



HILT IV: Research Report

Dennis Nicholson, Anu Joseph and Emma McCulloch

Centre for Digital Library Research (CDLR), University of Strathclyde, Glasgow

May 2009

Contents

1.0 Introduction	1
2.0 HILT Overview	1
3.0 Research Issues and Findings	2
3.1 System Architecture	2
3.1.1 Database design.....	2
3.1.2 Data handling issues	2
3.1.3 Collections database	2
3.1.4 Distributed approach.....	4
3.2 Terminologies and Terminology Mapping	4
3.2.1 DDC 22.....	4
3.2.2 Terminological spine	4
3.2.3 Need for 3 digits	4
3.2.4 Satellite schemes.....	5
3.2.5 Mapping of schemes to DDC 22.....	5
3.2.6 Mapping methodology.....	6
3.2.7 Mapping types.....	8
3.2.8 Storing mappings	8
3.3 Web Services and Standards	9
3.3.1 SRU/W	9
3.3.2 SKOS	10
3.3.3 BS 8723	10
3.4 Toolkit and Functions	10
3.4.1 Provision – different programming languages – Perl and PHP.....	10
3.4.2 HILT APIs.....	10
3.4.3 Spell checker	11
3.4.4 Wordnet.....	11
3.4.5 Portlet	11
3.4.6 XCRI/WikiWord/Umbel	11
3.5 Responsibilities of Services	11
3.5.1 Interface design.....	11
3.5.2 Quality of cataloguing	11
3.5.3 Use of standard subject schemes	12
3.5.4 Currency of schemes.....	12
4.0 Associated Work	12
4.1 Gold Dust.....	12
5.0 Dissemination	12
6.0 Evaluation	14
7.0 Future Service Issues	14
Appendix A: HILT IV: High-Level Mapping Study – UNESCO to DDC	15

1.0 Introduction

Within the HILT Phase IV project, a collection of research issues relating to the provision of an effective future entry-level service, or its further refinement, were investigated. This report forms the HILT Phase IV deliverable documenting research into a) any possible alternative approaches to spine provision and their implications; b) the identification of preferred spines for specific query types where options exist; c) many-to-many mappings; d) guidelines for others wishing to produce HILT-compatible mappings themselves; e) searching with compound terms; f) mapping types required for effective user services at different service levels; g) mapping grading and coding; h) a list of terminology or related service types likely to enrich user experience if encompassed within the HILT architecture; and i) the possible value of providing a HILT portlet (based on the JSR168 or WSRP standards) as a way of providing services with a relatively easy way of incorporating useful core user interface features into local services.

This deliverable reports on the research topics on the above list, together with an indication of other work identified during the project as necessary for the development of an effective future entry-level service or its further refinement and, where appropriate, a report of such research carried out during Phase IV and the HILT Embedding Project Extension. Research into areas included in the list above are reported on under the headings below:

- System architecture
- Terminologies and terminology mapping
- Web services and standards
- Toolkit (main deliverable)
- Responsibilities of services
- Associated work
- Dissemination
- Future service issues

Before documenting project findings in relation to the above list, a brief overview of the HILT model will be presented to contextualise the research issues being addressed.

2.0 HILT Overview

HILT is developing a pilot toolkit to help facilitate cross searching and browsing across information services using different subject schemes.

Since few services within the JISC information environment use the same subject scheme to describe their resources, or even any standard subject scheme at all in many cases, there is a need to provide tools to enable users to cross-search and browse distributed collections offered by JISC services and projects. This is what HILT is working towards.

HILT has a number of terminologies stored within its database, including subject headings, thesauri and classification schemes. These subject schemes are reconciled by mapping them to a Dewey Decimal Classification Scheme (DDC) 22 spine.

HILT uses SRU/W, a web services protocol, meaning that functions can be invoked on a machine to machine (M2M) basis. HILT provides results marked up in SKOS (Simple Knowledge Organisation System). Once the SKOS output is received it is then the responsibility of individual services as to how they want to parse SKOS results for subsequent presentation to their users.

The model means that much of the HILT functionality called upon by remote services takes place in the background, without the end user having to be heavily involved in specifying detailed information needs. So where, for example, a user search term does not match an index term within their service of choice, HILT should be able to provide an appropriate synonym, broader or narrower, or related term with which a relevant collection or service can be searched, where permissible, using the OpenURL protocol. So the 'user' of HILT will often be an information service accessing HILT and taking results back to its own (human) end users.

The architecture supporting this model is distributed and designed to provide an extensible service.

3.0 Research Issues and Findings

3.1 System Architecture

3.1.1 Database design

SQL server 2005 is used for storing data and there are fifteen schemes including DDC 22 available in HILT. Though SQL server seems to support loading large XML files into the database, it didn't prove feasible for some schemes due to their size. Programmes were written (using PHP) to split these large files into smaller chunks before loading them into the database. Normalising the tables affected the performance, especially when these tables are searched using JOIN. Flat tables were created to improve performance along with normalised tables knowing that data will not change that often. These flat tables have to be generated every time data changes in the main table.

A problem was encountered with SQL server – it periodically hibernates and subsequently takes 30 seconds to respond when a query is issued. The reason may be due to DNS resolution delays (either at the client or server side) or encoding mismatch between query and indexes, leading to table scans. The query response time after a long break was high and this issue was tackled by scheduling a job activity (run a stored procedure) every 10 minutes.

3.1.2 Data handling issues

Note: see also 3.2.1 DDC 22 for specific data handling issues relating to this classification scheme and 3.2.4 Satellite schemes for examples of some of the issues encountered in handling other schemes.

3.1.2.1 Foreign characters

SQL server driver for PHP5 cannot handle UTF-8 data properly, and resulted in storing question marks instead of Arabic characters. The SOAP server that interacts with the database is also written in PHP and the server couldn't pass a query with foreign characters properly. We resolved the issue by storing terms with translated characters as well as English characters in the database. We hope to solve this issue in PHP6.

Foreign characters also appeared in the upload of a selection of Welsh language mappings to DDC 22, obtained in XML format. These terms were amended manually within the database before making the Welsh mappings available for use within HILT.

3.1.2.2 Unique identifiers

Unique identifiers to represent a term are important from a database design point of view as well as from a SKOS representational point of view. Many schemes came without these unique identifiers, which generated problems while building relationships - especially with duplicate terms in UNESCO. Though LCSH provides identifiers for its data, there are not unique either. Automatic identifiers generated by the database created an unforeseen identifier clash issue, across different schemes, especially when represented in one SKOS file. This has been solved by including the scheme name before the identifier.

3.1.3 Collections database

The get_collections API currently queries a local collections database held at CDLR, containing a list of JISC collections and services catalogued by DDC number. The get_collections API uses this database to identify information services of potential relevance to a user query.

The intention within HILT has always been to use IESR to fulfil this role within the HILT toolkit. Due to various limitations of IESR in the earlier stages of HILT, the local collections and services database was set up within CDLR to simulate the functionality of IESR. At the end of this phase IV, IESR is not yet able to provide the functionality that HILT requires. As a result the switch from the local collections and services database has not yet been made.

