Terminology server for improved resource discovery: analysis of model and functions

George Macgregor[1], Emma McCulloch[2] and Dennis Nicholson[2]

[1] Information Strategy Group, Liverpool Business School, Liverpool John Moores University, Liverpool, UK
[2] Centre for Digital Library Research, University of Strathclyde, Glasgow, UK
G.R.Macgregor@ljmu.ac.uk; e.mcculloch@strath.ac.uk;
d.m.nicholson@strath.ac.uk

Abstract. This paper considers the potential to improve distributed information retrieval via a terminologies server. The restriction upon effective resource discovery caused by the use of disparate terminologies across services and collections is outlined, before considering a DDC spine based approach involving inter-scheme mapping as a possible solution. The developing HILT model is discussed alongside other existing models and alternative approaches to solving the terminologies problem. Results from the current HILT pilot are presented to illustrate functionality and suggestions are made for further research and development.

1 Introduction: Subject Interoperability Problem

One impediment to searching distributed digital collections is the difference in metadata standards used, particularly within subject or keyword fields [1]. By adopting different subject schemes, information providers may unwittingly prevent the widespread discovery of, and therefore access to, their resources. Unless the terminology employed by online collections and services is widely used and/or known to the user, search terms may not match those embedded within resource metadata. The likelihood of this depends on a variety of factors, including the knowledge of the user and the specificity of the resource. Figure 1 illustrates the problem simplistically. A indicates the subject(s) of retrieved documents and B indicates the subject(s) of those that may remain undiscovered in response to a user query for 'Lung disease' within a traditional information retrieval (IR) system. (There are a great many more terms, and from a wider range of schemes, that may feature in either A or B; Fig. 1 shows a selection of these only.)

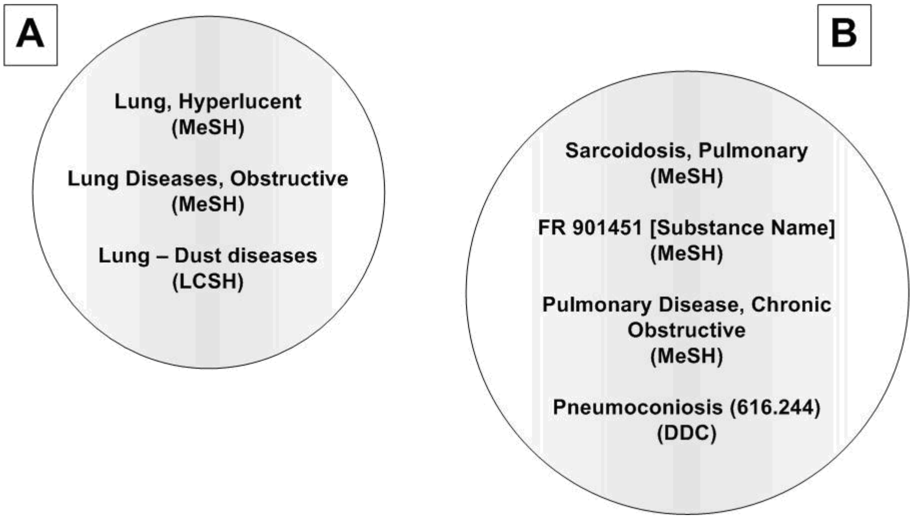

Fig. 1. Examples of documents retrieved in response to Lung disease (A), via assigned subject metadata, together with scheme information, and documents not retrieved (B).

Figure 1 shows that the user query will not retrieve documents indexed using specific terms, which may be conceptually equivalent to the user's search term 'Lung disease'. Depending on the user's perspective on any given topic therefore, vital documents may be missed. For example, amongst the potentially relevant material not retrieved are resources concerned with various aspects of lung disease including specific manifestations and treatments.

This 'translation' problem between subject schemes creates a barrier to discovery and access, and various methodologies to address this well-documented problem have been proposed over the years [2][3][4][5]. This paper will focus on the model adopted by the HILT project (http://hilt.cdlr.strath.ac.uk) and will discuss the potential of such a system to overcome, or at least minimise, the lack of interoperability afforded by collections and services' adoption of different schemes.

The paper describes and discusses a pilot terminologies service designed to facilitate resource discovery and access across distributed heterogeneous services by improving interoperability via inter-scheme mapping. Section 2 provides a general description of the HILT model. Section 3 reviews alternative models and their features, while section 4 pays particular attention to the use of SKOS Core. Section 5 presents HILT results sets and considers their ability to improve distributed information retrieval. Section 6 discusses the value of each of the functions in relation to the aim of improving resource discovery and section 7 presents conclusions and suggestions for further research.

2 The HILT Solution

The current instantiation of HILT [Fig. 2] demonstrates the model's functionality via the use of two (or more) independent SRW clients, a central SRW server and a SOAP server, described in Fig. 2 as the 'HILT pilot requests handler: SOAP server'. Non-proprietary standards including SRW [6] have been adopted enabling services to develop their own local user interfaces, capable of connecting to the HILT SRW server and employing HILT mappings within their local environment(s). Completing the model are two databases; one holding records of collections and services within the JISC (Joint Information Systems Committee) Information Environment [7] and the other holding terminologies data including mappings from satellite schemes to the central DDC spine. The response to a user query is wrapped in SKOS (Simple Knowledge Organization System) Core [8].

HILT's model involves inter-scheme mapping, whereby concepts/terms from a range of different schemes are mapped to a Dewey Decimal Classification (DDC) Scheme [9] [10] spine, which acts as a switching language [11][12]. The mapping of subject schemes is not problem free [4][1]. Schemes typically illustrate "'theoretical, conceptual, cultural and practical" [13] variations, often making the mapping process difficult, particularly if implemented via an intermediary switching language. The process has also been documented as costly and time consuming [1], as well as highly variable in its success according to subject area [3] due to differing structures, levels of specificity and, particularly, the varying proportion of single and compound terms within domain-specific schemes [14].

Despite its various drawbacks, the mapping approach does offer a practical solution to the interoperability issue, provided sound methodologies are adopted and that 'complete' mappings are implemented. Complete in this sense refers to the extent of mappings implemented between a term or concept in one scheme and any number of possibly equivalent terms or concepts in another. It is highly probable that "one-to-one relationships are certainly not sufficient" [13] for the purposes of an effective terminology server in a distributed information retrieval environment. HILT is piloting a mapping based system, investigating the value of high level mapping and more granular, complete mapping within specific subject areas. It is worth noting that the model also provides some generic terminological functionality, such as the provision of broader and narrower terms, related terms, non-preferred terms and so forth. Such terminological data can be used by services to implement retrieval tools such as interactive query expansion or hierarchical browsing of scheme data [15].

HILT currently holds XML versions of DDC 22 [9], AAT (Art and Architecture Thesaurus) [16], GCMD (Global Change Master Directory) [17], HASSET (Humanities and Social Science Electronic Thesaurus) [18], IPSV (Integrated Public Sector Vocabulary) [19], JACS (Joint Academic Coding System) [20] JITA (Classification Scheme used within E-LIS repository) [21], LCSH (Library of Congress Subject Headings) [22], MeSH (Medical Subject Headings [23], NMR (National Monuments Record Thesaurus) [24], SCAS (Standard Classification of

Academic Subjects)[25] and UNESCO Thesaurus [26]. The adoption of further schemes is currently under consideration due to the need to satisfy the requirements of two JISC services/collections within the remit of further research. An example of a scheme to be added in the near future is CAB Thesaurus [27]. By incorporating all schemes used by services and collections within the JISC Information Environment (or, indeed, any given realm) it is envisaged that individual services will be able to implement their own mappings between local collections and the centrally available HILT DDC spine. Appropriate documentation would be provided by the HILT project to facilitate this process and to ensure standardisation and consistency throughout.

Like the selection of individual schemes, the adoption of a DDC spine has been purposive. Not only is DDC a universal scheme covering most subject areas, it is also available in many languages, thus potentially facilitating multi-lingual as well as multi-KOS interoperability. Another advantage of adopting DDC as a spine is that there already exist many mappings to it from other schemes such as LCSH [22] and MSC (Mathematics Subject Classification) [28].

Preliminary research has been conducted [29] into the various types of mapping required within a system such as HILT. It is thought necessary to characterise the range of different types of equivalence imposed between terms/concepts from disparate schemes, partially to provide users with detailed relevance feedback but also as a basis for ranking results returned in response to any given search. For example, a plural version of a user's singular search term may, in some cases, be more valuable than a narrower term.

