68th IFLA Council and General Conference
August 18-24, 2002

Code Number: 006-122-E
Division Number: IV
Professional Group: Classification and Indexing
Joint Meeting with: -
Meeting Number: 122
Simultaneous Interpretation: -

Subject-based Interoperability: Issues from the High Level Thesaurus (HILT) Project

Dennis Nicholson
Strathclyde University
Glasgow, UK

Introduction: HILT Phase I to HILT Phase II

The subject-based interoperability issues covered in this paper arise from two projects, now called HILT Phase I and HILT phase II. HILT Phase I (previously referred to only as the HILT Project) reported in November 2001. It was funded jointly by the Joint Information Systems Committee (JISC) and the Research Support Libraries Programme (RSLP) and lasted just over a year. JISC funding for HILT Phase II is expected to be confirmed in April 2002. Phase II will last for 12 months, and will utilise the work of HILT Phase I, and the skills and experience of the team that carried it out, to build on the cross-community consensus achieved in HILT Phase I by creating a pilot terminologies mapping service or route map with a specific focus on current concerns in the developing Distributed National Electronic Resource (DNER), including – but not necessarily limited to – Higher Education (HE) and Further Education (FE) focused subject terminologies for collection level description in the JISC's planned Information Environment (IE). The user evaluation and cost benefit analysis of various levels of service will also be features of the project.

The Problem

Ensuring that FE and HE users of the IE can find appropriate learning, research and information resources by subject is one of the major challenges facing the JISC, the DNER, the Resource Discovery Network (RDN), and the various key information and learning service providers across the archives, libraries, museums, and electronic services domains. As HILT Phase I discovered, the various service providers use a range of subject schemes (from general schemes like LCSH, UNESCO, DDC, and AAT, to specific schemes like MeSH) to meet the requirement to adequately and consistently describe their resources for

accurate retrieval. If cross-searching and browsing is to function coherently for users of the IE, these schemes must be mapped to one another, perhaps using a common 'spine' such as DDC with international and multi-lingual application and the potential to facilitate machine to machine (M2M) interworking. More importantly, perhaps, the terminologies in the minds of different types of FE and HE users must be 'disambiguated'[1], then translated into the service-assigned terms the users need to cross-search or browse the group of services of relevance to their query. The aim of HILT Phase II is to build and evaluate a pilot service that will mediate this process as a DNER 'Shared Service' in the Information Environment.

HILT Phase 1

HILT Phase 1 found that:

- Many different subject schemes and practices are in use in UK services who believe that subject searching across their services is of value both to their users and their staff.
- There was a strong consensus across the Archives, Electronic Services, Library, and Museums communities in favour of a more practically focused follow-up pilot project that would develop, and accurately determine the full costs and benefits of, a networked, user and machine responsive, interactive route map to the terminologies used by these communities and the relationships between these terminologies (previously referred to within HILT as a 'pilot mapping service' - see Terminologies Route Map (TeRM) diagram in Appendix A for an outline description of what it is and how it would function)
- Further research was required into the effectiveness, level and nature of user need, practicality, design requirements, and costs against benefits of such an approach before a long term commitment to a possibly expensive service could be justified. This, it was determined, could best be done via a pilot project that would examine these and related issues.

Further details of HILT Phase 1 can be found on the HILT web-site[2] generally, and in the HILT Final Report[3] in particular.

HILT Phase II: Aims

HILT Phase II moves this process into the pilot project stage, focusing - as recommended by the HILT Phase I evaluator - on terminology and thesauri requirements at collection level, but also bearing in mind the need to extend this in due course to the needs of item level retrieval. It will utilise the work of HILT Phase I, and the skills and experience of the team that carried it out, to set up a pilot terminologies route map or TeRM service, similar to that proposed in HILT Phase I, aiming to:

a. Provide a practical experimental focus within which to investigate and establish subject terminology service requirements for the JISC Information Environment, with particular reference to DNER, RDN, User, Collection Level, International Compatibility, and local, regional, national and UK-wide access considerations.
b. Make recommendations as regards a possible future service, taking into account a range of factors, including the level and nature of user need, practicality, design requirements, effectiveness, functionality available in existing commercial software packages as against original development, and (above all) costs against benefits to FE and HE users of a full terminologies service focussed primarily on collection level needs

