?

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
A	1	1			2
B	4	3	1	3	11
C	3	1		2	6

D	1			1	2
E			1		1
F	8				8
Total	17	5	2	6	30

Table 2: Practical strategies of users by group

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
Same	8	3	1	4	16
Different	9	2	1	2	14
Total	17	5	2	6	30

Table 3: Number of users whose ideal and practical strategies are the same/different

Reasons for change of strategy:

- User would automatically go to services he is familiar with (intermediary)
- User would adopt a wide mix of methods depending on his timescale (student)
- User claimed he wouldn't use such services in practice; he would use the library OPAC or Google (student)
- User claimed she would go directly to her library OPAC in practice

3. Which search terms you would use or suggest using when searching for this topic?

'CAD'; 'Computer aided design in the UK'

'Computer assisted design'; 'CAD'

'Graphic design'; 'computer graphics'; 'media AND design AND graphics'

'Computer aided design'

'Archival description standards'

'graphic design'

'Communication AND information'; 'communication of information'

'Child development'

'Calcium'; 'microfiphages'

'lung disease'; 'heart disease'

'Air pollution health effects'

'Language linguistic death'

'Quartz exposure'; 'lung cancer'

'IM in UK'

'Analytical chemistry'; 'analytical chemistry AND UK'

'graphics AND technology'

'Virology research in the UK'; 'current status'

'Parasite'; 'modulation'

'Financial services'; 'financial development'; 'pensions'; 'financial service regulation'

'financial resources in UK'

'Architectural conservation'; 'historical buildings and restoration'

'Business admin'; 'UK business admin'

'Fashion marketing in the UK'

IM in UK

'Financial services in UK'

'Town and country planning'; 'planning journal'
 'Current state of publishing in the UK'
 'Construction management'; 'construction industry'

Table 4: Search terms given by users
 2 students were unable to provide subject terms for this questions.

TASK: (a) compiling a bibliography on publishing techniques

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
A	2			1	3
B	3	2			5
C	6	3	1	3	13
D	2				2
E	3				3
F	1		1	2	4
Total	17	5	2	6	30

Table 5: Ideal strategies of users by group

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
A	2			1	3
B	4	4		1	9
C	3			2	5
D	2	1		1	4
E	1		1		2
F	5		1	1	7
Total	17	5	2	6	30

Table 6: Practical strategies of users by group

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
Same	11	2	1	3	17
Different	6	3	1	3	13
Total	17	5	2	6	30

Table 7: Number of users whose ideal and practical strategies are the same/different

Reasons for change of strategy:

- no data

'Publication industry'; 'publishing techniques'
 'Publishing'
 'Publishing techniques'
 'publishing'
 'publishing'
 'publishing'; 'marketing'

'Publishing techniques'
 'Publishing techniques'
 'Publishing AND techniques'
 'Publishing'
 'Bibliography publishing'
 'Publishing techniques'
 'Publishing techniques'
 'Bibliography AND publishing techniques'
 'Publishing techniques'
 'Publishing techniques'
 'Publishing techniques'
 'Publishing techniques'; 'methods in publishing'
 'Publishing techniques'; 'Publishing'
 'Publication techniques'
 'Publishing techniques'
 'Publishing techniques'; 'faq of publishing'
 'Publishing techniques'
 'Publishing techniques'

Table 8: Search terms given by users

5 users were unable to provide subject terms for this questions; 4 students and one intermediary. The intermediary claimed she would use Harvard bibliography and referencing tools.

TASK: (b) finding a specific book about Robert Burns

Group	Students	Researchers	Lecturers	Intermediaries	Total
Strategy					
A	5	1		3	9
B	3	2			5
C	1	1			2
D	2				2
E					
F	6	1	2	3	12
Total	17	5	2	6	30

Table 9: Ideal strategies of users by group

Group	Students	Researchers	Lecturers	Intermediaries	Total
Strategy					
A	4	2		3	9
B	1	2			3
C					
D	2				2
E					
F	9	1	2	3	15
Total	16	5	2	6	29

Table 10: Practical strategies of users by group

Note: one student did not provide a response

Group	Students	Researchers	Lecturers	Intermediaries	Total
Strategy					
Same	12	3	2	6	23
Different	4	2			6
Total	16	5	2	6	29

Table 11: Number of users whose ideal and practical strategies are the same/different

Reasons for change of strategy:

- no data

'Burns, Robert'
 'Robert Burns'
 'Robert Burns'
 'Robert Burns'
 'Robert Burns'
 'Robert Burns'
 'Robert Burns'
 'Robert Burns'
 'Robert Burns'
 'Robert Burns AND Scotland AND poet'
 'Burns'
 'Robert Burns'; author name if known
 'Robert Burns'; 'Books about Robert Burns'
 'Robert Burns'
 'Robert Burns'
 'Robert Burns'
 'Robert Burns'
 'Robert Burns'; the book
 'Robert Burns'
 'Burns'
 'Robert Burns'; 'Poetry and Scotland'
 'Robert Burns'
 'Robert Burns'
 'Burns'
 'Robert Burns'
 'Burns'
 'Robert Burns'
 'Rober Burns'; 'Scottish history'

Table 12: Search terms given by users

Two users were unable to provide terms for this task. One of them claimed they would use the author or title, or ISBN if known.

TASK: (c) identifying key articles on the history of architectural conservation

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
A	4				4
B	4	2	1	2	9
C	5	2		2	9
D					
E	1			1	2
F	3	1	1	1	6
Total	17	5	2	6	30

Table 13: Ideal strategies of users by group

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
A	6	3	1	2	12
B	2	1		3	6
C	4	1			5
D					
E					
F	4		1	1	6
Total	16	5	2	6	29

Table 14: Practical strategies of users by group

Note: One student did not respond

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
Same	9	1	1	3	14
Different	7	4	1	3	15
Total	16	5	2	6	19

Table 15: Number of users whose ideal and practical strategies are the same/different

One student didn't give a response for what they would do in practice.

Reasons for change of strategy:

- User claimed they would ideally try to identify the most useful service but may not look at all in depth but in reality they would try to find a key service and check for recent articles and also possibly check reading lists of relevant university modules (intermediary).
- One lecturer claimed they should consult the maximum amount of services but would probably tend towards EBSCO and Ingenta in practice.
- One researcher's actual strategy would be different from her ideal one due to lack of familiarity with the subject area.

'history AND architecture AND conservation'; 'building conservation'

'Architectural conservation AND history'

'Conservation and buildings'; 'architectural conservations'

'architectural conservation'

'History AND architect* AND conserv*'

'Buildings AND conservation AND history'

'Architectural conservation'

'Key article in architectural conservation'
 'History of architectural conservation'
 'Architectural conservation'
 'Architectural conservation'; 'historical architecture'
 'Architectural conservation, history of'; relevant dates
 'History AND architectural conservation'
 'Architectural Conservation' 'History of'
 'Architectural Conservation'
 'Architecture AND conservation'
 'Key articles in Architectural conservation'
 'Architecture and conservation'; 'conservation and architecture'; 'architectural history'
 'Architectural conservation'
 'Architectural conservation'; 'subject report on architectural conservation'
 'Architectural conservation'; 'History of architectural conservation'
 'history of architectural conservation'
 'History of Architectural conservation'
 'History of architectural conservation'
 'architectural conservation'; 'architectural organisations'
 'architecture'; 'conservation of architecture'
 'Articles on architectural conservation'
 'Architectural conservation, history of'

Table 16: Search terms given by users

Two users did not provide search terms for this task. One of these, an intermediary, claimed he would use a citation index and review articles which he considered useful for overviewing an issue or area of research. The other, a student, felt that the search terms entered would depend on the brief of the task/assignment.

TASK: (d) general study of journalism software

Group	Students	Researchers	Lecturers	Intermediaries	Total
Strategy					
A	2	1		1	4
B	5	2		1	8
C	3			2	5
D	2				2
E	2				2
F	3	2	2	2	9
Total	17	5	2	6	30

Table 17: Ideal strategies of users by group

Group	Students	Researchers	Lecturers	Intermediaries	Total
Strategy					
A	2	3		1	6
B	1	1		2	4

C	3			1	4
D	2				2
E	2				2
F	7	1	2	2	12
Total	17	5	2	6	30

Table 18: Practical strategies of users by group

Group	Students	Researchers	Lecturers	Intermediaries	Total
Strategy					
Same	11	3	2	3	19
Different	6	2		3	11
Total	17	5	2	6	30

Table 19: Number of users whose ideal and practical strategies are the same/different

Reasons for change of strategy:

- One student thought she should look at all services and consider some in depth as the subject area is a very general one. However, in practice she would use the library OPAC or Google.

Search terms

'Journalism software'; 'IT'
 'Journalism software'; 'Journalism computers'; 'Journalism computer applications'
 'Journalism software'
 'Journalism software'
 'Journalism software'
 'Journalism AND software AND comput*'
 'reporting AND journalism AND software'
 'Journalism AND software'
 'Journalism software'
 'Journalism software'
 'Software, journalism'
 'Journalism software'
 'Journalism software'; 'journalism AND software'
 'Journalism software'
 'Journalism software'
 'Journalism and software'
 'Journalism software'; 'newspaper software'; 'publishing software'; 'software for journalists'
 'Journalism software'
 'Journalism software'
 'Journalism software'; 'journalism technology'
 'Journalism software'
 'Journalism software'; 'technology in journalism'
 'Desktop publishing software'; 'journalism software'
 'Journalism software'
 'Journalism software'; 'manual of journalism software'

Table 20: Search terms given by users

Five users did not provide search terms for this task; one lecturer, one researcher and three students.

TASK: (e) preparing for a test on statistical methods/tests

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
A	4			1	5
B	4	1		2	7
C	5	4			9
D					
E	1				1
F	3		2	3	8
Total	17	5	2	6	30

Table 21: Ideal strategies of users by group

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
A	2			1	3
B	4	3		1	8
C	2	1			3
D		1		2	3
E					
F	9		2	2	13
Total	17	5	2	6	30

Table 22: Practical strategies of users by group

Note: One student did not respond

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
Same	10	2	2	3	17
Different	7	3		3	13
Total	17	5	2	6	30

Table 23: Number of users whose ideal and practical strategies are the same/different

Reasons for change of strategy:

- One student claimed that ideally he would look at all services but in reality he would rely on previous knowledge as he dislikes the subject area.

'Statistics methodology'

'statistics'; 'statistical techniques'

'Statistics'; 'statistics made easy'

'statistical methods'; 'quantitative methods'

'numerical tests'; 'statistical tests'

'Statistical tests'

'General statistical methods'

'Statistical methods'; 'statistics'

'Statistical models AND tests'

'Statistical methods'; tests
 'statistical methods OR statistical tests'
 'statistical methods'
 'statistics AND (methods OR tests)'
 'statistical methods'; 'statistics AND tests'
 'Statistical methods'
 'Statistical methods'; 'regression analysis'
 'statistical books'
 'statistics'; 'statistical methods'
 'Statistical tests and methods'
 'statistical methods'
 'Statistical methods'
 'statistics'

Table 24: Search terms given by users

TASK: (f) preparing for a discussion based tutorial on article writing

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
A	8				8
B	4			2	6
C	2	3		2	7
D	2	2			4
E					
F	1		2	2	5
Total	17	5	2	6	30

Table 25: Ideal strategies of users by group

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
A	6				6
B	3	4		3	10
C	1				1
D	2	1		1	4
E					
F	5		2	2	9
Total	17	5	2	6	30

Table 26: Practical strategies of users by group

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
Same	9	1	2	4	16
Different	8	4		2	14
Total	17	5	2	6	30

Table 27: Number of users whose ideal and practical strategies are the same/different

Reasons for change of strategy:

- One student claimed that ideally he should only look at one service as he only has to contribute ideas to the discussion but in reality he would look at several books.

Search terms

'Article(s)'; 'Writing structure'; 'article structure'
 'Presentation'; 'essay writing'
 'Articles'; 'writing articles'; 'writing academic articles'; 'academic writing'
 'Business writing'; 'business presentations'; 'report writing'
 'Article writing'
 'Writing AND stud*'
 'Writing style'
 'Article writing'
 'Article writing'
 'Writing techniques'
 'Presentation skills'
 'Article writing'
 'Article writing'
 'Article writing, Journalism'
 'article writing'; 'how to write articles'
 'Article writing'
 'Tutorial on article writing' 'notes on...'
 'Article and writing'
 'Written English'; 'Writing articles'; 'readable English'; 'How to writes articles'
 'Good Article Writing'; 'proper English'
 'How to write articles'
 'Tutorial in article writing'
 'Essay writing'; 'Essay preparation'; 'Essay planning'
 'Article writing discussion'; 'article writing style'
 'writing articles'
 'Article writing'; 'methodology of article writing'

Table 28: Search terms given by users

g) planning a presentation to your tutorial group on poster design

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
A	3				3
B	6	1		2	9
C	3	2		2	7
D	1				1
E	1				1
F	3	2	2	2	9
Total	17	5	2	6	30

Table 29: Ideal strategies of users by group

Group Strategy	Students	Researchers	Lecturers	Intermediaries	Total
----------------	----------	-------------	-----------	----------------	-------

A	1				1
B	3	2		3	8
C	3	1		1	5
D	1				1
E					
F	9	1	2	2	14
Total	17	4	2	6	29

Table 30: Practical strategies of users by group

Note: One researcher did not respond

Group	Students	Researchers	Lecturers	Intermediaries	Total
Strategy					
Same	10	3	2	4	19
Different	7	1		2	10
Total	17	4	2	6	29

Table 31: Number of users whose ideal and practical strategies are the same/different

Note: One researcher did not respond in the case of their practical strategy

Reasons for change of strategy:

One student claimed he should use all of the services but in reality would do a Google image search.

Search terms

'Design AND presentation AND posters'

'Presentation'; 'power point'

'desktop publishing' 'design posters'

'poster design'

'poster design'

'Presentation design'

'Poster design'

'Poster design'

'Poster design'

'Poster design, general'

'Poster design AND articles'

'Posters AND design'

'poster design'

'Poster and design'

'Marketing'; 'Marketing design'

'Poster design'; 'design and posters'

'Poster design'; 'tips for poster design'

'Poster design'

'Presentation techniques'; 'poster design'

'Photoshop'; 'office projector'

Table 32: Search terms given by users

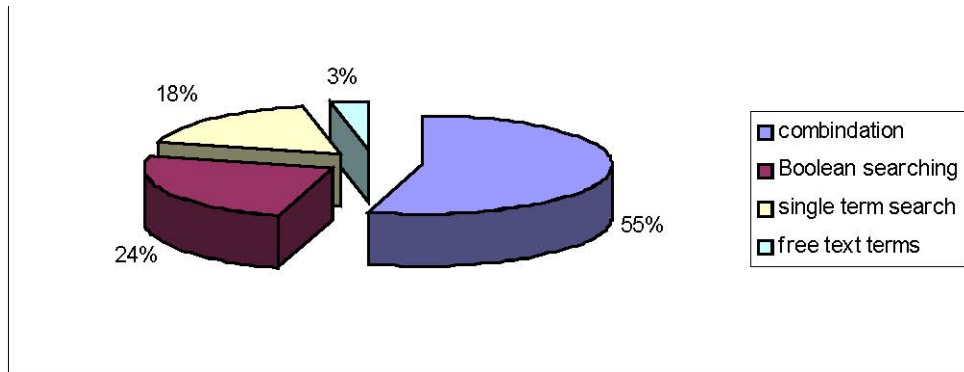
6. Detailed Results: Question 4

Search Techniques Proposed and Adopted by Users

Search techniques	Combination	Boolean searching	Single term search	Free text	No response
Search					
Tasks					
Compiling a bibliography	14	4	6	0	5
<i>Finding a specific book about Robert Burns</i>	10	2	17	0	1
Identifying key articles on the history of architectural conservation	15	10	1	3	1
General study of journalism software	17	8	2	0	3
Preparing for a test on statistical methods/tests	12	7	3	1	7
Preparing for a discussion based tutorial on article writing	17	6	1	2	4
Planning a presentation to your tutorial group or poster design	12	6	1	0	11
97	43	31	6	32	

Search techniques	Combination	Boolean searching
Search Tasks		

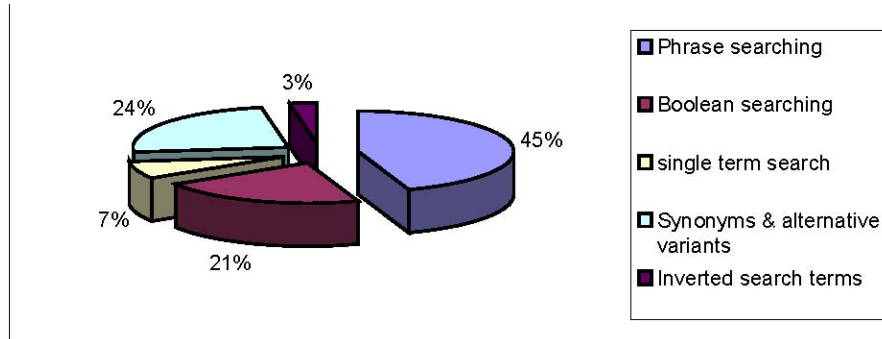
Table 33 Proposed search techniques (by task)



Search techniques	Phrase searching	Boolean searching	Single term search	Synonyms & alternative variants	Inverted Search terms	No response
Search Tasks						
Compiling a bibliography on publishing techniques	14	6	3	4	0	3
<i>Finding a specific book about Robert Burns</i>	18	1	4	5	0	2
Identifying key articles on the history of architectural conservation	4	8	5	4	4	2
General study of journalism software	11	5	0	8	1	5
Preparing for a test on statistical methods/tests	12	4	0	6	0	8
Preparing for a discussion based tutorial on article writing	11	6	0	9	0	4

Planning a presentation to your tutorial group or poster design	7	7	0	6	0	10
Total	78	37	12	42	5	34

Table 34 Search techniques implied by search formulated (by task)



Percentage of search techniques implied by search formulated

Glasgow : Cente for Digital Library Research, 2004

Appendix D.1 Workshop Overview

Workshop Aims:

- 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.

Workshop Environment

Elements:

- A room with 25 networked computers providing access to the HILT pilot
- The pilot with complete DDC21 schedules, subdivisions, relative index and LCSH from OCLC file plus some UNESCO and some MeSH for the medical area.
- An initial short talk describing what the workshop is seeking to learn (cast in general terms to avoid skewing results)
- Workshop run twice with different groups totalling 21 participants

- Aim was to explain the printed instructions up to a point to individuals but aim to avoid telling them anything that might skew our results
- Participants asked to bring a recent exam or essay or tutorial or lab question they have been given or have set or, failing that, to invent a reasonable facsimile
- An initial demonstration of the software was given to half of the participants, but not to the other half to test any effect of 'training'
- A short discussion with participants before they leave to discuss and clarify their responses (before payment in the case of students)

Participants

Forty-one participants were recruited from students and intermediaries (librarians and information professionals). The students came from University of Strathclyde in Glasgow and search intermediaries from a range of libraries and information institutions. E-mails and notices on library boards were used to approach the target groups. As an incentive the students were paid £10 each. Table 1 shows the distribution of intermediaries by subject.

Intermediaries

Archivist	1
Arts and social sciences	2
Nursing and midwifery	1
Engineering	1
General	4
Library systems administrations	1
Performing arts	1
Total	11

Table 1. Distribution of intermediaries by subjects

Table 2 shows the distribution of students by the subject of their course of research.

Students

Library and information science	11
Computer science	2
Chemical engineering and chemistry	2
Psychology	2
Food sciences	3
Law	2
Business, economics and marketing	5
History	1
Operational research	1
Public health	1
Total	30

Table 2. Distribution of students by subjects

Participants were asked whether or not they have made use of the computer applications such as search engines, OPACs, online databases and spreadsheets. The table 1 show users' responses based on the

application.