Based on an earlier study by Chaplan [30], McCulloch and Macgregor [29] determined a need for at least nine types of equivalence relationship and consider it necessary that mapped terms be encoded accordingly, in order to provide the user with information on whether or not a search term returned by the system is, for example, a synonym (i.e. concept match), a plural version or a broader or narrower term of that originally sought by the user. Dolin et al [31] have noted that "Because the relationship between two concepts can differ depending on the use case, it is possible that different cross map sets will contain the same source and target concept, but with a different relationship". This may suggest that a single mapping requires to be encoded to reflect multiple types of equivalence.

The SKOS MVS (Mapping Vocabulary Specification) has been proposed as a means of categorising the various types of relationship evident between mapped terms [32]. This has proven insufficient at its current stage of development and suggestions for extending the MVS have been submitted [29] [33].

Alternative models proposed for terminology services and as potential solutions to the interoperability problem will now be briefly presented. The HILT model will thereafter be described in further detail in relation to its functionality, with discussion of how such an intermediary system could be exploited within the distributed information environment to improve resource discovery.

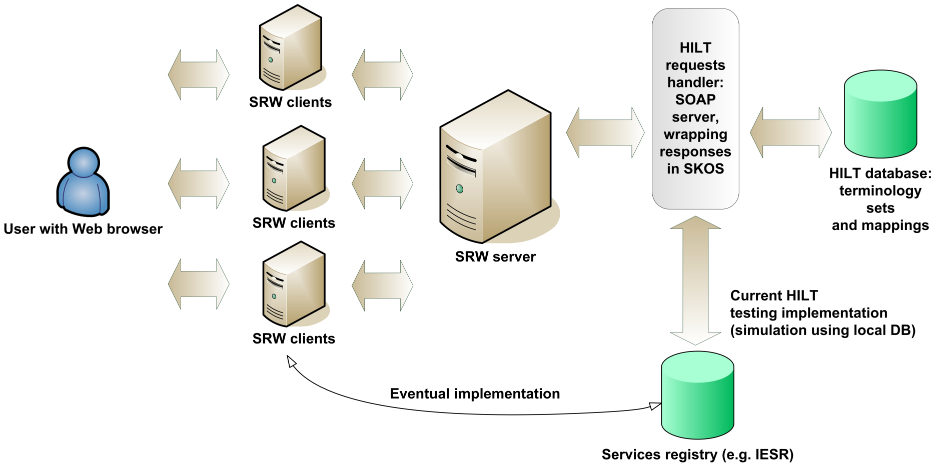
SRW clients
SRW clients
SRW clients
User with Web browser
SRW server
HILT requests handler: SOAP server, wrapping responses in SKOS
HILT database: terminology sets and mappings
Current HILT testing implementation (simulation using local DB)
Eventual implementation
Services registry (e.g. IESR)

Fig. 2. HILT pilot architecture.

3 Alternative approaches

Although there are many different approaches to solving the interoperability problem, for the purpose of this paper we will limit ourselves to reviewing those developing terminology servers. Many different examples of terminology servers and services have been proposed [34] [35], too numerous to review here. We will therefore further limit ourselves to discussing those that adopt mapping methodologies. We will consider one general model as well as looking at one within a specific subject domain - medicine, a domain in which much research and development has been conducted into the merging of, and switching between, standard terminologies in use.

The Aquarelle terminology service [36] exhibits the same basic components as HILT, namely "vocabularies in local databases, local thesaurus management systems of wider use and central Term Servers for retrieval". Although currently HILT holds terminologies centrally within the same site as the main terminology server the vision is that this element of the model will become distributed in due course, with individual collections and services able to plug their own local terminologies into the central model. The overview of the Aquarelle service shown in Figure 3 indicates a significant degree of similarity to the centralised HILT model.

The Aquarelle service was developed in the 1990s [36][37] but is no longer in operation. It is unclear whether the project was discontinued due to the viability of the model or for other reasons.

A second initiative worth noting is the GALEN programme, one component of which is the GALEN terminology server [38]. GALEN is an operational terminology service active within the clinical area. It offers the ability to provide clarification of concepts, e.g. do you know about the "leg"?; concept manage-

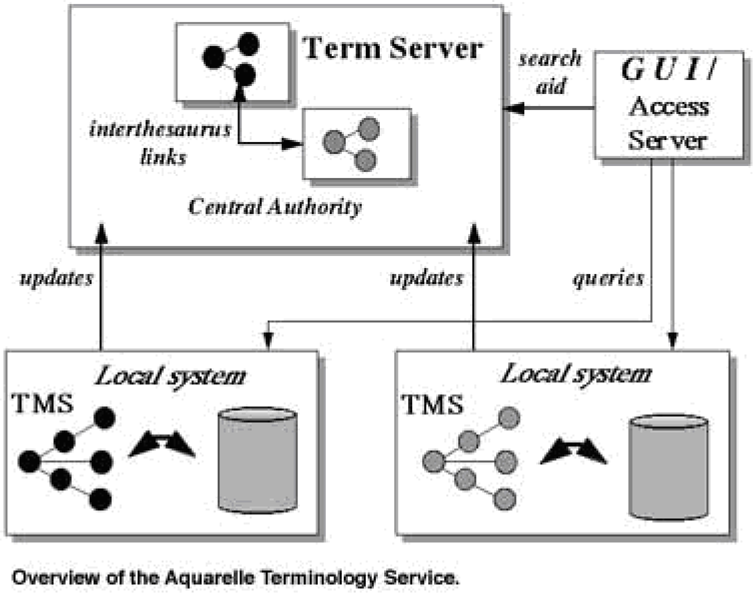

Fig. 3. Aquarelle Terminology Service architecture.

ment and specialisation, e.g. what is known about the leg? What bones does it contain? if they are broken, how might they be clinically described?; translation functionality, e.g. ”what is a French language phrase for the combination of a severe fracture of the neck of the left femur?”; identification of the preferred term for a particular concept; coding e.g. ”what is the closest ICD code for this concept?”; and extrinsic information e.g. ”is there any relevant literature known about this condition?”. Providing this range of functionality is an architectural model that fits ”very comfortably with the notion of client-server computing, and commercial implementations now use standard object component technologies to deliver their services” [38].

Contrary to the primary function of the HILT model (i.e. to switch between several different terminologies via inter-scheme mappings), the GALEN model is optimised for the answering of clinical questions and appears to provide a broad databank relating to various aspects of conditions and treatments and so on, as opposed to acting as an intermediary between the user and services or collections. In this respect it appears more closely related to the notion of an expert system. Although architecturally similar, the functionality of GALEN is very different to that of HILT. GALEN does map natural language to concepts and concept to classification schemes, but the purpose of doing so is more extensive than the provision of a switching mechanism.

It has been documented [39] that the key desiderata for a clinical terminology server are 1) word normalization, 2) word completion, 3) target terminology specification, 4) spelling correction, 5) lexical matching, 6) term completion, 7) semantic locality, 8) term composition and 9) decomposition [39]. These functions echo those identified as desirable by HILT. However, the purpose of such a clinical server is mainly to enable "clinicians to enter patient observations, findings, and events, such as procedures. It does not need to carry the weight of terminology updates, maintenance, or development and thus might be regarded as a server "lite"." Quite distinct from HILT's aim to improve mediated resource discovery and retrieval for the end user, it seems that the primary users of this type of model are professionals, who are likely to have a substantial degree of knowledge about the terminology and conditions being queried.

It seems therefore that although much of the functionality desired by HILT is also desirable in other domain specific terminology servers. HILT represents a novel implementation in that it aims to cover all areas of knowledge, by incorporating and mapping together schemes from all disciplines and (eventually) languages. It follows that HILT has a wider remit than other servers currently implemented. Although the Aquarelle service is similar to HILT in terms of architecture and functionality, its stage of development remains unclear.

