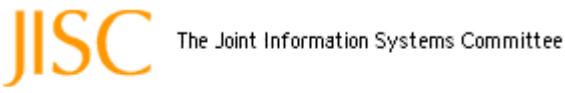
Delivering HILT as a JISC IE shared service

Rachel Heery
UKOLN, University of Bath

19 October 2003



Contents Page

1. Executive Summary

This report is intended to review issues related to delivering HILT terminology services within the JISC Information Environment (IE) as one or more machine to machine (m2m) services.

Recommendation 1: Specify in detail all HILT services as separate components.

Recommendation 2: Validate, in collaboration with portals and VLEs, scenarios for use of HILT terminology services. Develop detailed use cases to inform implementation of HILT terminology services.

Recommendation 3: Build scenarios to illustrate discovery and location of resources within a 'rich environment' of terminology services (to include, for example, Google, keyword searching, use of specialist and general controlled vocabularies, mapping services, ranlink analysis, annotations). Scenarios should be based on a variety of actors, using a variety of terminology services. Develop the most compelling scenarios as detailed use cases to inform implementation of HILT.

Recommendation 4: HILT should follow lead of JISC IE regarding alignment with Web Services architecture.

Recommendation 5: Investigate feasibility of collaborative development of open source task manager for use with JISC IE shared services, and in particular for use with HILT and other terminology services.

The delivery of controlled vocabularies in an automated fashion holds the promise of enhanced support for end-users, and new models of service delivery. However the standards on which such services might be based are not mature and are the subject of on-going research. HILT might aim at taking up innovative standards based initiatives (and risk early implementation) or accept that for now terminology services are best delivered using a proprietary based solution (acknowledging that migration to a standards based solution may be needed in the future). This report does not take a view on where HILT should be placed along this continuum from 'research project' to 'operational service'. Recommendations are required of this report so, with the above caveats, concluding recommendations are given below.

Concluding Recommendations*:*

- Provide m2m demonstrator services based on controlled vocabularies mapped within Wordmap. Develop SOAP based interfaces between JISC IE components and Wordmap APIs. Use these sevices in the short term as an aid to firm up use cases, in the longer term as a basis for pilot service if this approach is still appropriate at that stage.

- Carry out investigative implementation of Zthes based solution, whether data is exchanged using Z39.50 or OAI-PMH, with a view to taking advantage of standards based structured controlled vocabularies (particularly faceted vocabularies) as they become available from third party agencies.

- Track developments within the Semantic Web and eScience activities to ensure decisions made now concerning both syntax for structuring vocabularies, and data exchange protocols take account of forward compatibility.

2. Purpose of document

The HILT project is exploring the provision of terminology services. This report is intended to review issues related to delivering HILT terminology services within the JISC Information Environment (IE) as one or more machine to machine (m2m) services. There is potential for HILT to provide an 'infrastructural' or 'shared' service to other components of the JISC IE. The technical architecture of the JISC IE mandates that such services provide access on a 'machine to machine' basis, whilst allowing that a human Web based interface might be provided in addition.

This report will explore issues arising from delivering HILT as a shared service within the JISC IE. It will consider how HILT might interact on a m2m basis with other JISC components, the impact on the HILT technical architecture of delivering such services, and how related developments in the wider world might affect future plans for HILT.

Within this report a ***m2m service*** is characterised as an on-line service enabling:

- interaction between software components with no human intervention
- interfaces between one or more software applications or intelligent software agents

Within this report ***controlled vocabulary*** is used to encompass structured subject headings, thesauri and classification schemes