[1] The process of determining whether the user who types in 'lotus' is searching for information on the car, the software package, the flower, or one of the many additional meanings of this term
[2] http://hilt.cdlr.strath.ac.uk/
[3] http://hilt.cdlr.strath.ac.uk/Reports/FinalReport.html

HILT Phase II: Participants

HILT Phase II will last for 12 months, and will involve roughly the same mix of participants as HILT Phase I, but with the addition of more direct involvement from representatives from the DNER, the RDN, and FE. Specifically:

- The Centre for Digital Library Research (CDLR) at Strathclyde University – lead;
- DNER representative
- mda (formerly the Museums Documentation Association);
- National Council on Archives (NCA);
- National Grid for Learning (NGfL) Scotland;
- Online Computer Library Center (OCLC);
- RDN representative
- FE Representative
- Scottish Library and Information Council (SLIC);
- Scottish University for Industry (SufI);
- UK Office for Library and Information Networking (UKOLN).
- Terminology experts, Alan Gilchrist and Leonard Will (external evaluator)

Through its involvement in the CAIRNS[4] clumps project (which utilised collection strengths to landscape mini-clumps), the SCONE and SEED[5] projects which combined to build a cross-sectoral collections database[6], and HILT[7] Phase I, the lead site - Strathclyde University's Centre for Digital Library Research[8] - has extensive experience in the use of collection level descriptions in a dynamic distributed environment, and of associated terminology problems. It also has available a rich distributed information environment in which to study the operation of the pilot and its interaction with users and services. This includes the CAIRNS distributed catalogue with universities, National Library of Scotland (NLS), NGfL, SLAINTE, and Glasgow Digital Library (GDL) databases, a subject-based collection strengths landscaping mechanism, the SCONE named collections database, an Open Archives Initiative (OAI) e-prints server, New Opportunities Fund (NOF) and other digitisation project databases, and the potential to mount other Z39.50 databases. Other participants - particularly UKOLN, mda, NCA, the RDN and the DNER, and the HILT terminology experts, add additional depth and breadth to the team. In addition, OCLC has agreed to assist the study by providing access to a machine-readable mapping of LCSH to DDC and associated access to expertise. The CDLR also works closely with the ten Glasgow FE colleges within the RSLP GDL project.

Building the TeRM

For the purposes of this project, the pilot TeRM would be built using commercially available Wordmap[9] software. This is known (through HILT Phase I experience) to provide a good initial illustration of the kind of facilities needed for the pilot. This does not imply a preference for this software or supplier, nor even for a commercial as opposed to a 'home-grown' or open source approach. The project would aim to develop a full requirement specification through evaluative activities conducted by user and service focus groups and external experts. It would then compare *all* relevant packages available, having conducted an in-depth survey of all current commercial and other

[4] See http://cairns.lib.strath.ac.uk/ - Z39.50 catalogue including universities, NLS, NGfL, and others
[5] See http://scone.strath.ac.uk/ and http://seed.cdlr.strath.ac.uk/
[6] See http://scone.strath.ac.uk/service/index.cfm
[7] See http://hilt.cdlr.strath.ac.uk/
[8] See http://cdlr.strath.ac.uk/
[9] See www.wordmap.com

solutions. WordMap would be amongst those able to offer software that might meet a significant part of the specification, but would not be favoured. The question of whether or not a community-based open source approach is preferable to buying a commercial solution would also be examined.

There are good reasons for using a specific piece of commercial software at this stage of development. Experience within HILT Phase I suggests that project participants find it easier to discuss the requirements of such a service given a real illustrative example on which to focus. It is therefore believed essential that we mount an illustrative pilot early on in the project in order to help engage the interest and attention of users and other stakeholders and give them a practical environment within which to envisage and consider the problem. Wordmap is being used because we want to have a real working demonstrator at an early stage for users and service providers to interact with. Attempting to draw out the full requirement *before* implementing an illustrative pilot would, it is believed, result in a poorly researched requirement as users and service providers would not have been sufficiently stimulated by operation in a real context to allow a full specification to emerge. This approach is viewed as a pragmatic one that will enable us to evaluate the real uses and issues in a timely way, whilst also avoiding the potential waste and risk involved in development from scratch before a full requirement has been established.

