Digital directions

Multiple terminologies: an obstacle to information retrieval

Emma McCulloch

The author

Emma McCulloch is a Researcher at the Centre for Digital Library Research, University of Strathclyde, Glasgow, UK.

Keywords

Classification, Information retrieval, Digital libraries, Open systems

Abstract

An issue currently at the forefront of digital library research is the prevalence of disparate terminologies and the associated limitations imposed on user searching. It is thought that semantic interoperability is achievable by improving the compatibility between terminologies and classification schemes, enabling users to search multiple resources simultaneously and improve retrieval effectiveness through the use of associated terms drawn from several schemes. This column considers the terminology issue before outlining various proposed methods of tackling it, with a particular focus on terminology mapping.

Electronic access

The Emerald Research Register for this journal is available at
www.emeraldinsight.com/researchregister

The current issue and full text archive of this journal is available at
www.emeraldinsight.com/0024-2535.htm

Introduction

As it becomes increasingly difficult for users to satisfy their information needs due to the rapid expansion of the Web and its sprawling nature, it is also becoming progressively impractical for users to consult a wide range of sources to satisfy an information query. Consequently, it is of growing importance that users are able to search multiple online sources simultaneously. With such a wide variety of resources available, however, the feasibility of achieving interoperability between them is gradually diminishing. Not only do services employ different technical standards, indexing practices, search facilities and algorithms, but also the basic language on which retrieval systems are founded differs widely. It is no longer sufficient for users to make decisions on whether to use keyword or phrase searching, employ Boolean operators, or try their luck with truncation, they must also now give consideration to the terminology they use.

Terminology problem

The majority of online academic sources employ terminologies and/or classification schemes to assist with the organisation of material and its subsequent retrieval. It follows that user terminology must match that employed within a particular service in order to retrieve a complete and relevant set of results. Yet there are so many terminology sets in use that monitoring them has become inconceivable, let alone gaining an understanding of which are used in different services or collections and how they are applied. Hammond (2001) claims "it takes an expert searcher a year to become familiar with a new vocabulary and its use". If this is true for an "expert searcher", what chance does the average user have? Illustrating the extent of the problem, some services use standard schemes such as Dewey Decimal Classification (DDC) and Library of Congress Subject Headings (LCSH) while others use subject specific schemes like Medical Subject Headings (MeSH), Art and Architecture Thesaurus (AAT) and Social History and Industrial Classification (SHIC). It is also fairly common for indexing staff to modify these schemes to cater for local collections or broader/narrower subject areas than those covered within the standard versions (HILT, 2003).

Library Review
Volume 53 · Number 6 · 2004 · pp. 297-300
 · ISSN 0024-2535
DOI 10.1108/00242530410544376

Received: 23 February 2004
Reviewed: 27 February 2004
Revised: 1 March 2004
Accepted: 5 March 2004

Library Review
Volume 53 · Number 6 · 2004 · 297-300

Compounding the issue are those services that use in-house or home-grown schemes unique to that one service alone and the common practice of using uncontrolled keywords assigned by authors and content providers. This enormous variety of schemes has resulted in disparate terminologies being implemented throughout various sections of the online community, resulting in marked differences across sectors and subject areas, rendering any cross-sectoral or multi-disciplinary searching an arduous undertaking for the end-user.

It seems logical then that schemes be linked in some way to achieve semantic interoperability, enabling users to retrieve information effectively, particularly if it is not held within a single repository. Yet, the continuous expansion of this problem, due to the ongoing creation of new terminologies, means that the application of any one potential solution is becoming increasingly problematic.

Mapping approach

A considerable amount of research has been conducted into term mapping with the aim of promoting interoperability between terminologies. Doerr (2001) defines mapping as, "the process of identifying terms, concepts and hierarchical relationships that are approximately equivalent".

The HILT project (HILT, 2000-2003) studied the terminology issue, adopting a mapping approach to attain compatibility between schemes used within the Joint Information Systems Committee Information Environment, with the aim of improving cross-sectoral searching and browsing (JISC, 2003). A pilot "terminologies server" was developed with a large proportion of LCSH, and selected areas of UNESCO and MeSH, mapped to a central DDC spine within a centralised system. When a user enters a query, the meaning of their term(s) is disambiguated through an interactive process before being matched to DDC headings. DDC numbers associated with these headings are then continuously truncated until a corresponding DDC number is found in the metadata of collections held within a local database. Relevant collections are then returned to the user along with mapped terms from other schemes, which can be used to enhance retrieval. The project listed recommendations for the future design of a fully comprehensive terminologies server and has highlighted problematic areas, such as increasing compatibility between user terms and existing subject metadata, catering for the specificity of user queries and incorporating local variations to schemes.