Although dissimilar in architectural terms, Renardus [40] is similar to HILT in that it employs DDC as its central terminology. This service enables users to search by title, subject, description, creator, document type or DDC classification. In contrast to HILT, Renardus retrieves item level resources in response to the entry of a DDC number, without first clarifying what the user is intending to search for. This aspect of the model is not conducive to user interrogation since the average user is unfamiliar with DDC notation and is likely to experience difficulties in expressing an information need in this way. HILT, on the other hand, provides the user with DDC captions relating to a specific numerical notation, providing relevance feedback throughout the search process. The user is able to ensure he/she is within the correct discipline, determining the relevant focus of a given subject, since different aspects of the same basic concept may be located in various disciplines of a classification system. When browsing the DDC hierarchy for a subject in Renardus - thus accessing the more meaningful captions of the scheme - the service intends to link the user into gateways holding records on the subject of interest. At the time of writing it was noted that few gateway services have retained collaboration with Renardus, resulting in 'dead ends' for many of the browse trees.

Should the HILT architecture and general model prove effective, it may be that elements of the HILT model could be tackled in different ways. For example, is a DDC spine the best option in this context? The very nature of DDC (and indeed library classifications) has been questioned and undoubtedly causes problems relating to the mapping of schemes [41]; most obviously because the majority of schemes contain terms and/or concepts whereas the unique identifier conveying a concept in DDC is a numerical notation. Further difficulties stem from the analytico-synthetic properties of DDC, requiring a subject to be analysed before undertaking the synthesis of an appropriate notation by which it can be expressed. This means that all notations to which terms from a satellite scheme may require to be mapped will not necessarily be pre-coordinated; that is, the mapping process may also require an extensive process of number building to express concepts accurately. In conducting such number building it is common to add standard subdivisions to a basic concept, where rules tend to vary according to circumstance. For example, where a three digit notation ends in 0 e.g. 370, the 0 added to indicate the addition of a standard subdivision is omitted; in other circumstances there may be an instruction to add an extra 0. These types of practice are likely to have implications for the truncation process adopted by HILT, described in section 5.1. Standard subdivisions can only be added once, which means that subjects referring to multiple locations or dates cannot be expressed adequately. So, for instance, France and Belgium cannot be incorporated into a single notation to express, for example, French language usage in these two countries. One final difficulty worth mentioning is that not all areas of DDC reflect the superordinate or subordinate nature typical of hierarchical schemes. An example of this can be seen in the 900 section, where 900 denotes History, geography, and auxiliary disciplines [42]. One level down the hierarchy lies 970 denotes History of North America, while 973 relates to United States. Although, therefore, United States is subordinate to History of North America, this is not reflected in the DDC notation, with each number being of equal length.

Such limitations seem to warrant the investigation of alternative schemes, bearing in mind that an effective spine must be universal in nature since it should encompass all concepts expressed within all other schemes being mapped [12]. Although much work in the area features a central DDC spine [40] or mappings of individual schemes to DDC [28], several other projects have employed a central terminology other than DDC. UDC (Universal Decimal Classification) [43] has been adopted in this context due to its ability to offer "international notation, depth documentation, retrieval and mechanization facilities" [44] [45]. Other initiatives have implemented direct mappings between two disparate schemes [30] [33] devoid of the switching model favoured by HILT. Although clearly valid and likely to improve retrieval within a given subject discipline, it is unlikely that such an approach would prove universally effective or scalable.

4 SKOS: Modelling Terminological Data

SKOS Core [8] is a useful development within the context of M2M terminology service architectures. SKOS Core is an application of the Resource Description Framework (RDF) proposed by the W3C Semantic Web Deployment Working Group [46] and provides a flexible framework for representing the structure and content of KOS (or 'concept scheme') on the Web. SKOS Core essentially comprises a series of RDF properties and RDF Schema (RDFS) classes to encode the content and structural characteristics of KOS. As an application of RDF, SKOS data remains inherently adaptable and can be integrated with other RDF data on the Web using Semantic Web applications. A draft mapping specification has also been proposed by Miles et al [32] enabling the mapping of concepts between different KOS within the SKOS framework.

Although the primary objective of SKOS Core is to provide a means of publishing KOS for the Semantic Web, use of the specification for dynamic client-server interactions has attracted attention from those active in terminology service research and development [47] [48] [49] [50]. SKOS Core can prove particularly advantageous in such contexts since terminological data can be richly modelled and data structures can be maintained when communicating with clients, particularly when using web service protocols such as SOAP [15]. This can facilitate reliable, flexible and simple multipurpose reuse by client services.

Alternative frameworks are available to facilitate the aforementioned functionality. These can occasionally be inappropriate or less flexible, thus increasing the potential for low adoption among client services. Despite increased complexity, OWL [51] has been demonstrated as effective within similar technical architectures [52]. It also continues to be used successfully to represent some terminological data [53]; however, it remains unsuitable for other schemes [54]. For example, the OWL class-instance does not reflect the structure of all KOS, resulting in the need for unnecessary KOS reengineering [55].

Zthes [56] provides an abstract model and an XML schema for relational vocabulary representation (particularly thesauri) and is suitable for 'storing and transmitting' such terminological data. Use of Zthes can be advantageous as the specification also defines how queries to Zthes-compliant terminologies can be implemented using Z39.50 and/or SRU/W. Further experimentation with this approach has been undertaken by Vizine-Goetz et al [57]. However, Zthes remains less suited to handling disparate terminological data [58]. The flexibility of SKOS and its increased suitability with Web services and the Semantic Web community make it more conducive to the system we demonstrate here [59].

5 Functionality

Within the third phase of the project five distinct functions were implemented to simulate ways in which users may interact with HILT, based on a set of use cases [60]. It was deemed desirable to build a system which could, for example, 1) provide terminological data on any given term within a scheme held; 2) return all instances of a given search term within DDC, together with the appropriate hierarchical data and DDC notation; 3) return all terms across schemes related (predetermined via mapping) to the DDC notation matched to a given search term; 4) return combinations of 1), 2) and 3) as specified by the user.

Each of the functions developed (get_collections, get_all_records, get_ddc_records, get_non_ddc_records, get_filtered_set) will be discussed in turn, to help contextualise their purposes, with a view to aiding discovery and access across distributed digital collections. The purpose and mechanism of each function will be documented, before illustrating its value, or otherwise, by presenting HILT output in response to an example query. This will better explain the strengths and weaknesses of the system in its current instantiation. Examples will be given for queries sent to the HILT pilot requests handler: SOAP server via the test HILT SRW client [61].

5.1 get_collections

The get_collections function aims to provide the user with collection information relevant to the area of a subject query. It will return information and/or a link to and/or dynamic searching of any collection(s) classified under a specified DDC number or its stem. The process is carried out as follows:

1. A DDC number relating to a caption/hierarchy identified during the disambiguation stage (user enters term prior to this stage; this is then matched to appropriate notation(s)) is sent from the SRW client service to the SRW server.
2. The SRW server sends an appropriate request for get_collections via the SOAP server.
3. The get_collections function queries the database using successive truncations of the DDC number sent.
4. The SOAP requests handler receives back collections' connection details and scheme information.
5. The SOAP requests handler wraps the results in Dublin Core Collection Description Application Profile (DC CD AP) and sends the results back to the SRW server.
6. The SRW server sends the results back to the client service.
7. The client service processes the results to offer the user a set of collections relevant to their query.

On entering the query '371.07' (Education - Schools and their activities; special education - Religious schools) to the pilot demonstrator search box [61] (which simulates the processes of stages 1 and 2 above) the following result is returned. The result is expressed in DC CD AP within a SOAP envelope (envelopes have been edited out in all examples given). On development of a more advanced end-user oriented system, the result will be parsed by a client and presented to the end-user in a human readable format, dependent on how a given local service decides to present the information being returned.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-SOAP envelope -->
<metadata
xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xmlns:dc="http://purl.org/dc/elements/1.1/"
xmlns:dcterms="http://purl.org/dc/terms/"
xmlns:dcmitype="http://purl.org/dc/dcmitype/"
xmlns:iesr="http://iesr.ac.uk/terms/#usesControlledList"
xmlns:cld="http://purl.org/cld/terms/">
```

