Challenges and issues in terminology mapping: a digital library perspective

Abstract

Effective information retrieval within digital libraries is limited by the lack of semantic interoperability between subject schemes used by online services and collections. The use of multiple terminologies and ad hoc modifications to standard schemes prevents users from cross searching multiple repositories, cross-sectoral resources and interdisciplinary material. In order to overcome this, improved compatibility between schemes is required. This paper considers potential solutions to the terminology problem, with a particular focus on the mapping approach. Key aspects of the mapping technique are discussed with reference to practical applications and initiatives.

Introduction: Terminology problem

Achieving semantic interoperability in the digital information environment is severely impeded by the adoption of different terminology sets and subject schemes within online services and collections. Variation in the way they are applied serves to compound the issue further. The result is that users are unable to cross search multiple sources, cross-disciplinary and cross-sectoral material simultaneously. In order to make life easier for users, therefore, and increase their ability to retrieve a greater proportion of information relevant to their needs with no additional effort, it is essential to encourage compatibility between terminologies.

Significance of a solution

To prevent the problems caused by the use of disparate terminologies escalating, it is essential to identify an effective solution. If the issue is neglected, resources will continue to be classified in non-standard ways increasing the extent of the problem. The longer the problem continues, the more expensive it will be to resolve. Legacy metadata will thrive and the cost of modifying this to fit an agreed solution will increase in direct proportion. A rapid and effective solution is therefore highly desirable.

Proposed solutions

Over the last two decades, different approaches have been proposed to achieve subject interoperability and to provide more consistent access to information. The Open Archives Forum (2002) breakout session on subject interoperability suggested automatic, semi-automatic classification, crosswalks and mapping as potential solutions to the terminologies problem. In addition, cross-browsing, schema mapping, and coding vocabularies in an easily processable and machine-readable format such as RDF and XML have been suggested. Chan and Zeng (2002) identified a number of methods for achieving and improving interoperability. These include 1) derivation/modeling attained

by developing a specialized or simpler vocabulary with an existing, more comprehensive vocabulary as a starting point or model; 2) translation/adaptation whereby a controlled vocabulary is developed which consists of terms translated from one in a different language with or without modification; 3) mapping (intellectual) between equivalent terms in different controlled vocabularies or between verbal terms and classification numbers; 4) a mapping system partly or heavily reliant on computer technology; 5) linking – a list of terms linked with other terms that are not conceptual equivalents but are closely related linguistically; and 6) switching which makes use of an intermediary language or scheme to move among equivalent terms in different vocabularies.

Mapping research

Although a range of different techniques have been proposed as potential solutions to the problem, the mapping approach has received a considerable amount of attention in the research arenas of various subject disciplines. Doerr (2001) defines mapping as "the process of identifying terms, concepts and hierarchical relationships that are approximately equivalent".

The mapping approach proposes to combat the problem by imposing links between equivalent terms in different terminology sets. Users would then search for a subject term that would retrieve resources catalogued using a 'core' or 'central' scheme, in addition to retrieving resources catalogued or classified using mapped or associated terms. Thus, the success of the retrieval process is greatly increased both in terms of precision and recall. Users do not have to consider alternative terms on which to base a search; such terms would either be given to them or searched for automatically.

The linking of terminologies in this way seems logical, however even if the mapping approach was widely adopted as a solution to the terminologies problem, a considerable number of sub-issues remain to be addressed including those relating to the type and extent of mappings implemented. For instance, during research into mapping Laborline thesaurus terms to LCSH (Library of Congress Subject Headings) (Chaplan, 1995), a total of nineteen different match types were identified. Thus, it is a complex process in itself and requires a great deal of intellectual effort.

Highlighting the complexity of the approach, Chamis (1991) suggested three levels of compatibility which provide insight into the ways in which mapping can be carried out:

- consistency in spelling variants, singular and plural forms, verb tenses, and other grammatical variations (controlled word forms)
- equivalent and synonymous terms, cross-referenced to the preferred term, acronyms and antonyms, homographs and metaphors, multi-word and pre-coordinated terms (subject heading lists)
- semantic and generic relationships i.e. Broader Term, Narrower Term, Related Term

Complexities aside, if subject schemes were effectively mapped together it would mean that users could cross search multiple and interdisciplinary sources simultaneously. The level of effort users would expend on search activities would therefore be greatly reduced. For example, mapping initiatives in the medical field have greatly improved retrieval effectiveness for users. Medline and EMBASE databases provide links between free-text terms and MeSH (Medical Subject Headings) and EMTREE (EMBASE Thesaurus) terms. The value of this approach emerges when the "terms entered directly into MeSH do not retrieve relevant hits" (Levy, 2004), as illustrated by searching for 'lung cancer', for example. Although this is a non-MeSH term, the query is mapped to the standard term 'lung neoplasms', thus retrieving hits. So even when a user searches for a term not held within the standard medical terminology in use, the system is able to offer an equivalent term as a result of existing mappings.