Applications used	Number of users
Internet search engines (e.g. Google)	30
Library catalogues (OPACs)	27
Other web based or other online databases or services	27
General applications (e.g. word processing packages, spreadsheets, databases, email)	30

Table1. Computer applications used by participants

Glasgow : Centre for Digital Library Research, 2004

Appendix E: RDN Issues paper - current issues in relation to subject schema

Introduction

This brief paper outlines some of the current issues facing the RDN in offering a central, high level, interdisciplinary subject browse. The following sections of this document give an overview of the organisation and its technical architecture and go on to outline specific issues and related studies and project work before detailing several requirements in this area.

Background

The Resource Discovery Network (RDN) is a free Internet service dedicated to providing effective access to high quality Internet resources for the learning, teaching and research community. The service is primarily aimed at Internet users in further and higher education, although others will also find the service to be of value for personal and professional development. Funded primarily by JISC, the RDN provides

access to a series of Internet resource catalogues or “hubs” containing descriptions of high quality Internet sites, selected and described by specialists from within UK academia and affiliated organisations. Value-added services such as interactive web tutorials and alerting services are also provided to enable users to make more of their time on the Internet. Further information on the services available can be found at <http://www.rdn.ac.uk>

The RDN is a distributed service comprising 8 subject hubs based at UK universities and the Resource Discovery Network Centre (RDNC). The network has grown from initially independent projects based within UK universities, and the current distributed hardware and software structures in the network reflect this. Most of the hubs maintain local databases to run their subject/faculty level service using a range of web and database platforms. In delivering a central interdisciplinary search service, the RDN has moved away from its initial z39.50 based cross-searching model and records from subject-based services are now gathered (on a weekly basis) into a consolidated central interdisciplinary database using the OAI Protocol for Metadata Harvesting in the following architecture:

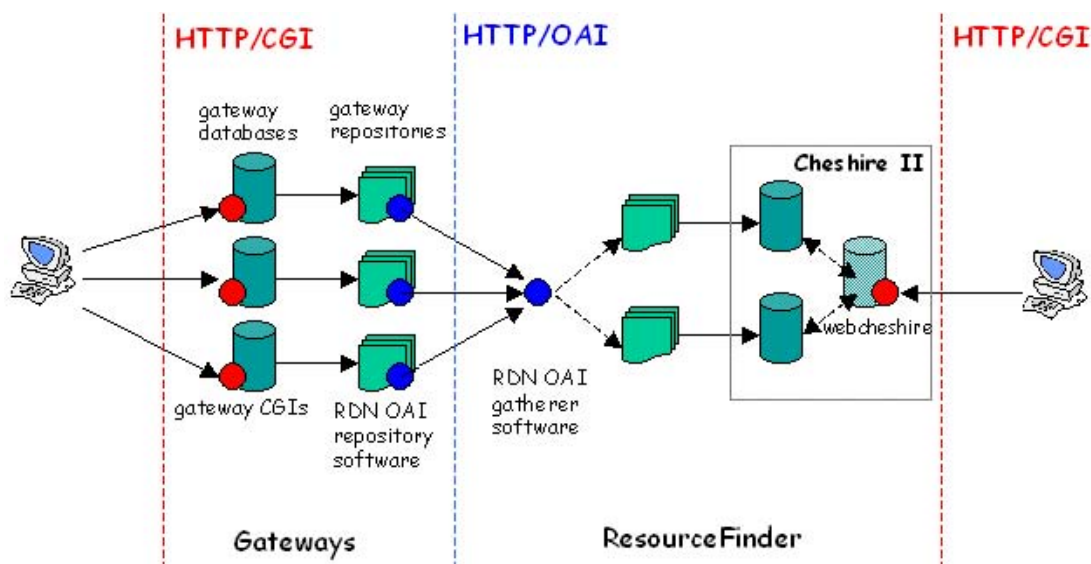


Figure 1 (Taken from *Building ResourceFinder*, <http://www.ariadne.ac.uk/issue30/rdn-oai/> by Pete Cliff)

Subject schema in the RDN

Due to the fact that the RDN brought together a number of initially independent and unrelated service providers, there is no single overarching classification scheme in use across the RDN. The browse tree currently shown at www.rdn.ac.uk is based on hubs, or major sections within hubs. The major headings are:

- ARTS & CREATIVE INDUSTRIES
- BUSINESS
- COMPUTING
- EDUCATION
- ENGINEERING
- GEOGRAPHY & ENVIRONMENT
- HEALTH & MEDICINE
- HUMANITIES
- LAW
- LIFE SCIENCES
- MATHEMATICS
- PHYSICAL SCIENCES

- REFERENCE
- SOCIAL SCIENCES
- SPORT, LEISURE & TOURISM

A pilot version of an interface to navigate these headings and those at the next level within each section is available at: <http://www.rdn.ac.uk/cgi-bin/browse> (it should be noted that this tool may not be up to date with current browse structures used at hubs).

A brief study of subject schema in used by RDN hubs was conducted last year and the results of this made available in October 2001 (<http://www.rdn.ac.uk/publications/browse/analysis-2001-10/>). This work showed 9 schema in use across the RDN. The nine schema were as follows:

- DDC
- NLM
- LC
- Ei*
- MSC*
- HESA*
- UDC
- APA
- Biz-Dewey*

** indicates local modification to a scheme*

Whilst the use of these numerous schema offers subject hubs the facility to provide detailed and appropriate subject navigation to their users, it creates an issue in the RDN's presentation of an interdisciplinary service at <http://www.rdn.ac.uk>. Due to the various classification schemes in use, this central service is currently unable to offer a consolidated interdisciplinary browse of records and instead links out to the various hub interfaces. We have received feedback that the current situation is confusing for some users would ideally like to be able to offer a broad browse view based at a relatively high level based on mapping or a similar solution. Further details on this area are available in the draft report at <http://www.rdn.ac.uk/publications/browse/requirements/>.

Applying subject areas to items in retrospect – new project work

RDN is engaged in new project work which will involve some autoclassification of metadata records describing e-print resources at UK universities. Again the OAI Protocol for Metadata Harvesting will be used and once gathered, metadata records will be passed to external Web services that will enhance the records, adding (or validating) authoritative forms of author names, automatically assigning a subject-classification to the metadata and parsing semi-structured citation information in the document text to form structured, machine-readable, citations in the form of OpenURLs.

It is intended to perform these enhancements using a copy of the full-text of the publication from the repository (as shown by the dashed line in the diagram below). If full-text is not available, automated subject classification may be possible based on the existing metadata, though the results are likely to be of much lower quality.

These external Web services will be developed by OCLC (subject classification and name authority) and the University of Southampton (citation analysis) based on existing technologies at those two organisations. We are using 'Web service' in the technical sense to refer to an application component that is available on the web behind a SOAP interface.

The OCLC Web services will be hosted by them and accessed remotely by the e-Prints UK service. Note that OCLC will not be receiving funding from this project proposal for the development of these Web services. This acknowledges the collaborative, exploratory nature of this work. The work will be carried

out by the Office of Research and is in line with existing research directions that are looking at providing individual knowledge organisation services in a web services environment. OCLC is interested in working with the project to consider emerging business practices in relation to exchange and reuse of OAI-available metadata.

The citation analysis Web service will be hosted by UKOLN, based on the software developed by the Open Citation project, with technical support and consultancy available from staff at the University of Southampton.

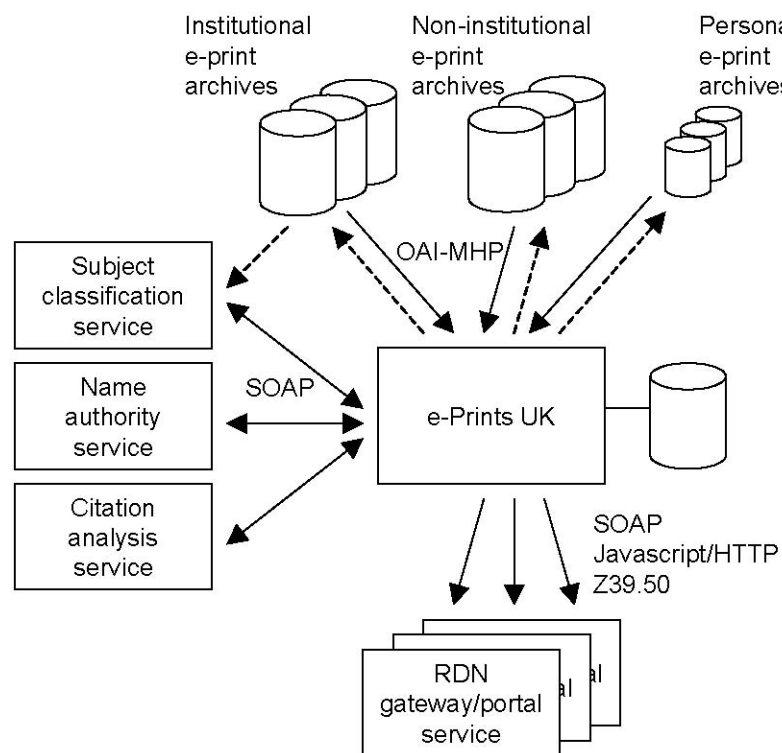


Figure 1

Note that a potentially interesting benefit of the e-Prints UK project is that, by offering the enhancement functionality as Web services, it will be possible (assuming appropriate business agreements are in place) to embed these services directly into the e-print archive cataloguing tools made available within participating institutions.

The e-Prints UK service will be available to end-users in a number of ways. Firstly, there will be a central Web site for the project, integrated with the current RDN Web site, providing a search interface to all the enhanced, harvested metadata. In addition, e-Prints UK will offer shared, configurable discovery services that enable the RDN hubs, UK academic institutions and other organisations to simply embed e-Prints UK within their services. This functionality will be based on three approaches. Firstly, a Z39.50 target supporting Functional Areas A and C of the Bath Profile will be developed. Secondly, a Simple Object Access Protocol (SOAP) interface will be provided, allowing sophisticated integration of e-Prints UK within other services. Thirdly, a simpler, less-sophisticated, approach will also be developed based on Javascript and HTTP linking, for those services not able to support SOAP. These approaches will be closely based on the RDN's existing RDN-Include and RDNi-Lite offerings. Initially the SOAP interfaces will be used to embed e-Prints UK services into the 8 subject-based discovery interfaces at the RDN's faculty level hubs. Enhancing records is crucial in this procedure in order to provide coherent subject access to a range of data drawn from repositories that do not share a common schema, controlled vocabulary for subject data or other cataloguing practice.

Summary of requirements

In relation to the details outlined above, RDN is seeking an overarching classification solution which enables the following:

1. The ability to classify RDN metadata records, at the next level of detail below the 15 broad terms currently used at www.rdn.ac.uk (i.e. breaking down each category into circa 10 sub-headings)
2. The ability to utilise such a classification scheme to provide a central interdisciplinary browse service by mapping each of these terms to points in hubs' existing browse tree structures.
3. The possibility of building a fully integrated browse tree at www.rdn.ac.uk using these terms
4. The facility to offer searches limited by subject area at www.rdn.ac.uk
5. The ability to classify externally created e-print archive records to the same scheme (or at a minimum to individual hubs) in order to provide subject sections of this data to hubs for presentation.
6. The possibility to consider applying the schema/mapping to other datasets in the RDN e.g. Behind the Headlines, Jobs.ac.uk news feeds, LTSN news feeds, Jobs, JISCmail lists.

Glasgow : Centre for Digital Library Research, 2004

Appendix F

Notes on Possible Clustering-Based Enhancements to User Tool Set

Overview:

Context

Notes on Cheshire clustering from Cheshire documentation

Examples: MerseyLibraries.Org and Archives Hub

Notes on RLG's RedLightGreen initiative

Questions a Follow Up Clustering Study Should Examine

Context

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. 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 has asked 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 core terminology server facilities. Appendix F details the current (very embryonic) state of HILT research and analysis as regards this area (The project conducted some desk research on the clustering function utilised in the Cheshire initiative and, very late on, RLG’s RedLightGreen initiative. It also made an early attempt to identify questions requiring investigation in respect of the possible use of clustering as a terminologies server tool. Notes on these are provided below.)

Notes on Cheshire

The following description is taken from the Cheshire Final report ¹ :

A brief overview of the Cheshire system

The development of *Cheshire system* is a joint JISC/NSF funded project with principal investigators from the University of Liverpool and the University of California, Berkeley. The Cheshire project is developing a next-generation online catalogues and full-text information retrieval system using advanced IR techniques. This system is being deployed in a working library environment and its use and acceptance by local library patrons and remote network users are being evaluated. The Cheshire II system was designed to overcome twin problems of topical searching in online catalogues, search failure and information overload as well as to provide a bridge between the purely bibliographic realm of previous generations of online catalogues and the rapidly expanding realm of full-text and multimedia information resources.

Main features:

¹ See <http://cheshire.lib.berkeley.edu/>

- Using advance information retrieval techniques such as probabilistic and Boolean retrieval models, which permits the combination of Boolean and probabilistic elements within the same search.
- A client/server architecture with implementations of current information retrieval standards including Z39.50 and SGML and XML
- It includes a programmable graphical direct manipulation interface under X on Unix and Windows NT. There is also CGI interpreter version that combines client and server capabilities. These interfaces permit searches of the Cheshire II search engine as well as any other z39.50 compatible search engine on the network
- It permits users to enter natural language queries and these may be combined with Boolean logic for users who wish to use it
- It supports open-ended, exploratory browsing through following dynamically established linkages between records in the database, in order to retrieve materials related to those already found. These can be dynamically generated “hyper searches” that let users issue a Boolean query with a mouse click to find all items that share some field with a displayed record.
- Stemming and relevance ranking algorithms
- Use of query reformulation, query expansion and relevance feedback techniques
- access different domains and information resources (text and document retrieval, numeric databases, and geographic information systems) through the support for *transverse searching* (in which data found in a text database can be used to find related data in a numeric or geo-spatial database)

The project aim

A primary aim of the project was to enable the enhanced retrieval of unfamiliar metadata across domains, e.g. constructing linkages between natural languages expressions of topical information and controlled vocabularies for geospatial, textual, and statistical. To this end, a number of methods developed using Z39.50 to automatically "cluster" together topics which may be semantically related for digital library projects; and have incorporated this technology in a number of national services some cross-domain.

Through this way, effort was made to develop a research-oriented method of providing access to subject headings, no matter how unfamiliar they may be to the end user, by automating the process of association between natural language and their subject headings. This capability appears to have been effective in enabling users to map their query to the controlled vocabularies (subject headings) used in descriptive metadata; it may be used to cross-search different thesauri and automate associations between them and the user's inquiry.

Search engine capabilities

The Cheshire II search engine supports several methods for translating the user's query terms into the vocabulary used in the database. These include support for field-specific stopword lists, field-specific query-to-key conversion functions, stemming algorithms that reduce significant words to their *roots* by converting suffix variations, such as plural forms of a word, to a single form, and support for mapping database and query text words to a standardized form based on the WordNet dictionary and thesaurus.

The search engine also supports direct probabilistic searching of any indexed field in the SGML records. The probabilistic ranking method used in the Cheshire II search engine is based on the *staged logistical regression* algorithms developed by Berkeley researchers and shown to provide excellent full-text retrieval performance in the TREC evaluation of full-text IR systems.

The techniques of "Classification Clustering" use natural language parsing software to identify phrases in the language of the users of bibliographic databases, taken from the titles and abstracts in the literature to be searched, and then apply statistical association techniques to associate these words and phrases with the metadata terms of the target.

This technique is currently used to facilitate automatic subject retrieval across any number of thesauri supported by a number of distributed datasets. The initial findings suggest that this functionality may facilitate access to metadata describing geospatial datasets. Specifically, methods of mapping geographic place names in text (natural language) to probable geographic coordinates; for mapping geographic coordinates to sets of nearby named places at different levels of geographic or political detail and of different place name types (e.g. city, country, state or province, country).

The Cheshire system is now able to map the searcher's notion of a topic to the terms or subject headings actually used to describe that topic in the database.

The system is able to provide direct connection between ordinary language queries ("query vocabularies") and indexing terms ("entry vocabularies") actually used to organize information in a variety of databases. These innovations are now implemented in a production environment as part of the Archives Hub, MerseyLibraries.org, etc., all of which support cross-thesauri retrieval without the expense associated with the development and maintenance of higher level thesauri. We are planning to implement this innovation as part of the JISC funded Information Environment Service Registry (IESR) which will be extended across all JISC datasets.

The project has extended development of these associative techniques to provide support for "subdomain" vocabularies, e.g. association dictionaries which will lead searchers to the appropriate term or cluster of subject access terms that are likely to satisfy their information needs for specialized topics ("subdomains") which may be non-textual or include cross-thesauri and trans-lingual support. The development and implementation of these techniques have enabled the system to develop automatically a "likelihood ratio

weighting" associated with each searching term and each metadata value which will may lead the searcher more quickly to required information.

Metadata Reuse: Entry Vocabulary Modules (EVMs)

One primary research objective of the JISC/NSF project is to enable the enhanced retrieval of unfamiliar metadata using what we call "Entry Vocabulary Modules", or EVMs. This capability, growing out of the Cheshire project, is really a method of constructing linkages between natural language expressions of topical information and controlled vocabularies automatically.

One of the more common challenges facing any end user is in navigating various data sources which might use different thesauri. The Archives Hub is a case in point: data contributors follow either the LCSH (Library of Congress Subject Headings) or UNESCO thesauri. How do users unaccustomed to using either thesauri find out the information of interest to them? A key objective was to develop more research-oriented methods of providing access to these subject headings, no matter how unfamiliar and bewildering they may be to the end user, by automating the process of association between natural languages and their subject headings.

To facilitate this, the project has used the Cheshire system's support for probabilistic information retrieval on any indexed element of the dataset(s). This means that we can use a natural language query (for example, plain English) to extract the most relevant entries in one or more databases. From this information the server can *automatically* present to the user a cluster of subject headings which might be relevant to their inquiry. The user then can select the subject heading or combination which is most appropriate and then use this as a basis for a more effective subject search across the different databases.

This capability has been effective in enabling users to map their query to the controlled vocabularies (subject headings) used in descriptive metadata; much more so than traditional Boolean methods. But a greater (and unanticipated) benefit may be that we are now able to cross-search different thesauri and automate associations between them and the user's inquiry.

It specifically addresses the critical issue of "vocabulary control" by supporting probabilistic "best match" ranked searching (as discussed below) and support for "Entry Vocabulary Modules" (EVMs) that provide a mapping between a searcher's natural language and controlled vocabularies used in the description of digital objects and collections.

Classification clustering technique

The techniques of "Classification Clustering" use natural language parsing software to identify phrases in the language of the users of bibliographic databases, taken from the titles and abstracts in the literature to be searched, and then apply statistical association techniques to associate these words and phrases with the metadata terms of the target.

In a two-stage search method developed in the Cheshire prototype, the system uses probabilistic "best match" techniques to match a user's initial topical query with a set of *classification clusters* for the database, so that the clusters are retrieved in decreasing order of probable relevance to the user's search statement. This aids the user in subject focusing and topic/treatment discrimination.

The classification clustering method involves merging topical descriptive elements (title keywords and subject headings) for all MARC records in a given Library of Congress classification. The individual records are clustered based on a normalized version of their class number, and each such *classification cluster* is treated as a single 'document' with the combined access points of all the individual documents in the cluster. "Normalisation" of the class number involves converting it into a standard format containing the topical portion of the LCC number, and removing individual "book numbers", dates, and copy-level information. The title and subject heading information for all documents in each normalised class are merged to provide the frequency information used to generate the probabilistic term weights, and the vector representation of the classification. The clusters can be characterised as an automatically generated pseudo-thesaurus, where the terms from titles and subject headings provide a lead-in vocabulary to the concept, or

topic, represented by the classification number. The method used to retrieve and rank the classification clusters is based on a probabilistic retrieval model.

classification clustering method developed for Cheshire system overcame one of the major problems of using MARC records with advanced retrieval methods, that is, the limited topical information available in the record (generally only a title and a small number of subject headings), by automatically grouping terms derived from the same classification area

Effectiveness of Cheshire Clustering Approach

There were two papers discussing issues relating to the efficiency and effectiveness of the Cheshire system.

² In the first paper Larson describes the retrieval evaluation of Cheshire in a test collection of 30,000 records mainly in the Library of Congress class Z (library and information science) and using 10 test queries. The results showed that the use of classification clusters for query expansion in conjunction with probabilistic partial-match techniques and full stemming was found to provide the best performance for the online catalogue database and test queries.