Changes are required to IESR in order to facilitate its integration with HILT functionality. If these changes cannot be accommodated, HILT will not be able to use IESR as its collections and services database and will have to consider other options to fulfil this function. This is not desirable since IESR has already been funded to provide this function and it is neither HILT's intention nor wish to duplicate work taking place elsewhere.

To integrate HILT with IESR, a number of functional requirements are necessary.

a) DDC uses ranges to denote coverage of certain concepts. In HILT, we are constructing scheme hierarchies using parent identifiers, which means that we have to include ranges as these parent identifiers, where appropriate. If a DDC caption is identified as useful, its parent may also be relevant to the user's query. Such ranges are in some cases expressed using greyed out zeros in DDC. For example, the 100 section shows:

DDC notation	DDC caption
100	Philosophy & psychology
100	Philosophy
100	Philosophy, parapsychology and occultism, psychology

Table 1: The 100 division of DDC [Source: WebDewey, OCLC]

The uppermost notation in the table above denotes coverage of 100-199. The middle notation in the table denotes coverage of 100-109. The third notation in the table does not denote a range, it indicates DDC 100 precisely. The representation of these three variations is difficult to accommodate within the HILT database since the DDC notation is used within the system as a concept's id. All ids within the database require to be unique.

As with the majority of ranges, and easily illustrated by 302-307 'Specific topics in sociology and anthropology' cataloguers are instructed to assign a single notation to a given resource, rather than assigning a range. For example, within 302-307 the instruction to "Class comprehensive works in 301" is given, along with the note stating that "Unless other instructions are given, class a subject with aspects in two or more subdivisions of 302-307 in the number coming last". The DDC rule of three¹ may also come into play, where resources covering three or more subjects equally are classed at the first higher number that includes all three. It may not be necessary to incorporate ranges within HILT to facilitate resource discovery therefore, since ranges should not be applied to individual resources. It is necessary however, to devise a way of handling the (seemingly) duplicate use of notations, in numerical terms, within HILT to enable the system to generate a browsable hierarchy of DDC using ids and parent ids and to display the broader hierarchy of any given search term (within the get_ddc_records function). Further research is required to fully examine the implications if incorporating DDC ranges into HILT.

b) HILT also requires that subject scheme information be available for collections and services listed in IESR. Without this, HILT will not be able to direct a user to an appropriate subject scheme for a given collection or service, meaning that potentially relevant mappings or browse hierarchies of the scheme in question will not be exposed to the user. HILT functions identify potentially relevant services and collections in response to a user query; the user should then be given details of the scheme used within that service/collection, and any terms mapped from that scheme to the DDC notation of relevance. This information is central to the HILT model to enable users to either enter the relevant mapped term within a service/collections search facility, or to search a service/collection remotely using the openURL protocol where available. Without the fundamental scheme information being made available via IESR, the HILT functions cannot be executed.

c) The OpenURL protocol enables HILT to query services and collections dynamically from within the HILT interface. The selection of a search term in relation to an Intute collection, for example, will invoke a new web browser window in which the relevant search will be performed without further action from the user. IESR does not currently store information on services' use of the OpenURL protocol.

d) Elimination of duplicate records and improved consistency of metadata. Progress has been made on this front but early in the project we were unable to delete duplicate records from IESR and many of the records varied hugely in their coverage.

Following detailed discussions with IESR staff Leigh Morris and Jo Lambert, as well as with Vic Lyte, we are

¹ http://www.oclc.org/dewey/resources/teachingsite/courses/choice_of_number_review.pdf

now working with IESR to resolve current barriers to service integration between HILT and IESR.

3.1.4 Distributed approach

HILT successfully connected to an OCLC server and an IESR server using the SRU/W protocol. This and the fact that the OCLC service is theoretically discoverable through a services registry like IESR is taken to show that a distributed approach encompassing a wider, non-JISC, environment, is feasible.

3.2 Terminologies and Terminology Mapping

3.2.1 DDC 22

A new version of DDC was released during the HILT programme of work. Datasets therefore required to be updated from DDC 21 to DDC 22 at the outset of Phase IV, to ensure currency of data. This required the HILT team to obtain an XML file containing the new version of DDC from OCLC, complete with LCSH to DDC mappings. Additional time was spent loading the data into the HILT database and undertaking data cleaning to standardise notations, eliminate foreign characters and the like. Changes to the arrangement of some sections of DDC introduced new data handling challenges.

For example, within the revised version a number of structural changes are evident. There is wider use of ranges to provide guidance to the structure of DDC, which is not currently being handled effectively by HILT.

Although WebDewey² presents various ranges, making use of greyed out zeros, as appropriate, this cannot be replicated within the HILT database since HILT requires all DDC numbers to have a minimum of three digits, in order to facilitate the truncation process, on which HILT relies for some of its functionality. Since DDC notations are used as ids within HILT, these notations also require to be unique. Within the HILT database, distinct notations are required for distinct captions in order to create a means of generating a browse hierarchy. Term identifiers (i.e. the DDC notations in the case of DDC) and parent identifiers are used to identify the structural arrangement of captions within the scheme.

The example shown in Table 1, which is echoed for each of the hundred divisions (i.e. those notations ending in 00) as well as all notations ending in a single zero, have so far been handled unsatisfactorily by HILT. The 'fullest' caption was retained and given the unique identifier of _00. The information in Table 1 was therefore collapsed, with 100 being represented by 'Philosophy, parapsychology and occultism, psychology' labelling this entire class, with no upper, or higher, levels beyond. This is clearly inadequate and does not represent DDC and its use of ranges accurately.

3.2.2 Terminological spine

HILT uses DDC as its terminological spine. DDC was decided upon as HILT's spine for a number of reasons. 1) The subject coverage of DDC is universal. 2) Its notational system means that it lends itself fairly well to truncation (and subsequent refinement of results sets), a key feature of HILT's disambiguation process where users are required to select the context of their search term from its (possible) various instances throughout DDC. 3) DDC has been translated into over 30 languages³, while new translations are being continually worked on.

Alternative spines have been, and will continue to be, considered. UDC is a possible alternative spine and pilot mappings are currently being created between UDC and DDC to enable HILT to assess its suitability, or otherwise, in relation to HILT's existing satellite schemes and in relation to different subject areas, services and collections.

Section 7.0 describes current thinking behind a possible sustainable future architecture. Such a model would allow many alternative spines to be used.

3.2.3 Need for 3 digits

The HILT system requires DDC notations to contain a minimum of three digits. DDC 22 has captions that

² <http://www.oclc.org/dewey/versions/webdewey/>

³ <http://staff.oclc.org/~dewey/dewey.htm>

overlap conceptually, and are denoted as ranges, shown using greyed out zeros. This arrangement is repeated for each of the hundred divisions (see Table 1 for example). Due to the need for unique identifiers this structural arrangement has not yet been satisfactorily represented within HILT. Further work is needed to devise a way of representing the notations that signify a range in such a way that is compatible with the HILT APIs and also with SKOS.