```
<dcmitype:Collection>
<dc:title>BUBL LINK: Education</dc:title>
<dc:identifier xsi:type="dcterms:URI">http://bubl.ac.uk/link/
</dc:identifier>
<dcterms:abstract>Catalogue of selected Internet resources.
</dcterms:abstract>
<dc:creator>BUBL Information Service</dc:creator>
<dc:type xsi:type="dcterms:DCMIType">Collection</dc:type>
<dc:subject xsi:type="dcterms:DDC">370</dc:subject>
<cld:isAccessedVia>http://hilt.cdlr.strath.ac.uk/bublsearch/
bubl.cfm?queryString=</cld:isAccessedVia>
</dcmitype:Collection>
<dcmitype:Collection>
<dc:title>Education-line</dc:title>
<dc:identifier xsi:type="dcterms:URI">http://www.leeds.ac.uk/
educol/</dc:identifier>
<dcterms:abstract>Project funded under the Electronic Libraries
programme to gather an electronic archive of preprints, grey
literature and texts in education and training. </dcterms:abstract>
<dc:creator>Leeds University</dc:creator>
<dc:type xsi:type="dcterms:DCMIType">Collection</dc:type>
<dc:subject xsi:type="dcterms:DDC">370</dc:subject>
</dcmitype:Collection>
</metadata>
```

Fig. 4. Result for get_collections function using query '371.07' (Education - Schools and their activities; special education - Religious schools).

Figure 4 shows two collections being returned in response to the query '371.07': BUBL LINK: Education and Education-line. The value of this function is illustrated by its flexibility. For example, Figure 4 above shows that both collections returned have been classified in the system's collections database at DDC 370. This is due to the ability of the system to truncate a DDC number successively in the event of no direct matches in response to a query. Since no match was found for 371.07, the system has searched upwards through the DDC hierarchy until a match was found at 370. This means that however specific the DDC number sent via point 1 above is, collections should always be returned, even if broadly classified at one of the ten main classes (i.e. 000 - 900). Once collections have been identified at any given point via the process of truncation, no further truncation will be invoked. This means that a query for 371.07 will return the two collections above classified at 370, but will not present more general collections relating to education, classified at 300.

For research purposes, experimentation for get_collections has been with a local collections database containing test data; however, the model has been designed to interact with distributed service registries as a source of accurate collection and service descriptions. To this end research testing HILT interaction with the Information Environment Services Registry (IESR) [62] is currently being pursued.

5.2 get_all_records

The get_all_records function retrieves records that include - or are mapped to records that include - the term or term phrase specified within a given query. This function operates as follows:

1. User enters term via the embedded SRW client service, and a resultant request is sent to the SRW server.
2. The SRW server parses the request to obtain search terms and uses these to call the SOAP get_all_records function.
3. The get_all_records function queries the database to find (1) all DDC records that either include the user term or that are mapped to from other non-DDC records that include the term (2) all non-DDC records mapped from the DDC records retrieved under (1) and returns these records to the SOAP server.
4. The SOAP requests handler wraps the results in SKOS Core with the SKOS Mapping Vocabulary Specification (MVS) and sends the results to the SRW server.
5. The SRW server sends the results back to the client service.
6. The client service processes the results to offer DDC and non-DDC records to the user.

The result of a query entered selecting the get_all_records function should contain DDC numbers, mapped terms and details of what scheme such terms belong to, and mapping match type information denoting the nature of the equivalence relationship imposed. The following code (Figure 5), embedded within a SOAP envelope, illustrates the result returned in response to a query for 'Natural hazards':

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-SOAP envelope -->
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core.rdf#"
xmlns:map="http://www.w3.org/2004/02/skos/mapping#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:Concept rdf:about="#363.34">
<skos:prefLabel xml:lang="zxx">363.34</skos:prefLabel>
<skos:altLabel xml:lang="en">Disasters</skos:altLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/DDC.rdf"/>
<map:exactMatch>
<skos:Concept rdf:about="#16117"/>
</map:exactMatch>
<map:exactMatch>
```

```
<skos:Concept rdf:about="#16118"/>
</map:exactMatch>
<map:narrowMatch>
<skos:Concept rdf:about="#16119"/>
</map:narrowMatch>
<map:narrowMatch>
<skos:Concept rdf:about="#2256"/>
</map:narrowMatch>
<map:narrowMatch>
<skos:Concept rdf:about="#762"/>
</map:narrowMatch>
<map:exactMatch>
<skos:Concept rdf:about="#2696"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#143"/>
</map:exactMatch>
</skos:Concept>
<skos:Concept rdf:about="#16117">
<skos:prefLabel xml:lang="en">Disasters</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#16118">
<skos:prefLabel xml:lang="en">Emergency management</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#16119">
<skos:prefLabel xml:lang="en">Natural disasters</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#2256">
<skos:prefLabel xml:lang="en">Natural disasters</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/UNESCO.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#762">
<skos:prefLabel xml:lang="en">Natural Hazards</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/GCMD.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#2696">
<skos:prefLabel xml:lang="en">HAZARDS, ACCIDENTS AND DISASTERS
```

```
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#143">
<skos:prefLabel xml:lang="en">Civil emergencies</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/IPSV.rdf"/>
</skos:Concept>
</rdf:RDF>
```

Fig. 5. Result for get_all_records function in SKOS RDF/XML, using the query 'Natural hazards'.

Figure 5 shows that following the initial query for 'Natural hazards', DDC 363.34 (Disasters) was selected as an appropriate match. In addition to the DDC record returned, a number of mappings to DDC 363.34 from other satellite schemes were returned as shown in Table 1.

Term	Source Scheme	Type of Equivalence
Disasters	DDC	Exact match*
Disasters	LCSH	Exact match
Emergency management	LCSH	Exact match
Natural disasters	LCSH	Narrow match
Natural disasters	UNESCO	Narrow match
Natural hazards	GCMD	Narrow match
Hazards, accidents and disasters	HASSET	Exact match
Civil emergencies	IPSV	Exact match

Table 1. Summary of results for 'Natural hazards', selecting get_all_records function *note that exact match in this sense (in line with SKOS MVS) encompasses a concept match.

The encoded result and Table 1 indicate the range of related terms available within the loaded terminologies. These enjoy some form of equivalence relationship with the original query. By offering synonymous and narrower terms to the user query, HILT is providing the opportunity to explore matched concepts in other schemes and by extension interrogate alternative repositories using the correct query to match local indexes. It also allows users to conduct a more specific search by opting to use those terms returned as having a narrower foci than the original query.

5.3 get_ddc_records

The get_ddc_records function retrieves any DDC record that includes the term(s) specified, or that is mapped to by a record from another scheme that includes the term(s) specified. This function is handled as follows:

1. User enters term via embedded SRW client service, and a resultant request is sent to the SRW server.
2. The SRW server parses the request to obtain search terms and uses these in a call to the SOAP get_ddc_records function.
3. The get_ddc_records function queries the database for DDC records that include the user term entered or that are mapped to by non DDC records that include the term.
4. The SOAP requests handler receives DDC numbers and associated DDC captions, wraps the results in SKOS Core, and sends them back to the SRW server.
5. The SRW server sends the results back to the client service.
6. The client service processes the results to offer the user terms possibly relevant to their query from DDC with corresponding DDC numbers.

Figure 6 illustrates functionality in response to a search for a DDC caption, 'Shore protection'.

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-SOAP envelope -->
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core.rdf#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:ConceptScheme rdf:about="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/DDC.rdf"/>
<skos:Concept rdf:about="#627.58">
<skos:prefLabel xml:lang="zxx">627.58</skos:prefLabel>
<skos:altLabel xml:lang="en">Shore protection</skos:altLabel>
</skos:Concept>
<skos:Concept rdf:about="#333.91716">
<skos:prefLabel xml:lang="zxx">333.91716</skos:prefLabel>
<skos:altLabel xml:lang="en">Shore protection, . . .
</skos:altLabel>
</skos:Concept>
</rdf:RDF>
```

Fig. 6. Result for get_ddc_records in SKOS RDF/XML for the query, 'Shore protection'.

The result shows two distinct incidences of the caption 'Shore protection' within the DDC schedules; one instance resides in the 600 section (Technology) with the other dealing with social aspects of 'Shore protection' in the 300 section (Social sciences). No results are returned from any scheme other than DDC in response to this function. Part of the added value offered as a result of the mapping based methodology adopted by HILT in relation to the get_ddc_records function is that DDC records will be returned following matches to terms in other schemes, which are mapped to DDC. An example whereby 'Plant genetics', a known term from the HASSET scheme, was searched for using the get_ddc_records follows (Figure 7):

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-SOAP envelope -->
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core.rdf#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:ConceptScheme rdf:about="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/DDC.rdf"/>
<skos:Concept rdf:about="#631.5233">
<skos:prefLabel xml:lang="zxx">631.5233</skos:prefLabel>
<skos:altLabel xml:lang="en">Agricultural genetics</skos:altLabel>
</skos:Concept>
<skos:Concept rdf:about="#581.35">
<skos:prefLabel xml:lang="zxx">581.35</skos:prefLabel>
<skos:altLabel xml:lang="en">Genetics</skos:altLabel>
</skos:Concept>
<skos:Concept rdf:about="#631.53">
<skos:prefLabel xml:lang="zxx">631.53</skos:prefLabel>
<skos:altLabel xml:lang="en">Plant propagation</skos:altLabel>
</skos:Concept>
</rdf:RDF>
```

Fig. 7. Result for get_ddc_records in SKOS RDF/XML for the query, 'Plant genetics'.