In another evaluation, Larson examined the performance of the Cheshire system in comparison with a control system called ZPRISE as part of the TREC (Text Retrieval Conference) investigations ³. The results indicated that the Cheshire system showed poorer performance in terms of recall and precision as compared to the control system i.e. ZPRISE. However, it should be noted that this evaluation was based on TREC test collections. No report was found about the effectiveness or efficiency of the system in a distributed resource discovery or in relation to collection level or item level retrieval from users' point of view.

Cheshire system availability

Cheshire source code is freely available on the Web for use by academic or non commercial organisations. The set up instructions and tutorials are also available on the web. (<http://cheshire.lib.berkeley.edu/>)

Web-based service using the Cheshire system

In order to gain an insight into the ways in which the subject searching techniques used in the Cheshire system in particular the 'classification clusters', two web-based services which use the system were examined. These two are MerseyLibraries.org and Archives Hub.

Examples: (1) MerseyLibraries.org

MerseyLibraries.org is a website developed and maintained by the Libraries Together: Liverpool Learning Partnership. The website allows for searching across 12 academic and public libraries in Merseyside.

² Larson, Ray R. (1992). Evaluation of Advanced Retrieval Techniques in an Experimental Online Catalog. *Journal of the American Society for Information Science*, 43(1), p. 34-53.

³ Larson, R. (2001) TREC interactive with Cheshire II. *Information Processing and Management* 37(3): 485-505.

When a user enters a term such as 'Thesaurus' and chooses the subject search option from the drop down list, the system brings back a number of hits from different libraries.

MerseyLibraries Results - Microsoft Internet Explorer

Address: http://www.merseylibraries.org/~cheshire/cgi-bin/distibsearch.cgi

MerseyLibraries Results

Search: Title Keywords [Home](#)

Liverpool John Moores University:53 [more](#)

Full	Title	Author
1	Moys classification and thesaurus for legal materials	Moys, EM
2	Thesaurus of archaeological site types	Royal Commission on Historical Monuments (England)
3	LCSH and PRECIS in library and information science, a comparative study	Tonta, Y
4	Indexing from A to Z	Wellisch, HH
5	Subject control of film and video, a comparison of three methods	Mallet, L

Liverpool Community College:13 [more](#)

Full	Title	Author
1	Collins Roget's International Thesaurus	Chapmar, Robert L
2	The Oxford School Thesaurus	Spooner, Alan
3	Roget's International Thesaurus	Chapmar, Robert L
4	Collins wordfinder: the ultimate thesaurus from A to Z	Gilmour, Lorna
5	In other words: a thesaurus of euphemisms	Neaman, Judith

Liverpool Institute for Performing Arts:4

Full	Title	Author
1	Chambers combined dictionary/thesaurus	Manser, Martin
2	Roget A to Z	Chapmar, Robert L
3	Roget's international thesaurus	Chapmar, Robert L
4	The Oxford thesaurus, an A - Z dictionary of synonyms	Urdang, Laurence

If the user click on any of the retrieved item, details of the item together with a list of subject headings or 'classification clusters' on the left side of the page is offered to the user to choose form.

MerseyLibraries Results

Search: Title Keywords [Home](#)

Browse Author
Tonta, Y

Browse Title
[LCSH and PRECIS in libr...](#)

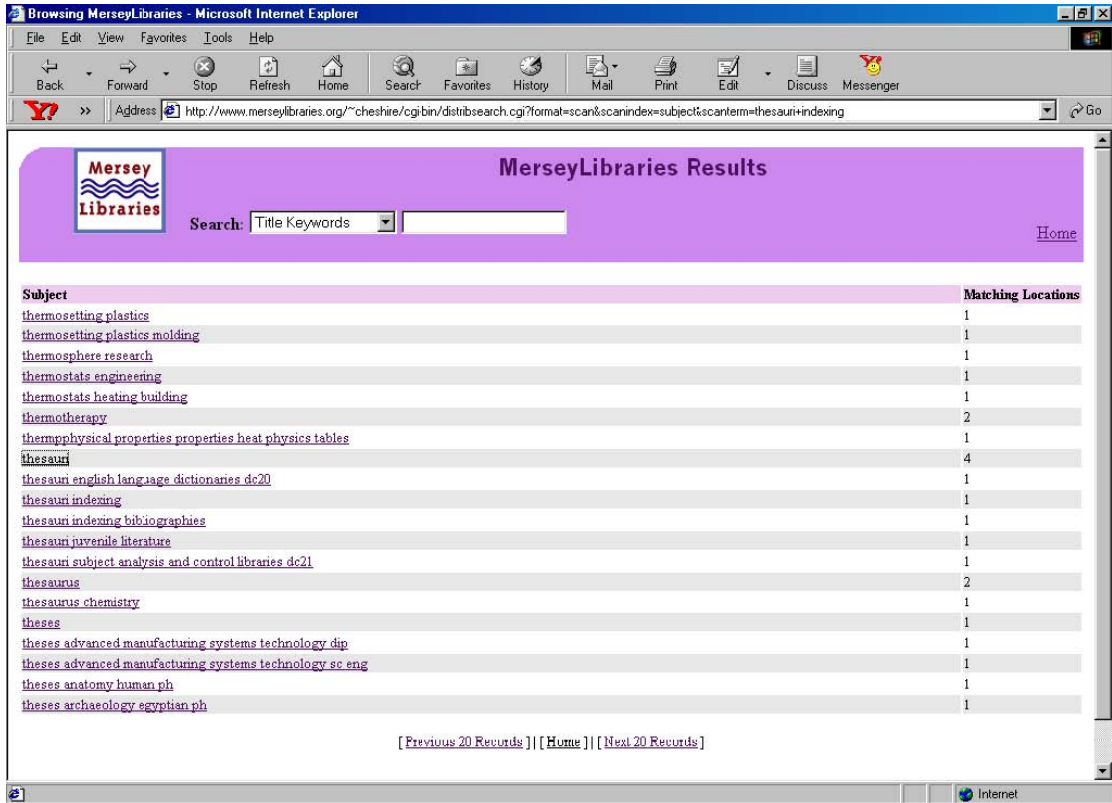
Browse Subjects
[Subject indexing](#)
[CITATIONS: INDEXING](#)
[INDEX LANGUAGES: INDEXING](#)
[INDEXING](#)
[INTERMEDIATE LEXICON:](#)
[INDEXING](#)
[KEYWORDS: INDEXING](#)
[LIBRARIANSHIP: KEYWORDS:](#)
[INDEXING](#)
[MULTILINGUAL THESAURI:](#)
[THESAURI: INDEXING:](#)
[INFORMATION RETRIEVAL](#)
[SEARS LIST: SUEJECT HEADINGS:](#)
[INDEXING](#)
[THESAURI: INDEXING](#)
[URBAN INFORMATION:](#)
[THESAURUS](#)

LCSH and PRECIS in library and information science; a comparative study
by Tonta, Y

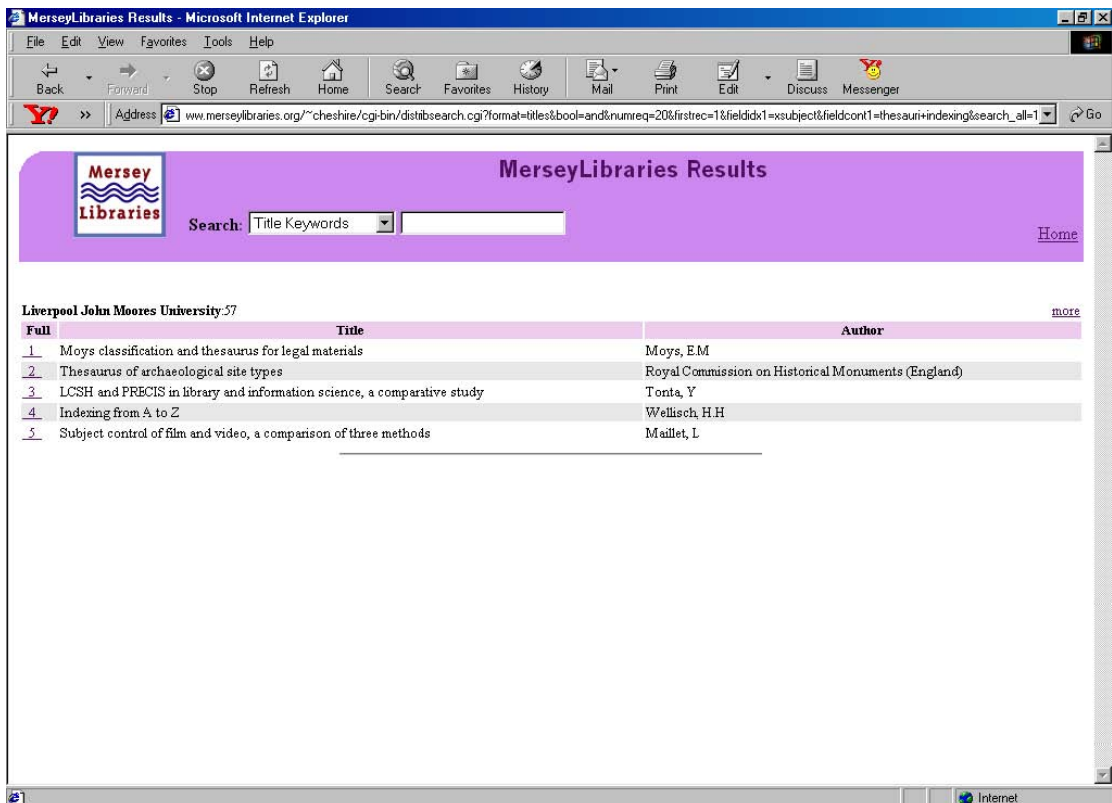
Series: Illinois Univ. Graduate School of Library Science. Occ. pap., 194
Publisher: University of Illinois (1992)
Subjects: Subject indexing
 CITATIONS: INDEXING
 INDEX LANGUAGES: INDEXING
 INDEXING
 INTERMEDIATE LEXICON: INDEXING
 KEYWORDS: INDEXING
 LIBRARIANSHIP: KEYWORDS: INDEXING
 MULTILINGUAL THESAURI: THESAURI: INDEXING: INFORMATION RETRIEVAL
 SEARS LIST: SUBJECT HEADINGS: INDEXING
 THESAURI: INDEXING
 URBAN INFORMATION: THESAURUS

Shelfmark: 029.5 TCN
Location: Liverpool John Moores University -- Aldham Roberts LRC

Upon clicking on any of the left hand side terms, the user is provided with an alphabetical list of subject headings in which the selected term is located.



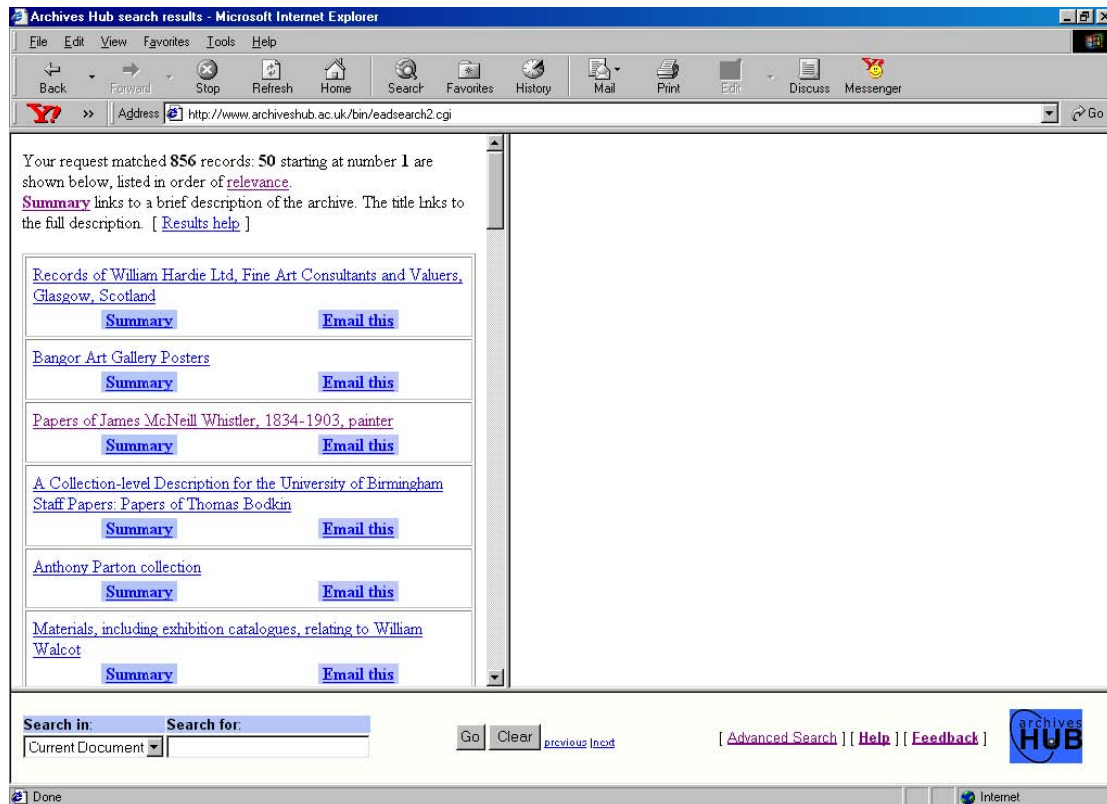
Clicking on any of the subject headings will lead the user to results related to that particular subject heading.



Examples: (2) Archives Hub

Archives hub is a national gateway to descriptions of archives in UK universities and colleges.

When a user inputs a term for instance 'art gallery', the system brings back a list of retrieved items.



The screenshot shows a Microsoft Internet Explorer browser window displaying search results from the Archives Hub website. The address bar shows the URL: <http://www.archiveshub.ac.uk/bin/eadsearch2.cgi>. The main content area displays the following text:

Your request matched **856** records: **50** starting at number **1** are shown below, listed in order of [relevance](#).
[Summary](#) links to a brief description of the archive. The title links to the full description. [[Results help](#)]

The search results are listed in a table-like format with the following entries:

Records of William Hardie Ltd, Fine Art Consultants and Valuers, Glasgow, Scotland	Summary	Email this
Bangor Art Gallery Posters	Summary	Email this
Papers of James McNeill Whistler, 1834-1903, painter	Summary	Email this
A Collection-level Description for the University of Birmingham Staff Papers: Papers of Thomas Bodkin	Summary	Email this
Anthony Parton collection	Summary	Email this
Materials, including exhibition catalogues, relating to William Walcot	Summary	Email this

At the bottom of the page, there is a search bar with the following elements:

- Search in:
- Search for:
- Go Clear [previous](#) [next](#)
- [[Advanced Search](#)] [[Help](#)] [[Feedback](#)]
- Archives HUB logo

If the user chooses one of the titles, details of that particular record will appear on the right hand side of the interface. At the end of each record there is section called 'access points' where other terms related to the user's query appear.

Archives Hub search results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss Messenger

Address http://www.archiveshub.ac.uk/bin/eadsearch2.cgi

Your request matched **856** records: **50** starting at number **1** are shown below, listed in order of [relevance](#).
[Summary](#) links to a brief description of the archive. The title links to the full description. [[Results help](#)]

Records of Wilham Hardie Ltd, Fine Art Consultants and Valuers, Glasgow, Scotland	Summary	Email this
Bangor Art Gallery Posters	Summary	Email this
Papers of James McNeill Whistler, 1834-1903, painter	Summary	Email this
A Collection-level Description for the University of Birmingham Staff Papers: Papers of Thomas Bodkin	Summary	Email this
Anthony Parton collection	Summary	Email this
Materials, including exhibition catalogues, relating to William Walcott	Summary	Email this

Whistlers and further family: an exhibition of portraits and pictures, manuscripts and mementos relating to the family of James McNeill Whistler (Glasgow, 1980)

Whistler and Mallarmé (Glasgow, 1973) Exhibition catalogue, Hunterian Museum

Notes

Date(s) of Description

Compiled by David Powell, Hub Project Archivist, 22 March 2002

No alterations made to date

Access Points

[Arts](#)
[Artistic creation](#)
[Artists](#)
[Letter writing](#)
[Library collections](#)
[Whistler, James Abbot McNeill. \(1834-1903 \) Painter](#)

Search in: Search for:

Current Document [previous](#) [next](#) [[Advanced Search](#)] [[Help](#)] [[Feedback](#)]

Internet

If the user then selects one of the terms, he will be led to a page called 'subject browsing' where he can see and choose from an alphabetical list of terms including the term selected in the previous stage. It also allows the user to jump to previous or next page of subject terms.

Archives Hub search results - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Back Forward Stop Refresh Home Search Favorites History Mail Print Edit Discuss Messenger


Address http://www.archiveshub.ac.uk/bin/eadsearch2.cgi?format=sgmlscan&bool=AND&maxrecs=20&firstrec=1&fieldid1=ssubject-a&fieldcont1=Artists Go

Subject Browsing

Your request was submitted and has matched the following subjects. Click on the name to see the finding aid, or to list the finding aids if it occurs in more than one.

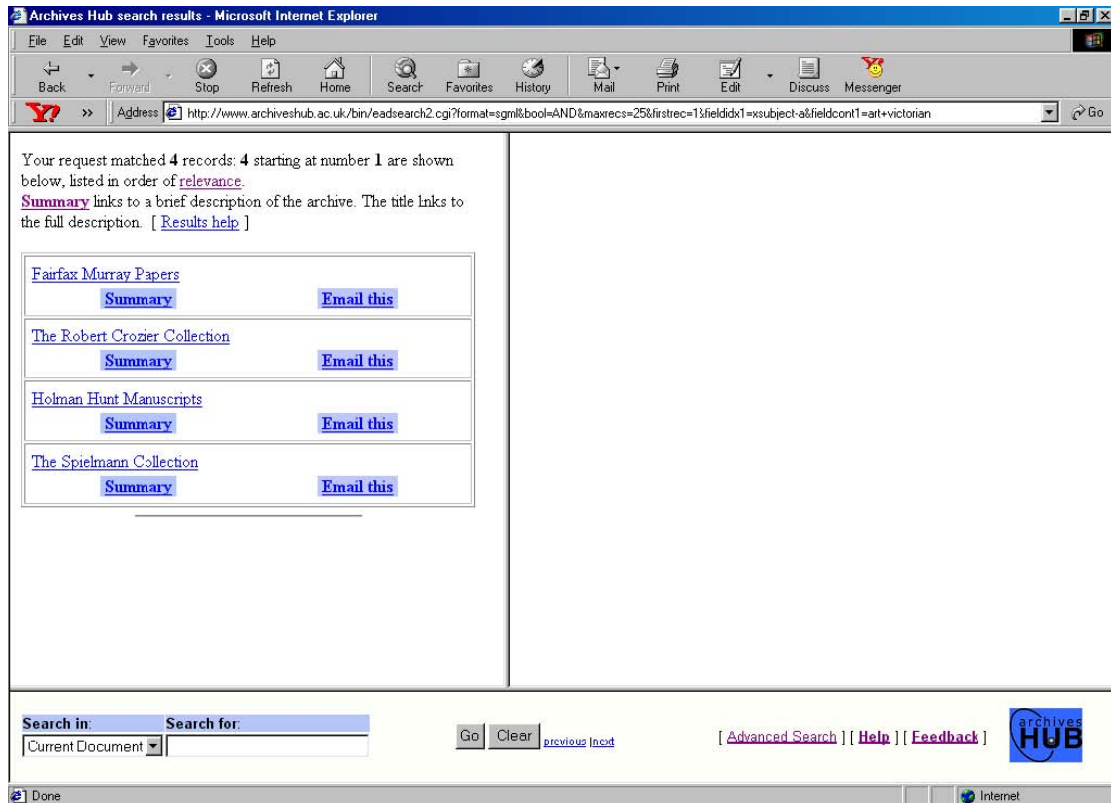
art patronage	(1 match)
art russian	(12 matches)
art theory	(2 matches)
art victorian	(4 matches)
arthur miller centre for american studies	(1 match)
artificial satellites	(1 match)
artificial satellites russian	(1 match)
artillery	(2 matches)
artillery drill and tactics	(2 matches)
artistic creation	(8 matches)
artists	(43 matches)
artists british	(2 matches)
artists european	(1 match)
artists irish	(1 match)
artists russian	(2 matches)
arts	(45 matches)
arts and crafts	(2 matches)

Search in: Search for:

Go Clear [previous](#) [next](#) [[Advanced Search](#)] [[Help](#)] [[Feedback](#)] 

Done Internet

Finally if the user selects one of those terms, he will be shown a page with retrieved items by that particular subject term.