3.2.4 Satellite schemes

Fifteen schemes other than DDC have been uploaded to the HILT database. These are: AAT (Art and Architecture Thesaurus), GCMD (Global Change Master Directory), NMR (National Monuments Record), HASSET (Humanities and Social Science Electronic Thesaurus) JACS (Joint Academic Coding System), IPSV (Integrated Public Sector Vocabulary), UNESCO Thesaurus (United Nations Education, Scientific and Cultural Organisation Thesaurus), MeSH (Medical Subject Headings), CAB Thesaurus, JITA (subject scheme used within the ELIS⁴ repository), LCSH (Library of Congress Subject Headings), RAE units of assessment, SCAS (Standard Classification of Academic Subjects - replaced by JACS in 2002), SPEIR (in-house scheme used by the CDLR SPEIR project), XCRI (eXchange of Course-Related Information). Selected portions from selected schemes are mapped to areas of DDC, as appropriate. Schemes were chosen based on their use within collections and services within the JISC Information Environment (IE). For example, an email-based survey revealed that both HASSET and MeSH were used within various areas of the Intute service.

3.2.4.1 Scheme variation

Typically, schemes are of different sizes, they exhibit different levels of granularity, they are structured differently, they cover different subject areas and so on. It follows that what works well whilst mapping one particular scheme to DDC, may not work well for any other scheme. As HILT incorporates fifteen different schemes, it is difficult to impose generic mapping rules to be followed in every case. Some of the issues encountered with specific schemes are documented elsewhere in this report.

3.2.5 Mapping of schemes to DDC 22

The decision(s) on which portions of chosen schemes to be mapped was based on the research premise that high and deep level mappings could be used to facilitate browsing and searching, respectively, within potentially relevant collections and services.

3.2.5.1 What to map

HILT maps concepts from several subject schemes to a DDC spine. This will facilitate a process of vocabulary switching to improve interoperability within, and across, services employing different schemes to describe their resources. A range of high and deep level mappings will be implemented in order to help the user to 1) identify hierarchical information including broader, narrower and related terms associated with a concept in a given scheme and 2) identify appropriate concepts in given schemes with which to search specific services employing those schemes.

This dual-approach of providing high and deep level mappings was taken to 1) offer the user search and browse services, as mentioned above, and 2) to research the value of mappings at different levels of granularity.

The subject coverage of mappings to be included was decided on the basis of JISC collections and services with whom we were able to work during phase IV. What schemes to focus on and the extent of mapping implemented was largely dependent on this. HILT worked with CAIRNS (<http://cairns.lib.strath.ac.uk/>) and the Social Sciences section of Intute (<http://intute.ac.uk/socialsciences/>). An additional consideration was to ensure schemes selected covered the same subject area(s), to enable a direct comparison to be made between the effectiveness of search versus browse using the same set of mappings.

It followed that HASSET, UNESCO and IPSV would be mapped to the DDC spine at a high level. Deep level mapping would be focused within the subject areas of mental health and psychology within MeSH and HASSET.

Mapping work is also time consuming and costly. Whilst mapping portions of three different schemes – HASSET, MeSH and UNESCO - to DDC, HILT researchers calculated that the average mapping takes 7 minutes to create. Cost and resourcing issues may also influence decisions taking on what schemes, and what

⁴ <http://eprints.rclis.org/>

portions of schemes, should be mapped.

For the purposes of the HILT IV follow-up, the Embedding project, JACS has also been mapped to DDC in its entirety. To fit with the high and deep-level mapping model, all 1300 JACS codes/terms were mapped to the top 919 DDC notations, since this meets functional requirements of Edina and Intute-based services. Not all of these 919 notations have been mapped to, but the targets for mappings from JACS are restricted to these 919 notations. The project has also created mappings from the RAE subject headings to DDC, with a view to investigating the possibility of using HILT functionality to improve subject access in institutional repositories.

3.2.6 Mapping methodology

There is a fairly substantial list of reasons why, especially when we get down to individual examples, mappings are difficult to create (not least because of scheme variation, mentioned in 3.2.4.1.). Careful attention needs to be paid to the scope of concepts in individual schemes to ascertain the coverage of a term and the nature of its equivalence to a DDC notation.

In HILT, we have adopted a pragmatic approach, creating mappings that we consider useful to the user. So, where we can't attain an exact match to a particular DDC concept, we may opt to include a range of broader, narrower and related, including Boolean combinations, to try to give the user a range of potentially useful options.

As previously noted, HILT incorporates a range of high and deep level mappings.

3.2.6.1 High level mapping

The provision of high level mappings will enable users to enter an appropriate point of a browse hierarchy of a given scheme within the area relating to their subject interest. As such, HILT will undertake a mapping exercise from HASSET, IPSV and UNESCO to the top three levels of DDC. The top thousand DDC notations, and corresponding captions have therefore been identified. Equivalent concepts from HASSET, IPSV and UNESCO will then be identified and mapped to each DDC notation individually. The optimum mapping will be an exact match. If there is no exactly equivalent concept within a satellite scheme (in this case HASSET, IPSV and UNESCO), a combination of concepts that collectively constitute an exact match should be sought. Where combinations are adopted + and | symbols will be used to represent AND and OR respectively. Where an exact match cannot be identified within a satellite scheme (either directly or via a Boolean combination), a narrower or broader match may be sought, with a view to pinpointing the next best match. Judgement should be used to decide which of narrower or broader is the closest match to the concept being mapped to. Where both are considered useful, both should be included. Where neither narrower nor broader concepts can be identified a related match may be identifiable.

3.2.6.2 Deep level mapping

It has been decided that the areas relating to psychology within HASSET and MeSH will be mapped to DDC, as this area has been identified as useful to intute.

HASSET has the following terms in this area:

```
PSYCHOLOGY
|..APPLIED PSYCHOLOGY
|  ..CLINICAL PSYCHOLOGY
|    ..PSYCHOANALYSIS
|    ..PSYCHOTHERAPY
|      ..DRUG-PSYCHOTHERAPY COMBINATION TREATMENT
|      ..HYPNOTHERAPY
|  ..EDUCATIONAL PSYCHOLOGY
|  ..OCCUPATIONAL PSYCHOLOGY
|    ..MANAGERIAL CHARACTERISTICS
|    ..LEADERSHIP
|  ..SOCIAL PSYCHOLOGY
|..DEVELOPMENTAL PSYCHOLOGY
|  ..ADOLESCENT PSYCHOLOGY
|  ..CHILD PSYCHOLOGY
```

```

| ..EMOTIONAL DEVELOPMENT
|   ..EMOTIONAL IMMATURITY
|   ..EMOTIONAL MATURITY
| ..INDIVIDUAL DEVELOPMENT
| ..MENTAL DEVELOPMENT
|   ..LANGUAGE DEVELOPMENT
| ..PERSONALITY DEVELOPMENT
|   ..PERSONALITY CHANGE
| ..PARAPSYCHOLOGY
| ..EXTRASENSORY PERCEPTION

```

Figure 1: Section of HASSET hierarchy [Source:

<http://www.data-archive.ac.uk/findingData/thesaurusInfo.asp?keyword=PSYCHOLOGY>]

Within MeSH section F relates to psychology, as available at

<http://www.nlm.nih.gov/mesh/2008/MeSHtree.F.html>

The methodology for establishing deep level mappings is the same as that for high level mappings, although reversed since we are unable to identify DDC notations/captions in advance. Terms to be mapped will be identified within each of the satellite schemes (HASSET and UNESCO), before actively seeking an appropriate match within DDC. The optimal outcome is to establish mappings in the following order of preference: exactMatch; narrowMatch/broadMatch; relatedMatch

3.2.6.3 Need for many-to-many mappings

At the uppermost levels of the DDC hierarchy, subjects often appear inter or multidisciplinary. In cases such as 000, for example, where the associated caption is ‘computer science, information & general works’, it is highly probable that many-to-one mappings will be required to capture equivalence for the caption as a whole. It is also probable that a ‘better’ match will be identified for corresponding UNESCO terms at a lower level of DDC, making many-to-many mappings necessary.