3. Summary of HILT functionality

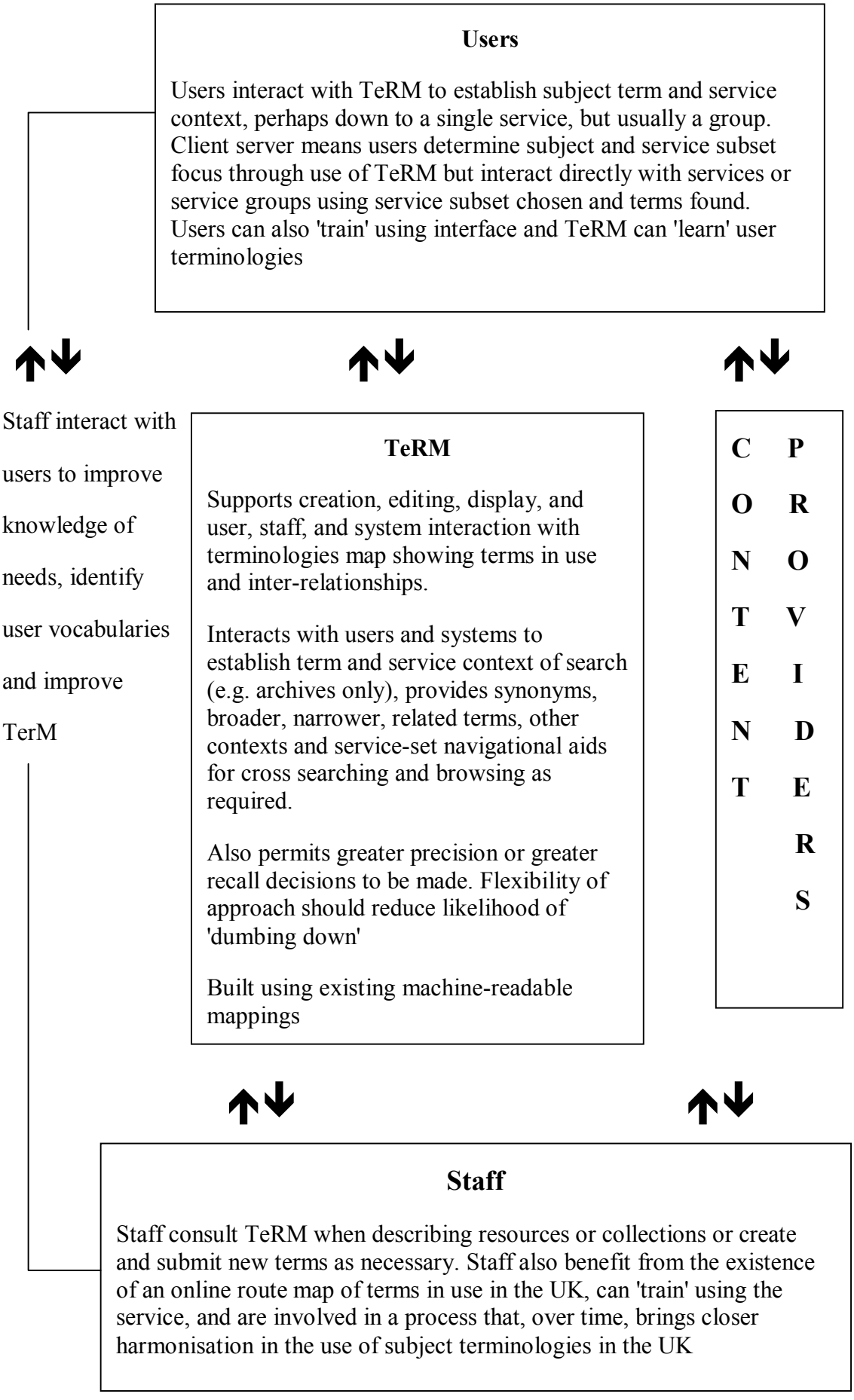
3.1. The problem area HILT addresses

HILT is addressing the difficulties facing an end-user who wants to carry out a subject search or browse across a number of resources that are indexed using different controlled vocabularies. The HILT view is that different controlled vocabularies will continue to be used to describe resources located in different subject domains and different curatorial traditions. A variety of controlled vocabularies exist to meet the requirements of specific communities of use. This is a view shared by many terminology and taxonomy experts, and is borne out by experience. However there is also an overarching requirement to provide a subject based facility for search and browse across the boundaries of discipline and institution. This has been a topic of discussion within the JISC IE since the MODELS Terminology workshop in 2000 [1].

3.2. HILT Phase One

HILT Phase 1 reviewed approaches to improving cross-searching and cross-browsing by subject. The project reported a consensus across the Archives, Electronic Services, Library, and Museums communities in favour of taking forward a pilot project to implement and evaluate an interactive 'route map' to the terminologies used by these communities. The project recommended that JISC fund development and cost benefit analysis of a pilot mapping service. The recommended approach was illustrated in the Phase 1 Final Report by means of an Interactive Terminologies Mapping Roadmap (TerM) [2], here updated to incorporate recent changes in the JISC IE terminology.

Figure 1: HILT Terminology service (TeRM) - Web based interface



The HILT Phase 1 Final Report recommended a solution based on mapping between existing large scale controlled vocabularies, whilst acknowledging the need to accommodate specialist subject-specific schemes and multilingual requirements into the route map in a cost-effective way.

3.3. HILT Phase II

HILT Phase II was funded as a short pilot project running from mid-2002 to late 2003. The aim of HILT Phase II is to build on the work completed in the previous phase, moving to the 'Pilot Project' stage. In order to ensure a realistic scope, it was decided that Phase II would focus on the provision of terminology services at the collection level, whilst recognising the need to extend this in due course to the requirements of item level retrieval. There was no commitment to software development within the project, however the project has developed software to support a pilot HILT Web Interface to Wordmap, a commercially available taxonomy management system.

The project aims were to:

- build an initial pilot focussed primarily on collection level needs.
- determine requirements, costs, and benefits to FE and HE users of the HILT terminology service
- investigate services based on mappings between controlled vocabularies.

In order to inform the pilot, a series of mapping exercises has been carried out as outlined in the HILT Final Report. Using DDC numbers as a central 'spine', HILT maps data between various controlled vocabularies. For example, it can take a DDC number and map this to an associated UNESCO term, or take a DDC number and map to Library of Congress Subject Headings (LCSH).

The HILT pilot interface interacts with human users via a Web interface, however this report will explore the future potential for HILT as a m2m 'shared service'.

4. HILT services

HILT Phase II offers a number of terminology services. These services might in future be delivered as m2m services, invoked by user facing services which provide functionality direct to the end user e.g. portals or Virtual Learning Environments (VLEs) or invoked by by brokers which in their turn interact with portals. In this section we characterise the HILT services in a relatively abstract manner. In the following section, to make the discussion more concrete, we suggest a number of scenarios based on the services HILT offers.

In the HILT context it is useful to distinguish different ways in which m2m services might be delivered:

- automated use of HILT service not requiring user intervention. In this mode services might be invoked 'behind the scenes' by a portal or VLE to add value. The service is requested and delivered in a completely automated way with no intervention from the user e.g. automatic mapping of terms from one scheme to another

- user access to HILT services via an intermediary 'user facing service', such as a portal or VLE, with the portal or VLE providing an interface to HILT e.g. if portal initiates dialogue with end-user to map terms

The same service might be delivered in either way, depending on the particular application. Both of these modes of m2m delivery are in contrast to a human oriented Web based interface direct to HILT.

4.1. Term mapping

HILT will map a term or notation from one scheme to another scheme by means of a central DDC 'spine'.

4.1.1. Issues connected with m2m provision of term mapping

This service relies on creating mappings to the DDC notation for all registered controlled vocabularies. HILT is ambitious in that it is attempting to provide multi-dimension mapping between multiple vocabularies based on a central 'spine' vocabulary. This means that

- the relationships required to fulfil such mappings are likely to be more complex than those provided by ISO 2788: *Guidelines for the establishment and development of monolingual thesauri* [3]. This means that consensus needs to be reached as to how mapping relationships may be expressed in a standard way

- when the result of a query is a group of possible mappings rather than a definitive mapping then disambiguation and contextualisation will be necessary. HILT will need to provide functionality to enable disambiguation (see Leonard Will's evaluation report for HILT Phase 1 [2, p56]).

- such mapping will have inherent limitations as classification and subject heading schemes are inherently pre-coordinate

There are implications in using pre-coordinate and post-coordinate schemes in any mapping service and these need to be taken into account. As noted in Leonard Will's evaluation report for Phase 1, whereas many classification schemes are inherently pre-coordinate, thesauri are often used in a post-coordinate fashion.

Given the interest in faceted retrieval, HILT may want to investigate the possibilities offered by a faceted approach, in particular that being offered in the FAST project by LCSH (see section 8.1.6).

4.2. Disambiguation

If mapping a term to DDC results in more than one result then HILT responds with alternative DDC headings. Thus term mapping may result in a single result or multiple results.

In some applications return of a multiple results may be acceptable. However in other applications the return of multiple results will indicate unwanted ambiguity. In this case, a disambiguation process will be required. Within a m2m context, the disambiguation dialogue may be carried out by the portal or VLE interacting with the end-user.

The disambiguation process might involve the portal or VLE requesting from HILT more context for a particular term e.g. an extended extract from the DDC hierarchy surrounding the term.

4.2.1. Issues concerned with m2m provision of disambiguation

The disambiguation service relies on HILT returning all instances of matching DDC headings to the portal with sufficient context to allow the user to choose appropriate term(s).

4.3. Collections Finder service

Given a specific DDC number (typically derived from term mapping), HILT, in interaction with another JISC IE shared service (such as the JISC IE Service Registry), will truncate the DDC number and send the truncated DDC number as a query to the JISC IE Services Registry in order to match DDC numbers in collection descriptions. The truncation is carried out one number at a time in an iterative process until a hit is found. HILT returns details of the collections that are described by this DDC number to the requesting software.

4.3.1. Issues concerned with m2m provision of collections finder service

This service relies on collection descriptions including DDC notation to describe the subject content of the collection. There would need to be consensus amongst collection description creators as to how DDC is used to characterise subjects within a collection.

4.4. Additional services

The following HILT services have been proposed, but have not yet been fully defined.