Notes on RLG's RedLightGreen project (<http://www.rlg.org/redlightgreen/>)

This project investigates the ways in which the RLG union catalogue can be tailored to suit the needs of undergraduates and the public. Its aim is to The catalogue covers over 126 million bibliographic records representing 42 million descriptions of books, maps, films, recordings, and manuscripts from 300 countries, in over 370 languages.

Funded by a grant from The Andrew W. Mellon Foundation awarded in March 2002, this RLG initiative seeks to:

- support the discovery of authoritative sources for students, scholars, and researchers
- create an entry point to the larger range of Web and library resources
- increase the presence of library resources on the Web

To simplify record retrieval for Web users, RLG has adapted the Functional Requirements for Bibliographic Records established by the International Federation of Library Associations and the Library of Congress, which distinguishes between a *work*, an *expression*, a *manifestation*, and an *item*. Data manipulation using this approach will aggregate what can be an overwhelming number of editions into a manageable set of works that match a user's search terms.

To take advantage of the RLG Union Catalog's depth and breadth of content, RLG is also using Recommend Inc.'s MindServer technology to find more subject correlations between works and increase retrieval. Through subject heading variations, relationships, frequency in collections, international content, and

searching assistance, users will discover information that has been difficult or impossible to find.

An example of the functionalities of RedLight Green is as follows. A student might enter a search for the keywords "Civil War" without specifying the American, Spanish, or other civil wars. Using Recommind, RedLightGreen can organize the results in clusters of related items, letting the student pick which civil war interests her. At the same time the application can insert more specific, scholarly subject classification terms into the search that have been derived from the MindServer data.

Questions a Follow Up Clustering Study Should Examine

An Item Level Tool

It is assumed that clustering might be a useful tool for item level searching within collections. It would, for example, potentially have the following useful advantages:

- Would help user retrieve by subject where the user's term was in the metadata describing the collection but not in the UK term set
- Would help user retrieve by subject where the user's term was in the UK term set but the subject scheme used by the collection was not mapped to in the terminologies server
- It would improve the user's search strategy by suggesting similar and alternative terms to use
- To what extent the disambiguation and contextualisation of subject terms can be incorporated in the clustering approach?

Questions:

- How successful is it as a method of improving subject retrieval?
- How often would the user term be there?
- To what extent the clustering technique contributes to the retrieval of subject terms not similar to those of users?
- How relevance and recall and precision of subject terms would be improved?

A Collection Level Tool

Could clustering type approach replace HILT approach altogether?

If you had a single clustering index for all services in the world (or a distributed system that behaved as a single index would), user term would often (but not always) be there and provide user with an entry to the subject clusters. But a number of questions need to be answered by hard research:

1. Is the idea feasible?
2. How often would the user term be there?
3. Would the technique scale to cover all of these services, or would the results become confusing to the user?
4. Would the end result be a more interoperable subject universe? Or would reliance on the technique reduce attempts to keep subject descriptions compatible?
5. If it did undermine subject interoperability generally, would the effectiveness of the technique deteriorate as general subject interoperability deteriorated, so that the technique itself ultimately failed?
6. Does the technique work better if subject structure exists already?
7. If it does, does its efficacy vary from subject area to subject area (is it as effective in the arts as it is in an area like medicine?). Is it better for databases with well structured standard scheme subject descriptions and poorer for DIY approaches
8. To what extent the clustering approach can be used to assess the collection level strength?

Appendix G: HILT and the IE Registry

1. Points to note:

1. What we are trying to do in telling you about these 'HILT requirements' is warn you that the HILT Phase II final report is likely to note that a JISC Terminologies Service will need the operational JISC IE Registry to meet these requirements. We do not need the IE *pilot* to do these things. We are simulating this stuff for the purposes of the HILT pilot. These are possible future requirements for you to note rather than necessarily functionality you need in the IER pilot.
2. HILT is still in progress and will not end until September 2003. Some of the requirements we specify may change and others may be added. There is nothing we can do about this except keep you informed about changes as soon as we know for sure.

2. Requirements

1. We are looking to have the collections in the registry classified by DDC to a granularity level appropriate to the subject coverage of the collections included (e.g. a collection covering chemistry would be classified at the number for chemistry rather than the number for science).
2. In addition to having the collections classified in this way, we need information held on which subject schemes and classification schemes the service uses to describe items in the collection. Sometimes there can be more than one scheme and we should allow for more than two subject schemes and more than two class schemes just in case.
3. We believe we may also need to be able to specify which version of a scheme is in use (e.g. DDC20, DDC21 and so on).
4. A unique identifier for a collection is another probable requirement and is a recommendation of the CC-Interop report compiled by Gordon Dunsire.
5. We may need to be able to distinguish sub-collections of a collection to provide an optimal level of service to users in respect of some services. Treating these as individual collections is another way round this but this may have implications for how the IE Registry operates in other contexts. On the other hand, treating sub-collections as individual collections *might* obviate the need specified at 2(2) above to have more than one scheme recorded against a collection.

Appendix H: Cost-Benefit Analysis Report

Appendix H has three sub-sections:

Section 1: HILT Cost Benefit Analysis Exercise: Framework and Notes **Page 2**

This is the text used to guide the participants of the cost-benefit analysis process and describes the approach taken in carrying out this exercise at the HILT Steering Group meeting of 18th September 2003.

Section 2: HILT Cost-Benefit Analysis Exercise: Tables **Page 7**

These are the tables used when carrying out the process. Costing were added by the HILT team prior to the meeting of 18th September. The mapping of high level benefits to JISC objectives had been agreed at an earlier SG meeting. The basis of the estimated costs is included at the end of the tables.

Section 3: HILT Cost-Benefit Analysis: Results and Conclusions **Page 18**

This is a report on the exercise and the conclusions drawn

Note:

An early draft of the Final Report of the project was provided to the HILT Steering Group to provide background information during the cost-benefit analysis process.

Section 1: HILT Cost Benefit Analysis Exercise: Framework and Notes

1 Using the INSIGHT Model For HILT Purposes: Overview

It was agreed at earlier meetings of the SG and the PMG that the INSIGHT¹ model be used to carry out a cost-benefit analysis of HILT options.

The value of the model – and its likely usefulness to HILT – lies as much in its ability to provide a framework for helpful discussion of the issues as in any final outcome. The likely endpoint for project usage of it is two or three illustrative outcomes (possibly with one or other given particular support by the project and project groups) and a spreadsheet or similar mechanism to permit JISC to explore the effects of varying some of the variables. This is sensible since the values we put on many of them will be based on informed but nevertheless subjective stakeholder judgements.

It is important to the process that evaluators (in this case the HILT SG) are aware of what decision the evaluation will be supporting – in this case ‘Assessment of future options for investment’ on the terminologies server front. As agreed by the Steering Group, HILT will adopt the marginal costing rather than the full costing method but will note instances where evaluators feel there may be wider effects from changes to marginal costs. INSIGHT recommends the full costing approach, but the marginal costing was agreed to be appropriate for the HILT Cost-benefit analysis (CBA). Marginal costing looks only at those costs the evaluator feels are affected by the decision – so, for example, we have assumed below that the evaluation would not include the cost of supporting JANET which would be required regardless of which HILT option were adopted.

2 Two Stage Process: INSIGHT and Instantiation

Subsequent to the Steering Group meeting that discussed the approach to be taken in the CBA, the HILT team realised that a two-stage process was necessary:

- INSIGHT CBA of different levels of terminology server functionality, each associated with different cost and benefit levels (does it have a DDC to LCSH mapping, does it have a disambiguation function, and so on)
- Subsequent analysis of different approaches to the instantiation of the terminology server (developed from HILT pilot or from scratch, developed commercially, developed by every local portal and so on)

A section describing an additional post-INSIGHT process has therefore been added to this document. This two stage process is not only logically necessary to a sensible decision making process, it also reflects what was proposed in the bid document and elsewhere.

3 Summary of Steps to be Undertaken by HILT SG

INSIGHT CBA

INSIGHT evaluation steps:

Determine the Evaluation Period

It was agreed that this should be five years (starting with the 2003/04 financial year). It was also suggested that there be a breakdown into specific phases (e.g. Building phase, incorporation by portals etc. within this period). The team were not clear how to implement this and have not done so, although a certain amount of phasing is built into the structure of the cost elements table (table 1).

Take a Decision on the Staff Categories to Use

It was agreed that one category was sufficient and suggested a salary level of 33K be used.

¹ See the project web-site at See <http://www.mis.strath.ac.uk/predict/projects/insight/index.htm>.

Take a View on Overheads

It was agreed that the marginal costing method would be adopted but that the report would note any area where the SG members felt there might be significant implications for wider costs if there was a variation in marginal costs. It is suggested that this be done in conjunction with the SG during the cost-benefit analysis process.

Identify the ‘Value Added Activities’ to be Assessed

The value added activities are the options to be evaluated, options A-H in the final row of table 1. As

indicated above, the best approach here appears to be to compare costs and benefits of differing levels of functionality in the first instance then compare likely effects on costs and benefits of differing instantiation methods. The options and their estimated costs are also in table 1.

Identify the Related Support Activities

These are the activities required to support the options to be evaluated – the cost elements listed in table 1.

Identify Costs

Actual costs associated with these activities or cost elements - to include staff, revenue (including recurrent) and capital. These are also detailed in table 1.

Identify benefits and relate benefits to strategic objectives

See table 2 for a list of high-level benefits and a mapping of these to JISC strategic objectives. See table 3 for a breakdown of the high level benefits into benefit elements. Benefit elements have a direct association with functionality levels in table 1. Shading in the benefit elements column of table three shows how benefit levels group together in functionality level groupings.

Carry Out Benefits Evaluation

This process is dealt with in section 4 below. Table 3 is utilised for this.

Conduct INSIGHT Evaluation (Calculation of cost-benefit ratios)

This process is dealt with in section 4 below. Table 4 is used to record results.

Additional Process: Comparison of Instantiation Options

This process is dealt with in section 5 below

4 Stage One Details: INSIGHT CBA Process

The assumptions that underpin HILT's analysis of required design elements on the one hand (see table 1 – cost elements) and associated benefits and benefit elements on the other (see table 3) are stated in Section 3 of the draft Final Report. This should be read before conducting the INSIGHT cost-benefit analysis

The aim of the INSIGHT process is to establish a cost-benefit ratio for each of the options (e.g. versions with or without LCSH mapping or the disambiguation functionality). The higher the final cost-benefit ratio (for a particular costs/ benefits/ strategic objectives scenario) the better the investment (at least for that scenario). ***The aim of the process described below is to calculate the cost-benefit ratios for the various options and their associated levels of functionality (and to look at some variations on this).***

Experiment 1: Compare INSIGHT Cost-benefit ratios of Options A-H 'as is'

Notes: Options A-H vary according to the mix of functionality groupings they include (numbered 0-4 in Table 1). The mix for groupings for A-H is specified at the end of table 1.

Process:

- Examine table 1. Discuss and agree cost elements and costs as detailed by HILT. Copy results into table 4.
- Briefly discuss table 2 which lists high-level benefits and maps benefits to JISC objectives. Benefits must be agreed with the decision making group and should reflect the strategic objectives of the organisation seeking to make the decision (in this case JISC).
- Discuss the breakdown into benefit elements and shaded/unshaded functionality groupings in table 3.
- Calculate total weighted benefit scores for each of the options A-H 'as is' using table 3:
 - o Discuss the 'do nothing' option (particularly table 3)
 - o Give high-level benefits a weighting of 1-1000 (Gw), but ensure that the total sum of weightings against all benefits adds up to 1000. If possible, come to a collective

- o agreement on weights.
- o Agree weightings for shaded/unshaded functionality *groupings* within each high level benefit (Fw) (sum of Fws under each high level benefit should sum to Gw weighting for that benefit and sum of all Fws should add to 1000 (Gw sum)). If possible, come to a collective agreement on weightings for functionality groupings under each high level benefit.
- o Assign score (Es) of 1 for each functionality *grouping*
- o Score options A-H accordingly (Fw*1 for each functionality grouping included in an option, Fw*0 otherwise). Since we are dealing with levels of functionality, Fw x Es for any given functionality grouping be the same for every option that shows the benefit
- o Calculate the sum of weighted scores for each option and record in last row of table 3.
- o Copy the results into table 4. Calculate cost-benefit ratio for options A-H 'as is'
- If necessary, agree basis of sensitivity analysis and apply sensitivity analysis (Look at various variations of the above based on different assumptions on benefit and weighting and costs scores – the point is to facilitate decision-making rather than necessarily to arrive at one fixed cost-benefit ratio and treat it as the only possible result)

Notes:

- Options vary according to the functionality groupings they entail, so a grouping is either in an option or out, making a score of 1 sufficient for each grouping. Options with the grouping score Fw*1, options without it score Fw*0 (for that grouping).
- Gw weightings necessary because one high level benefit may be assessed as being more important to JISC objectives than another
- Fw weightings necessary because one functionality grouping may be assessed as being more important to a particular high level benefit than another is (e.g. disambiguation group may be more important to user searching benefit than MeSH grouping) and because they each have different associated costs

Experiment 2: Compare INSIGHT Cost-benefit ratios of Options A-H with selected costed benefit elements removed

Process:

- Use same Gw as above
- Agree benefit element weightings (Ew) for benefit elements within each functionality grouping under each high level benefit (these sum to Fw)
- Explore the effects of removing selected benefit elements and their associated costs out of options A-H (reduce Fw by Ew and redo calculation with appropriate reduction in cost)
- Perhaps also examine cost-benefit ratios of selected individual elements

Notes:

- Ew weighting necessary because individual benefit elements in a functionality grouping may be assessed as more or less important than each other and because they have associated costs

General Points To Note:

1. **This is an exploratory exercise to see what we can learn about costs against benefits of different functionality levels. Our assessment of benefit weightings will be 'fuzzy' and our assessment of costs is very rough. The ratios we finally hit on won't be that meaningful in themselves, but the ranking may be useful. In general what we learn through the exercise will be more important than the numbers we come up with.**

2. I am aware that some of the cost elements listed under what I am calling ‘functionality groupings’ relate to research work or project costs. However, the term ‘functionality groupings’ is still the best label for these groupings, even if some of what they entail are not functionality elements.

Notes (from minutes of last SG meeting):

It was agreed to consider the use of normalised weightings that sum up to (e.g.) 100. This was thought to make the figures easier to process. It was also thought that it might be necessary to split some benefits into more detailed assessment criteria, but noted that if benefits are split up into individual assessment criteria in this way, there is a tendency to weight them higher than if they were treated as a single benefit. Care needed to be taken to ensure this did not skew results. It was thought that dealing with different elements of a benefit in a separate spreadsheet and feeding the results back into the core analysis would help prevent this. By handling elements as a separate process, differences in agreement and/or understanding would be highlighted. There may also be a need for ‘sensitivity analysis’ on occasion – an attempt to investigate what is needed to alter the outcome significantly (eg. Dropping off of highest and lowest scores to see the effect on the outcome). If, for example, small changes in the assigned weights completely change the result, the analysis is not secure. It would be deemed robust if significant changes were required to change the outcome. It is also important to avoid stacking the questions to get the desired result. It was decided that criteria and weightings be presented before asking people to score options independently. It was also noted that costs are not always financial, with time take for a task being one example. Since CR had some experience in the area, it was agreed that DN should consult him on the final approach (this was done).

5 Stage Two Details: Comparison of Instantiation Possibilities

This will examine likely effects on costs and benefits of adopting one of the following methods of instantiation. 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

[NB: The following is a ‘guesstimate of the likely conclusion from the HILT team]

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 term 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

Section 2: HILT Cost Benefit Analysis Exercise: Tables

Table 1: Options, Cost Elements, and Costs in £K

Options	Cost elements	Exclude?	Staff Costs	Capital Costs ²	Revenue Costs	Total Element Cost	Total Option Cost
'Do Nothing'³ Option [0]	Cost of fixing deteriorating interoperability if nothing is done		£600,000 ⁴				
	Base Mapping Option [1]			£25,000	£15,000		
	Server and other equipment						
	Software licensing			£65,000	£75,000		
	Mappings database		£8,250				
	DDC Licensing			£6,000	£4,800		
	DDC Processing		£8,250				
	LCSH Licensing (included with DDC)					£0	
	LCSH mapping (included with DDC)					£0	
	UNESCO Licensing			£31			
	UNESCO mapping		£17,600				
	UK term set creation		£57,000				
	UK terms mapping		£76,000				
	RDN terminologies harmonisation study		£40,000				
	RDN-based clustering tool study		£60,000				
	Interface needs user study (enhanced pilot with clustering)		£40,000				
	Term match facility		£8,250				

	Staff amend maps facility	£16,500	
	Staff training module	£11,000	
	Online user training module	£33,000	
	Ability to host and map other schemes	£8,250	
	Ability to interact with other mapping services	£16,500	
	Processes to cope with scheme updates		£13,750
	Project management costs	£250,000	
	Training	£3,760	
	Publicity	£1,100	
	Marketing	£1,285	
	Redevelopment		£20,625
Base Services Option [2]	Disambiguation facility	£16,500	
	DDC collection identifier	£5,500	
	Any hits test/rank facility	£5,500	
	User terms monitor	£5,500	
	Additional Training	£3,760	
	Additional Publicity	£1,100	
	Additional Marketing	£1,285	
	Additional Redevelopment		£5,000
Other schemes option [3]	Mesh Licensing	£0	
	Mesh mapping	£87,892	
	AAT Licensing		£460
	AAT Mapping	£500,000	
	Additional Training	£3,760	
	Additional Publicity	£1,100	
	Additional Marketing	£1,285	

	Additional Redevelopment		£5,000
UK extensions option [4]	Regional terms creation	£57,000	
	Regional terms mapping	£76,000	
	Additional Training	£3,760	
	Additional Publicity	£1,100	
	Additional Marketing	£1,285	
	Additional Redevelopment		£0
Option A= 0			
Option B= 1			
Option C=1+2			
Option D=1+3			
Option E=1+4			
Option F=1+2+3			
Option G=1+2+4			

Option
H=1+2+3+4

² It was noted that capital costs have to consider depreciation, but this should not be an issue over a five year life cycle.

³ Leave things as now; users can cope and another service would confuse

⁴ Assumes 30 services, 1000 records a year, £4 a record to do subject stuff, 5 years

Table 2: High Level Benefits Associated with Relevant JISC Objectives

BENEFIT TYPE⁵	JISC Objectives	JISC Recommendations
Improved user ability to formulate and execute successful subject searches	All five high level benefits are relevant to all 6 of the selection of JISC objectives listed below	Applies to recommendations 3, 11, 12, 23 below
Improved staff ability to provide quality subject descriptions appropriate to needs of JISC users	Applies to recommendations 3, 11, 15 below	
Improved mapping of user subject queries to staff subject descriptions of items within and beyond UK HE and FE	Applies to recommendations 3, 11, 12, 15, 23 below	
Ongoing basis for a process that will halt deterioration in, and begin to monitor and improve, interoperability in respect of subject description of resources	Applies to recommendations 3, 12, 13, 15 below	
General improvement in JISC collection development and utilisation activities	Applies to recommendations 3, 12, 13, 16 below	

Relevant Key objectives:

- a. Build an online information environment providing secure and convenient access to a comprehensive collection of scholarly and educational material
- b. Ensure the continued provision of, and wide access to, a world-leading network to support education and research in the UK
- c. Promote innovation in the use of ICT to benefit learning and teaching, research and the management of institutions
- d. Improve staff and student skills in the exploitation of ICT, particularly in their use of the Internet
- e. Provide a focus for collaboration between UK educational IT initiatives to help create a wider information-literate society
- f. Promote and facilitate international collaboration in the exploitation of ICT

Relevant Recommendations:

3. The JISC will encourage international collaboration between organisations and agencies building significant Internet information environments to promote coherence in the global exchange of such resources.