A decision has been taken to encode ‘computer science’ and ‘information’ as individual NTs of DDC ‘computer science, information & general works’. Likewise for ‘psychology’ and ‘philosophy’ in relation to DDC ‘psychology & philosophy’. This is because each of the terms form part of the subject being mapped to, but only partially. In other words the DDC term is broader than either of the mapped terms, individually.

Since this group of narrowMatches are sufficiently different from a narrowMatch dictated by hierarchical structure, another argument is that we should introduce a match type signifying ‘partial exact match’. It is arguable that each of the terms psychology and philosophy are more closely matched than other, perhaps more typical, NTs.

The need for Boolean operators (or SKOS classes) within queries is also relevant here for combined search, as is the issue of pre/post coordination of mapped terms. Such issues have been subject to discussion within the SKOS community and have not yet been resolved.

3.2.6.4 Boolean mapping

Boolean mappings are required to express equivalence relationships between satellite schemes and DDC. Since concepts rarely overlap in their entirety between schemes, it follows that AND and OR are useful for combining concepts, to determine a ‘closer’ match with the concept being mapped to/from.

3.2.6.5 Mapping study: UNESCO to DDC

A high-level mapping study was undertaken whilst mapping UNESCO to DDC. Findings of this exercise are documented as Appendix A. Note that this study took place early on in the project schedule; many of the issues have now been superseded by subsequent research. The study is included here, intended to be illustrative of the types of specific issue that may be encountered while undertaking terminology mapping.

3.2.7 Mapping types

To help prioritise the usefulness of mappings implemented SKOS mapping types⁵ are being adopted within HILT to indicate the type of relationship between a particular term and a DDC notation.

A detailed research study was conducted in HILT III (McCulloch & Macgregor, 2008)⁶ to determine the appropriate range of mapping types or equivalence relationships required, with which to categorise mappings from satellite schemes to the DDC spine. The outcome of this study was that five mapping types would be used within HILT, in line with SKOS standards. These are: exactMatch; broadMatch; narrowMatch; majorMatch and minorMatch. Subsequent to this study, major and minor match were deprecated, being replaced with overlappingMatch. Further discussion then concluded that relatedMatch should be used instead of overlapping, even although a significant number of contributors felt there was a clear need for both overlapping and related, which they viewed as significantly distinct (see also 3.3.2.1).

So the four mapping types used at the time of undertaking mapping work are exactMatch, broadMatch, narrowMatch and relatedMatch. Since the completion of this work a new SKOS mapping type has been introduced - closeMatch. This, and future changes, will require to be accommodated within HILT but the research team feel it would be better to wait for the standard to stabilise in the area of mapping before incorporating changes throughout the HILT APIs.

3.2.7.1 JACS to DDC: mapping type issue

Consultant terminology expert Leonard Will⁷ observed, when creating high-level mappings (see 3.2.6.1) from JACS to the top three levels of DDC (as defined within the HILT project), very few cases of exact match were found. This was mainly because the limited range of DDC numbers used did not allow the exact JACS topics to be expressed. More precise numbers are available in many cases in DDC. Most of the JACS concepts are therefore shown as narrower than the DDC number to which they are mapped. This was deemed unreliable, though, because in many cases there was some overlap - the JACS concepts included things that are not in the DDC classes and vice versa. The lack of an 'overlap' match is a serious limitation, and guesses had to be made as to the direction in which most of the overlap occurred in deciding whether to show the match as broader or narrower.

3.2.8 Storing mappings

When undertaking terminology mapping in line with HILT's methodology, an Excel Spreadsheet should be established for each scheme being mapped, whether at high or deep level. DDC notations should be listed in the leftmost column, with DDC captions in the column alongside. Mapped terms from a satellite scheme should be included in the third column, as single terms or combinations, as appropriate. The fourth column will show the type of mapping equivalence between the concepts being mapped. A notes field is also useful to accommodate any information that might justify the assignation of a particular mapping type, which may not be obvious otherwise. For example, if there is a USE FOR instruction to a term which, if preferred, would constitute an exact match.

DDC Class	DDC Caption	IPSV Term	Mapping Type	Notes
005	Computer programming, programs, data	Programming	exactMatch	
020	Library and information sciences	Library and information services	narrowMatch	

⁵ <http://www.w3.org/TR/2009/WD-skos-primer-20090317/#secmapping>

⁶ McCulloch E. & Macgregor G. Analysis of equivalence mapping for terminology services, *Journal of Information Science* 2008 34(1) pp.70-92. Available at <http://strathprints.strath.ac.uk/3173/>

⁷ <http://www.willpowerinfo.co.uk/>

070	Documentary media, educational media, news media; journalism; publishing	Journalism + Newspapers + Communications industries	exactMatch	Communications industries: UF publishing
-----	--	---	------------	--

Table 3: Example mappings from IPSV to DDC (1)

Where different equivalence relationships are expressed for different mappings to the same DDC notation/caption, these should be stored on different rows of the Excel file, to facilitate parsing of results. For example, if a broadMatch and a narrowMatch are both identified as valid mappings to the same DDC notation/caption this data should be stored as follows:

DDC Class	DDC Caption	IPSV Term	Mapping Type	Notes
130	Parapsychology and occultism	Psychology	broadMatch	
130	Parapsychology and occultism	Occultism	narrowMatch	

Table 4: Example mappings from IPSV to DDC (2)

Where the match type is the same e.g. all narrowMatches, as many terms can be stored on a single row of the file, using Boolean operators as appropriate. For example, three distinct narrow matches should be expressed using the OR (|) operator.

3.3 Web Services and Standards

3.3.1 SRU/W

Implementation is based on Index Data's SimpleServer – a simple Perl module intended to develop Z39.50, SRU and SRW servers. SimpleServer is based on popular YAZ toolkit which is robust, efficient, portable and inter-operates well with different Z39.50 and SRU/W servers. There is also a SOAP envelope involved, though it is transparent to the clients. See also 3.1.4.

3.3.1.1 Explain response

The EXPLAIN operation returns a ZeeRex (<http://explain.z3950.org/>) XML file that allows a client to find out the functional capabilities of an SRU/W server, and which indexes are available to use in CQL queries. It was hoped that more detail about the CQL structure, including controlled vocabulary values, could be given in the ZeeRex file. However, the ZeeRex format does not allow for this possibility, and so the considered recommendation is to create a Context Set reference document (a human readable document) explaining the CQL structure. The ZeeRex maintainers are considering adding a value to the index definition in ZeeRex to indicate whether the index contains controlled values. If this comes into operation then clients could use an SRU/W scan operation to find the controlled vocabularies.