Figure 7 shows the DDC notation, and corresponding captions, to which the HASSET term 'Plant genetics' is mapped. Three mappings have been implemented; one to DDC 631.5233 'Agricultural genetics'; one to DDC 581.35 'Genetics' and a third to DDC 631.53 'Plant propagation'. Clearly the value of such results is user dependent, and reliant on the completeness of mappings implemented.

5.4 get_non_ddc_records

The get_non_ddc_records function retrieves any non-DDC record that includes a mapping to the DDC number sent. That is, the system retrieves records from other schemes (non-DDC) that have been mapped to an input DDC number. Only the non-DDC records mapped to the DDC number sent are retrieved, as follows:

1. User chooses DDC number on screen and embedded SRW client service sends an appropriate request to the SRW server.
2. The SRW server parses the request and sends an appropriate query to the SOAP get_non_ddc_records function.

3. The get_non_ddc_records function searches the database to find non-DDC records containing a mapping to the DDC number sent and returns the results to the SOAP server.
4. The SOAP server wraps the results in SKOS Core and SKOS MVS and returns them to the SRW server.
5. The SRW server sends the results back to the client service; results comprise DDC number entered, terms from other schemes mapped to that DDC number, with the name of the scheme and match type information defining the relationship between a scheme's term and the DDC number entered.
6. The client service processes the results and provides the user (via the service interface) with information on which term to use for individual schemes used by individual JISC collections.

The DDC notation 631.53 will form the search query to illustrate the get_non_ddc_records function. We saw from the get_ddc_records result above that this notation relates to 'Plant propagation'. The result for this query is presented below (Figure 8):

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-SOAP envelope -->
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core.rdf#"
xmlns:map="http://www.w3.org/2004/02/skos/mapping#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:Concept rdf:about="#631.53">
<skos:prefLabel xml:lang="zxx">631.53</skos:prefLabel>
<skos:altLabel xml:lang="en">Plant propagation</skos:altLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/DDC.rdf"/>
<map:exactMatch>
<skos:Concept rdf:about="#36011"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#36012"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#36013"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#36014"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#36015"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#36016"/>
```

```
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#2539"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#17"/>
</map:exactMatch>
<map:exactMatch>
<skos:Concept rdf:about="#4712"/>
</map:exactMatch>
</skos:Concept>
<skos:Concept rdf:about="#36011">
<skos:prefLabel xml:lang="en">Plant breeding</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#36012">
<skos:prefLabel xml:lang="en">Plant cell culture</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#36013">
<skos:prefLabel xml:lang="en">Plant micropropagation
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#36014">
<skos:prefLabel xml:lang="en">Plant mutation breeding
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#36015">
<skos:prefLabel xml:lang="en">Plant propagation</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#36016">
<skos:prefLabel xml:lang="en">Vegetative propagation
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/LCSH.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#2539">
```

```
<skos:prefLabel xml:lang="en">Plant genetics</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/UNESCO.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#17">
<skos:prefLabel xml:lang="en">Plant Breeding and Genetics
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/GCMD.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#4712">
<skos:prefLabel xml:lang="en">PLANT GENETICS</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
</rdf:RDF>
```

Fig. 8. Result for a get_non_ddc_records query for DDC 631.53 (Plant propagation), in SKOS RDF/XML.

The system has retrieved nine results, summarised in Table 2:

Term	Source Scheme	Type of Equivalence
Plant breeding	LCSH	Exact match
Plant cell culture	LCSH	Exact match
Plant micropropagation	LCSH	Exact match
Plant mutation breeding	LCSH	Exact match
Plant propagation	LCSH	Exact match
Vegetative propagation	LCSH	Exact match
Plant genetics	UNESCO	Exact match
Plant breeding and genetics	GCMD	Exact match
Plant genetics	HASSET	Exact match

Table 2. Summary of results for get_non_ddc_records.

Table 2 indicates that terms have been retrieved from a total of four distinct schemes, relating to the search for DDC 631.53. This notation and corresponding caption is shown at the beginning of the result set, before listing all terms mapped to this notation from other schemes. As mentioned before, work continues into establishing mapping types and appropriate coding of such equivalence relationships. The indication that all terms are 'exact matches' to the original query is therefore misleading. Where explicit relationships have not yet been established within the HILT research programme, the default is to express any relationship as an exact match; this will be rectified as the project progresses.

5.5 get_filtered_set

get_filtered_set is a more generic terminological function, not employing the use of mappings. get_filtered_set retrieves records that meet the specified parameters;

that is, the search term entered but 'filtered' by scheme name(s) and /or field name(s). Functionality to filter a search by scheme, and/or to search preferred and non-preferred terms will be in-built. This enables a user to search one scheme directly, or to incorporate multiple schemes in the scope of his/her search. The get_filtered_set function operates as described below:

1. User enters term via embedded SRW client service, and a resultant request is sent to the SRW server.
2. The SRW server parses the request and uses the results to send an appropriate query to the SOAP get_filtered_set function.
3. The get_filtered_set function queries the database for records that match the terms and the specified filters and the results are sent back to the SOAP server.
4. The SOAP server wraps the results in SKOS Core and returns them to the SRW server.
5. The SRW server sends the results back to the client service; results comprise terms together with information about each term's source scheme, notation (DDC) or ID (other schemes), and broader, narrower and related terms, where applicable.
6. The client service processes the results to provide the service interface with terms from specific schemes relevant to the query and with any relevant additional data on the terms (e.g. related terms).

To illustrate the functionality of the get_filtered_set function, 'Plant genetics' will be searched for, selecting HASSET as the preferred scheme to be searched. Results are detailed in Figure 9:

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<!-SOAP envelope -->
<rdf:RDF
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:skos="http://www.w3.org/2004/02/skos/core.rdf#"
xml:base="http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/concepts.php">
<skos:Concept rdf:about="#2465">
<skos:prefLabel xml:lang="en">GENETICALLY MODIFIED CROPS
</skos:prefLabel>
<skos:broader rdf:resource="#1389"/>
<skos:broader rdf:resource="#2463"/>
<skos:related rdf:resource="#110"/>
<skos:related rdf:resource="#2466"/>
<skos:related rdf:resource="#4712"/>
<skos:altLabel xml:lang="en">GM CROPS</skos:altLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#4712">
```

```
<skos:prefLabel xml:lang="en">PLANT GENETICS</skos:prefLabel>
<skos:broader rdf:resource="#624"/>
<skos:broader rdf:resource="#2467"/>
<skos:related rdf:resource="#2465"/>
<skos:altLabel xml:lang="en">PLANT BREEDING</skos:altLabel>
<skos:altLabel xml:lang="en">PLANT REPRODUCTION</skos:altLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#1389">
<skos:prefLabel xml:lang="en">CROPS</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#2463">
<skos:prefLabel xml:lang="en">GENETIC ENGINEERING
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#110">
<skos:prefLabel xml:lang="en">AGRICULTURAL PRODUCTION
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#2466">
<skos:prefLabel xml:lang="en">GENETICALLY MODIFIED FOOD
</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#4712">
<skos:prefLabel xml:lang="en">PLANT GENETICS</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#624">
<skos:prefLabel xml:lang="en">BOTANY</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
<skos:Concept rdf:about="#2467">
<skos:prefLabel xml:lang="en">GENETICS</skos:prefLabel>
<skos:inScheme rdf:resource="http://hiltm2m.cdlr.strath.ac.uk/
```

```
hiltm2m/schemes/HASSET.rdf"/>
</skos:Concept>
</rdf:RDF>
```

Fig. 9. Results for a get_filtered_set query (set to HASSET) for the term 'Plant genetics'. Result in SKOS RDF/XML.

