Summary of High-level Thesaurus (HILT) mapping work – principal results

George Macgregor and Emma McCulloch - June 2006

This brief report summarises terminology mapping and equivalence issues found as part of the Jisc funded High-level Thesaurus (HILT) project (phase III), in particular the match types required to support machine-to-machine (M2M) terminology services.

One continuing problem inherent in the terminology mapping process - whether intellectual or automated - is accurately characterising the type of mapping match found between terminologies. The assumption underpinning mapping is that equivalence can exist between disparate Knowledge Organization Systems (KOS) and their respective terminologies; however, exact equivalence is rarely attainable. The existence of linguistic inconsistencies across terminologies (e.g. synonyms, homonyms, antonyms, etc.), grammatical variations (e.g. singular / plural forms, alternative spellings or punctuation, verb tenses, etc.), variations in subject coverage, and the relative specificity with which terminologies accommodate like concepts, render any exact equivalence problematic. Disparity in the semantic structure of the terminologies being mapped can also be particularly acute across different KOS; for example, classifications have radically different structures to that of relational vocabularies. Consequently mapped terms may not exemplify exact equivalence, but only partial equivalence.

Given that exact equivalence between terminologies will be rare, it is necessary to accurately characterise the degree of equivalence by assigning match types during the mapping process. This is often necessary to enable advanced search functionality and to provide users with sufficient information to make relevance judgements.

There is much research in the area of mapping match types. The most significant and comprehensive contribution has been proposed by Chaplan (1995). She proposes 19 match types to characterise equivalences between terminologies for vocabulary switching. These are listed in the table below (Table 1).

Table 1: Chaplan Match Types

Match type code	Definition
1	Exact match
2	Exact cross-reference match
3	Exact match, but with intervening characters
4	Plurals
5	Subordination, in the form of a species-genus relationship
6	Superordination, in the form of genus-species relationship

7	Part-of-speech difference
8	Word-order variation
9	Further specification
10	Spelling variation
11	Suffix variation
12	Abbreviation or acronym
13	Subdivision
14	Concept match
15	Homograph
16	Translation
17	Date or numerical variation
18	No match
19	Opposite or negative

Chaplan suggests that these match types could be used in conjunction with a variety of terminologies; however, this – until now – has never really been tested. A simple program was therefore written to extract 50 random terms from various HILT terminologies: LCSH, UNESCO, AAT and MeSH. These terms were then mapped to DDC and appropriate Chaplan match type codes were assigned. This work was duplicated by both authors to increase validity. When the mapping and the assignation of match types was complete, the results of each author were compared. Inconsistencies were resolved via discussion and further consulting terminology schedules. The match types assigned to each mapping were then totalled in order to indicate which match types were likely to be required across all terminologies.

Match Types Required for HILT III

The match types identified for HILT III as a result of this test are listed in table 2. In addition to these match types, it is clear that match type 19 ('Opposite or negative') needs to be employed also. Although they are extraordinarily rare, anecdotal evidence indicates that such matches do exist between terminologies. Instances of such matches were not found in our test because the data set was simply too large to encounter them.

Table 2: Chaplan Match Types: HILT Requirements

Match type code	Definition
1	Exact match
3	Exact match, but with intervening characters
4	Plurals

5	Subordination, in the form of a species-genus relationship
6	Superordination, in the form of genus-species relationship
7	Part-of-speech difference
9	Further specification
10	Spelling variation
14	Concept match
PLUS	
19	Opposite or negative

Points to note

- It should be noted that the inclusion of radically different terminologies could see this match type list increasing. This would be particularly likely should HILT decide to serve multi-lingual terminologies or include terminologies employing unconventional semantic structures.
- However, the above list adequately accommodates mappings to DDC from LCSH, UNESCO, AAT and MeSH – all of which are comprehensive, detailed and complex terminologies (this is especially the case for LCSH and MeSH). There is therefore good reason to assume that similarly structured and/or detailed terminologies – whether they are relational vocabularies, classifications or term lists - will also be accommodated.
- Term lists (e.g. authority files, glossaries, gazetteers, dictionaries, etc.) were not selected for testing. HILT has several term list terminologies (e.g. JACS). Such terminologies were excluded from our test on the basis that they assume radically simple structures when compared to relational vocabularies and classification. Any match types capable of accommodating the later two forms of KOS should be more than capable of accommodating term lists.
- Of the Chaplan match type **not** selected, most actually suffer from definition inconsistencies. Though this does not affect current HILT work, it may emerge at a later date when (or if) HILT wants to integrate numerous disparate terminologies (as mentioned above, e.g. multi-lingual terminologies). Problems with these definitions are due to be documented via a published research paper which may propose alternative and more robust definitions for these match types.

References

Margaret A. Chaplan, Mapping Laborline Thesaurus terms to Library of Congress Subject Headings: implications for vocabulary switching, *Library Quarterly* 65 (1) (1995) 39-61.