3.3.1.2 Namespace declaration in individual records

SRU/W responses are XML documents containing an SRU/W specific wrapper or envelope. In the case of SRW there is a further SOAP envelope involved but this should be invisible to the application using the SRW client. In particular, the SRU/W searchRetrieve response may contain records that are transmitted in XML. These are part of the XML searchRetrieve response and so any XML namespaces used within these XML records must be declared so that the record data is within scope in the SRU/W XML document as a whole.

The most efficient, and conventional, place to declare these namespaces would be within the root element opening tag (along with other namespace declarations required for the SRU/W XML document). However, due to limitations of the software being used to front the SRU/W server (Indexdata's perl Net::Z3950::SimpleServer based on their popular yaz library) the namespace declarations for the records can only be inserted at the record data level. This means that for every record's namespaced tags to be in the scope of a declaration the namespace

declaration must be inserted into the opening tag of every record. This is less efficient than desired, but it is not invalid XML. It makes no technical difference to the XML document, only to its size.

3.3.1.3 Caching

An SRU/W search request can specify the start record and the number of records to retrieve from the result set. This is so that records from a large result set can be retrieved a handful at a time for manageable browsing by the end-user. Unlike Z39.50, SRU/W does not give identifiers to the result sets, but rather uses the original CQL query issued in the request. Since this query won't change between requests for different pages of records it is suitable for caching the SOAP response (from the HILT SOAP server) in the SRU/W server ready for the next page request. In fact, the caching is now keyed on the SOAP method that the CQL query is translated into since the SRU/W client can request the hilt:matches or hilt:concepts part of the SOAP response by altering the CQL query. This provides a huge performance boost and greatly improves the response time for users subsequently requesting further results from an initial CQL query, or indeed the hilt:concepts.

3.3.2 SKOS

A working draft of the SKOS primer or, user guide, can be found at <http://www.w3.org/TR/2009/WD-skos-primer-20090317/> SKOS is a developing standard and the accompanying reference document is located at <http://www.w3.org/TR/skos-reference/>. Both are published by the Semantic Web Deployment Working Group as part of the W3C Semantic Web Activity.

3.3.2.1 Limitations

SKOS hasn't solved the issue with representing Boolean mappings, which force us to find a way to represent our data. HILT uses + sign to represent AND and | sign to represent OR in SKOS records. SKOS validation fails because of these signs in the output, but can easily adapt to any SKOS accepted solution in the future.

The change from SKOS mapping type overlappingMatch to relatedMarch (see also 3.2.7 and 3.2.7.1) is potentially limiting since, although thesaurus standards don't provide a way to distinguish between these, it may be useful to have both since overlapping concepts, by definition, must be from the same facet of a scheme while related concepts can be taken from different facets of a scheme.

See also 3.1.2.2 Unique identifiers.

3.3.3 BS 8723

BS8723 Parts 1-4 were adhered to.

3.4 Toolkit and Functions

A toolkit has been developed to illustrate how different HILT functions can be usefully embedded within a service. The Perl based toolkit along with documentation is available to download in the wiki at http://linuxserv.cdlr.strath.ac.uk/hiltwiki/index.php/Hilt4_Toolkit

A demonstrator of the same is available at <http://hilt4.cdlr.strath.ac.uk/toolkit/intro.cgi>

3.4.1 Provision – different programming languages – Perl and PHP

3.4.2 HILT APIs

The HILT Toolkit uses the following APIs, accessible for testing at http://hilt4.cdlr.strath.ac.uk/hilt_SRU/W.cgi:

get_ddc_records

- Returns DDC captions and numbers related to a subject term. The user can then choose the most appropriate to his/her interest.

get_collections

- Returns collections classified under a specified DDC number or its stem, including subject scheme used. We will look at this function in more detail shortly.

get_non_ddc_records

- Returns terms from schemes other than DDC by matching user terms to DDC notations, before identifying mappings to those particular notations.

get_all_records

- Combines the functions of get_DDC_records and get_non_DDC_records.

get_filtered_set

- Allows specified fields from specific terminologies or combinations of terminologies to be searched.

3.4.3 Spell checker

A spell checker based on an index created from HILT's local database has provided added-value in HILT's search functionality (get_sp_suggestions). The Lucene spell checker is implemented in Java and the implementation is based on David Spencer's code using the n-gram method. In order to access Java classes in PHP (SOAP server is built using PHP), PHP-Java bridge is enabled in the server and the function is available as a web service. The existing class has also been extended to accept compound queries.

3.4.4 Wordnet

WordNet® is a large lexical database of English language terms, developed under the direction of George A. Miller. Nouns, verbs, adjectives and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept. WordNet has also added value in HILT searches, using get_wordnet_suggestions, by helping users to choose their search term. WordNet suggestions are available as a web service and the implementation is Java based.

3.4.5 Portlet

After a brainstorming session with Edina staff, it was determined that the portlet (based on the JSR168 or WSRP standards) did not have the level of functionality that HILT required. In particular, it did not appear to be possible to pass parameters between two portlets on a service screen.

3.4.6 XCRI/WikiWord/Umbel

The possibility of integrating these projects with HILT was explored. Discussions with XCRI are ongoing; WikiWord data is not yet in a position to be used; Umbel can add values to HILT search functionality.

3.5 Responsibilities of Services

Optimum results from HILT are dependent on a range of external factors, some of which lie directly with the collections and services being consulted/searched by HILT. Responsibilities of individual collections and services that will have a direct effect on the value of HILT to users, as well as elements over which HILT had no control such as the way HILT is presented within, or incorporated into, a service include:

3.5.1 Interface design

The way in which HILT functionality is presented to users within services and collections, once embedded in their local websites, is largely the responsibility of the service itself. HILT offers a downloadable toolkit for integration but will not offer assistance in redesigning service websites to accommodate this.

As such, little work has been done in this phase of HILT relating to interface design since it is probable that individual services and collections will want to incorporate HILT in different ways, perhaps only using part of the toolkit in some cases or perhaps only using elements of the toolkit relating to a specific scheme or schemes, in line with how its resources have been classified.

See also 6.0 Evaluation.

3.5.2 Quality of cataloguing

The value of terms returned to HILT depends on how effectively they have been applied to individual records within services and collections. HILT can be used to identify exact matches, narrower/broader terms and so on,

but the user will only retrieve useful resources from a service or collection if they have been effectively catalogued using appropriate terms.

3.5.3 Use of standard subject schemes

It is also crucial that services and collections apply subject schemes exactly according to the instructions/scope notes within a given scheme. Standard schemes should be used ‘as is’, that is, they should not be adapted to suit local needs if services and collections wish to be fully integrated with HILT. HILT maps standard schemes to a DDC spine. Unless a service or collection informs HILT directly, provides their adapted scheme for upload into the HILT database and requests that additional mappings be created from this locally adapted scheme to the DDC spine, that collection or services terminology will not be reconciled with those used in HILT.

3.5.4 Currency of schemes

In addition to employing a standard scheme (unless effort is to be made to incorporate local or in house schemes into the HILT infrastructure) it is also essential that the most recent version of any given scheme is used. HILT will update its terminological content in line with that of scheme providers so it is essential that services and collections do the same to maintain interoperability with HILT data to give users the best results.

4.0 Associated Work

See also 3.4.6 XCRI/WikiWord/Umbel.