Figure 9 shows how HILT can provide extremely specialised terminological data, in this case from a single scheme selected using the get_filtered_set function. Any individual scheme or any combination of schemes within the system can be accessed in this way. Further, a user / client service can specify whether they wish to retrieve preferred, non-preferred or related terms within a search for terms in any scheme(s). Such terminological data can be used in a variety of ways; however, it is expected that get_filtered_set will be used most by those services wishing to extend the retrieval tools available to users. For example, using get_filtered_set to implement forms of interactive query expansion or hierarchical scheme browsing to improve local repository interrogation or to aid query formulation. GoGeo! has implemented a keyword search demonstrator employing HILT get_filtered_set functionality [63]. This provides a real-life example of how HILT could be integrated within an existing service in order to mediate searching of associated collections and as a means of providing query expansion opportunities for users.

The SKOS result in Figure 9 indicates that searching for 'Plant genetics' within HASSET retrieves terms including 'Genetically modified crops', 'GM crops', 'Plant genetics', 'Plant breeding', 'Plant reproduction', 'Crops', 'Genetic engineering', 'Agricultural production', 'Genetically modified food', 'Botany', 'Genetics'. Not all of these are likely to be directly relevant to a user requesting information on 'Plant genetics'. The current search parameters within HILT first search for an exact phrase match i.e. Boolean AND; thereafter conducting further searches in line with the Boolean OR principle. It follows that single terms within the search query 'Plant' and 'Genetics' are retrieved individually, which may or may not prove relevant in every instance.

5.6 Function summary

HILT currently enables users / client services to retrieve DDC only terms, non-DDC only terms, a combination of both DDC and non-DDC terms, or to specify an individual scheme or a selection of schemes, which they wish to search. In the latter case, functionality also extends to the switching on or off of preferred, non-preferred or related terms, enabling yet greater search specificity.

The perceived effectiveness or otherwise of four out of the five functions (excluding get_collections) is heavily reliant on inter-scheme mapping. Results for get_ddc_records, get_non_ddc_records, get_all_records and get_filtered_set, where more than one scheme is selected, is dependent upon an effective mapping infrastructure. It is therefore necessary to ensure valid and robust mappings are implemented. Such mappings should also be complete. That is, one-to-one mappings are likely to be insufficient for the types of scenarios presented above, even between one individual scheme and another.

6 Discussion

The preceding examples relating to each of HILT's five functions currently implemented indicate that the system does indeed have the potential to improve distributed information retrieval where different services/collections employ disparate terminologies.

The classification of services/collections by DDC enables the get_collections function to retrieve details of services holding resources covering the user's chosen subject area. One limitation of this function in its current instantiation is that within the local collections database searched by HILT, each service/collection is only assigned one DDC number. For general and multidisciplinary collections it would be pertinent to extend this to as many DDC numbers as required to convey subject coverage adequately. This would facilitate the retrieval of collections, with only a subset of items relevant to a user's needs. It is thought that the assignation of multiple class numbers in this way would greatly enhance the get_collections function by opening up more potentially relevant information sources to the user. It should be noted that IESR [62] already offers multiple DDC numbers for any given collection.

A further limitation relates to the process of truncation implemented. The example in 5.1 above shows that a search for 371.07 will retrieve collections classified at 370 but nothing beyond that. It is proposed to extend the process of truncation beyond the decimal point so that a general collection will be returned if nothing more specifically relevant is returned. The retrieval of a general social science collection classified at 300, for example, is considered to have greater value to the user than a scenario where they retrieve no hits. By extending truncation beyond the decimal point users will retrieve collections classified at one of the ten main DDC classes.

In some of the current examples, scheme information is missing from the DC CD AP result returned. It should be noted that this is due to incomplete information within the collections database. This issue should be ameliorated with the incorporation of relevant collection and service registries to the HILT model, as noted in 5.1. In line with the architecture of the JISC Information Environment, it is intended that the collections database ultimately be maintained externally and independently by the IESR [62].

The additional four functions described in section 5 illustrate how users can retrieve exact matches for terms across schemes, synonyms or concept matches, along with broader or narrower terms. Such functionality will aid improved retrieval performance for users by lowering the cognitive load experienced by the user during query formulation [64]. Where in general search engines a user may retrieve no directly relevant hits, or relevant hits may be buried a considerable way down a long results list, HILT provides alternative search terms with a view to expanding users' queries, and where no exact or concept matches exist, related terms in the form of more general, more specific and so on will be presented.

The dynamic element of the system, whereby selected terms trigger a search within a relevant collection, further improves the level of information retrieval for the user. This process miminises the number of clicks and limits the need for the user to re-enter search terms into a number of different services' search boxes.

The success of these types of functions is heavily reliant on the appropriateness of mappings implemented, as well as the accuracy of repository resource indexing (particularly in distributed subject resource discovery contexts). Users will only benefit from the retrieval of synonyms and the like if they have been correctly identified and encoded as such within the HILT model. The cost and time consuming nature of implementing mappings has already been discussed. Due to such constraints, HILT proposes to first consider a fairly broad set of mappings, likely to be imposed between satellite schemes and DDC's top 1000 captions, or most frequently used numbers, before piloting an area of more in-depth mapping within a more detailed subject area. This work is likely to inform how to proceed with fuller-scale mapping exercises. It is hoped that patterns will emerge to enable some degree of automation to be implemented, although manual verification of the appropriateness or otherwise of relationships will still be required. It will also be necessary to review existing mappings within the current instantiation. OCLC provided an XML version of DDC 22 with mappings to LCSH, many of which appear inappropriate for the purpose of HILT. Function testing has revealed that many of the DDC-LCSH mappings are not considered of potential benefit to users retrieving information from distributed sources. This may be a result of such mappings having been derived statistically.

Progression towards a more precise system depends on refinement of search parameters. Results sets presented in section 5, particularly that for the get_filtered_set function, indicate that fairly imprecise results are currently being retrieved due to the broad nature of the current search parameter. It is thought likely that this will require refinement, perhaps to only search using Boolean AND in the first instance. The OR operator could potentially be invoked if requested by the user. This will maintain transparency enabling the user to keep track of the results provided. Otherwise, some of the terms returned may not appear directly relevant to the user's search, giving the impression of an ineffective system.

It is considered of interest to investigate the suitability of other universal schemes with a view to replacing DDC as a spine, although the full extent of the advantages of using DDC have not yet been fully explored. HILT will continue to work with DDC, whilst considering how alternatives may improve or degrade the level of success for the user in relation to the functions implemented.

The range of schemes incorporated into the current HILT model should clearly be reviewed and extended as necessary. The selection of schemes was originally purposive since the project largely depended on those schemes it could obtain free of charge for research purposes and in a suitable format for uploading into a terminologies database with minimal intervention. Depending on the nature of HILT's growth, and the community it requires to serve, the inclusion of schemes will be heavily modified. It is also of interest to incorporate folksonomies into the HILT model. The inclusion of folksonomies, or folksonomy-type terms is likely to create a range of additional access points for users unfamiliar with formal terminology used to express certain concepts. Less formal terms in everyday usage could be mapped to the DDC spine in the same way as standard schemes and it is possible that tag clouds characteristic of Web 2.0 folksonomy driven services could have a role to play in the expression of synonymous concepts, as well as broader and narrower equivalence relationships. HILT has done some preliminary work in incorporating user terms taken from search logs, to ascertain whether or not this improves the hit rate for users following the translation process afforded by mapping such terms to DDC, which can then, in turn, be translated to any other scheme providing relevant subject coverage. Folksonomies or folksonomy-type terms are likely to be incorporated as research proceeds, in addition to mappings being established from the standards schemes included.

The validity of an ontological approach to developing a terminology server is also of interest. Sanchez-Alonso and Garcia-Barriocanal [65] investigated the feasibility of mapping SKOS Core metadata to an upper ontology. Various difficulties were encountered as a result of the lack of formalisation in the current instantiation of SKOS and the need for mapping criteria to promote semantic interoperability. The authors endeavour to find a way "to map a concept in a SKOS scheme to a term in an upper ontology that provides a formal definition". Their investigation found that an intermediate model was required to do so. At present, there is no immediate remit to pursue this type of approach within HILT, although the progress of others working in the area will be followed with interest.

For the purpose of creating further and more advanced functionality within the system, it will first be necessary to survey the JISC community to determine the types of features they would find useful in a system such as HILT. Such a survey is planned for the current phase of the project and is likely to inform the design of additional functions. User evaluation is also necessary to assess the appropriateness and usefulness of such functions. The functions already described in the current paper will also be assessed by users in the near future.