. Any hits test/rank facility

. User terms monitor

. User training module

. Clustering facility

It may be cost effective to provide at least some of these as m2m services, offering them as shared services rather than developing duplicated functionality across portals and other user facing services. In order to consider the feasibility of this, a more detailed definition is required for each service including user requirements.

Recommendation 1: Specify in detail all HILT services as separate components.

5. Application scenarios

The following scenarios consider the terminology services offered by HILT. The scenarios illustrate a range of uses for HILT, and are not exhaustive. The scenarios are indicative of the way applications might access HILT in a (more or less) automated fashion, and show how HILT services might be combined with other shared services to offer varied functionality. It would be incumbent on an application designer to identify the most effective way to incorporate HILT services within their application.

5.1. Query enhancement

5.1.1. Query enhancement : human Web interface

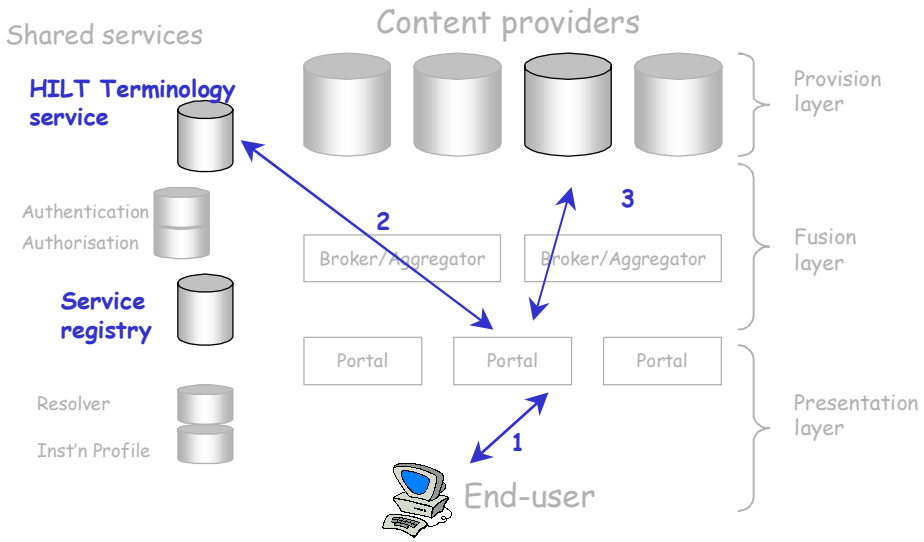
Student wishes to search across JISC IE collections for information about earthquakes (images, journal articles, material held in archives, maps). Aware that material will be indexed using different schemes and vocabularies, the student clicks on the HILT terminology service bookmarked in their browser. The student enters the term 'earthquakes' as a query to HILT. HILT returns subject headings and classification numbers relevant to 'earthquakes' from a variety of subject schemes. Student can link to other JISC or institutional services that they know about, then use the subject headings and classification numbers suggested by HILT as entry terms for searching individual collections, or for broadcasting a cross-search.

5.1.2. Query enhancement : m2m

Student wishes to search across specified JISC IE collections for information about earthquakes (images, journal articles, material held in archives, maps). The user clicks on the earth sciences portal bookmarked on their browser. User selects collections of interest by means of collection descriptions listed within the portal display. User enters 'earthquakes' as a search term. The portal accesses HILT automatically and enhances the term 'earthquakes' with mappings from other thesauri, classification number, synonyms. The original search term plus mappings are used automatically by the portal as the basis for searching the collections already specified by the user.

One possible workflow for this scenario follows:

Figure 2: Query enhancement: m2m workflow



1. User selects collections of interest within portal by means of collection descriptions (previously downloaded by portal from IE Service Registry). User enters 'earthquakes' as a search term.

2. The portal accesses HILT automatically and enhances the term 'earthquakes' with mappings from other thesauri, classification number, synonyms.

3. The eenhanced search terms are used to search collections already specified by the user.

5.2. Disambiguation of terms

5.2.1. Disambiguation of terms : human Web interface

User enters term into HILT terminology service in order to perform query enhancement. HILT responds with information regarding different meanings, or different contexts of the term. User can select which meaning(s) or context(s) are relevant before proceeding to use these terms within a search.

5.2.2. Disambiguation of terms: m2m

User enters term 'wireless' into portal search, requesting query enhancement. The portal passes the term to the HILT mapping service which returns multiple hits. The portal displays the hits to the user and asks if they want to choose any or all of the terms, or whether they want more context from the hierarchy around each term. The user requests more context for one of the terms. HILT returns an extract from the

DDC hierarchy. The user then selects terms for query and the portal sends the enhanced terms to a content provider.

5.3. Collection finder

5.3.1. Collection finder: human Web interface

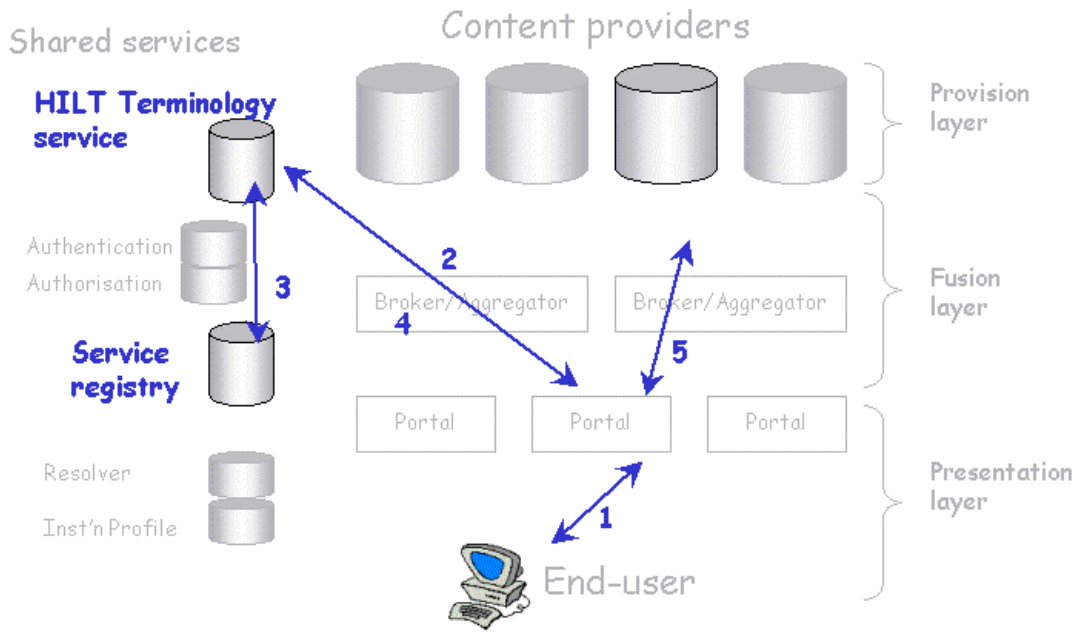
Student wishes to locate appropriate collections within the JISC IE which will provide information about Georgian architecture. Student enters search phrase into HILT. HILT searches all registered vocabularies to find match. The term is mapped to DDC notation. HILT searches IE Service Registry by DDC notation, if no matches found DDC notation is truncated by one digit and search is repeated. This process is carried out in iterative way until a specified number of matching collections are found. HILT returns details of collections to student.