A similar approach was adopted within the Aquarelle Terminology Service. Aquarelle chose to implement a system whereby terms are held and thesauri are managed locally, with a central term server(s) handling the retrieval (Doerr and Fundulaki, 1998). The technique was considered fairly labour intensive since "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" (Doerr and Fundulaki, 1998). This was also found within the HILT project which noted that, once established, an effective system requires a facility for practitioners to add their own mappings (HILT, 2003), which, in turn, raises questions of how to encourage this, and how to ensure mappings are applied accurately and consistently.

Renardus, an EU-funded project (Renardus, 2002) implemented mappings between terminologies used by Resource Discovery Network hubs (RDN, 2003) and DDC to provide a centralised browse interface to subject gateways. Like the Multilingual Access to Subjects (MACS) project (Infolab, 2000), Renardus demonstrates that mapping can be used to tackle retrieval problems caused by language barriers by imposing links between multilingual schemes. This demonstrates 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 hierarchy, users are directed to the relevant part of an individual subject hub's terminology. This means users still require to access a number of different sources before finding associated terms due to the distributed nature of schemes in use.

Considering specific subject disciplines employing multiple terminologies, initiatives involving medical term mapping illustrate the technique's viability. Medline and Embase provide links between free-text terms and MeSH and EMTREE headings. One advantage 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. However, it has been reported that user terminology, "while medical, is often not found directly in medical terminologies" (McCray *et al.*, 1999). It follows that extensive

Library Review
Volume 53 · Number 6 · 2004 · 297-300

mapping of individual user terms would have to be undertaken before such a system could be truly effective in the wider community or commercial sector. It is also likely that some sort of disambiguation or contextualisation phase, as proposed by the HILT project (HILT, 2003), will be required to clarify the exact user requirement. McCray *et al.* (1999) reported that they will continue investigating this area and "intend to explore the development of a terminology server whose goal it is to mediate between user terminology and terminology as it is reflected in a variety of medical information resources". This is likely to be an onerous task as misspellings, and other idiosyncrasies evident in user terminology, will also have to be considered to provide a fully functional systems.

It seems there is strong support for the mapping approach as a solution to the terminology problem and it is widely recognised that mapping does improve retrieval (CARMEN, 2000; Saeed and Chaudhury, 2002), although difficulties with the approach's general efficacy remain. The labour intensiveness of the mapping work, the maintenance demands of a terminology server, particularly the implementation of scheme updates and local variations, along with the complex nature of user searching, all serve to complicate the issue and to inflate the cost of an effective solution.

Will mapping prevail?

There is no doubt that research into terminology mapping has significantly contributed to the investigation of semantic interoperability, yet many aspects of the approach remain to be studied and the implementation of a widely accepted solution is not imminent. It remains unclear whether mapping is the way forward, particularly due to the high level of human input required to implement mappings and the costs associated with the development and maintenance of such a system. As such, a number of alternative solutions have been proposed which deserve consideration.

Could Hammond's (2001) proposal to use Smartlogik's technology to develop "SignPost" terms be a more appropriate solution than the "master thesaurus" implied by mapping? She describes how "digests of each and every index term, in every controlled vocabulary", are generated directly from the text of abstracts, removing the human element of identifying equivalences in different terminologies. These SignPosts then search multiple terminologies simultaneously to return associated terms to the user. To date, this technique has been implemented within the medical field; might some degree of human intervention be necessary to verify relationships between terms in other disciplines?

Clustering, a technique adopted by the Cheshire system, returns semantically related resources to the user by creating links between metadata fields of material associated by subject coverage (Cheshire, 1999). It does not, therefore, tackle the problem of improving compatibility between terminologies directly, but rather serves to create connections between catalogue records. A completely scalable and universally adoptable solution surely has to impose connections between related subject areas irrespective of specific content. Larson (1999) has considered the clustering approach in conjunction with automatic categorisation/classification, the latter being an alternative also investigated by HILT.

Finally, while discussing how to improve retrieval effectiveness in general terms, Koch (2000) has suggested that service providers hope to tackle the issue by "increasing the size of their databases and offering more powerful searching and ranking features". If this is the case, and no attempt is made to tackle the issue of cross searching or to address the terminology problem, it will surely lead to users becoming increasingly dependent on powerful computers and having to define their searches yet further to draw out the specific information they require – an outcome at odds with the allure of semantic interoperability.