7 Conclusion and further research

Some areas for future research were discussed in the previous section. In addition to these, further research into match types should be conducted to establish how best to express the nature of equivalence relationships between terms. Currently, five mapping types are in use, in line with the SKOS MVS. These are exact match, narrow match, broad match, major match and minor match. It is thought likely that further match types may prove useful although this theory must be considered in the context of user testing.

It is considered likely that a range of additional use cases, and therefore functions, will prove valuable within the HILT service. A survey of potential users of HILT (both services/collections and individuals) should be undertaken to inform the HILT team on what these use cases might be. Appropriate functionality can then be designed and built in to the system.

To assess the more robust measures of retrieval, precision and recall, precision being the proportion of relevant documents retrieved within the retrieved set and recall being the proportion of relevant documents retrieved from the total number of relevant documents available, rigorous testing is required within a controlled environment. It is necessary to build a document collection and run robust tests in order to assess such measures of success.

In conclusion, effective resource discovery can only be realised if the means of access becomes more transparent. If users are unable to locate relevant resources on the web due to lack of awareness and openness, the success of digital publishing is compromised. Users require to be made aware of the existence of resources relevant to their needs and require metadata to be sufficiently penetrable to conduct effective and efficient information retrieval. In an environment where subject metadata varies from collection to collection or service to service, in an increasingly fragmented digital world, such efficiency cannot be realised. Terminologies need to be brought together to improve interoperability between services, thus making disparate collections cross-searchable. It is the authors' belief that a system like HILT can go some way to improving the openness of resources and therefore widening access to material held in heterogeneous collections across the web, which would otherwise be hidden, and that HILT's architecture and mapping based infrastructure will, in time, prove an efficient means of reaching this goal.

References

1. McCulloch, E., Shiri, A., Nicholson, D.: Challenges and issues in terminology mapping: a digital library perspective. The Electronic Library. Vol. 23 No. 6 (2005) 671-677
2. Open Archives Forum: Breakout Session [online]. Lisbon. (2002) 19 Available at: http://www.oaforum.org/otherfiles/oaf_d43_workshop2.pdf [cited 31 August 2007]
3. Chan, L., Zeng, M.: Ensuring Interoperability among Subject Vocabularies and Knowledge Organization Schemes: a Methodological Analysis [online]. 68th IFLA Council and General Conference, (2002) Glasgow. Available at: http://www.ifla.org/IV/ifla68/papers/008-122e.pdf [cited 31 August 2007]
4. Doerr, M.: Semantic problems of thesaurus mapping [online]. Journal of Digital Information. Vol. 1 No. 8 (2001) Available at: http://jodi.tamu.edu/Articles/v01/i08/Doerr/ [cited 31 August 2007]
5. Koch, T., Neuroth, H., Day, M.: Renardus: Cross-browsing European subject gateway via a common classification system (DDC). In: McIlwaine, I., C. (ed.):

Proceedings of the IFLA Satellite Meeting Held in Dublin, Ohio, 14-16 August 2001 and Sponsored by the IFLA Classification and Indexing Section, the IFLA Information Technology Section and OCLC. K. G. Saur, Munchen (2003) 25-33

6. SRW/SRU. Information available at: http://www.loc.gov/standards/sru/ [cited 31 August 2007]
7. JISC. JISC Information Environment. Information available at: http://www.jisc.ac.uk/whatwedo/themes/information_environment.aspx [cited 31 August 2007]
8. Miles, A., Brickley, D.: (eds). SKOS Core guide: W3C working draft 2 November. (2005) Available at: http://www.w3.org/TR/2005/WD-swbp-skos-core-guide-20051102/ [cited 29 August 2007]
9. OCLC. DDC 22 [online]. Available via OCLC Connexion: http://connexion.oclc.org [cited 31 August 2007]
10. Nicholson, D., Dawson, A., Shiri, A.: HILT: A pilot terminology mapping service with a DDC spine. Cataloging & Classification Quarterly. Vol. 42 No. 3/4 (2006) 187-200
11. Coates, E. J.: Switching languages for indexing. Journal of Documentation. Vol. 26 No. 2 (1970) 102-110
12. Horsnell, V.: The Intermediate Lexicon: an aid to international co-operation. Aslib Proceedings. Vol. 27 No. 2 (1975) 57-66
13. Koch, T.: Desire project handbook: 2, 5 subject classification, browsing and searching [online]. Available at: http://www.desire.org/handbook/2-5.html [cited 29 August 2007]
14. NISO (National Information Standards Organization), Report on the Workshop on Electronic Thesauri, November 4-5, 1999. Available at http://www.niso.org/news/events-workshops/thes99rprt.html [cited 20 July 2007]
15. Macgregor, G., Joseph, A., Nicholson, D.: A SKOS Core approach to implementing an M2M terminology mapping server. International Conference on Semantic Web and Digital Libraries (ICSD-2007 Proceedings of the), 21-23 February. Bangalore, India. Bangalore: Documentation Research & Training Centre, Indian Statistical Institute (2007) 109-120 Available at: http://eprints.cdlr.strath.ac.uk/2970/ [cited 29 August 2007]
16. J. Paul Getty Trust, Art and Architecture Thesaurus Online. Available at: http://www.getty.edu/research/conducting_research/vocabularies/aat/ [cited 31 August 2007]
17. GCMD, Global Change Master Directory. Available at: http://gcmd.nasa.gov/ [cited 31 August 2007]
18. HASSET, Humanities and Social Science Electronic Thesaurus. Available at: http://www.data-archive.ac.uk/search/hassetSearch.asp [cited 31 August 2007]
19. IPSV, Integrated Public Sector Vocabulary. Available at: http://www.esd.org.uk/standards/ipsv/ [cited 31 August 2007]
20. JACS, Joint Academic Coding System. Available at: http://www.hesa.ac.uk/jacs/jacs.htm [cited 31 August 2007]
21. JITA Available at: http://eprints.rclis.org/jita.html [cited 31 August 2007]
22. LCSH, Library of Congress Subject Headings. Available at: http://www.loc.gov/cds/lcsh.html [cited 31 August 2007]
23. MeSH, Medical Subject Headings, Available at: http://www.nlm.nih.gov/mesh/ [cited 31 August 2007]

24. NMR, National Monuments Record, Available at: http://thesaurus.english-heritage.org.uk/ [cited 31 August 2007]
25. SCAS, Standard Classification of Academic Subjects. Available at: http://www.ucas.com/higher/courses/scascode.pdf [cited 29 August 2007]
26. UNESCO Thesaurus, Available at: http://www2.ulcc.ac.uk/unesco/ [cited 20 July 2007]
27. CAB Thesaurus, Available at: http://www.cabi.org/DatabaseSearchTools.asp?PID=277 [cited 20 July 2007]
28. Iyer, H. and Giguere, M.: Towards designing an expert system to map mathematics classificatory structure. Knowledge Organization. Vol. 22 No. 3/4 (1995) 141-147
29. McCulloch, E., Macgregor, G.: Analysis of equivalence mapping for terminology services. Journal of Information Science. Vol. 33 No. 5 (2007)
30. Chaplan, M. A.: Mapping Laborline Thesaurus terms to Library of Congress Subject Headings: implications for vocabulary switching. Library Quarterly. Vol. 56 No. 1 (1995) 39-61
31. Dolin, Robert, H., Mattison, John, E., Cohn, S., Campbell, Keith, E., Wiesenthal, Andrew, M., Hochhalter, B., LaBerge, D., Barsoum, R., Shalby, J., Abilla, A., Clements, Robert, J., Correia, Carol, M., Esteva, D., Fedack, John, M., Goldbert, Bruce, J., Gopalarao, S., Hafeza, E., Hendler, P., Hernandez, E., Kamangar, R., Khan, Rafique, A., Kurtovich, G., Lazzareschi, G., Lee, Moon, H., Lee, T., Levy, D., Lukoff, Jonathan, Y., Lundbert, C., Madden, Michael, P., Ngo, Trongtu, L., Nguyen, Ben, T., Patel, Nikhilkumar, P., Resneck, J., Ross, David, E., Schwarz, Kathleen, M., Selhorst, Charles, C., Snyder, A., Umarji, Mohamed, I., Vilner, M., Zer-Chen, R., Zingo, C.: Kaiser Permanente's Convergent Medical Terminology. MEDINFO. Vol. 11 No. 1 (2004) 346-50 Available at: http://square.umin.ac.jp/DMIESemi/y2004/20041129_3.pdf [cited 31 August 2007]
32. Miles, A., Brickley, D.: (eds). SKOS Mapping Vocabulary Specification. (2004). Available at: http://www.w3.org/2004/02/skos/mapping/spec/ [cited 29 August 2007]
33. Liang, A., Sini, M., Chun, C., Li, S. J., Lu, W. L., He, C. P., Keizer, J.: The mapping schema from Chinese Agricultural Thesaurus to AGROVOC, 6th Agricultural Ontology Service (AOS) Workshop on Ontologies: the more practical issues and experiences, July 25-28, Vila Real, Portugal, 2005 (Food and Agriculture Organization, Rome, 2005). Available at: ftp://ftp.fao.org/docrep/fao/008/af241e/af241e00.pdf [cited 31 August 2007]
34. LexGrid, The Lexical Grid: Shared Terminology Resources. Available at: http://informatics.mayo.edu/LexGrid/index.php?page=aboutlg [cited 31 August 2007]
35. OCLC Terminologies service. Available at: http://www.oclc.org/terminologies/default.htm [cited 31 August 2007]
36. Doerr, M., Fundulaki, I.,: The Aquarelle Terminology Service, ERCIM News, 1998, no. 33. Available at: http://www.ercim.org/publication/Ercim_News/enw33/doerr2.html [cited 31 August 2007]
37. Christophides, V., Doerr, M., Fundulaki, I.: The Aquarelle Folder Server, ERCIM News, no. 33. Available at: http://www.ercim.org/publication/Ercim_News/enw33/doerr1.html[cited 31 August 2007]