5.3.2. Collection Finder: m2m

Student wishes to locate appropriate collections which will provide information about Georgian architecture within the JISC IE. Student enters search phrase into institutional portal requesting collection finder service. Portal searches HILT mapping service to obtain DDC notation. HILT searches IE Service Registry collection descriptions for a match to DDC notation. If no matches found, DDC notation is truncated by one digit and search is repeated. This process is carried out in iterative way until a specified number of matching collections are found. HILT returns details of collections to portal. Portal directs original search term, enhanced with term mappings, to these collections.

One possible workflow for this scenario follows.

Figure 3: collection finder m2m



1. User enters search phrase into portal requesting collection finder service.

2. Portal searches HILT mapping service to obtain DDC notation.

3. HILT searches IE Service Registry collection descriptions for a match to DDC notation. If no matches found, DDC notation is abridged by one digit and search is repeated. This process is carried out in iterative way until a specified number of matching collections are found.

4. HILT returns details of collections to portal.

5. Portal directs original search term to identified collections.

5.4. Enriching metadata creation

5.4.1. Enriching metadata creation: human Web interface

A librarian is creating metadata, using Dublin Core, to describe a journal article about earthquakes. The librarian wishes to include a variety of subject terms identified by subject scheme. The librarian clicks on the HILT terminology service bookmarked in their browser, and enters the term 'earthquakes' as a query to HILT. HILT returns subject headings and classification numbers relevant to 'earthquakes' from a variety of subject schemes. The librarian enters the various terms and classification numbers as subject properties in the metadata.

5.4.2. Enriching metadata creation: m2m

An author is creating metadata using the DC-dot tool for a paper they have written. The author enters a number of terms in the dc:subject field. DC-dot automatically send these terms to HILT to obtain DDC notation and mappings to LCSH. DC-dot adds these values to the metadata record. If disambiguation is required DC-dot will enter a dialogue with the author to clarify which terms are appropriate.

5.5. Collection finder with scheme identification

5.5.1. Collection finder with scheme identification: human Web interface

A researcher wants to carry out a literature search for information on climate change. The researcher enters terms into HILT and requests collection finder service with scheme identification option. HILT locates appropriate DDC notation, and initiates the collection finder service. Once appropriate collections have been identified, HILT queries the JISC IE metadata schema registry which provides access to the schemas in use in JISC IE collections. HILT requests details of the subject schemes in use in each of the identified collections. HILT then maps the DDC notation to those schemes and returns the names of collections and schemes to the researcher.

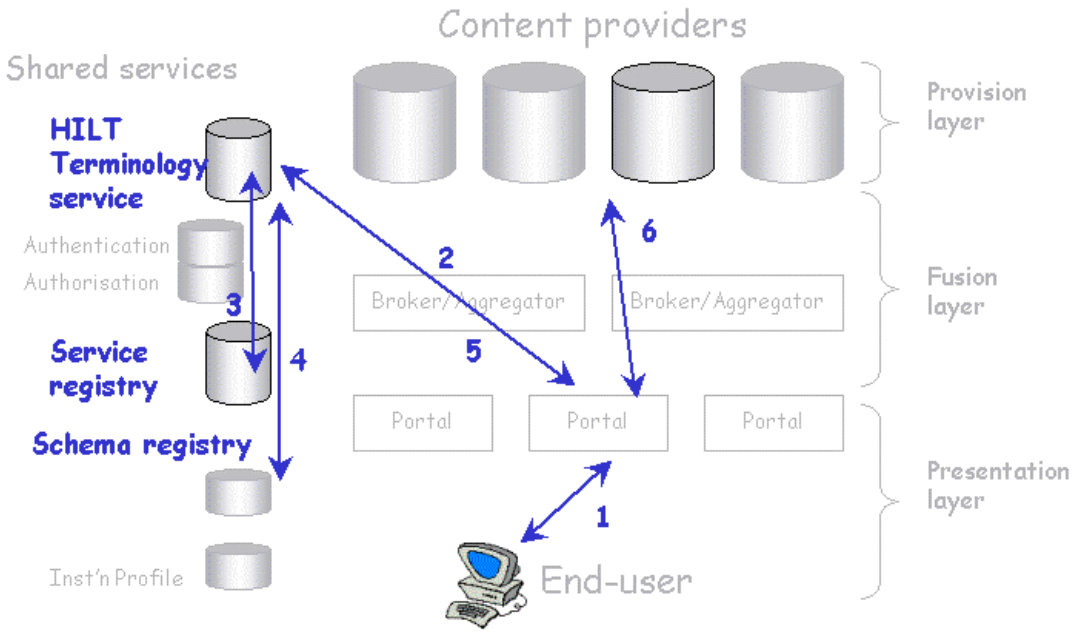
5.5.2. Collection finder with scheme identification and mapping: m2m

A researcher wants to carry out a literature search for information on climate change. The researcher enters terms into Earth Sciences portal requesting the 'comprehensive search' option. The portal sends a request to HILT for the collection finder service. HILT locates appropriate DDC notation, and initiates the collection finder service. Disambiguation process is carried out if appropriate. Once relevant collections have been located, HILT queries the JISC IE metadata schema registry which provides

access to the schemas in use in JISC IE collections. HILT requests details of the subject schemes in use in each of the identified collections. HILT then maps the DDC notation to these schemes and returns the names of collections and schemes to the portal. The portal then identifies appropriate content providers for the identified collections by searching its own sub-set of favourite service providers, previously downloaded from the JISC IE Registry. The portal then sends the appropriate search terms to each service provider.

One possible workflow for this scenario follows.

Figure 4: m2m Collection finder, scheme identification, mapping



1. A user enters terms into portal requesting the 'comprehensive search' option.

2. The portal sends a request to HILT for the collection finder service. HILT locates appropriate DDC notation. Disambiguation process is carried out if appropriate

3. HILT initiates the collection finder service, sending DDC notation to JISC IE Service Registry, iteratively truncating DDC number and matching against DDC within collection descriptions until relevant collections have been located,

4. HILT queries the JISC IE Metadata Schema Registry which provides access to the schemas in use in JISC IE collections. HILT requests details of the subject schemes in use in each of the identified collections. HILT then maps the DDC notation to the various schemes that have been identified.

5. HILT returns to portal the identifiers of relevant collections and appropriate terms for searching each collection.

6. The portal then locates appropriate content providers for the identified collections by searching its own sub-set of favoured service providers, previously downloaded from the JISC IE Registry. The portal then sends the appropriate search terms to each of these content providers.