Users/services can
conduct own
disambiguation
with information
from system
System has training
module on the
above
System
development in this
area informed by
RDN study
System
development in this
area informed by
clustering study
System
development in this
area informed by
user study
Service offers
disambiguation
facility
Service offers find
collections facility
Service offers any
hits test/rank
facility
Service offers
additional training
on disambiguation,
find collections and
any hits facilities
System can
recognise terms
used by UK users
not in standard
schemes and advise
directly and via
M2M on terms
used by services
using MESH and
AAT
System can
recognise terms
used by UK users
from MESH and
AAT and advise
directly and via
M2M on terms
used by services
using these
schemes

System can recognise terms from MESH and AAT and advise on UK terms that may have been used in legacy metadata
Service offers training on MESH, AAT
System can recognise terms used by UK regional users not in standard schemes and advise on terms used by services using DDC, LCSH, or UNESCO terms
System can advise on regional UK terms possibly used in legacy metadata
Service offers regional terms training

Improved staff ability to provide quality subject descriptions appropriate to needs of JISC users

Online source for DDC, LCSH, and UNESCO standard terms and their use

Source for DDC classification scheme
UK level source of terms needed for UK HE and FE retrieval but not in standard schemes
Mechanism for adding to this in a coordinated (and, hence, interoperable) way at UK level
System capable of assisting staff in local services with legacy metadata

problems in service
subject descriptions

Staff training in all
of the above
System
development in this
area informed by
RDN study
System
development in this
area informed by
clustering study
System
development in this
area informed by
user study
Feedback on how
other staff have
used the term sets
and class numbers
in the above group
to describe items
System provides
feedback on terms
users use to search
for subject
information
Staff training on
feedback
mechanisms
Online source for
MESH and AAT
standard terms and
their use
Feedback on how
others have used
AAT and MESH
term sets to
describe items
MESH and AAT
training
Regional level
source of terms
needed for UK HE
and FE retrieval
but not in standard
schemes.
Mechanism for
adding to this in a
coordinated (and,
hence,
interoperable) way
at regional and UK
levels

Improved mapping of user subject queries to staff subject descriptions of items within and beyond UK HE and FE

Mapping of terms used by users to terms used by collections utilising DDC, LCSH or UNESCO where user terms are in UK or DDC, LCSH or UNESCO term sets and service uses schemes without change

Mapping of terms used by users to terms used by collections whose legacy metadata follows the pattern of the UK terms set

Information on user-driven disambiguation process

Training in the use of these facilities

System development in this area informed by RDN study

System development in this area informed by clustering study

System development in this area informed by user study

Online disambiguation function

Automated JISC collections identifier

Feedback on which terms to use in identified services where these use DDC, LCSH or UNESCO term sets

Sample retrieval from these services based on the recommended terms

Training in the use of these facilities

Above facilities
extended to cover
collections utilising
MESH and AAT
and terms from
these used by users
Training in the use
of these facilities

Above facilities
extended to cover
user terms in UK
regional terms set
and the use of such
terms in services
legacy metadata
Training in the use
of these facilities

Ongoing basis
for a process that
will halt
deterioration in,
and begin to
monitor and
improve,
interoperability
in respect of
subject
description of
resources

Central service that will
facilitate improved
interoperability in respect
of the use of DDC,
LCSH, UNESCO, and
UK variations in their
use, and that will begin
to solve legacy metadata
problems whose source is
the creation of UK
variations on standard
terms generally

Extension of
interoperability to
other schemes
through ability to
interact with other
mapping services

Extension of
interoperability to
other schemes
through provision
of facility to allow
addition of other
schemes

Additional training
where necessary

Improved ability to
influence
development of
subject and class
schemes to meet
JISC needs

Extension of above interoperability improvements to MESH and AAT usage and variations
Training in these extensions
Improved ability to influence development of MESH and AAT subject schemes to meet JISC needs
Extension of above interoperability improvements to cover regional usage and variations to the various schemes
Training in these extensions

General improvement in JISC collection development and utilisation activities

Improved ability of JISC and other staff to monitor or sample subject coverage and identify and deal with collection weaknesses

Improved ability to identify duplication
Improved ability of JISC and other staff to monitor user subject needs as reflected in user searches
Improved ability of lecturers and librarians at institutions to identify and utilise useful materials on behalf of the students they serve
Improved value obtained from expenditure on JISC collections because users alerted to their existence and

subject contents

Final weighted total score for each option (add up totals under columns A-H)

⁶ NB All elements are available via direct user interface and M2M

Table 4: INSIGHT Evaluation of Cost-benefit Ratios

Option	Mix	Description	Five Year Cost (£ K)	Benefits Score	Cost-benefit ratio
A	A	Do nothing option			
B	1	Basic interoperability process created; staff services to support creation of UK terms set, mapping to DDC, LCSH, UNESCO; Direct and M2M user advice on terms in these schemes; staff and user training			

C **1+2** Option B plus direct and M2M
disambiguation, collection finder, sample hits
and collection ranking, user term monitoring,
training

D	1+3	Option B extended to AAT and MESH but without option C
E	1+4	Option B extended to regional variations to UK terms set, but without option C or AAT and MESH
F	1+2+3	All 5 schemes, UK terms set without regional variations, but with disambiguation, collection finder etc
G	1+2+4	DDC, LCSH, UNESCO, and UK terms set with regional variations, disambiguation and related services, but no AAT or MESH
H	1+2+3+4	Everything: all 5 schemes; UK with regional variations, disambiguation and related services

Final task: Instantiation Approach Discussion

Notes on Basis for Costings

Cost types

Software

Perpetual license for Wordmap Enterprise Taxonomy Management System (single server)	costs	65,000
User licenses (Minimum 200)		0
Maintenance (five years)		75,000
Total		140,000

Hardware

Server purchase	Costs	£20,000
maintenance and updates over five years		£3,000
Total		£23,000

License Purchase

DDC		£2,000
UNESCO		£31
AAT		£460
Total		£2,491

Terminology mapping

Costs

UNESCO to DDC	0
MeSH to DDC	0
AAT to DDC	0
Total	0

Staff

Programmer (2 Years)	66000
terminology expert (2 years)	60000
Researcher (2 years)	40000
Administrative	15000
Total	181000

Promotional costs

Staff training (4workshop)	1,880
User training (4workshop)	1,880
Travel (for all workshops)	2,160
Publicity and marketing (1000 brochures)	385
publicity and marketing (100 posters)	2,000
Total	8,305

License	65,000
License per user	50
Maintenance(annual)	15000
Years	5

Cost of mapping per hour	£45
Cost of mapping per term	£4
Time spent per term(minute)	7

UNESCO	4400(64 days)
MeSH	21,973 (320 days)
AAT	125000(7 years)

No of staff workshops	4
No of user workshops	4
Cost of venue per workshop	£150
Cost of lunch per person	£8
Number of participants	320
No of participants per W	40

Travel cost per person	180
number of travelers	12

poster	£20
No of posters	100

Cost of record creation by distributed services

cost of record	£18.00
Checking and maintenance of record	£2.00
Location and assessment (acquisition)	£3.00
Record creation	£3.00
Collection management	£3.00
Development of Subject Specialised Classification (SSC)	£3.00
Management of cataloguing staff	£1.00
Staff training and development	£2.00
Technical support and maintenance	£1.00
Total	£18.00

Average number of records by subject gateways (per year) 2274

Number of JISC participating services and collections 30

SSC for all records in one year £6,822.00

SSC for all records for 30 services/collections £204,660.00

Cost of all records created by all services/collections in 1year £1,227,960.00

Section 3: HILT Cost-Benefit Analysis: Results and Conclusions

The cost benefit analysis was conducted at the steering group meeting on 18th of September. The group began by considering six high level benefits to ensure unanimity on their inclusion in the benefits assessment process. 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.

Benefit Weightings

The group members were asked to weight five high level benefits, summing to a total of 1000. The average of each group member weights was then calculated, ensuring the average weightings retained a total of 1000.

Table 5 shows the five high level benefits with average weightings assigned and ranked from 1 to 5.

Table 5. Ranked benefits with weightings

Ranking	Benefit	Weightings
1	Improved user ability to formulate and execute successful subject searches 1	390
2	Ongoing basis for a process that will halt deterioration in, and begin to monitor and improve, interoperability in respect of subject description of resources 4	190
3	Improved staff ability to provide subject descriptions appropriate to needs of JISC users 2	180
4	Improved mapping of user subject queries to staff subject descriptions of items within and beyond UK HE and FE 3	150
5	General improvement in JISC collection development and utilisation activities 5	90

The benefit judged most favourably by the group (irrespective of cost) was ‘Improved user ability to formulate and execute successful subject searches’, earning a weighting of 390; more than double that of any other benefit.

Functionality Group Weightings

Table 6 shows the average benefit weights previously discussed and how the group distributed these weights between the functionality groups within that benefit. For example the benefit “Improved user ability to formulate and execute successful subject searches” was given an average weighting of 390. The group then reassigned the 390 between the four functionality groupings within that benefit to give functionality weightings. Average functionality weightings were then calculated. These are recorded in the table below.

Table 6: Functionality group weightings

Benefit	Benefit weights	Benefit rankings	Option label ([1]-[4])	Group Elements	Functionality group weightings
----------------	------------------------	-------------------------	-------------------------------	-----------------------	---------------------------------------

Improved user ability to formulate and execute successful subject searches	390	1	[1]	System can recognise terms used by UK users from DDC, LCSH or UNESCO and advise directly and via M2M on terms used by services using these schemes	140
--	-----	---	-----	---	-----

System can recognise terms from either of these sets and advise on UK terms that may have been used in legacy metadata

Users/services can conduct own disambiguation with information from system

System has training module on the above

System development in this area informed by RDN study

System development in this area informed by clustering study

System development in this area informed by user study

[2]	Service offers disambiguation facility	123
-----	--	-----

Service offers find collections facility

Service offers any hits test/rank facility

Service offers additional training on disambiguation, find collections and any hits facilities

[3]	System can recognise terms used by UK users not in standard schemes and advise directly and via M2M on terms used by services using MESH and AAT	67
-----	---	----

System can recognise terms used by UK users from MESH and AAT and advise directly and via M2M on terms used by services using these schemes

System can recognise terms from MESH and AAT and advise on UK terms that may have been used in legacy metadata

Service offers training on MESH, AAT

[4]	System can recognise terms used by UK regional users not in standard schemes and advise on terms used by services using DDC, LCSH, or UNESCO terms	60
-----	--	----

System can advise on regional UK terms possibly used in legacy metadata

Service offers regional terms training

Online source for DDC, LCSH, and UNESCO standard terms and their use

Improved staff ability to provide quality subject descriptions appropriate to needs of JISC users	180	3	[1]	75
---	-----	---	-----	----

Source for DDC classification scheme

UK level source of terms needed for UK HE and FE retrieval but not in standard schemes

Mechanism for adding to this in a coordinated (and, hence, interoperable) way at UK level

System capable of assisting staff in local services with legacy metadata problems in service subject descriptions

Staff training in all of the above

System development in this area informed by RDN study

System development in this area informed by clustering study

System development in this area informed by user study

[2]	Feedback on how other staff have used the term sets and class numbers in the above group to describe items	50
-----	--	----

System provides feedback on terms users use to search for subject information

Staff training on feedback mechanisms

[3]	Online source for MESH and AAT standard terms and their use	30
-----	--	----

Feedback on how others have used AAT and MESH term sets to describe items

MESH and AAT training

[4]	Regional level source of terms needed for UK HE and FE retrieval but not in standard schemes.	25
-----	---	----

Mechanism for adding to this in a coordinated (and, hence, interoperable) way at regional and UK levels

Improved mapping of user subject queries to staff subject descriptions of items within and beyond UK HE and FE

150

4

[1]

Mapping of terms used by users to terms used by collections utilising DDC, LCSH or UNESCO where user terms are in UK or DDC, LCSH or UNESCO term sets and service uses schemes without change

58

Mapping of terms used by users to terms used by collections whose legacy metadata follows the pattern of the UK terms set

Information on user-driven disambiguation process

Training in the use of these facilities

System development in this area informed by RDN study

System development in this area informed by clustering study

System development in this area informed by user study

[2]

Online disambiguation function

49

Automated JISC collections identifier

Feedback on which terms to use in identified services where these use DDC, LCSH or UNESCO term sets

Sample retrieval from these services based on the recommended terms

Training in the use of these facilities

[3]

Above facilities extended to cover collections utilising MESH and AAT and terms from these used by

23

users

Training in the use of these facilities

[4]	Above facilities extended to cover user terms in UK regional terms set and the use of such terms in services legacy metadata	20
-----	--	----

Training in the use of these facilities

Ongoing basis for a process that will halt deterioration in, and begin to monitor and improve, interoperability in respect of subject description of resources

190

2

[1]

Central service that will facilitate improved interoperability in respect of the use of DDC, LCSH, UNESCO, and UK variations in their use, and that will begin to solve legacy metadata problems whose source is the creation of UK variations on standard terms generally

124

Extension of interoperability to other schemes through ability to interact with other mapping services

Extension of interoperability to other schemes through provision of facility to allow addition of other schemes

Additional training where necessary

Improved ability to influence development of subject and class schemes to meet JISC needs

[2] Extension of above interoperability improvements to MESH and AAT usage and variations 33

Training in these extensions

Improved ability to influence development of MESH and AAT subject schemes to meet JISC needs

[3] **Extension of above interoperability improvements to cover regional usage and variations to the various schemes** 33

Training in these extensions

General improvement in JISC collection development and utilisation activities

90

5

[1]

Improved ability of JISC and other staff to monitor or sample subject coverage and identify and deal with collection weaknesses

90

Improved ability to identify duplication

Improved ability of JISC and other staff to monitor user subject needs as reflected in user searches

Improved ability of lecturers and librarians at institutions to identify and utilise useful materials on behalf of the students they serve

Improved value obtained from expenditure on JISC collections because users alerted to their existence and subject contents

The column headed “option label” signifies which functionality groups are included in the following options:

- Base mapping option [1]
- Base services option [2]
- Other schemes option [3]
- UK extension option [4]

i.e. each of the functionality groups labeled [1] denotes the elements included within the base mapping option, and so on.

Cost-Benefit Ratios

Totalling functionality weights for [1]-[4] enables us to consider the mix of functionality groups to be included in each of the options A-H shown below. For example to consider option C we must total functionality weights for all functionality groupings labeled [1] and [2] (see table 1). These results comprise the overall benefit scores for options A-H and are recorded below.

The five year costs were calculated by the HILT team prior to the steering group meeting. The final stage of the cost benefit analysis process was to divide the benefit score by the five year costs for each of the options A-H, resulting in a cost benefit ratio for each.

These ratios have been ranked from 1 to 7 where 1 indicates the most favoured option and 7 is the least preferred. The closer the ratio is to 1, the better relationship between proposed benefits and cost effectiveness.

Table 7: Cost-benefit analysis ratios

Option	Mix	Description	Five Year Cost	Benefits Score	Cost-benefit ratio	Ranking
A	A	Do nothing option				

B	1	Basic interoperability process created; staff services to support creation of UK 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
C	1+2	Option B plus direct and M2M disambiguation, collection finder, sample hits and collection ranking, user term monitoring, training				
			£926,096	742	801	1
D	1+3	Option B extended to AAT and MESH but without option C				
			£1,481,448	640	043	7
E	1+4	Option B extended to regional variations to UK terms set, but without option C or AAT and MESH				
			£1,021,906	592	058	5
F	1+2+3	All 5 schemes, UK terms set without regional and service variations, but with disambiguation, collection finder etc				
			£1,525,593	895	587	4
G	1+2+4	DDC, LCSH, UNESCO, UK terms set with regional and service variations, disambiguation and related services, but no AAT or MESH				
			£1,065,241	847	795	2
H	1+2+3+4	Everything: all 5 schemes; UK with regional and service-specific variations, disambiguation and related services				
			£1,664,738	1000	601	3

7

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

Discussion and Conclusion

As option A was eliminated at the beginning of the exercise, the results comprise options B-H.

As table 7 shows, option C emerged as the most favoured option in terms of the cost-benefit ratio (it has the fourth highest benefit score and the second lowest cost).

The cost of this option has been calculated at £926,096 which is only more expensive than one other option, option B, which represents a less sophisticated system. Option C offers greater functionality in the areas of disambiguation, collection finding, result ranking, feedback on terms, staff use of term sets, and more extensive training.

Option G, ranked second in terms of cost-benefit ratio, includes regional terms creation and mapping. 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.

The third most highly ranked cost-benefit ratio is for option H. This option comprises all benefit elements ie. [1] + [2] + [3] + [4] so is the optimum solution in terms of functionality. In contrast, the cost of this option is the highest at £1,664,738; at least £139,145 greater than any of the other variations. However, despite this high cost, the greatly increased level of benefits results in option H being ranked third most favourably.

All other options (D, E, F, H) have a lower cost-benefit ratio due mainly to the high cost of additional mappings. For example, the inclusion of AAT and MeSH adds a further £599,497, leading to a lower cost-benefit ratio for options including these schemes.

Option D is the lowest ranked ratio. This is due to the fact it has a wider coverage of schemes but lacks a user oriented approach (disambiguation and collection finding have been removed from option D).

Considering the difference between cost-benefit ratios for each of the options from B to H, options C and G emerge as the most favoured options. Common to each is the inclusion of functionality groups labelled [1] and [2] in table 2, the Base Mapping [1] and Base Services [2] options. Option G adds regional terms to the equation.

It was of interest to discover how the addition of MeSH, a specialist thesaurus would affect the cost-benefit ratios of the first two highest ranked options. Thus, options I and J were considered as shown in table 8 below.

Table 8: Cost-benefit analysis ratios for options I and J

Option	Mix	Description	Five Year Cost	Benefits Score	Cost-benefit ratio	Ranking
I	C+ MeSH	1+2+MeSH	£1,013,988.00	753	743	3
J	G+ MeSH	1+2+4+MeSH	£1,153,133.00	855	741	4

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.

Glasgow : Cente for Digital Library Research, 2004

Appendix I: Initial and Interim Service Specifications

Appendix I.1 (1): TeRM Diagram

Staff
interact
with
users
to
improve
knowledge
of needs,
identify
user
vocabularies
and improve
TeRM



f
ract
i
s

rove
wledge
eeds,
tify

abularies
improve
M

*Note: Examples can be seen at www.wordmap.com with www.oingo.com and vivisimo.com

Users

Users interact with TeRM to establish subject term and service context, perhaps down to a single service, but usually a group. Client server means users determine subject and service subset focus through use of TeRM but interact directly with services or service groups using service subset chosen and terms found. Users can also 'train' using interface and TeRM can 'learn' user terminologies

**S
E
R
V
I
C
E
S**

TeRM

Supports creation, editing, display, and User, staff, and system interaction with terminologies map showing terms in use and inter-relationships.

Interacts with users and systems to establish term and service context of search (e.g. archives only), provides synonyms, broader, narrower, related terms, other contexts and service-set navigational aids for cross searching and browsing as required.