Mapping issues

Clearly, the primary advantage of mapping is that users' retrieval effectiveness is enhanced as a result of links imposed between terms of different subject schemes, enabling resources from multiple digital repositories and using different schemes to be retrieved simultaneously using a single search string. It seems there is strong support for mapping as a potential solution to the terminology problem and it is widely recognised that mapping does improve retrieval (CARMEN, 2000; Smith, 2004; Wilkie, 2003).

Another key benefit of mapping has been highlighted by the Renardus (2002) and MACS (Multilingual Access to Subjects) projects (Infolab, 2000). Both initiatives have demonstrated how mapping can be used to combat information retrieval barriers caused by the use of multilingual schemes. Such research illustrates the scalability of the approach, suggesting that mapping could be a universally acceptable solution. One limitation of the Renardus approach, with regard to achieving total interoperability, however, is that following identification of areas of interest within the DDC (Dewey Decimal Classification) hierarchy, users are taken beyond the Renardus interface into the relevant part of an individual service/collection's terminology. This means users still require to access a number of different sources before finding associated terms.

On the downside, however, a crucial point to consider is the labour intensiveness of the mapping work itself. Even if mappings are implemented on an automated or semi automated basis, it may be necessary for associations to be validated manually. In subject areas where automation methods are not practical due to highly specific terminology or varying levels of granularity between schemes, it is likely that mappings would have to be carried out completely on a manual basis. This is a subjective process and so potentially problematic to control. Even if guidelines were implemented to improve consistency there would remain an element of subjectivity impossible to standardise.

Koch (2001) confirms the complexity of the approach, in the DESIRE Project handbook, pointing out that mapping can be a very lengthy and complex process as it involves "theoretical, conceptual, cultural, [and] practical" differences between the controlled

vocabularies. The Aquarelle Terminology Service (Doerr and Fundulaki, 1998) reflects the arduous nature of such a service explaining that "a Term Server must be fed with equivalence expressions between the meaning of terms in different authorities, either by an expert team or by linguistic methods and subsequent human control". Thus, the procedure should be undertaken by experienced professionals. This means that in addition to proving time consuming, the mapping technique is also an expensive one.

Much research has been conducted which stresses the difficulties associated with the structure of terminologies. Neville (1970) and Milli and Rada (1988) identified a number of problems when mapping between two thesauri, some of which are due to the level of specificity and exhaustivity of thesauri and the problem of mapping terms of different hierarchical status. In addition, Whitehead (1990) has reported a number of problems while mapping AAT (Art and Architecture Thesaurus) to LCSH relating to the complex structure of LCSH subject headings, confusion caused by subdivisions, the issue of pre-coordination, matching LCSH compound headings to terms from different AAT facets, and different approaches in controlling synonyms and homonyms. Chaplan (1995) encountered a number of difficulties when mapping Laborline thesaurus to LCSH, mainly due to the large number of homographs in LCSH, unevenness in levels of coverage and its jargon, for instance ‘management by exception’.

An additional problem with the mapping approach arises from the need to amend existing mappings due to scheme updates and local variations. It would be necessary to modify existing mappings when an updated version of a scheme is issued. A procedure for doing so seamlessly is crucial to the success of the mapping approach. In the case of the HILT II project (2003), a key deliverable was to build a pilot terminologies server holding the complete set of LCSH, and selected areas of UNESCO and MeSH thesauri, mapped to a central DDC spine within a centralised system. This was to be implemented within the JISC IE (Joint Information Systems Committee Information Environment) (2003) with the aim of improving cross searching and browsing among JISC collections and services. Project findings and recommendations were made for the design of a full scale terminologies server. A number of features were recommended including a facility for individual cataloguers and indexers to add their own mappings. This is an essential element of such a system as it would not be practical for a central body or agency to implement all necessary mappings, maintain them and make amendments. However, this set up could lead to inconsistencies and lack of standardisation.

The difference between user and standard terminologies creates a further challenge. Any successful implementation of mapping would need to account for misspellings and other idiosyncrasies common in user searches. In the medical field, McCray et al (1999) have noted that terms typically entered by users do not tend to match standard medical terminologies. This confirms that extensive mapping work, again of an intellectually demanding nature, would have to be undertaken to ensure an effective system could retrieve appropriate resources even where non-standard terms are entered by users. A mechanism is required whereby user terms are efficiently matched to those held in standard subject schemes.