5.6. m2m interaction between JISC IE components

Consultation with portal and VLE designers (and users) is required to validate these scenarios and to work up more detailed use cases. It can be seen from the above scenarios that the interaction between components in the JISC IE can be complex. In addition it should be noted that some transactions between shared services will be carried out in a dynamic fashion whilst in other cases the portal or VLE will 'embed' a sub-set of data from a shared service. As currently envisaged, the scale and complexity of relationships within HILT would indicate that m2m access to HILT would be dynamic. In contrast, recent discussions (within JISC meetings) of the portals' use of the JISC IE Services Registry have not focused on dynamic access. It is assumed that most portals and VLEs will download collection and service descriptions from the IE Services Registry, store the descriptions locally and query their own local database.

Note that the metadata schemas registry is at proposal stage and is mentioned for illustrative purposes only within the above scenarios.

Recommendation 2: Validate, in collaboration with portals, VLEs, and brokers, scenarios for use of HILT terminology services. Develop detailed use cases to inform implementation of HILT terminology services.

6. Placing HILT in wider context

When considering scenarios, the HILT project acknowledges that it will not exist as a terminology service in isolation. Far more likely is that there will be a number of distributed terminology services based on both specialist controlled vocabularies, and large-scale general controlled vocabularies. Research is being carried forward in a number of communities (digital library, Semantic Web, e-science) to enable existing individual controlled vocabularies to be accessed in m2m fashion, whether as 'standalone' vocabularies or as more complex structured mappings between vocabularies.

If one considers user requirements, services attached to a particular controlled vocabularies may be appropriate within a particular service settings, and services providing access to mapping between vocabularies in other settings. In addition we need to keep in mind that subject access is just one approach to resource discovery and needs to be positioned alongside other services, for example services based on web based search engines such as Google, free text indexing, link analysis, annotation

or other forms of metadata. A workshop of the ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis held in 1999 [4] recommended that use of subject headings from standard schemes needs to be used in conjunction with for example:

- A combination of keywords and controlled vocabulary should be used to allow users the choice of simple free-text indexing as well as complex controlled vocabulary indexing.
- Use of multiple vocabularies should be accommodated. For a general vocabulary covering all subjects, the Subcommittee recommends the use of LCSH or Sears with or without modification.
- In order to achieve the desired level of specificity, controlled vocabulary terms assigned to the metadata record could be supplemented and complemented by keywords and other subject-related elements, such as title, abstract, statement of content, etc.
- Synonyms should be handled by system design implementation of the controlled vocabulary or thesaurus. If this is not available, an alternative is to include all identified synonyms and related terms, along with the keywords, in the metadata record.
- Tools such as online thesaurus display should be developed to provide access to controlled vocabulary structures, showing both hierarchically (broader and narrower) and horizontally related terms.

Constructing scenarios for such a rich environment is considered outside the scope of this report, but is recommended as a necessary step to inform further work on HILT.

Recommendation 3: Build scenarios to illustrate discovery and location of resources within a 'rich environment' of terminology services (to include, for example, Google, keyword searching, use of specialist and general controlled vocabularies, mapping services, link analysis, annotations). Scenarios should be based on a variety of actors, using a variety of terminology services. Develop the most compelling scenarios as detailed use cases to inform implementation of HILT.

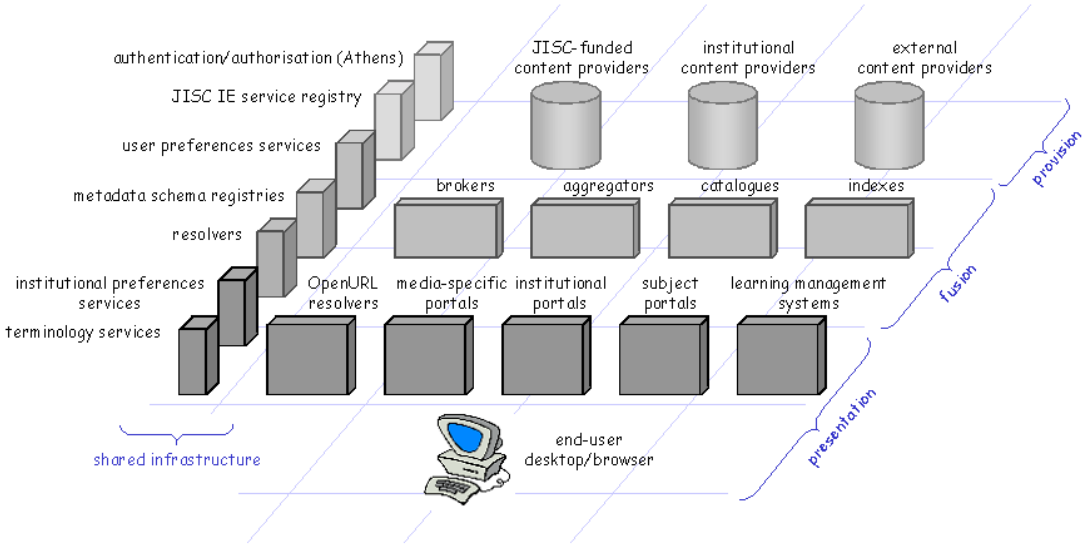
7. Placing HILT in context of JISC IE

The JISC IE technical architecture specifies a set of standards and protocols that support the development and delivery of an integrated set of networked services that allow end-users to discover**,** access**,** use and publish digital and physical resources as part of their learning and research activities

7.1.1. The JISC IE layers

The JISC IE architecture is made up of 'layers' as illustrated in the following diagram [5].

Figure 5: JISC IE architecture



The HILT service is intended to form part of the ***shared service infrastructure*** of the JISC IE. Infrastructural Services are a range of shared structured network services that are called on by content providers, brokers, aggregators, indexes, catalogues and portals. Such services might include authentication, authorisation, service registry, user profiling, resolver, institutional profile, metadata schema registry and terminology services. Some of these services already exist (authentication and authorisation), some are in development (service registry) whilst others are still at proposal stage (e.g. metadata schema registry).

Within the definition of the JISC IE, HILT would be defined as a 'transactional network service'. Transactional network services are those that are not primarily concerned with the provision of access to a 'content collection', and might include format conversion, printing, authentication or e-commerce services. Transactional services are distinguished from 'informational network services'. Informational network services include those that provide access to, or metadata about, items or collections at a digital location. Examples include Web sites, document supply services, abstracting and indexing services, data archives, online catalogues, databases, email archives, etc.

The key standards and protocols specified in the technical architecture are listed in the JISC IE Architecture Standards Framework [5]. Transactional services are covered by the following clause:

All JISC IE structured network services not covered by the specific cases listed above ***should*** *be offered using the Simple Object Access Protocol (SOAP) 1.1. Alternatively, the use of HTTP 1.1 GET or POST requests to return XML documents* ***may*** *be appropriate.* [5]

This means, according to the JISC IE framework, HILT would need to be accessed either using the SOAP protocol, or using HTTP GET/POST.

However in order to be delivered in a m2m way, HILT also needs to meet constraints of a Structured Network Service in JISC IE terms.