4.1 Gold Dust

We analysed Personal Interest Profiles (PIPs) data collected as part of the Gold Dust project (<http://www.hull.ac.uk/golddust/>) and explored the possibility of matching these terms with any specialised subject area ontology in HILT. The outcomes are

Out of 5180 key phrases collected using two methods devised by Gold Dust, A and B, only 86 distinct HILT terms matched the terms identified using method A and 127 in method B⁸. There was overlap of these terms across terminologies and detailed matching of terms in different HILT schemes are listed in the table below.

HILT scheme	HILT terms matched in Method A (out of 2730)	HILT terms matched in Method B (out of 2450)
CAB	55	69
Dewey	16	25
GCMD	7	6
HASSET	5	12
IPSV	2	7
JACS	3	0
LCSH	27	29
Mesh	24	28
NMR	1	54
SCAS	1	2
UNESCO	7	19

Table 5: Number of terms matched in individual schemes in HILT

5.0 Dissemination

An email list HILT-collaborators was established early on in the project to establish potential stakeholders’

⁸ The exercise here was not to compare methods A and B as defined by the GoldDust project, but to investigate whether HILT would be of any value in identifying terms that may feature in PIPs.

functional needs from a terminology service such as HILT. What do they want it to be able to do? What would they find useful? What level of technical expertise is available to them to embed HILT functionality in their local services? These, and other, questions were discussed and the HILT team received positive feedback on the APIs being developed. An archive of the HILT-collaborators mailing list is available at <https://www.jiscmail.ac.uk/cgi-bin/webadmin?A0=HILT-COLLABORATORS>

HILT IV work with the members of the HILT Collaborators email list showed that there was a good deal of interest from the community in these services. Twenty-nine people joined the list which was set up specifically to look at embedding work of this kind. A questionnaire asking what kinds of uses were of interest was answered by sixteen people. All indicated an interest in at least one potential way of using HILT, most indicated an interest in three or more ways.

A wiki was established to facilitate exchange between the HILT project team and interested parties. The wiki was used to disseminate elements of the toolkit during development stages, for testing and feedback. Demonstrations of completed elements were also made available by this means. The wiki was later used to elicit discussion regarding technical embedding of HILT's APIs in EDINA-based, intute-based and CDLR-based services within the remit of the embedding extension project.

In addition, several talks, demonstrations and publications have been delivered throughout the project lifetime:

Presentations/demonstrations:

- HILT: Enhancing subject search through embedded web services, JISC Conference 2009, Edinburgh International Conference Centre, Edinburgh, 24th Mar 2009.
- Metadata and Scotland's information environment: potential benefits of Web 2.0, Metadata Issues and Web Services, CIGS Seminar, 30 Jan 2009.
- HILT IV Pilot Toolkit Demonstration, CIG 2008, University of Strathclyde, Glasgow, 3-5 Sep 2008.

Scheduled presentations/demonstrations:

- Nicholson, D. Signposting the crossroads: terminology web services and classification-based interoperability, Classification at a Crossroads – Multiple Directions to Usability, International UDC Seminar, The Hague, 29-30 Oct 2009.
- Nicholson, D. Looking at the Past and preparing for the Future, IFLA Satellite Meeting, Florence, August 2009.
- Embedding terminology web services to improve subject access, Where will it all end? – Emerging Technology in the Library, MmIT North West, Liverpool John Moores University, 17th Jun 2009.

Conference Papers:

- Nicholson, D. 'Optimising Interoperability in Multi-KOS Subject Searching: Framework for a Collaborative Approach? The Challenge of the Electronic Environment to the Organization of Knowledge - Second International Seminar on Subject Access to Information, Helsinki, Finland, 29-30 Nov 2007.
- Macgregor G. & McCulloch E. & Nicholson D. Terminology server for improved resource discovery: analysis of model and functions. Second International Conference on Metadata and Semantics Research. Corfu, Greece 11-12 Oct 2007. Available at: <http://strathprints.strath.ac.uk/3435/>
- 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). Bangalore, India 21-23 Feb 2007. Available at: <http://strathprints.strath.ac.uk/2970/>

Journal Articles:

- Nicholson, D. A Common Research and Development Agenda for Subject Interoperability Services? Signum, Issue 5, 2008.
- McCulloch, E. & Macgregor, G. Analysis of equivalence mapping for terminology services, Journal of Information Science, 2008 34(1) pp.70-92. Available at: <http://strathprints.strath.ac.uk/3173/>

Submitted Journal Articles:

- Nicholson, D., McCulloch, E. & Joseph, A. HILT IV: Subject Interoperability through Building and Embedding Pilot Terminology Web Services. World Digital Libraries.

Practitioner Articles:

- Nicholson, D. & Menzies, K. BUBL, HILT, and the Scottish Information Environment: potentials of Web 2 and Web 3, WIDWISAWN, 7(1) Available at:

6.0 Evaluation

An evaluation of HILT was undertaken during Phase IV by Ian Ruthven of the Computer and Information Sciences Department at the University of Strathclyde. This study is included as Appendix G of the final report, which will be available at <http://hilt.cdlr.strath.ac.uk/hilt4/documents.html>.

7.0 Future Service Issues

Toward the end of HILT IV and the embedding project, it became evident that a sensible path to follow to sustain the future of an effective terminologies service to improve subject interoperability was to:

- Identify the likely elements and architecture of an inclusive subject interoperability service that could, in time, incorporate not only HILT and a mapping based approach built around a DDC spine but other terminology services across the UK and beyond applying either different approaches to mapping or different approaches to interoperability.
- Identify the research requirements associated with this architecture and its elements.
- Work towards an agreed multi-initiative approach to the associated research (and development) with partners elsewhere in the UK and in the world.

Although work in this area is still at an early stage, efforts have nevertheless been made to begin work towards agreeing a collaborative approach with other ‘players’ in the terminologies field. A paper on the architecture was presented at an ontologies conference in Helsinki in November 2007, a paper on the topic published in the international version of the Signum journal in November 2008, and, in December 2008, steps were taken to contact major European projects in the terminologies area to begin the process of talking about collaboration and about applying for FP7 funding to carry the work forward.

A range of issues would require further research and development should a collaborative approach prove favourable. These are fully documented in the project’s final report, which will be made available at <http://hilt.cdlr.strath.ac.uk/hilt4/documents.html> following its acceptance by JISC.

Appendix A: HILT IV: High-Level Mapping Study – UNESCO to DDC

1. Introduction:

This document describes a HILT IV investigation into the mapping of UNESCO thesaurus to the top three levels of the Dewey Decimal Classification (DDC) Scheme hierarchy, or the top thousand DDC numbers (actually 919). Several satellite schemes are to be mapped to DDC at this level; the project has begun with UNESCO. In addition to these top level mappings, a deeper level mapping exercise is planned and is dependent on schemes in use within intute and across the JISC Information Environment.

Equivalence relationships identified between satellite scheme terms and DDC notations will be encoded using the SKOS Mapping Vocabulary Specification (MVS). It is of interest to consider Chaplan's mapping types in addition to the SKOS MVS. An initial attempt to reconcile the two approaches has been made, with a view to incorporating both approaches into the HILT mapping database. This may form part of an evaluative exercise undertaken later in the project. It will be of interest to consider whether the five SKOS mapping types are sufficient to characterise the nature of mappings between schemes or whether Chaplan's more detailed set is more valuable in the context of retrieval.