Terminologies and Terminology Related Issues

The initial illustrative TeRM would be based on the RDN terminologies[10], on terminologies available as part of the Wordmap taxonomies set, which include, in particular, a set of terms used by general internet users, and on selective subsets of LCSH, DDC, UNESCO, and AAT. OCLC will provide an LCSH – DDC mapping, and may also be able to provide a DDC to Conspectus subject headings[11] mapping. The UNESCO thesaurus is available online[12] and we will look to obtain AAT selections from manual sources. The aim would be a *selective* mapping sufficient for the purposes of the pilot in the first instance – i.e. not a comprehensive terminologies map. Consideration would also be given to the various issues raised by the HILT Phase I evaluator, Leonard Will (HILT Final Report, Section 10), and two additional questions:

1. The question of whether or not the TeRM needs a central spine

   A key element in the provision of such a pilot will be to 'translate' the user's subject retrieval 'problem' as couched in the user's own terminology to the various terminologies used in the distributed environment, and to do so in an intelligent and helpful way. This will usually require a certain amount of user-TeRM interaction to 'disambiguate' the term or terms used by the user (e.g. does she mean lotus, the flower, or the car, or the software package, or what?). There is then a question as to whether it is:

   a. Feasible
   b. Best in terms of good resulting retrieval for the user (note that this includes a need to retrieve across language barriers)

   for this interaction to take place between the user and a single central scheme to which all other schemes in the environment are mapped in the TeRM, or between the user and each individual scheme in turn. Following this, if the best answer is a single spine scheme, there is a question as to which existing scheme, if any, would best serve this purpose, the most likely possibility being DDC (a well-

[10] See, as an indication, the list created by Andy Powell at http://www.rdn.ac.uk/cgi-bin/browse
[11] Conspectus subject headings are used in the CAIRNS collection strengths database
[12] See http://www.ulcc.ac.uk/unesco/thesaurus.htm

structured, hierarchical scheme already translated into a significant number of major world languages).

2. The question of whether or not the best long term solution to the subject terminologies problem in a distributed environment might not be the identification and adoption of a single scheme accepted as adequate to cover all purposes in all domains - either an entirely new scheme, or an existing scheme, possibly amended to suit an accepted model requirement.

   This is, in essence, an extension of any cost-benefit analysis of the idea of a terminologies mapping service – an external reference point against which to assess the value to the community of the TeRM approach as against other possible approaches.

Building the Research Environment

This would be achieved by adding a range of DNER and other collections, including RDN collections, Archives collections, Museums collections, and a local OAI collection, to a *copy* of the SCONE Collections database[13] to create a HILT Phase II testbed collections database and CLD-based landscaping and cross-searching environment using the CAIRNS dynamic landscaping mechanism and broadcast search facility. The aim would be to utilise 'native subject schemes' for the collections in the environment, and to use the pilot TeRM to 'disambiguate' user terms and resolve differences between schemes. A range of user base-landscapes would be utilised, roughly associated with subject hubs as regards subject interests, but representing a variety of user circumstances, local, regional, national, UK-wide (general) and UK-wide (subject hub)[14]. The aim would be to link the TeRM to the landscaping mechanism if possible (CAIRNS experience suggests it should be), or to simulate this aspect if it is not (this would be less elegant, but sufficient for project research purposes).

HILT Phase II Deliverables

The specified deliverables for HILT Phase II are:

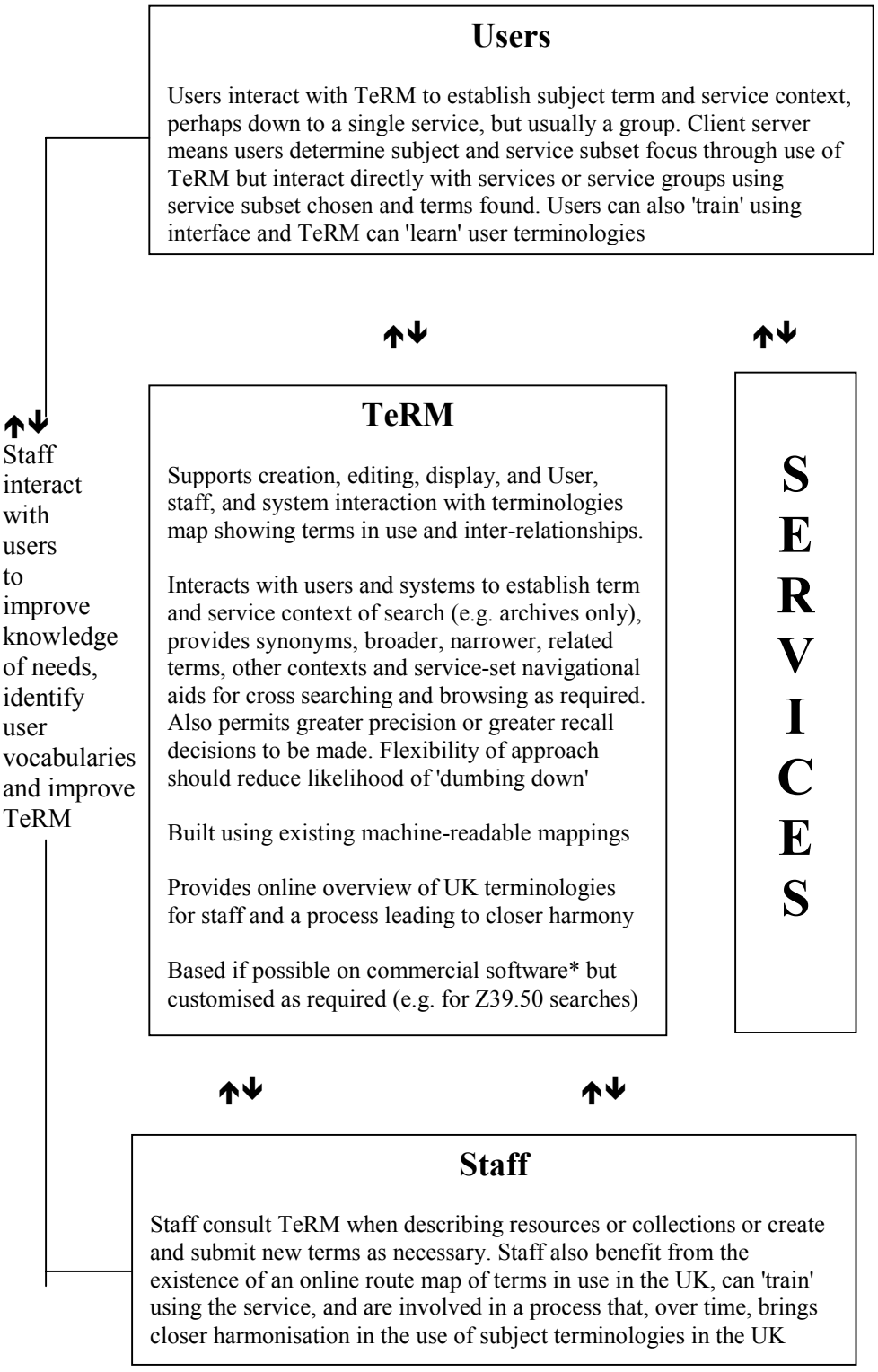
1. Greater understanding of the problem and of the needs of FE and HE users in respect of subject retrieval in the projected JISC Information Environment, both within JISC, JISC services, and - though dissemination activities - in the community as a whole.
2. An in-depth understanding of terminology mapping requirements in the DNER and associated UK services, taking local, regional, national, international, subject-hub, FE and HE, and archives, libraries, museums, and electronic services considerations into account.
3. A working pilot terminologies demonstrator service for the JISC IE (with limited functionality and with a full service possibly requiring a change of software).
4. Requirements, set up and maintenance costs, and costs against benefits, for a future service, including both user and M2M terminological and functional requirements.
5. Final Report on the project, together with appropriate recommendations.

Provided the expected funding is forthcoming (only informal notification received at time of writing), HILT Phase II will begin in May 2002.

[13] http://scone.strath.ac.uk/Service/Index.cfm

[14] Specifically, a university, an FE college, HE, FE in Glasgow landscape, HE, FE in Scotland landscape, HE, FE DNER level landscapes, HE, FE landscapes at an RDN subject hub

Appendix A Interactive Terminologies Route Map (TeRM) Diagram



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