*A **Structured Network Service** is a network service that provides structured access to structured resources. Structured network services are intended for use by software applications. Examples of structured network services are those based on Z39.50, the OAI-PMH, RSS/HTTP and SOAP. Note that an HTML-based Web site is not 'structured', in the sense that it does not provide structured access to structured resources.* [5]

This means, to comply with the JISC IE framework, HILT would need to be accessed in a structured way, for example using SOAP wrappers around structured query semantics. However in addition the resource itself, the controlled vocabularies and mappings, all need to be structured.

It would be possible for HILT to use SOAP or HTTP GET/POST to APIs (such as those available from Wordmap or other commercial products), in order to comply with this framework. However it would be preferable in addition for the query protocol, and the structure of the controlled vocabularies, to follow widely agreed standards. In effect the structure of the query language, the target data and the structure of returned records preferably should follow an agreed standard. This is particularly the case given the intellectual investment required for vocabulary mapping.

As yet there is no widely agreed standard for structuring thesauri/classification schemes/subject headings in a machine readable way. Similarly there is no widely agreed standard protocol for querying a controlled vocabulary. The JISC IE Standards Framework has not specified standards to cover terminology services as yet. There are a number of alternative approaches which whilst 'experimental' might be considered for implementation at the present time. These approaches are considered below, and should be contrasted with using Wordmap APIs as described in a later section. One possible option is that, until standards are widely agreed, HILT might best rely on a proprietary product.

7.1.2. Web Services within the JISC IE

The Web Services architecture (in terms of the triangular relationship between service requestor, service provider and service registry) layers onto the JISC IE architecture [6]. In addition the use of SOAP and XML within the JISC IE fits with the Web Services approach. However the 'self-description' of service components using Web Service Description Language (WSDL), with descriptions made available through a Universal Discovery, Description and Integration (UDDI) registry has not been implemented as yet within the JISC IE.

Recommendation 4: HILT should follow lead of JISC IE regarding alignment with Web Services architecture.

7.2. m2m workflow within JISC IE

In a m2m context the interactions between IE components and the HILT services (as outlined in the illustrative workflow diagrams in section 5) would need to be managed by a 'Task Manager'. There needs to be consideration as to which components within the JISC IE would carry out task management. The task manager might be located at the 'client side' e.g. as part of the portal functionality, or alternatively as part of a shared service, or indeed as part of a broker; or the task management function might be split between components. Within the JISC IE development programme there may be options for developing 'generic' task management components, or for exploiting an open source model to collaborate on development of task management modules within existing service components.

Recommendation 5: Investigate feasibility of collaborative development of open source task manager for use with JISC IE shared services, and in particular for use with HILT and other terminology services.

8. Terminology services

There is growing interest in exploiting controlled vocabularies (whether traditional controlled vocabularies or 'new vocabularies') in the context of knowledge management in a Web environment. Controlled vocabularies are being explored as a basis for subject access within large scale resource discovery systems, and for individual Web site architectures. NKOS is an informal group providing a focus for exchange of information regarding activity in this area [7]. Some of this activity is research oriented, elsewhere the commercial sector is involved in developing new products. Any longer term planning regarding terminology services needs to track activity within a number of research communities, including the Semantic Web, eScience, knowledge management, and digital library communities; as well as tracking activity within the commercial sector.

Exploitation of controlled vocabularies in an interoperable manner depends on the adoption of standards. Agreed standards covering the structure and syntax of controlled vocabularies will enable adoption of standard protocols for accessing vocabularies, and agreed formats for data exchange.

Emerging standard activity related to controlled vocabularies will be considered under the broad headings of

- Legacy standards
- Structure and syntax
- Protocols

8.1. Legacy standards

8.1.1. International standards for thesauri

The structure of thesauri are covered by existing standards, however these are largely oriented towards the print world and there are various plans for revising these standards.

ANSI/NISO Z39.19-1993 ***Guidelines for the construction, format and management of monolingual thesauri***

NISO launched an initiative in 2003 to revise this standard in order to address the needs of a changing information environment and an audience which now includes system developers and metadata creators, as well as librarians and information professionals. The revised standard is intended to

- Introduce more user-friendly language and include justifications and explanations of important concepts and principles.
- Update the content to reflect new technology, the current electronic information environment, the ways that users search or browse, and the types of content they will find.
- Expand the scope to a wider variety of producing organizations and content - beyond the traditional abstracting & indexing services - and add examples that are relevant to business and industry.

BS5723:1987 (ISO2788-1986) British standard guide to establishment and development of monolingual thesauri */ British Standards Institution. 1st rev.*

BS6723:1985 (ISO5964-1985)British standard guide to establishment and development of multilingual thesauri / British Standards Institution.

These British Standards are in the process of being revised to form BS 8723: a new British standard for structured vocabularies . The intention is to update the standard in view of changes in technology, in particular so that it can be applied to electronically stored vocabularies, to consider mapping between vocabularies, and to consider formats and protocols for exchanging data with other applications. The aim is to release the first two parts of the draft standard (General and Thesauri) for comment in 2003. Future work will cover: vocabularies other than thesauri (such as classification schemes, taxonomies, subject heading lists, ontologies), interoperability between vocabularies (mapping between same language vocabularies, mapping between multilingual vocabularies), and interoperability with applications.

Some researchers in the field have suggested that the relationships within thesauri, particularly the several relationships covered by 'Related Term' should be more precisely differentiated and that this work should be reflected in standards revision. Such refinements would allow for more effective automated use of thesauri. However this contrasts with the view of those who are concerned that any such move would be prohibitively expensive for large-

scale vocabularies. Certainly expressing deep semantic relationships is more likely to be feasible in the context of small scale ontologies, in particular vocabularies intended to support inference engines.

8.1.2. MARC 21 formats for authority and classification data

The MARC 21 formats are standards widely used within the library world for the representation and exchange of authority, bibliographic, classification, community information, and holdings data in machine-readable form. They consist of a family of five coordinated formats: MARC 21 Format for Authority Data; MARC 21 Format for Bibliographic Data; MARC 21 Format for Classification Data; MARC 21 Format for Community Information; and MARC 21 Format for Holdings Data.

The MARC 21 Format for Classification Data [8] supports description of classification numbers and captions.

The MARC 21 Format for Authority Data [9] is designed provide information concerning the authorized forms of names and subjects. The format allows for the expression of simple relationships between terms.

Neither of these formats is designed to express deep semantic relationships nor to accommodate mapping relationships across different vocabularies. However these are familiar record formats with wide deployment in bibliographic systems, and as such have potential for a library user base.