Also permits greater precision or greater recall decisions to be made. Flexibility of approach should

reduce likelihood of 'dumbing down'

Built using existing machine-readable mappings

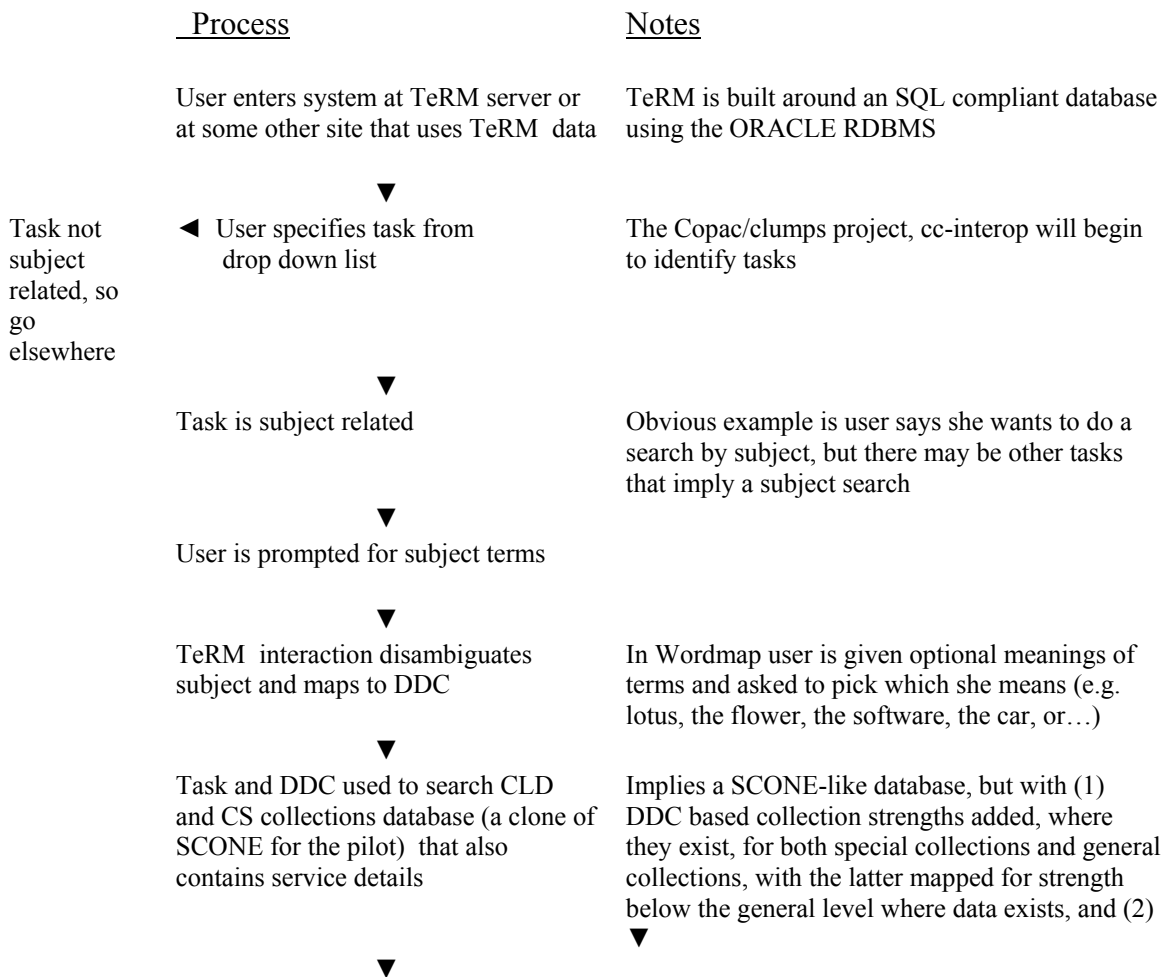
Provides online overview of UK terminologies for staff and a process leading to closer harmony

Based if possible on commercial software* but customised as required (e.g. for Z39.50 searches)

Staff

Staff consult TeRM when describing resources or collections or create and submit new terms as necessary. Staff also benefit from the existence of an online route map of terms in use in the UK, can 'train' using the service, and are involved in a process that, over time, brings closer harmonisation in the use of subject terminologies in the UK

Appendix I.1 (2): TeRM Diagram: Initial Service Specification Draft



Collections service identifies subgroup of collections relevant to task and to subject using DDC number, truncating it to find collections described at higher level of granularity



User 'trims' list of collections to suit her purpose, deselecting those she decides not to use initially



Collections service details include subject scheme(s) used in each collection, and connection details



(2) Task relevance specified. As in SCONE, the top level of any collection hierarchy will, if it is an online service, have connection details attached. Services returned will be ranked according to the granularity level of ▼ their subject strength coding, with those with lower granularity levels ranked highest

User may spot services not relevant to her or simply wish to search less than the full list. Option to test results from highest ranked service may be available

Specifying subject schemes necessary so that only appropriate terms are sent to any specific service, thereby avoiding the false drops that will occur if all terms from all schemes are sent to all services

Further interaction with TeRM based on 'trimmed' list of services provides terminologies set needed to search for the user's subject in each chosen service, including standard terms from scheme(s) and (possibly) common UK alternative terms or service specific alternative terms

User clicks button which says something like 'get terms', TeRM sends back appropriate terms for each service. Different levels of service are possible, beginning with terms from standard unchanged scheme used by service, to this plus common UK alternatives, with or without standard terms, to service specific alternatives. The addition and maintenance of UK alternatives will raise costs, and the addition of service specific terms will raise costs even further. Costs against benefits of each approach will be measured by HILT and compared against alternative approaches such as clustering

▼
User gets option of switching off all but one of the various schemes used by a given service and the user's own terms before searching

User will have the option to use her own terms, plus the standard term, plus UK or service alternatives for each service, or to switch any of these alternatives off.

▼
User searches services, either singly or in groups, using preferred term sets

This is assumed to be a broadcast search using Z39.50 plus, where facilities exist, other alternative protocols.

DMN
2/08/02

Appendix I.1 (3): Interim Specification

HILT Terminologies Server Pilot Specification (Version 3.0)

0. Status of this Document

Under construction: Specification is probably correct in general outline at this stage but may change slightly if ongoing consideration of the factors listed in section 1 dictates this. Consideration of the proposed workshop programme (section 3) may also have an influence, as will answers to the questions in section 4. Practical concerns as the pilot develops may also influence the design and all of these various factors may impinge on the content of sections 1 (factors influencing design), 3 (workshop), and 4 (questions). Also, at some point, this HILT Terminologies Server Pilot Specification document needs to be 'aligned' with the HILT Methodologies document and (possibly) the Project Plan. Note that this document has primarily been written to guide work by the project team and the management group. However, it also describes the approach being taken and the reasons behind it in a way that should be helpful to the Steering Group and JISC and others.

1. Introduction: Factors Influencing Pilot Design

The design of the pilot as outlined in section 2 below is influenced by the following factors:

Informed Assumptions

The informed view within HILT is that the following are valid assumptions:

- i. For users, the ideal terminologies server for the JISC IE, serving both collection level and item level requirements, would, at minimum, permit any user to identify all JISC collections relevant to any and every subject query he or she might have and reliably retrieve from each of them all items relevant to the query and only all items relevant to the query. It would also ideally filter out duplicates reliably and filter out collections the user was not permitted to access. For staff, it would provide facilities to enable the creation and maintenance of a the coherent standards-based subject environment necessary to support such reliable retrieval. Since a number of different subject and class schemes are in use across JISC services, this implies a service able to deal with a range of standard schemes and to reliably map any user term to any scheme in an appropriate way. Mapping all schemes to a DDC spine and a 'user term set' to the same spine is probably the most efficient way of doing this. Since UK services tend to amend schemes for UK use, the best first

- stab at a 'user term set' mapped to DDC is probably a set of subject terms and associated DDC numbers taken from a general UK collection such as a University Library.
- ii. Within this 'ideal' distributed system comprising terminologies server and associated JISC collections, retrieval – both precision and recall – would be optimized if (1) standard subject schemes and class schemes were used 'as is' to describe resources, (2) it was always clear to all assigning terms how the scheme should be used for a given resource, and (3) all users seeking to retrieve resources had a perfect knowledge of the scheme and how it would be used to describe resources and applied that knowledge in retrieval.
 - iii. None of these three sets of ideal circumstances are likely to be met in practice
 - iv. In designing a terminologies server to meet the collection level requirements in the JISC Information Environment, our task is to determine what factors determine the extent to which each of these three 'ideal circumstances' can be met and to use our knowledge of the factors, together with a cost-benefit analysis and a comparison of other approaches such as clustering, to allow us to specify a requirement for a full-blown terminologies server.
 - v. Even if a terminologies server and associated agreements on subject description practice could optimize the extent to which these ideal circumstances could be met in future, the problem of legacy metadata still has to be considered. The requirement for a full-blown terminologies server must also take this into account
 - vi. Factors likely to influence ii(1) include (a) Availability of a central server where (for example) 'UK versions' of terms used in the standard scheme are mapped to the standard scheme terms making it unnecessary for the service to use these alternatives on the service itself (b) Availability of a central server where extensions perceived as necessary to a scheme for local use can be reflected, standardized across services, and shown to the user
 - vii. Factors likely to influence ii(2) include (a) Online Training for staff and users (b) A 'resources like this' facility that shows both staff and users examples of the kind of resources that should be described in particular ways using a particular scheme
 - viii. Factors likely to influence ii(3) include (a) The extent to which staff changes to standard schemes are minimized (b) The extent to which changes made are known to users (c) General user knowledge of the schemes and changes to the schemes through training (d) System facilities that allow users to see and explore the scheme used and any amendments to the scheme used (e) The extent to which the terminologies server can 'recognize' terms used by users and map them intelligently to the scheme used by a collection
 - ix. All of the terminologies used in the JISC IE – and indeed in the world beyond this – cannot be encompassed in the HILT pilot. The aim must therefore be to focus on only a few terminologies and to aim to infer or deduce general principles that will apply to all. This assumes that factors ii(1), ii(2), and ii(3) operate in essentially the same way across all schemes. This is probably a safe assumption. For example, ii(3) – 'retrieval would be optimized if all users seeking to retrieve resources had a perfect knowledge of the scheme and how it would be used to describe resources and applied that knowledge in retrieval' – may vary from scheme to scheme in the sense that the difficulty in realizing it will be greater in some schemes, but it is safe to assume that the factor will operate in all schemes and that general rules on optimization can be identified

HILT Phase II Scope

HILT Phase II is not charged with preparing a specification for the ideal system described above. As indicated above, this would address both collection level and item level requirements of a terminologies server for the JISC IE. Item level requirements are beyond the scope of HILT Phase II. This is understood to mean that points a-c below are within scope, but that point d is not:

Within Scope:

- a. Enabling the user to identify relevant collections via browse and search functions, with the latter implying an ability (1) to 'recognise' a subject and map it to DDC via comparison with a 'user term set' and subsequent user-driven disambiguation and contextualisation, and (2) to subsequently process the DDC number and use the result to identify collections relevant to the user query

- b. Providing the user with information on subject or class schemes used to describe items in the collection
- c. Providing the user with access to the collection itself and to some 'collection level' guidance on how best to use the scheme in question, together with the user's own terms and related 'UK terms', to optimise retrieval from any given identified collection.

Beyond Scope:

- d. Providing facilities to enable the creation and maintenance of a the coherent standards-based subject environment necessary to support reliable retrieval from the collections themselves. This is not only out of scope for Phase II, it is a difficult problem that will entail significant work on legacy metadata in the services themselves, the creation of terminologies server facilities to support and facilitate such change, and additional research to inform these developments. It will be one thread of any HILT Phase III proposal. An interim solution may be to optimise the extent to which the terminologies server's UK terms set reflects the use of added and amended terms used in individual collections.

Barriers

Even within the scope of HILT Phase II, there are limitations on what is possible within the pilot. In particular:

- Optimizing the extent to which a JISC IE terminologies server can 'recognize' user queries and map them to standard schemes is a long term process that would have to take place within the context of any HILT Phase III. The best that can be achieved in the pilot is a 'first pass' at this.
- Detailed mappings of subject schemes to the DDC spine are impossible within the resources of the project, except to the extent that the process of creating them can be automated. Where this is impossible, only selective mappings can be provided

In addition, some of the more minor facilities that might be available in an operational service might not be worth including in the pilot.

Likely Design of Version 1 Operational System Post HILT II and of Associated Project

Although largely determined by factors already noted above, the likely shape and form of both a 'first pass' operational service as envisaged post-HILT Phase II and of the project associated with it are also a consideration in the design of the pilot. This is likely to have three elements:

- a. The relatively short term creation of an operational service offering the best 'first pass' at the facilities described above as within scope
- b. A longer term process aimed at optimising these facilities for individual services
- c. A longer term process aimed (1) at optimising the 'UK user term set' and its mapping to key schemes heavily used in the JISC IE – LCSH, DDC, UNESCO and a few others (2) at using this term set to improve interoperability in respect of cross-searches of JISC IE collections. (these are terms that staff add to improve UK retrievability and should in theory improve interoperability to some extent, although the problem of service specific false drops will need to be dealt with).
- d. Research into the item level requirements of building an optimal terminologies server for the JISC IE. Research that would, for example, aim to specify what is required to gradually make legacy metadata in distributed collections inter-compatible.

Function of Pilot

The final factor influencing the design of the pilot is its function. As currently envisaged, the Pilot will illustrate both what is feasible in Version 1 Operational System and also, as far as possible, illustrate the

various barriers and (maybe) possible solutions. It may also allow assumption testing to some extent, and the examination of alternatives such as clustering.

2. General Description of Pilot

The pilot will illustrate the Version 1 System, and some of the problems. It may also permit testing of assumptions. In essence, it will comprise:

- A collections finder interface based around a database describing JISC collections and in some cases sub-collections. The finder would offer two options: A browse collections by subject grouping or hierarchy option and a search option based on the user subject query resolver (see b below)
- A user subject query resolver, which will 'map' user terms to the HILT terminologies mapping, offer the user alternative DDC numbers to permit disambiguation and contextualisation, and identify a DDC number associated with the user's final choice. This will be built around a DDC spine which has at least the University Library UK terms set, DDC captions, LCSH terms, and UNESCO terms mapped to it. These mappings will be comprehensive if possible to automate but selective if manual work required. All terms from all schemes should be available for mapping to user terms but it must be possible to distinguish between terms from one set and terms from another. In some instances, we will wish only to display UK terms or only LCSH terms, or to search sample collections with one set or another, or to log which set the user term has been found in. It may also be necessary in some instances to identify terms in one terminology such as LCSH as not used in a particular collection or only used in a particular collection. The possibility of adding to the UK terms set by adding BUBL terms somehow is also worth considering.
- A query to collections mapping function based on processing the DDC number chosen at ii above in various ways with a view to identifying collections appropriate to the user query. Processing will involve number truncation but also (possibly) identifying standard subdivisions and acting on these. The latter can be illustrative rather than comprehensive and could possibly be based on identifying the end of a DDC number and looking for recognisable strings representing particular sub-divisions.
- A display and interact with identified collections screen arising as a result of either the browse option (see a) or the search option (see a, b, c). Ideally, this would show all of the following on one screen:
 - Browsable list of retrieved collections with helpful information about content to help users
 - Information on subject scheme used with mapping from unprocessed DDC number to appropriate term in subject scheme used by highlighted collection and route map showing (e.g.) broader and narrower terms in the scheme. The mapped term would be clicked on. It would be possible to click on the broader and narrower terms. Only one of the terms could be clicked on at a given time.

NB: Mappings between schemes will be of up to 19 types and the user interface will have to cope with all of them. One likelihood is that terms will have to be shown in the context of the scheme in question, so that the user can navigate between schemes rather than just use an alternative (and possibly) inappropriate mapped term. This should be possible. We only need to codify each mapping type uniquely in the mapping between terms and make user interface responses conditional on which mapping type is specified.

- Sample retrieval boxes. One showing items retrieved by searching collection using highlighted term from host service scheme, the other showing items retrieved by searching it using users terms and UK terms from the HILT mapping. The possibility of using LINK sub-sections to simulate all subjects cover might be worth examining if time and resources allow. It might even be possible to simulate (say) a UNESCO service by mapping in some small area manually or doing
- A link to host service search screen option
- The screen would also need a button to enable users to gather results from different services together, but this need not actually work. At best, a faked illustration for a specific search might be nice, but not a priority.
- A clustering service option. Again, this might only work for one or two services.
- A section of the screen where the user could be warned (various standard warnings) of the limitations

of the advice given and possible ways of determining whether the approach could be refined

- A screen (entirely separate?) showing how above functions might be an integral part of another service
- If the above can be based on Wordmap, we'd also want to be able to show a staff interface to the DDC spine mapping that would eventually permit collections specific updates to be done. If not an *illustration* screen showing that such a function is planned would be useful. It need not do much except illustrate what updating might mean.

3. HILT Workshop and Related

Questions to Examine

In essence, the function of the pilot terminologies server is to identify JISC IE collections likely to be relevant to any given subject query brought to the system by any UK-based JISC IE user and to then provide that user with information on how best to search each collection by subject in order to optimise relevant retrieval. The HILT team has an informed view of how best to achieve this described elsewhere in this document and is building a pilot server that will instantiate some aspects of this and illustrate others. The exact balance has yet to be determined and may not be known in full until late May when the pre-workshop pilot is complete.

Given this context, the questions HILT II needs to examine with students, teaching and research staff, and intermediaries such as librarians, archivists and others, either at a workshop or by other means, are:

- a. What is the best method of optimising the ability of a terminologies server addressing collection level needs to 'recognise' all subject queries input by users, looking at the effect of things like:
 - The different effectiveness levels of DDC captions and relative index terms used alone as compared with DDC captions and relative index terms enhanced by adding a set of 'UK-centric' terms
 - The use of a browse interface
 - The effect of training
- b. What is the best method of identifying collections, comparing things like
 - The DDC truncation method
 - A browse approach
- c. What facilities, interface features, information do users, teaching and research staff, and intermediaries think are needed in a terminologies server designed to meet collection level requirements. What do they think of the features provided in the HILT pilot.

A further interesting question is:

- d. How can we best optimise item-level retrieval from the various collections by optimising mapping to the user terms set, to legacy metadata, and to UK 'DIY' subject schemes.

As an item level requirement, this is beyond scope but it would nevertheless be useful to be able to find out something about this question if we could – if only because collection-level requirements should ideally be functionally 'in step' with item-level requirements. As indicated elsewhere in this document, however, this is really a question for any HILT III project.

Problems

Provided that a working pilot is actually available in time, which is now fairly certain, examining question c with users, staff, and intermediaries at a workshop should not present a difficulty. It is also fairly certain that useful information can be gleaned from such a workshop as regards the views of users, staff and intermediaries on questions a and b. Conducting definitive and reliable empirical tests of questions a and b,

however, is less likely to be possible. Definitive and reliable tests of question a would require that we had access to:

- A very large, representative group of users, staff and intermediaries with representative queries covering a representative range of subjects and enough time to conduct a representative set of queries
- A set of 'UK-centric' and DDC terms comprehensive enough to cover at least this representative set of users and queries

At present, the DDC terms are available (at least for one edition of Dewey), the question of whether or not a sufficiently comprehensive set of UK terms is available has yet to be determined, and there is no possibility whatever of the queries, users, staff and intermediaries, and time available at a workshop meeting the requirement specified above – which means that any results we might obtain from empirical tests at a workshop would be very suspect on a number of counts (for example, it might well be the case that the success or failure of the pilot to 'recognise' a query is due to unrepresentative users, unrepresentative queries, or an incomplete UK terms set.

Definitive and reliable tests of question b entail the additional problem that we can only determine the best method of identifying collections empirically if we know both what the enquirer was looking for (not necessarily the same thing as the terms he used) and what items in what collections he would have had to retrieve for a 'perfect' result.

Proposed Approach

Obviously, it is essential that we take these circumstances into account when determining our approach to examining these questions in the project. This being so, our proposed approach is as follows:

- Set about arranging a workshop in June involving users, teaching and research staff, and intermediaries more or less immediately. Attendees, as agreed previously with JISC and the project groups, to be mainly from the Glasgow area institutions, but with some externals if this is possible
- Aim to examine question c at the workshop, and questions a and b to the extent that this is possible. Include any elements of question d it might be feasible to examine only if time and resources allow
- Begin to plan the programme for the workshop in late April or early May but keep it under review as far as is practical until the likely specification of the pilot server is entirely clear.
- Aim to examine those aspects of questions a, b and c that cannot be addressed at a workshop by other means - interviews, online tests, tests conducted by the HILT team
- If the workshop fails to attract sufficient attendees, use these 'other means' as a backup approach for those aspects that would otherwise be tested at the workshop.