2. Aim:

Within the HILT project, terms from a number of subject schemes will be mapped to a central DDC spine. Mappings will then be used to provide subject interoperability by offering terms mapped to DDC, together with details of the type of mapping relationship evident. Subject access via scheme hierarchies entered at a point appropriate to the user's subject interest will be provided.

3. Methodology:

Notations corresponding to the top 3 levels of DDC (tens, hundreds, thousands) have been identified, together with captions. Exact matches are to be sought for each within the UNESCO thesaurus, taking into consideration the hierarchical context of each scheme and any associated scope notes and instructions. Where no exact match exists, the next 'best' match (likely to be narrower or broader) will be identified.

It is important to consider the hierarchical context of notations/terms since, for example, for DDC 192, the HILT database lists the term to map to as 'British Isles'. Only from the broader hierarchy or from detailed knowledge of DDC's structure do we know that this instance of 'British Isles' in fact relates to Modern western philosophy.

Mapped DDC numbers will then be input to the database, against mapped terms from each of the satellite schemes. In addition to the mapped number being entered, the type of relationship characterised by the mapping will be encoded in line with the SKOS Mapping Vocabulary Specification (MVS), as follows⁹:

exactMatch

broadMatch

narrowMatch

majorMatch¹⁰: Concept Match (CM) with significant overlap

minorMatch²: CM with slight overlap. If not exact, narrower, broader or major, it's probably minor i.e. some type of relationship, but none of them quite fit.

Definitions and examples:

exactMatch: applies to semantically equivalent concepts with 100% overlap. Such terms need not be exact character-by-character matches. That is, they may exhibit spelling variation, intervening characters and so on.

⁹ See section 5 (SKOS MVS developments) for details of recent development of the SKOS MVS and proposed resolutions to SKOS ISSUES.

¹⁰ Deprecated; proposal made to change to overlappingConcept

broadMatch: applies to a match where a term from a satellite scheme is broader in scope than the equivalent DDC term. e.g. UNESCO: science; DDC: natural sciences & mathematics.

narrowMatch: applies to a match where a term from a satellite scheme is narrower in scope than the equivalent DDC term e.g. UNESCO: history; DDC: history & geography.

majorMatch²: conceptually equivalent terms with significant overlap (>50%), but not exactly equivalent. e.g. spheres – balls

Not all spheres are balls e.g. globes

Not all balls are spheres e.g. rugby ball

Since the majority of spheres are balls, but exceptions exist (in both directions), this example constitutes a MajorMatch.

minorMatch²: conceptually equivalent terms with some degree of overlap, although not significant (<50%). If terms are not deemed exactly matched or majorly matched, but there is some level of overlap between them. It is unlikely that terms exhibiting a minor match will exist within the same discipline/hierarchy. For example: DDC

DDC number	DDC caption	UNESCO terms	Mapping type
005.8	Data security	CRIME	minorMatch

In the above example, Data security and CRIME are not exactly matched and one is neither narrower nor broader than the other. In some contexts however it is possible that there may be a degree of conceptual overlap. The extent of this overlap is likely to be encountered in limited circumstances so the terms may be deemed as a minorMatch. In practice, it is thought likely that minorMatch will be applied very infrequently.

4. Findings:

In undertaking the mapping of 919 UNESCO terms to the top thousand DDC captions the following issues were noted. One issue related to the nature of UNESCO itself is also documented here, after those relating more specifically to the mapping of UNESCO to DDC.

1. Need for many-to-many mappings

At the uppermost levels of the DDC hierarchy, subjects often appear inter or multidisciplinary. In cases such as 000, for example, where the associated caption is ‘computer science, information & general works’, it is highly probable that many-to-one mappings will be required to capture equivalence for the caption as a whole. It is also probable that a ‘better’ match will be identified for corresponding UNESCO terms at a lower level of DDC, making many-to-many mappings necessary.

A decision has been taken to encode ‘computer science’ and ‘information’ as individual NTs of DDC ‘computer science, information & general works’. Likewise for ‘psychology’ and ‘philosophy’ in relation to DDC ‘psychology & philosophy’. This is because each of the terms form part of the subject being mapped to, but only partially. In other words the DDC term is broader than either of the mapped terms, individually.

Since this group of narrowMatches are sufficiently different from a narrowMatch dictated by hierarchical structure, another argument is that we should introduce a match type signifying ‘partial exact match’. It is arguable that each of the terms psychology and philosophy are more closely matched than other, perhaps more typical, NTs.

NB. The need for Boolean operators (or SKOS classes) within queries is also relevant here for combined search, as is the issue of pre/post coordination of mapped terms.

2. Indirect exact matches

Where USE/UF relationships are evident within a scheme, mappings may be implemented as exact yet may not be immediately apparent. For example:

UNESCO: Palaeontology
 UF Palaeobiology, Palaeobotany, Palaeozoology
 DDC: 560

DDC number	DDC caption
560	Paleontology Paleozoology

In the absence of the UF instruction UNESCO term Palaeontology would be considered a narrowMatch to DDC 560, since it constitutes a sub-area of the DDC caption. However, the UF instruction indicates that the term palaeontology would also be used for instances where palaeozoology might be adopted. As a result of this instruction, the UNESCO term is deemed to match DDC 560 exactly.

When it comes to DDC 561 however, this approach becomes problematic.

DDC number	DDC caption
561	Paleobotany; fossil microorganisms

The relevant UNESCO terms to map to DDC 561 are palaeontology (since it is UF palaeobotany) and fossils. Palaeontology is an exactMatch to paleobotany in DDC, however since 'fossil microorganisms' is part of the same heading HILT mapping methodology dictates that this is in fact a narrowMatch. Fossils is a narrowMatch in the sense that it is a subset of palaeontology, yet is broader than fossil microorganisms. There is therefore a degree of conflict here. The term microorganisms in UNESCO could also be mapped here. This term is located within UNESCO's biology hierarchy as follows:

Microbiology
 NT1 Bacteriology
 NT2 Microorganisms

and is already mapped to DDC 628.536 within the HILT database. The complete hierarchy for DDC 628.536 (below) indicates however, that this does not appear to be an appropriate match for the context of microorganisms denoted by DDC 561.

[DDC 628.536: Technology > Engineering and allied operations > Sanitary and municipal engineering
 Environmental protection engineering > Pollution control technology and industrial sanitation
 engineering > Microorganisms]

3. Equal 'narrowness'/'broadness'

Where narrower matches are identified, comparable levels of granularity may not be evident, both within and across mapped schemes. Considering a mapping to DDC 000 computer science, information & general works, from a range of satellite schemes, including UNESCO we see the following:

AAT: computer programming
 CAB: computers
 UNESCO: computer science; information

Clearly, the three examples above are not equivalently narrower matches of computer science, information & general works. The UNESCO terms are immediately narrower, the CAB term is a narrower term of the AAT (2 levels narrower than DDC?) and the AAT term is a narrower term of the CAB term (i.e. 3 levels narrower than DDC?).

Within HILT, each of the above examples will be encoded as a narrowMatch, since each term is narrower than the DDC notation to which it is mapped. In retrieving a result set however, this aspect of the methodology does mean that the mapped terms presented in response to a query will not necessarily be equivalently narrower. This is really down to the nature of schemes themselves, their structures,

levels of granularity and so on. Do we foresee problems with this approach?