8.2. Structure and syntax

It is important to be aware that much of the activity described in the following section is still at the 'research stage', some of the emerging formats have only had test implementations with no operational deployment. In addition much of the current activity focuses on the use of thesauri rather than classification schemes. In general the various initiatives might also be applied to classification schemes, but there is even less implementation experience for such applications,

8.2.1. VDEX

Emerging from the IMS community, VDEX [10] specifies a 'mark up language' (or grammar) for controlled vocabularies. VDEX is designed to facilitate the exchange of controlled vocabularies (value lists). VDEX also permits additional information to be included with the definition of the value domain to allow the user of the terms to receive scope notes to help them apply the correct term. VDEX is not intended to support the expression of all possible vocabularies:, it is targeted at the light-weight expression of simple value lists. VDEX is not intended as a modelling language for vocabularies.

VDEX aims to enable expression of simple machine-readable lists of human language terms together with information that may aid end-users to understand the meaning or applicability of the various terms by means of scope notes

8.2.2. The Vocabulary Markup Language (Voc-ML)

The Vocabulary Markup Language (Voc-ML) is a draft XML DTD [11]. The DTD includes Dublin Core metadata that would describe the knowledge organisation systems being encoded. It also defines tags and syntax for uniquely identifying each term, its relationship to other terms, and descriptive information like scope notes and definitions. It is intended for use with a range of different types of system, including authority files, hierarchical thesauri, classification schemes, digital gazetteers, and subject heading lists. There is no evidence of deployment.

8.2.3. Thesaurus Interchange Format for the Semantic Web (TIF)

TIF [12] takes a concept-oriented approach to modelling thesauri, defining relationships between concepts rather than terms, and allowing for equivalence relationships as well as hierarchical relationships. This model is hospitable to mapping equivalent concepts between thesauri where the same concept may be described by different terms. An RDF schema has been specified based on the concept-oriented model, which is being proposed a a candidate interchange format for the Semantic Web. A TIF Simple (TIFS) schema has also been defined for properties within a term-oriented thesaurus in order to provide a 'term view' of a thesaurus. It is recommended that term-oriented properties be derived from concept-oriented properties.

Work is in progress to refine relationships expressed within the TIF schema using the Ontology Web language (OWL).

8.2.4. Faceted approaches

The faceted approach to indexing and retrieval has a long history going back to Ranganathan's COLON classification scheme, with more recent implementation within the Bliss classification scheme and the PRECIS indexing system. A faceted approach can be contrasted with the largely enumerative approach in many traditional controlled vocabularies such as DDC and LCSH. In such pre-coordinate schemes compound terms are explicitly listed within the hierarchical scheme. Whereas in a faceted system, terms are divided into high-level categories, or facets. Faceted systems are synthetic; they do not attempt to include the vast number of possible multi-concept headings or descriptors in a domain, but combine terms from a limited number of fundamental facets, as needed when indexing or querying [13, 14].

The development of FAST (Faceted Application of Subject Terminology), is underway at OCLC [15]. FAST is a faceted adaptation of LCSH with a simplified syntax that retains the very rich vocabulary of LCSH while making it easier to understand and apply. FAST is derived from LCSH, redesigning LCSH as a post-coordinated faceted vocabulary for an online environment. The first phase of the FAST development includes the development of facets based on the vocabulary found in LCSH topical and geographic headings and is limited to six facets: topical, geographic, form, period, with the most recent work focused on faceting personal and corporate names.

8.3. Protocols

8.3.1. Zthes profile

Zthes [16] is a Z39.50 profile for thesaurus navigation. The profile describes an abstract model for representing and searching thesauri (hierarchies of terms as described in ISO 2788) and specifies how this model may be implemented using the Z39.50 protocol. The protocol suggests a Zthes DTD for XML which is provided as an appendix to the profile. Other data formats such as MARC21 might also be used. The profile has been stable since 2001but has been used for experimental purposes only, and as yet has not been widely deployed.

The OCLC Metadata Switch project has undertaken experimental work providing terminology Web Services using Zthes over SOAP (SRU/W) [17].This work has been undertaken using controlled vocabulary structured as MARC21 authority records.

8.3.2. Alexandria Digital Library (ADL) protocol

The ADL protocol [18] specifies an XML- and HTTP-based protocol for accessing thesauri. The protocol enables software to access and utilize thesauri and provides for both querying thesauri and navigating within thesauri. The protocol does not support creation, maintenance, or sharing of thesauri, or mapping between thesauri.

The protocol provides five independent, stateless services which allow for

- Returning the thesaurus's properties.
- Returning a list of all terms in the thesaurus.
- Queryng the thesaurus by term name and returns a list of the matching terms.
- Returning the hierarchy of terms above (broader than) a given starting term.
- Returning the hierarchy of terms below (narrower than) a given starting term.

There is no evidence of deployment outside the project.

8.4. Use of OAI-PMH to harvest authority records

OCLC, in collaboration with OAI, have been experimenting with provision of thesaurus services for both m2m and human end-user use. The OCLC Office of Research Terminology Services project [19] is investigating cross-thesaurus linking alongside other means of improving access to thesauri. Under the auspices of the American Liibrary Association ALCTS group, machine-readable authority records have been created for the form/genre headings in the first chapter of GSAFD (*Guidelines on Subject Access to Individual Works of Fiction, Drama)*. The file contains 153 records, so is modest in size compared to general classification schemes, but enables experimentation and proof of concept demonstrators.

The GSAFD records have been enriched with information about mappings to LCSH, converted to XML and stored in an OAI-PMH compliant repository. An "MT" label has been introduced as an extension to the Z39.19 standard to indicate mappings to different subject schemes [20].

An OAI-PMH item exists per thesaurus term, and three OAI-PMH metadata formats have been made available per item:

- Simple Dublin Core
- MARC21 authority file record
- Z39.19 thesaurus format

The GSAFD Thesaurus can then be accessed in various ways using the OAI-PMH protocol:

- Interaction with the end-user via a Web browser to a Z39.19 record. Note that this relies on introducing a reference to an XSLT stylesheet into the OAI-PMH protocol request.
- Interaction by machines through OAI-PMH-based Web Services mechanisms.
- Harvesting by OAI-PMH service providers gathering thesaurus records

If more mapping data was made available in this way there would be potential to use this approach to gather records mapping other vocabularies.

9. Wordmap

HILT has based its pilot Web user interface on the WordmapTaxonomy Management System [21]. This is a commercially available product that supports management of multiple controlled vocabularies in a single user interface. It also supports management of partial views of controlled vocabularies, and mapping between different controlled vocabularies.

Wordmap is designed to make controlled vocabularies accessible to other applications using an Application Programmer's Interface (API). HILT has built its own customised Web interface to Wordmap based on available APIs.

The APIs may be exploited by any application (or mediating software) that can make calls to an ODBC datasource. Wordmap is a client server application, with a Java client calling an embedded Oracle 9i Standard Edition Database using JDBC. It serves taxonomy infrastructure and data through APIs, or XML/Web Services using SOAP to applications.

In summary the API's offered cover:

- Basic wordset information
- Wordset hierarchical information
- Information about the controlled vocabulary as a whole
- Miscellaneous utilities

Further development of terminology services based on Wordmap (or other commercially available products) is one potential way forward for HILT. Wordmap fits with the basic JISC IE architecture mandate for transactional services to be offered using SOAP. What Wordmap does not offer is a standard structure for controlled vocabularies as a basis for interoperable query and data exchange.