4. Questions for HILT to Answer Soon

- Build pilot using Wordmap or in own SQL database? Or do selectively to ensure performance? Or use Wordmap for (slow) full coverage and selective/own database for speed/illustration?
- Does OCLC DDC file have identifiable mappings of (1) 'Standard LCSH' (intellectual) (2) 'Standard LCSH' (statistical) (3) DDC captions and standard subdivisions? (4) Anything else of use? If something not there, can OCLC supply?
- Are electronic mappings of DDC to UNESCO, DDC captions and standard subdivisions, LCSH, AAT, Wordmap large taxonomy available anywhere?
- What is the status of the University Library to DDC mapping. Can it be loaded into Wordmap and/or our own database?
- Is a BUBL terms to DDC mapping feasible? What about from the browse index? What about a statistical approach – could we assume that the DDC number associated with a term highest number of terms is probably the appropriate term for that number?
- Will July JISC workshop include users? How many users do we need? Is ten enough?
- What free services use DDC, DDC captions, UNESCO, LCSH and provide browse access? Can we connect any of these readily to the pilot? In route map fashion?
- If g is either impossible or not comprehensive or just difficult, can we do something with BUBL

instead, either manually or by automated mapping of some UNESCO or LCSH or DDC captions to DDC numbers in BUBL?

- What can Wordmap provide in respect of user terminologies tests etc based on large taxonomy set? What mappings of this to other schemes can they/would they make available?
- A list of software packages that might be used for an operational service.
- Talk to OCLC about the idea of them providing the basis for the user terminology recorder and resolver and the associated staff updates to the mappings. Is recording other schemes like UNESCO possible?
- How does this document impinge on the Methodologies document?
- Can OCLC supply DDC standard subdivisions and numbers in electronic form?
- Information on what percentage of user queries are subject queries – lis-link survey?

5. New HILT Schedule

HILT Phase 2 Pilot	Month =	6	7	8	9	10	11	12
(March 2003 to end September 2003)		M	A	M	J	J	A	S

Scope, Design, Project Plan, Methodologies
 Project Plan, Methodologies Document WP
 Create research environment, other contexts
 Identify software solution sources
 Install pilot software and alternatives
 Terminology map modelling
 Implement and develop stage 3 pilot
 Improve stage 3 pilot
 Conduct human and M2M tests
 Plan, execute tests of pilot, alternatives
 M2M Requirements Study
 Detailed Functionality Survey
 Plan user / staff workshop
 Hold user / staff workshop
 Full service specification
 Examine costs and do cost benefit analysis
 Estimate costs, conduct cost benefit analysis
 Evaluation, Quality Assurance and Review
 Professional level evaluation process
 Agree, refine, monitor approach, progress
 Evaluate project and pilot (summative)
 Disseminate, consult and report
 Web-site activities
 Dissemination activities
 Compile draft final report
 Circulate draft final report
 Finalise report
 Submit/disseminate Report
 Project closedown activities

6. Draft Specification (Version 2) – Amendments not yet done. Is this level of detail necessary?

	<u>Process</u>	<u>Notes</u>
	User enters system at TeRM server or at some other site that uses TeRM data	TeRM is built around an SQL compliant database using the ORACLE RDBMS
	▼	
Task not subject related, so go elsewhere	◀ User specifies task from drop down list	The Copac/clumps project, cc-interop will begin to identify tasks
	▼	
	Task is subject related	Obvious example is user says she wants to do a search by subject, but there may be other tasks that imply a subject search
	▼	
	User is prompted for subject terms	
	▼	
	TeRM interaction disambiguates subject and maps to DDC	In Wordmap user is given optional meanings of terms and asked to pick which she means (e.g. lotus, the flower, the software, the car, or...)

And so on...

Appendix I.2: Development Requirement for an Operational Server

Overview

1. Introduction and Overview
2. Summary List of Requirements (with references to further details as appropriate)
3. Illustrative Description of Pilot Server
4. Functional Description of Pilot Server
5. Illustrative Mappings from Pilot Server

1. Introduction and Overview

This appendix is based on the Project's view of 'The Development Requirement for an Operational Server' prior to the cost-benefit analysis exercise, but also makes reference to changes made, both as a result of that

exercise, and as a result of further analysis of the user workshop outcomes on the usability of the interface to the pilot server. In particular:

The fact that AAT mapping was dropped from the requirement and MeSH and Regional modifications mappings made dependant on whether or not partnerships could be formed with funding partners who had an interest in these term sets. To a lesser extent, AAT is also in this latter category, but the cost is much higher for AAT and there is probably less clear benefit to JISC

The fact that functionality associated with the Base Services Option (disambiguation, collection identification etc) was noted as an area requiring further sophistication based on additional research to be carried out during the two years proposed as the timescale for developing the Base Mapping Option

Requirements specified fall into five categories:

- a. Controlled vocabularies and mappings proposed for a baseline operational server
- b. Other possible controlled vocabularies and mappings
- c. Functionality requirements known in detail, at least insofar as they were implemented in the pilot server (it is assumed that some of those relating to end user interface facilities will be extended and sometimes changed in the light of research proposed in the first two years of a project to develop an operational server)
- d. Functionality Requirements not known in detail
- e. Additional research proposed to inform development of the end user interface

2. Summary List of Requirements (with references to further details as appropriate)

Base Mapping Option [1]

Mappings database – See section 4 below under ‘DDC data importing and mapping’ and further information on mappings in sections 3 and 5

DDC spine and associated entry vocabularies – See illustration in section 5 below

LCSH and LCSH mapping – See illustration in section 5 below

UNESCO and UNESCO mapping – See illustration in section 5 below

UK oriented modifications registry terms set creation – See Full Report section 4

UK oriented modifications registry term set mapping – See Full Report section 4

Term match facility – See section 4 below under ‘search algorithm’

Processes to cope with scheme updates – See section 4 below under ‘DDC data importing and mapping’

Staff amend maps facility – See illustrative staff interface screen in section 5 below showing native Wordmap facility. Note that detailed examination of how this will be used in practice has yet to be carried out.

Staff training module – Further detail to be determined during development project

Online user training module – Further detail to be determined during development project

Ability to host and map other schemes – Further detail to be determined during development project

Ability to interact with other mapping services – Further detail to be determined during development project

RDN terminologies harmonisation study– See Full Report section 5 under ‘Other Issues...’

RDN-based clustering tool study – See Full Report section 5 under ‘Other Issues...’

Interface needs user study (enhanced pilot with clustering) – See Full Report section 5 under ‘User Interface Considerations’

Base Services Option [2]

Disambiguation facility – See section 4 below under ‘HILT search algorithms’

DDC collection identifier – See section 4 below under ‘HILT search algorithms’

Any hits test/rank facility – See section 4 below under ‘HILT search algorithms’

User terms monitor – Further detail to be determined during development project

Other schemes option [3]

MeSH and MeSH mapping – See illustration in section 5 below

AAT and AAT Mapping – No further information available

UK extensions option [4]

UK regions oriented modifications term set creation – Further detail to be determined during development project

UK regions oriented modifications term set mapping – Further detail to be determined during development project

3. Illustrative Description of Pilot Server

Pilot Terminologies Server Specification

The following components constitute the structure and function of the pilot server:

1. Subject schemes
2. JISC collections database
3. Mapping functions
4. User interface

Subject schemes

Four subject schemes have been incorporated into the pilot terminologies server. These are: DDC, LCSH, UNESCO, and MeSH. The DDC and LCSH mapping has been provided by OCLC. An illustrative mapping of UNESCO and MeSH terms to DDC has been conducted to provide examples of mapping in practice.

JISC collections database

One of the components of the server is a collections database covering JISC collections and services. The database consists of URL links and brief description of each collection or service, or service collection strength. Each collection, service, or service collection strength is classified by DDC. The database also records which subject scheme collections and services use to describe the resources in their collections.

Mapping functions

A relatively comprehensive literature review (Appendix B.2) was conducted to investigate the problems and issues in integrating and mapping thesauri and classification schemes and the different types of mapping reported in the literature. A list of 19 match types was provided in the review. In order to explore further the problems and issues of mapping in practice some mapping exercises were carried out and reported in Appendix B.3. Based on the literature review and the mapping exercise some examples of mapping were selected to build into the server. The examples and their match types are provided below:

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

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

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

User interface

There are two user interfaces to the pilot terminologies server namely search user interface for end-users and the staff user interface.

The end-user interface consists of three screenshots i.e. the homepage, disambiguation page, collection identification page. On the home page or *term input* stage users enter a search term and activate the search. On the *disambiguation* page users will be provided with a list of terms and their DDC context for the user to choose from. Since the service uses DDC as the main backbone, in most cases there are chances that users are presented with more than one option, in that case they require to disambiguate or contextualise the search term. On the *collection identification* page users will be presented with a list of JISC collections relevant to the search term selected at the previous stage as well as the Dewey Hierarchical path in which the search term appears. In addition to the collection titles and their brief description, users can find out about the associated subject scheme the particular collection uses for organisation of their contents and the

mapped search terms for that scheme to DDC together with Dewey number for that collection. At the final stage users can get access to collections through clicking on each collection's link.

The staff interface allows the librarians and indexers to create, amend or manipulate different versions of subject schemes held within the pilot terminologies server. For instance, staff can make changes to local versions of DDC or UNESCO through the interface. (Screenshots from both end-user and staff interfaces are included in the final report).

Figure 1 shows the homepage of the HILT pilot interface. The homepage consists of a search bar, a brief description of the service, a link to search tips and Dewey Decimal Classification (DDC) specific subject categories for browsing.

Figure 2 depicts the disambiguation stage of the HILT pilot service with possible options retrieved from DDC. At this stage the user contextualise their search terms and decide which option to choose. This page consists of a number of possible options and their context i.e. DDC hierarchy and a button labelled "more results" using which the user is able to search for more similar results.

Figure 3 shows the screenshot of the collection selection stage. The features on this page include: a search bar, DDC hierarchy for the selected term, a browsable list of JISC collections retrieved, Information on subject scheme used with mapping from unprocessed DDC number to appropriate term in subject scheme used by highlighted collection, DDC number for the collection, and a link to host service search screen option.

Figure 4 shows a JISC collection found as a result of selecting the term teeth.

Figure 5 shows the HILT pilot terminologies staff interface where different terminology instances can be created, modified or manipulated.