Conclusion

The mapping approach, or indeed any of the proposed solutions hoping to solve the terminology issue, will undoubtedly be costly. Substantial investment, both in terms of finance and time, will be required to tackle the problem effectively. The HILT II project conducted a cost benefit analysis on the development of a terminologies server, confirming that significant financial commitment is required, and that the development of such a system will be a gradual process, probably taking place over several years.

Even so, if the issue is not tackled in the very near future, costs are likely to spiral as legacy metadata builds up and yet more schemes are introduced. It seems crucial therefore that significant investment is made to improve the situation, irrespective of whether or not mapping is decided upon as the most attractive way forward. It is important that agreement is reached on the best approach to take, however, because if individual groups begin to implement alternative solutions, the situation will not improve and subject interoperability may become yet more difficult to attain. Perhaps the time has come to take stock, on an international basis, of research conducted to date concerning the challenge of subject interoperability.

References

CARMEN, 2000, http://www.mathematik.uni-osnabrueck.de/projects/carmen/index.en.shtml [Accessed 15 April, 2004]

Chamis, Alice Yanosko, 1991, *Vocabulary Control and Search Strategies in Online Searching*. Greenwood Press: Westport, Connecticut. pp. 14-15.

Chan, L. and Zeng, M., 2002, *Ensuring interoperability among subject vocabularies and knowledge organisation schemes: a methodological analysis*. 68th IFLA Council and General Conference, August 18-24, Glasgow, Scotland, UK. Available from: http://www.ifla.org/IV/ifla68/papers/008-122e.pdf [Accessed 14 April, 2004]

Chaplan, M. A., January 1995, *Mapping Laborline thesaurus terms to Library of Congress subject headings: Implications for vocabulary switching*. Library Quarterly, Volume 65, Number 1, pp. 39-61.

Doerr, M., 2001, *Semantic Problems of Thesaurus Mapping,* Journal of Digital Information, Volume 1, Number 8. Available from: http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/ [Accessed 14 April, 2004]

Doerr, M. and Fundulaki, I., 1998, *The Aquarelle Terminology Service*, ERCIM News Online Edition, Number 33. Available from: http://www.ercim.org/publication/Ercim_News/enw33/doerr2.html [Accessed 14 April, 2004]

HILT, 2003, HILT Phase II Final Report, http://hilt.cdlr.strath.ac.uk/hilt2web/finalreport.htm [Accessed 14 April, 2004]

Infolab, 2000, *MACS: Multilingual Access to Subjects*, http://infolab.kub.nl/prj/macs/ [Accessed 14 April 2004]

JISC, 2003, http://www.jisc.ac.uk/ [Accessed 14 April 2004]

Koch, T., 2001, *Desire Project Handbook: 2,5 Subject classification, browsing and searching*. Available from: http://www.desire.org/handbook/2-5.html [Accessed 15 April, 2004]

Levy, R., Jan/Feb 2004, *Thesaurus Mapping in Medline and Embase*, Chronolog, p 8.

McCray, Alexa, T., Loane, Russell, F., Browne, Allen, C. and Bangalore, Anantha, K., 1999, *Terminology Issues in User Access to Web-based Medical Information*. Available from: http://www.amia.org/pubs/symposia/D005626.PDF [Accessed 14 April 2004]

Milli, Hafedh and Rada, Roy., March 1998, *Merging thesauri: principles and evaluation*. IEEE Transactions on Pattern Analysis and Machine Intelligence, Volume 10, Number 2.

Neville, H. H., 1970, *Feasibility study of a scheme for reconciling thesauri covering a common subject*. Journal of Documentation, Volume 26, Number 4. pp. 313-336.

Open Archives Forum, 2002, http://www.oaforum.org/otherfiles/lisb_bs_SubjInterop.ppt [Accessed 14 April 2004]

Renardus, 2002, http://www.renardus.org/ [Accessed 14 April 2004]

Smith, A., M., 2004, *An examination of PubMed's ability to disambiguate subject queries and journal title queries*, Journal of the Medical Library Association, Volume 92, Number 1, pp. 97-100. Available from: http://www.pubmedcentral.gov/articlerender.fcgi?artid=314110 [Accessed 14 April 2004]

Whitehead, Cathleen., 1990, *Mapping LCSH into Thesauri: The AAT Model*, In Beyond the Book: Extending MARC for Subject Access. Ed. by Toni Petersen & Pat Molholt. Boston, Massachusetts: G.K. Hall, pp. 81-96.

Wilkie, F., 2003, *Thesaurus Mapping, Lincolnshire NHS Library and Information Services*. Available from: http://www.hello.nhs.uk/training/thesaurus_map.pdf [Accessed 14 April 2004]