Vocabularies need to be exported according to a standard to facilitate interworking, and at some stage one might expect the JISC IE to advise on a preferred standard for controlled vocabularies. However we have seen in the previous section as yet there is little consensus on standards in this area, and even less deployment.

If HILT progresses provision of a pilot m2m service using Wordmap then the project needs to consider

- assuring intellectual effort invested in defining mappings between vocabularies is not lost. It would be advisable to investigate and test export facilities from Wordmap, and plan possible migration routes to a more standard based system (keeping in mind that Wordmap might implement a standards based solution iself when consensus emerges on a particular standards based solution). In the medium term this might involve introducing protocol 'gateways' to accommodate variant use of protocols.
- Monitoring emerging 'candidate standards' for thesaurus navigation, taking these into account in current decisions regarding structuring data, thereby ensuring future compatibility as far as is possible.

10. Future directions

The delivery of controlled vocabularies in an automated fashion holds the promise of providing enhanced support to end-users, and new models of service delivery. However the standards on which such services might be based are not mature, and are the subject of on-going research. Some initiatives are taking existing standards (MARC 21, Z39.50) and doing new and interesting things with them, other initiatives are looking to current Semantic Web developments (RDF, OWL) as a way forward. There is little, if any, operational deployment of these initiatives at this early stage. The commercial sector meanwhile if fulfilling the requirements for better knowledge management (particularly in the corporate sector) by developing products that deliver terminology services in a more proprietary way, albeit using SOAP and XML.

Within the JISC IE, HILT is very much a pioneer project, certainly as regards terminology services, and also as regards being one of the first candidate shared services. As such it is somewhat unclear whether HILT should aim at taking up innovative standards based initiatives (and risk early implementation) or accept that for now terminology services are best delivered using a proprietary based solution (acknowledging that migration to a standards based solution may be needed in the future). This report does not take a view on where HILT should be placed along this continuum from 'research project' to 'operational service'. Recommendations are required of this report so, with the above caveats, concluding recommendations are as given below.

Concluding Recommendations:

- Provide m2m demonstrator services based on controlled vocabularies mapped within Wordmap. Develop SOAP based interfaces between JISC IE components and Wordmap APIs. Use these sevices in the short term as an aid to firm up use

cases, in the longer term as a basis for pilot service if this approach is still appropriate at that stage.

- Carry out investigative implementation of Zthes based solution, whether data is exchanged using Z39.50 or OAI-PMH, with a view to taking advantage of standards based structured controlled vocabularies (particularly faceted vocabularies) as they become available from third party agencies.

- Track developments within the Semantic Web and eScience activities to ensure decisions made now concerning both syntax for structuring vocabularies, and data exchange protocols take account of forward compatibility.

11. Further information

The Networked Knowledge Organization Systems/Services (NKOS) working group is concerned with the creation of interactive Knowledge Organization Systems (KOS) accessible over the Web (http://nkos.slis.kent.edu/). Workshops on KOS were held at JCDL and ECDL in 2003 [ref].

SWAD-Europe Thesaurus Activity: A comprehensive list of resources relating to thesauri, including XML formats, RDF formats, standards, online thesauri, papers and presentations iss maintained by Alistair Miles at CCLRC.
http://www.w3c.rl.ac.uk/SWAD/thes_links.htm

12. References

[1] MODELS 11: UKOLN/mda Terminology Workshop. Bath, January 2000.
http://www.ukoln.ac.uk/dlis/models/models11/

[2] Dennis Nicholson *et al. HILT: High-level Thesaurus Project: Final report to RSLP and JISC*. December 2001.
http://hilt.cdlr.strath.ac.uk/Reports/Documents/HILTfinalreport.doc

[3] International Organization for Standardization. ISO 2788: *Guidelines for the establishment and development of monolingual thesauri*, 2nd ed. Geneva: ISO, 1986.

[4] ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis. *Subject data in the metadata record: recommendations and rationale.* Illinois, July 1999.
http://www.govst.edu/users/gddcasey/sac/MetadataReport.html

[5] Powell, Andy. JISC Information Environment Architecture
http://www.ukoln.ac.uk/distributed-systems/jisc-ie/arch/

[6] Powell, Andy and Lyon, Liz**.** *The JISC Information Environment and Web services*. Ariadne Issue 31, April 2002. http://www.ariadne.ac.uk /issue31/information-environments/intro.html

[7] Networked Knowledge Organization Systems/Services. http://nkos.slis.kent.edu/

[8] *MARC21 Concise format for classification data.* Concise ed., Library of Congress, 2002. http://www.loc.gov/marc/classification/eccdhome.html

[9] MARC21 *Concise format for authority data*. Concise ed., Library of Congress, 2002. http://www.loc.gov/marc/authority/ecadhome.html

[10] IMS Vocabulary Definition Exchange. http://www.imsglobal.org/vdex/

[11] Vocabulary Markup Language (VocML), 2000. http://xml.coverpages.org/vocML.html

[12] Matthews, Brian, *et al. Modelling Thesauri for the Semantic Web*, SWAD-Europe Deliverable 8.1 http://www.w3c.rl.ac.uk/SWAD/deliv81.htm

[13] Tudhope D., Binding C., Blocks D., Cunliffe D. 2002. Compound Descriptors in Context: A Matching Function for Classifications and Thesauri. Proceeding of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL 2002), Portland. ACM Press. 84-93. http://www.glam.ac.uk/soc/research/hypermedia/publications/jcdl02.pdf

[14] FACET project page http://www.glam.ac.uk/soc/research/hypermedia/facet_proj/index.php

[15] Dean, Rebecca J. *FAST: development of simplified headings for metadata.* Paper presented at Authority Control: Definition and International Experiences conference, Florence, Italy, Feb. 10-12, 2003. http://www.oclc.org/research/projects/fast/

[16] Taylor, Mike. *Zthes: a Z39.50 Profile for Thesaurus Navigation*, Version 0.5 November 2001. http://zthes.z3950.org/profile/current.html

[17] Vizine-Goetz, Diane. The Terminology Services Project. Presentation at NKOS workshop at ECDL 2003. http://www.glam.ac.uk/soc/research/hypermedia/NKOS-Workshop.php

[18] Janée, Greg., Ikeda, Satoshi., Hill Linda L. *The ADL Thesaurus Protocol Alexandria Digital Library Project.* Version 1.0, 2002-12-09

[19] OCLC Office of Research. *Metadata Switch: Terminology Services.* http://www.oclc.org/research/projects/mswitch/4_termservs.htm

[20] Van de Sompel, Herbert., Young, Jeffrey A., Hickey, Thomas B. *Using the OAI-PMH ... Differently*. D-Lib Magazine, July/August 2003, Vol.9 No. 7/8. http://www.dlib.org/dlib/july03/young/07young.html

[21] Wordmap. http://www.wordmap.com.