4. Treatment of notes in DDC

To improve the consistency of mappings from satellite schemes to DDC, decisions have been taken regarding ‘class here’ and ‘include’ notes as they appear in the DDC schedules. Class here is to be treated as a concept match, encoded as either majorMatch/minorMatch², in line with the SKOS MVS, depending on the degree of conceptual overlap. Where notes under an area of DDC state ‘include TERM A’ TERM A from a satellite scheme will be deemed a narrowMatch of the DDC notation in question.

5. Consistent use of mapping types

In addition to the potential problem resulting from the assignment of the narrowMatch mapping type, further evidence has been uncovered that suggests the assignment of mapping types may be contradictory, depending on specific examples.

DDC 576:

DDC number	DDC caption	UNESCO terms	Mapping type
576	Genetics and evolution	GENETICS	narrowMatch
		EVOLUTION	narrowMatch

Each of the above terms is mapped as a narrowMatch since each form a subset of the complete DDC heading. However, both UNESCO terms ‘Genetics’ and ‘Evolution’ when considered independently, are clearly broader than DDC 576 (in contrast to the ‘psychology & philosophy’ case discussed in 1. above). This makes documentation of mapping guidelines problematic since many issues will have to be handled on a case-by-case basis.

For DDC 579:

DDC number	DDC caption	UNESCO terms	Mapping types
579	Microorganisms, fungi, algae	MICROORGANISMS	narrowMatch
		FUNGI	narrowMatch
		AQUATIC PLANTS	narrowMatch

Aquatic plants is also mapped as a narrowMatch since in UNESCO there is an instruction to USE aquatic plants for Algae. However, in the DDC schedules under 579 there is an instruction to class here microbiology and various other concepts. According to HILT mapping methodology, class here instructions (see 4. above) constitute a concept match (majorMatch/minorMatch²) between terms. This means that for the above example, microbiology would be mapped to microorganisms, fungi, algae as a majorMatch² or minorMatch². Depending on how results are to be ranked within HILT therefore, a concept match may be treated as a ‘better match’ than a narrowMatch. Clearly when looking at the terms however, this may not necessarily be the case.

If there are seemingly more appropriate narrowMatches for a particular caption, should we leave it at that? Or, should all instructions (from both classification scheme and thesaurus in this case) be considered? Can class here and include notes be treated consistently across all cases? Should we exert more flexibility in our interpretation of these?

If such notes are to be treated consistently there is greater scope for a degree of automation being introduced to the mapping process; or at least the potential to speed up the intellectual mapping process in these cases. Will this lead to mapping errors or less value for the user, however?

6. Lack of qualifiers in DDC

Not so much a mapping related issue but a potential problem for the disambiguation phase of HILT.

Although, as already mentioned, the complete hierarchical information available will be considered in the mapping of schemes, and indeed in the presentation of initial matches to a user search term, until now we have largely assumed that duplicate terms in DDC i.e. at the end point of hierarchies are likely to be located in different disciplines or, if within the same discipline, for example, two instances of ‘teeth’ appear in the technology hierarchy, that the nature of the hierarchy itself will provide sufficient means to enable the user/client to differentiate between them.

The process of mapping UNESCO to DDC however, has uncovered instances where identical terms belong to the same discipline, within very close proximity. For example, DDC 218 and DDC 233 both have the caption ‘humankind’ and both are located within the discipline of religion. The first instance relates to philosophy and theory of religion and the second to Christianity, Christian theology. How can HILT provide further information to the user in order to help the user choose between such instances? Is it a problem for HILT? What happens if a user wants to select two or more instances of a term? Will it be documented as a problem arising purely from DDC’s structure?

7. Need for compound searching of mapped terms

Even when mapping the highest levels of DDC we can see the need for pre-coordinated terms to be searched. For example, in the case of ‘psychology & philosophy’ as discussed earlier. This illustrates that UNESCO does not always have a suitable term to map to a DDC number that covers the full extent of the concept represented by that number. Often, terms mapped from UNESCO cover one aspect of the concept denoted within DDC. Looking at more granular subjects, for example:

DDC 193:
 100
[Philosophy & psychology](#)
 180-190
[Historical, geographic, persons treatment of philosophy](#)
 190
[Modern western philosophy](#)
 193
 *Germany and Austria
 Source: WebDewey

DDC number	DDC caption	UNESCO terms	Mapping type
193	Germany and Austria	PHILOSOPHY	broadMatch
		GERMANY	broadMatch
		AUSTRIA	broadMatch

In the above example, searching for either one of ‘Germany’ or ‘Austria’ or ‘Philosophy’ is unlikely to retrieve resources relevant to DDC 193. The SKOS classes – AND, NOT and OR¹¹ – may be applicable here.

8. Issues specific to UNESCO

UNESCO uses microthesauri. There are a total of seven microthesaurus headings within UNESCO as a whole. These have been included in the HILT database as top level terms. It has emerged however that some of these microthesaurus headings have duplicate preferred terms within lower levels of the microthesauri. If this was a consistent approach adopted by UNESCO it would be reasonable to delete the microthesaurus headings, using the duplicate preferred terms as terms to map to DDC as appropriate. It appears that some of the microthesaurus headings are unique however, creating uncertainty in how to handle them for HILT purposes. In addition, relationships exist between microthesaurus headings and preferred terms. Such relationships would be lost if deleted and if not duplicated between preferred terms. Should we encode microthesaurus headings (MTs) as such? How would HILT handle these?

¹¹ skos:Intersection; skos:Union; skos:Negation

5. SKOS MVS Developments

During November 2007 the following developments have occurred:

1) majorMatch and minorMatch have been deprecated.

SKOS ISSUE-39 (<http://www.w3.org/2006/07/SWD/track/issues/39>) proposes to introduce skos:overlappingConcept

skos:overlappingConcept may be expanded to accommodate specific weightings if required. Whereas majorMatch and minorMatch indicated semantic overlap of <50% and >50%, further specification of the extent of overlap can be expressed using skos:overlappingConcept e.g. 0-30%.

2) SKOS ISSUE-39 also proposes to introduce skos:equivalentConcept in place of exactMatch.

The status of skos:related remains open.

3) Classes:

skosm: AND

skosm: OR

skosm: NOT

have been replaced by skos:Intersection; skos:Union; skos:Negation respectively.

4) ISSUE-39 states that instances like the use case proposed by HILT are probably out of SKOS core scope, that is “mapping links focused more on the conceptual mapping process than the essence of the conceptual mapping result”.

5) Summary:

New set of mapping properties:

broader

narrower

related?¹²

equivalentConcept – replacing exactMatch?⁹

overlappingConcept – replacing major/minorMatch

New set of mapping classes:

skos:Intersection

skos:Union

skos:Negation

In light of the above developments, previous HILT mapping work will be revisited. Mappings previously denoted as major/minorMatch will be replaced by overlappingConcept.

6. Discussion

Issues documented in this paper will be tabled for discussion at the HILT IV Steering Group meeting scheduled for February 2008.

3/12/07

¹² Awaiting clarification from the SKOS mailing list on whether related is to be introduced and whether exact Match has been replaced