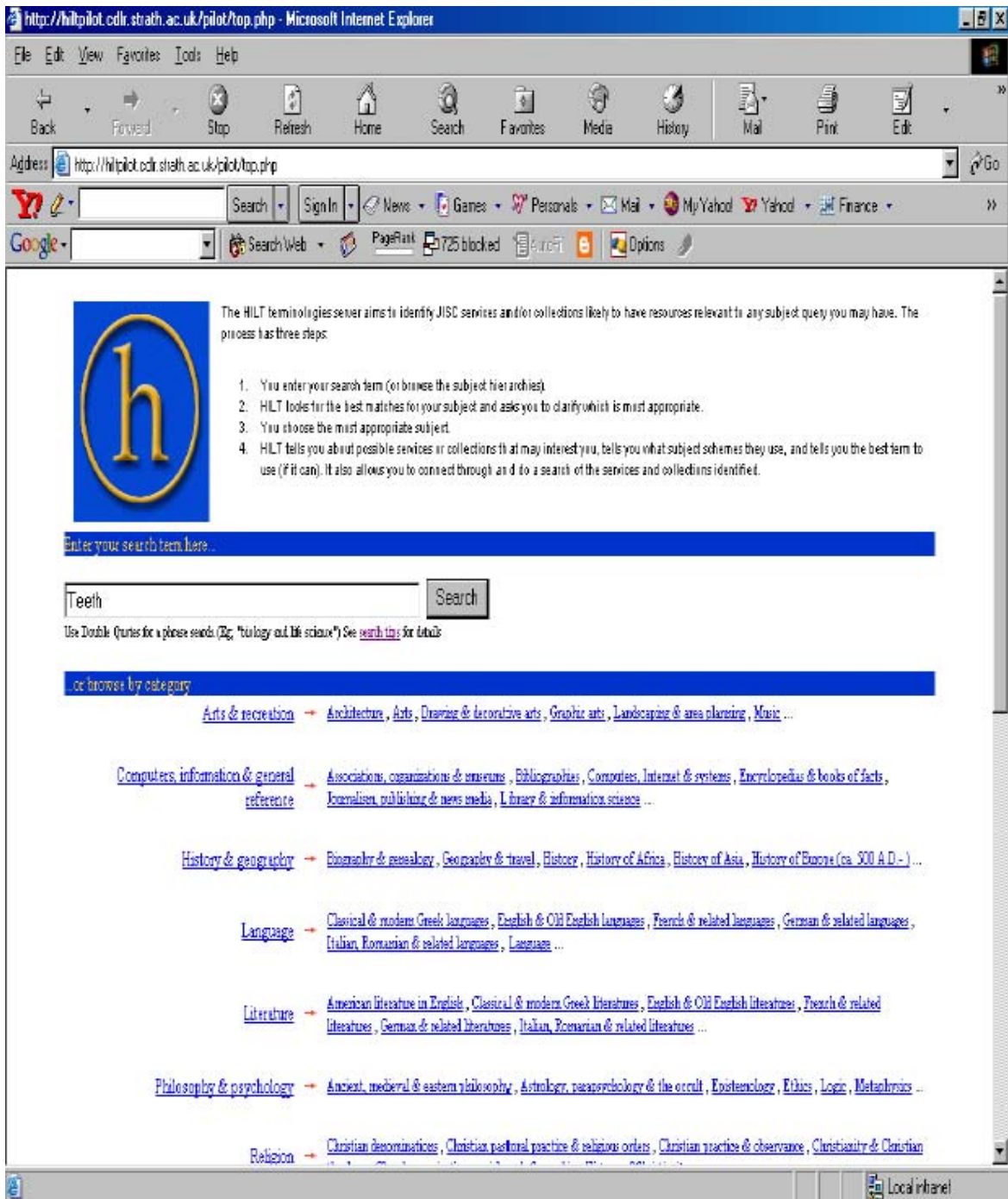


Figure 1. Homepage of the HILT Pilot Terminologies Service

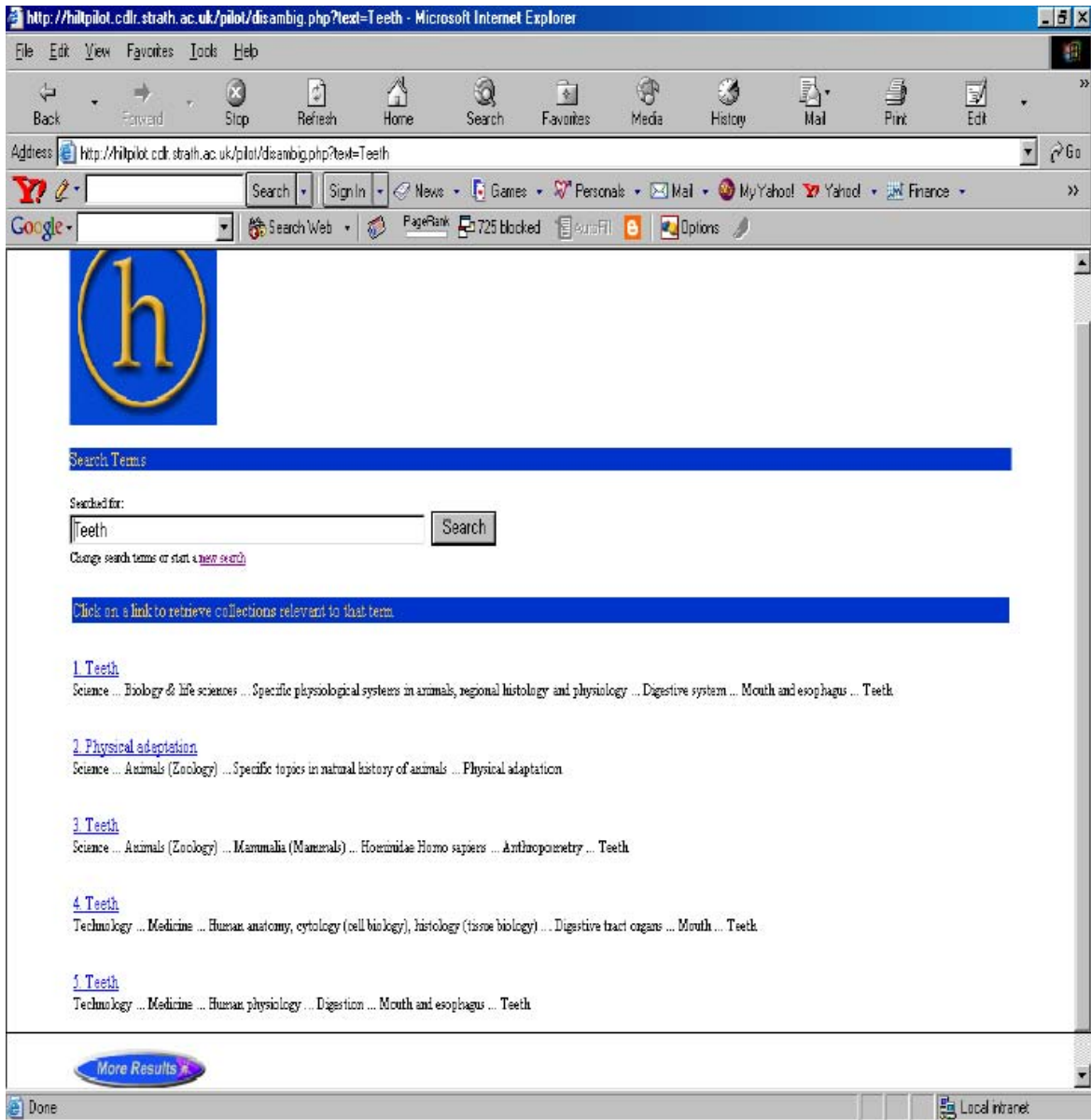


Figure 2. Disambiguation page of the HILT Pilot Terminologies Service

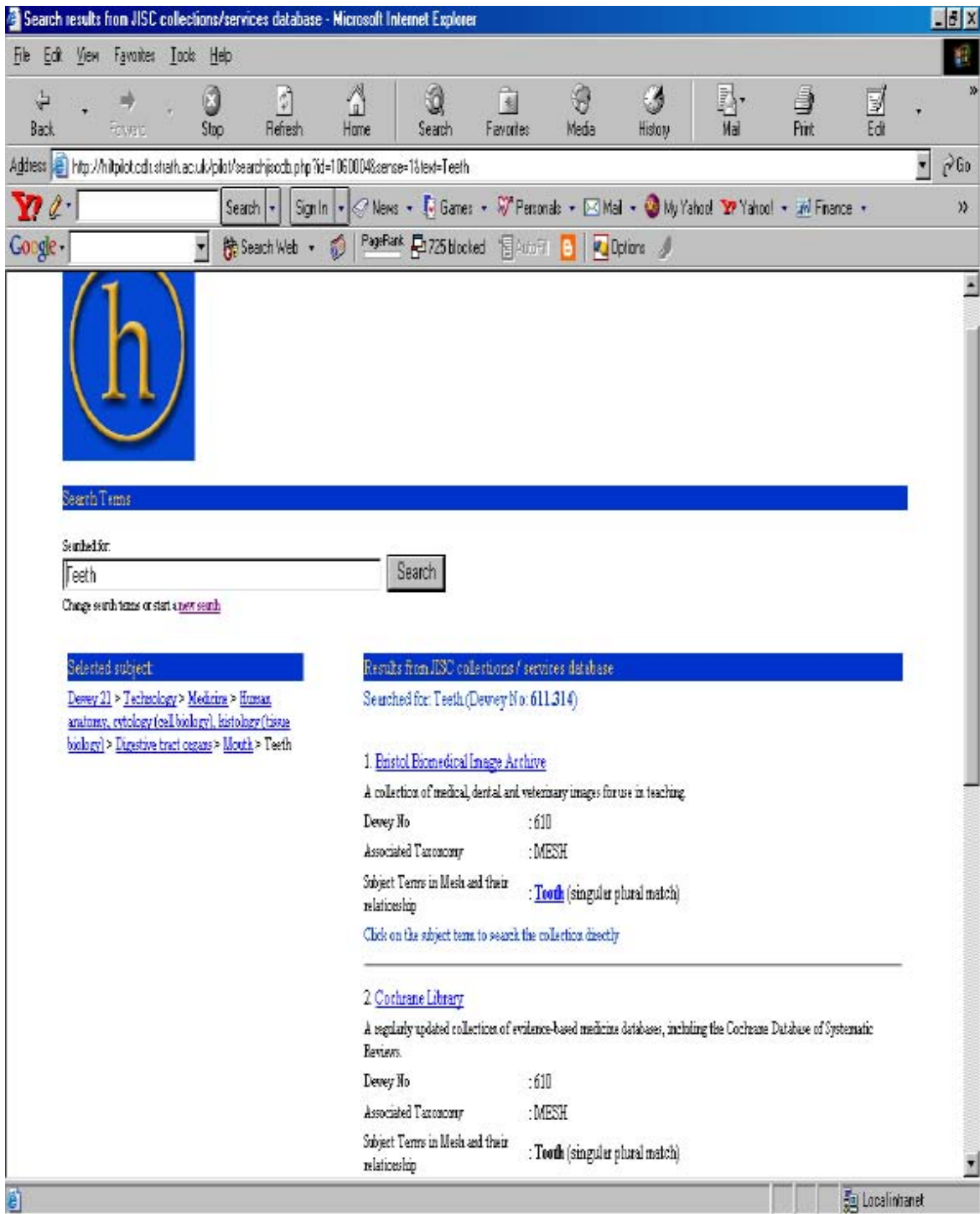


Figure 3. Collection selection page of the HILT Pilot Terminologies Service

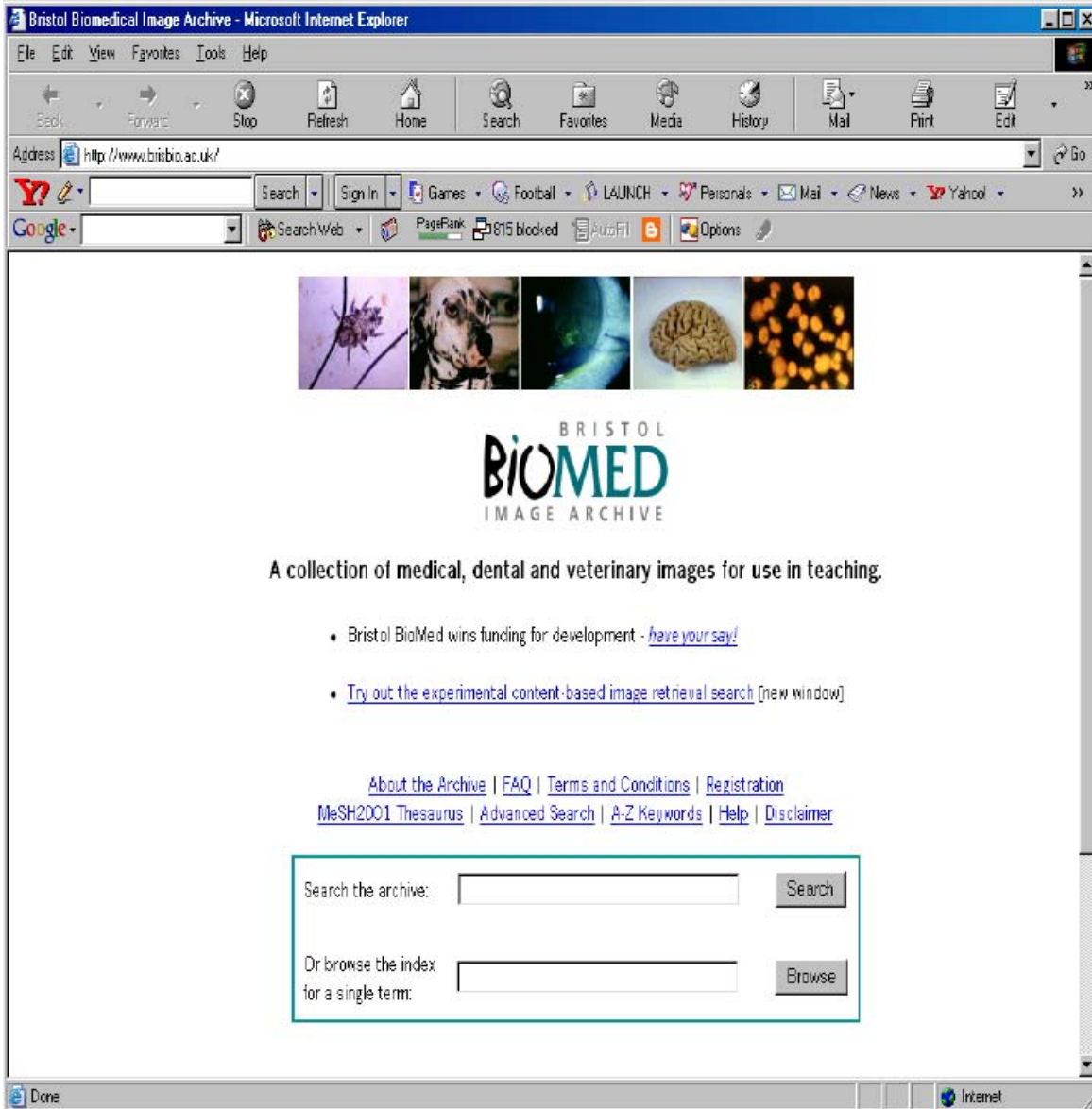


Figure 4. JISC collection found by the search term “Teeth”

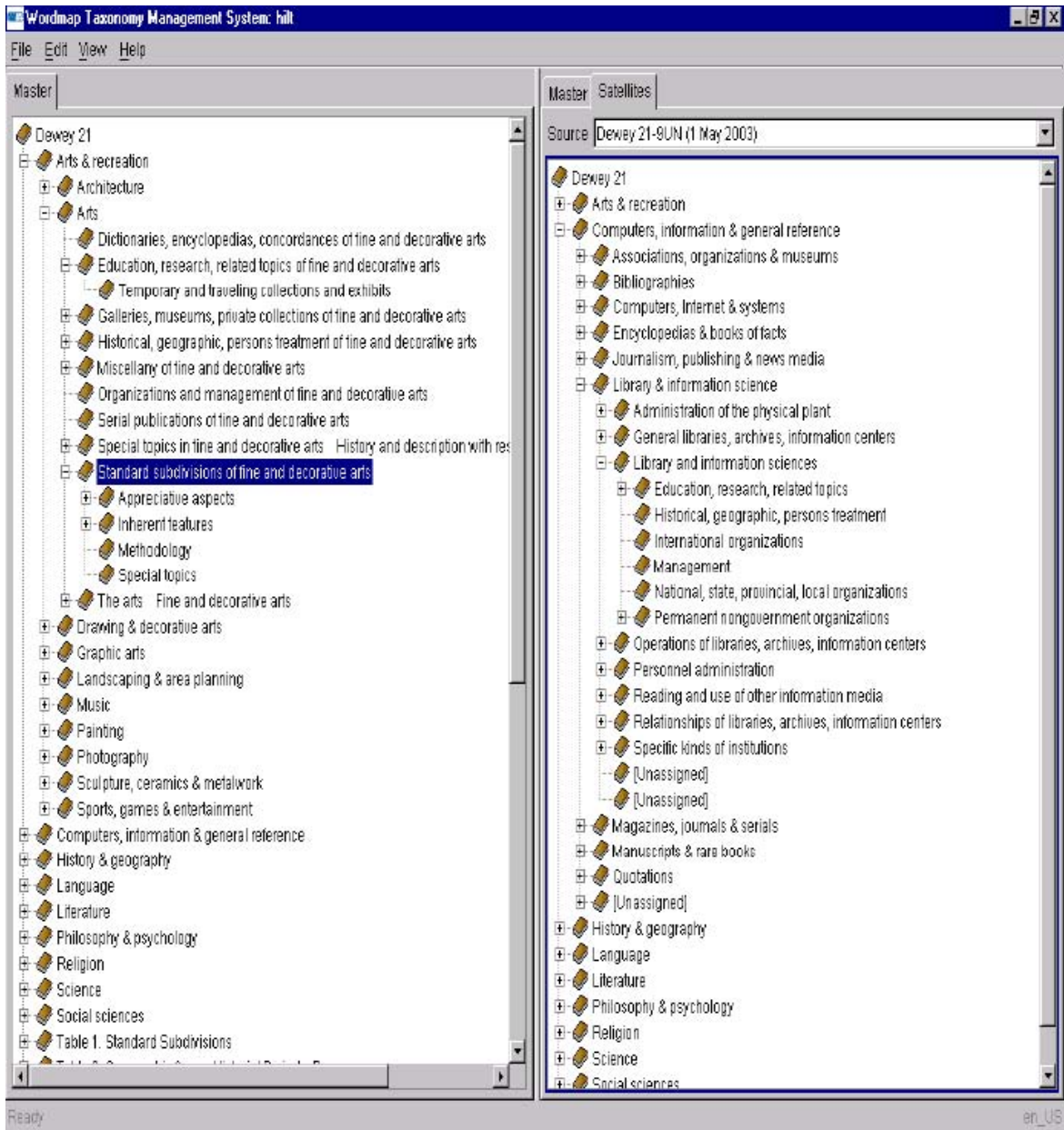
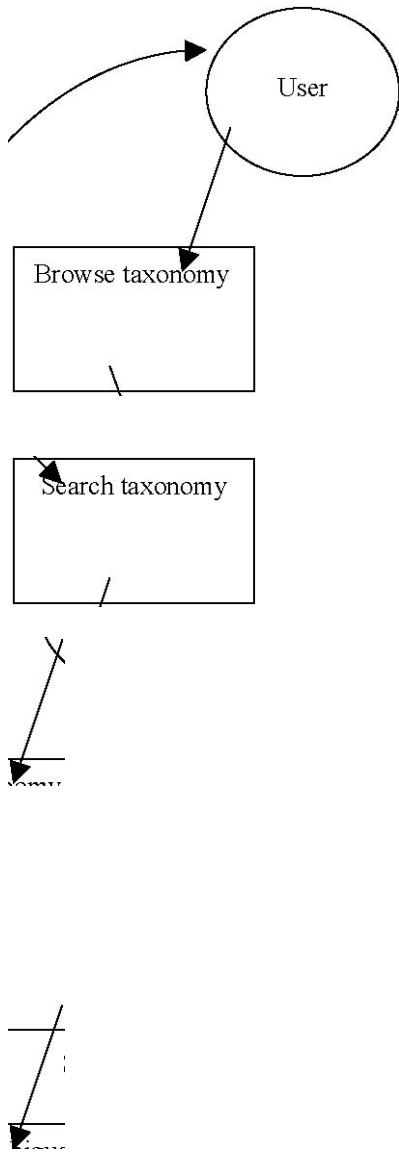


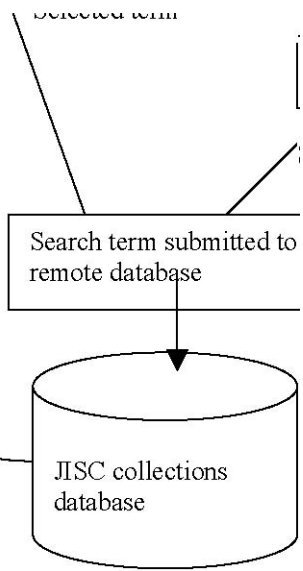
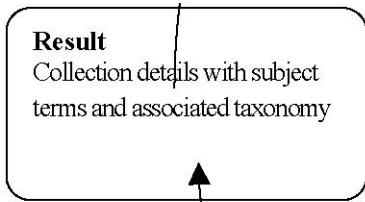
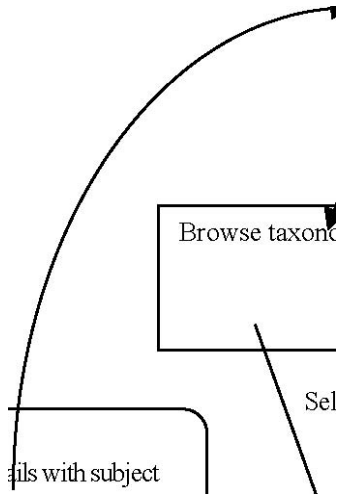
Figure 5. HILT pilot terminologies staff interface

Functional Description of Pilot Server

HILT search mechanisms

Users can either browse the taxonomy or search by typing in a query. The overall architecture of the system is shown below.





Search algorithm

The search process is quite complex and not intuitively obvious. During testing it became clear that no single algorithm could give the best results for all search terms. The resulting process appears to give useful results for most search terms tried, but is not guaranteed to give the best results for all possible search terms.

The system goes through the following steps when it receives a query.

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

User

Query

Search taxonomy

Browse taxonomy

Set of results

Selected term

Result

Collection details with subject terms and associated taxonomy

Disambiguation

Selected term

Search term submitted to remote database

JISC collections database

Multi-word queries

Before displaying results from multi-word queries, the system removes duplicates and assigns weights to individual items. If the same item has been retrieved as a result of a search with different words, that item gets a higher weighting (ranking) in the display of search results. The merging and ranking of search results gives the same effect as an AND search followed by an OR search.

If the user types in boolean terms such as AND or OR these will be stripped out as stop words and ignored. Each remaining word is then handled individually.

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

JISC collections database

The collections database is stored on a separate server and uses different database software to the HILT server. Once the user has selected a term (either by browsing or by search and disambiguation), the system identifies any collections relevant to that term by searching the collections database, then displays the collection name and description along with any subject terms relevant to that collection.

Searching the collections database involves the following steps:

1. The system retrieves the DDC number of the selected term along with the features (subject terms, taxonomy and relationship) corresponding to that term stored in the Wordmap database.
2. DDC numbers can be a single number (371.11) or a range (371.12-18). In the case of a range of DDC numbers, the system retrieves all the collections in that range. Otherwise, it retrieves collections relevant to the single DDC number.
3. The system also retrieves some broader collections. For example, if the DDC number of the term is 371.2134, the system retrieves collections with DDC number 371.2134, 371.213, 371.21, 371.2, 371, and 370. If there are no results for all these searches the system adopts a pattern searching to retrieve related collections, e.g. all collections with DDC number starting 371.
4. If the selected term is from DDC table 2 or table 6 (standard subdivisions) rather than the main DDC schedule, the table number is converted to a DDC number (table 2 maps to the DDC 900s and table 6 maps to the DDC 400s) and then treated as a DDC number when searching the collections database. For example, a search for 'Cairo' will retrieve T2-621.6 from table 2, which will retrieve any collections with DDC number 962.
5. If the collection allows remote searching by appending a variable search term to a fixed partial URL (as in the OpenURL standard), and if it uses one of the recognised taxonomies, then the system offers the user the option of dynamically searching the remote collection using the appropriate term provided by the terminology server. In order for this function to operate, the collections database has to include the partial URL to which search terms can be appended and remotely submitted (as well as the URL of home page of the collection), e.g.:

<http://www.data-archive.ac.uk/search/indexSearch.asp?ct=xmlKeywords&q1=>

At present this option is possible with only a small numbers of collections.

DDC data importing and mapping

The DDC 21 schedules are supplied by OCLC as a single large (50Mb) XML file. DDC tables are supplied in a separate (10Mb) XML file. A Perl program was written by Wordmap to convert this to Wordmap's own XML file format, so that it could be loaded into the Wordmap software. Various problems with this program were identified and fixed and the data reloaded several times, to correct hierarchies, character sets etc. This was a slow and uncertain process, and a more flexible and manageable method was required. A Visual Basic program was therefore written by CDLR to parse the OCLC XML file and load it into a relational database, then export it in Wordmap XML format. This had several advantages over the original method for data import:

- It gave far more precise control over which components of the DDC data were loaded into Wordmap.
- It ensured correct hierarchies were created.
- It improved database performance by removing numerous notes and other parts of the DDC schedules not being used for the HILT server.
- It allowed standard subdivisions to be parsed and imported.
- It allowed other taxonomies to be handled in a similar manner, e.g. Unesco.

- It offered a means of applying mappings between terms to the XML file, before loading it into Wordmap, enabling possible automation.
- It offered a means of applying existing mappings to any new versions of DDC.

Wordmap data structure

The structure of a Wordmap taxonomy has three major components: *terms*, *synonyms* and *features*. The DDC schedules are represented in a Wordmap taxonomy as follows:

Wordmap	DDC
Leadword	Heading
Synonyms	Relative index entries
Features	DDC numbers, LCSH terms, mappings, notes etc
Wordset (all the above)	Entry (all the above for a given DDC number)

Synonyms may be in any language and must have an associated language code. Only English is used at present. There are numerous types of feature, and new features types may be added. Wordmap provide a number of APIs (application programming interfaces) which allow various types of predetermined search, and retrieval of different elements of a wordset. Only leadwords and synonyms are currently searchable, not features. Therefore, in order for LCSH terms and index entries to be searchable, they have to be added as synonyms as well as features.

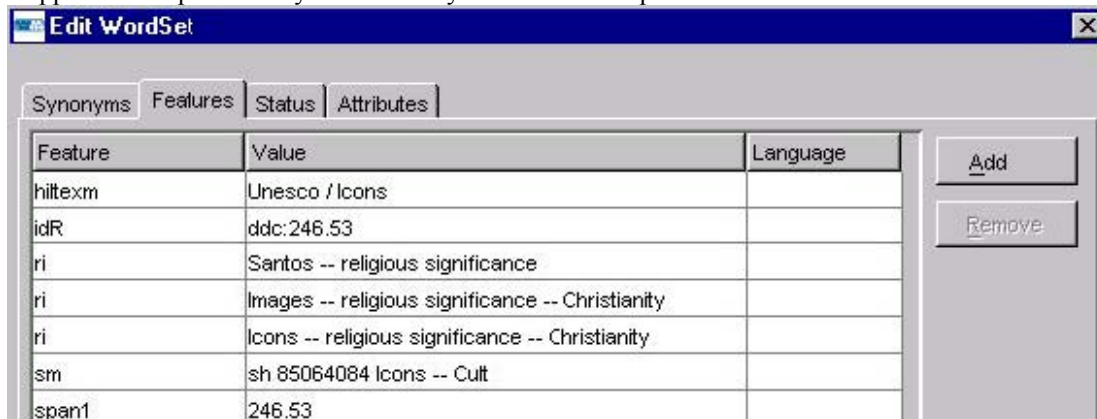
Mappings

All mappings held in Wordmap are currently implemented as features. The Wordmap software provides a function called inter-taxonomy links, which in theory may be used for mappings, but these do not currently provide the functionality required for the pilot terminology server.

A successful mapping between DDC and another taxonomy requires four pieces of information:

- DDC number, e.g. 246.53. This functions as a unique concept identifier.
- Taxonomy name, e.g. Unesco
- Mapped term, e.g. Icons
- Match type, e.g. exact, singular/plural, concept

In order to implement mappings in Wordmap, the match type is implemented as a feature type, and the mapped term is preceded by the taxonomy name. For example:



Here the *span1* feature is the DDC number, *sm* refers to a statistical mapping between DDC and LCSH (provided by OCLC) and *ri* refers to relative index entries, which are currently stored as both synonyms and features.

Although several match types have been identified, only three are currently used:

- hiltexm*: exact match, e.g. Icons / Icons
- hiltspm*: singular / plural match, e.g. Tooth / Teeth
- hiltctm*: concept match, e.g. Medical ethics / Ethics of medicine

All other feature types are either defaults provided by Wordmap or represent XML tags held in the source DDC file supplied by OCLC. (Single DDC numbers are identified by *span1*, while ranges are identified by *span1* and *span2*.)

Mappings can be added manually, using the Add button in the Wordmap interface, as shown above, or can be applied to the XML file before importing the data file to Wordmap. Both methods have been used successfully. Using the Wordmap interface is simpler and more user-friendly, while applying mappings to the XML file enables automation and aggregation of mappings (in a known format) from multiple sources, as well as providing a means for mappings to be applied to a different source data file (e.g. DDC22).

4. Illustrative Mappings from Pilot Server

See <http://hiltpilot.cdlr.strath.ac.uk/pilot/examples/>