38. OpenGALEN, Information available at: http://www.opengalen.org/faq/faq5.html [cited 31 August 2007]
39. Chute, C. G., Elkin, P. L., Sheretz, D. D., Tuttle, M., S.: Desiderata for a Clinical Terminology Server. American Medical Informatics Association. Available at: http://www.amia.org/pubs/symposia/D005782.PDF [cited 31 August 2007]
40. Renardus, Available at: http://www.renardus.org/ [cited 20 July 2007]
41. Svenonius, E.: Use of Classification in Online Retrieval. Library Resources and Technical Services. Vol. 27 No. 1 (1983) 76-80
42. Bowman, J., H.: Essential Dewey. Facet Publishing London (2005) 15
43. UDC Consortium, Available at: http://www.udcc.org/ [cited 31 August 2007]
44. Lloyd, G., A.: The Universal Decimal Classification as an International Switching Language. International symposium on UDC in relation to other indexing languages. Herceg Novi, Yugoslavia, June 28-July 1 (1971)
45. Balikova, M.: Multilingual Subject Access to catalogues of National Libraries (MSAC): Czech Republic's collaborations with Slovakia, Slovenia, Croatia, Macedonia, Lithuania and Latvia. In: Proceedings of the World Library and Information Congress: 71st IFLA General Conference and Council - Classification and indexing with cataloguing, Oslo, Norway, August 14-18 (2005) (IFLA, The Hague, 2005) Available at: http://www.ifla.org/IV/ifla71/papers/044e-Balikova.pdf [cited 30 August 2007]
46. W3C.: Semantic Web Deployment Working Group. (2007) Available at: http://www.w3.org/2006/07/SWD/ [cited 29 August 2007]
47. Vizine-Goetz, D., Houghton, A., Childress, E.,: Web services for controlled vocabularies. Bulletin of the American Society for Information Science and Technology, Vol. 32 No. 5 (2006) Available at: http://www.asis.org/Bulletin/Jun-06/vizine-goetz_houghton_childress.html [cited 29 August 2007]
48. Tudhope, D., Binding, C.: Toward terminology services: experiences with a pilot web service thesaurus browser. Bulletin of the American Society for Information Science and Technology, Vol. 32 No. 5 (2006) Available at: http://www.asis.org/Bulletin/Jun-06/tudhope_binding.html [cited 29 August 2007]
49. Nicholson, D., McCulloch, E.: Investigating the feasibility of a distributed, mapping-based, approach to solving subject interoperability problems in a multi-scheme, cross-service, retrieval environment. Proceedings of International Conference on Digital Libraries, 5-8 December, New Delhi, India. (2006) Available at: http://eprints.cdlr.strath.ac.uk/2875/ [cited 29 August 2007]
50. Svensson, Lars. G.: National libraries and the Semantic Web: requirements and applications. Proceedings of the International Conference on Semantic Web and Digital Libraries, Documentation Research and Training Centre, Bangalore, India. (2007) 101-108
51. W3C.: OWL Web ontology language guide, W3C, Massachusetts Institute of Technology, European Research Consortium for Informatics and Mathematics, Keio University. (2004) Available at: http://www.w3.org/TR/owl-guide/ [cited 29 August 2007]
52. Zhao, Y.: Combining RDF and OWL with SOAP for Semantic Web Services. Proceedings of the 3rd annual Nordic Conference on Web Services (NCWS'04), Vxj, Sweden 22-23 Nov (2004) 31-45 Available at: http://www.ida.liu.se/ yuxzh/doc/ncws-041002.pdf [cited 29 August 2007]

53. Liang, A., C., Sini, M.: Mapping AGROVOC and the Chinese Agricultural Thesaurus: Definitions, tools, procedures. New Review in Hypermedia and Multimedia. Vol. 12 No. 1 (2006) 51-62
54. Miles, A., Matthews, B., Wilson, M., Brickley, D.: SKOS Core: Simple knowledge organization for the Web. Proceedings of the International Conference on Dublin Core and Metadata Applications (DC-2005), Madrid, Spain, 12-15 Sep (2005) Available at: http://isegserv.itd.rl.ac.uk/public/skos/press/dc2005/dc2005skospaper.pdf [cited 29 August 2007]
55. Miles, A., Matthews, B., Beckett, D., Brickley, D., Wilson, M., Rogers, N.: SKOS: A language to describe simple knowledge structures for the web. Proceedings of XTech 2005: XML, the Web and beyond, Idealliance, Amsterdam, Netherlands (2005) Available at: http://www.idealliance.org/proceedings/xtech05/papers/03-04-01/ [cited 29 August 2007]
56. Zthes.: The Zthes specifications for thesaurus representation, access and navigation (2006) Available at: http://zthes.z3950.org/ [cited 29 August 2007]
57. Vizine-Goetz, D., Hickey, C., Houghton, A., Thompson, R.: Vocabulary Mapping for Terminology Services. Journal of Digital Information, (2004) Vol. 4 No. 4 Available at: http://jodi.tamu.edu/Articles/v04/i04/Vizine-Goetz/ [cited 31 August 2007]
58. Will, Leonard. RE: Zthes and DDC. Zthes - Development of the Zthes model for thesauri (mailing list), Index Data, Denmark (2005) Available at: http://lists.indexdata.dk/pipermail/zthes/2005-February/000020.html [cited 29 August 2007]
59. Nicholson, D., McCulloch, E.: Interoperable subject retrieval in a distributed multi-scheme environment: new developments in the HILT project. Ibersid, Zaragoza, Spain 2-4 Nov (2005) Available at: http://eprints.cdlr.strath.ac.uk/2317/01/Nicholson_ZaragosaPaperFinal.pdf [cited 31 August 2007]
60. HILT: M2M Final Report. Appendix D: Assessment: Use Cases, Protocols and Mark-ups. (2005) Available at: http://hilt.cdlr.strath.ac.uk/hiltm2mfs/0HILTM2MFinalReportRepV3.1.pdf [cited 29 August 2007]
61. HILT: Demonstrator. (2006) Available at: http://hiltm2m.cdlr.strath.ac.uk/hiltm2m/hiltsoapclient.php [cited 29 August 2007]
62. IESR: Internet Environment Services Registry. Available at: http://iesr.ac.uk/ [cited 31 August 2007]
63. Go-Geo! Demonstrator. (2006) Available at: http://nevis.ed.ac.uk:9200/gogeo-hilt2.html [cited 20 July 2007]
64. Ethimiadis, E.,N.: Interactive query expansion: a user-based evaluation in a relevance feedback environment. Journal of the American Society for Information Science. Vol. 51 No. 11 (2000) 989-1003
65. Sanchez-Alonso, S., Garcia, E.: Making use of upper ontologies to foster interoperability between SKOS concept schemes. Online Information Review. Vol. 30 No. 3 (2006) 263-277