Way forward?

It has already been established within the UK LIS community that doing nothing in respect of the terminology issue is not an option (HILT, 2000-2003). It is completely unfeasible, therefore, to continue relying on users to access multiple sources, gain an understanding of different terminologies, or to assume that extensive computing power will solve the problem. Albeit essential, given the investment an effective solution to the terminology issue demands, both in terms of time and money, a great deal of further research is required into potential ways forward. Will the mapping approach be the path to follow? Will HILT's proposal of a centralised term server prove effective, considering the high degree of maintenance and cooperation required? Is Renardus's methodology a more feasible option as the user is taken outside the centralised system to individual subject gateways, even although this requires extensive navigation? What about proposed solutions that abandon the idea of mapping altogether?

Library Review
Volume 53 · Number 6 · 2004 · 297-300

This problem has wide reaching implications and is not one that will be overcome with ease. Agreement between information providers, database creators, academics, practitioners and users alike will be required to ensure that all parties support and conform to the "way forward". The issue is so crucial to cross-sectoral and multidisciplinary retrieval, that international standards must be implemented where multiple terminologies are in use, or if resources are to be accessible through multiple gateways. There have already been steps in this direction within the medical field with the ALCTS/CCS/SAC/ Subcommittee on Metadata and Subject Analysis (1999), stating that "the problem is widely recognized as one which must be solved before the situation becomes intolerable", and that wherever cataloguing activity is undertaken "the development and refinement of methods for harmonization of subject terms from different controlled vocabularies should be undertaken". Is it not about time other disciplines followed suit and addressed the problem directly rather than continuing to create new terminologies, increasing the existing disparity and making an already difficult problem almost impossible to solve?

References

ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis (1999), "Subject data in the metadata record recommendations and rationale", available at: www.govst.edu/users/gddcasey/sac/MetadataReport.html (accessed 2 March 2004).

CARMEN (2000), "Content analysis, retrieval and metadata: effective networking", available at: www.mathematik.uni-osnabrueck.de/projects/carmen/index.en.shtml (accessed 2 March 2004).

Cheshire (1999), "Cross domain resource discovery project) final report", available at: http://cheshire.lib.berkeley.edu/FINALDLI.pdf (accessed 2 March 2004).

Doerr, M. (2004), "Semantic problems of thesaurus mapping", *Journal of Digital Information*, Vol. 1 No. 8, available at: http://jodi.ecs.soton.ac.uk/Articles/v01/i08/Doerr/ (accessed 2 March).

Doerr, M. and Fundulaki, I. (1998), "The Aquarelle Terminology Service", *ERCIM News*, No. 33, online edition, available at: www.ercim.org/publication/Ercim_News/enw33/doerr2.html (accessed 2 March 2004).

Hammond, R. (2001), "Negotiating the medical maze", *The Library Association Record*, Vol. 103 No. 4, pp. 218-20.

HILT (2000-2003), "HILT project", available at: http://hilt.cdlr.strath.ac.uk/ (accessed 2 March 2004).

HILT (2003), "HILT final report", available at: http://hilt.cdlr.strath.ac.uk/hilt2web/finalreport.htm (accessed 2 March 2004).

Infolab (2000), "MACS: multilingual access to subjects", available at: http://infolab.kub.nl/prj/macs/ (accessed 2 March 2004).

JISC (2003), available at: www.jisc.ac.uk/ (accessed 2 March 2004).

Koch, T. (2000), "JoDI noticeboard: networked knowledge organization systems workshop", available at: http://jodi.ecs.soton.ac.uk/noticeboard/nkos.html (accessed 2 March 2004).

Larson, R. (1999), "Cheshire II and automatic categorization", available at: www.sims.berkeley.edu/academics/courses/is245/f98/Lecture21/ (accessed 2 March 2004).

Levy, R. (2004), "Thesaurus mapping in Medline and Embase", *Chronolog*, January/February, p. 8.

McCray, A.T., Loane, R.F., Browne, A.C. and Bangalore, A.K. (1999), "Terminology issues in user access to Web-based medical information", available at: www.amia.org/pubs/symposia/D005626.PDF (accessed 2 March 2004).

RDN (2003), available at: http://rdn.ac.uk/ (accessed 2 March 2004).

Renardus (2002), available at: www.renardus.org/ (accessed 2 March 2004).

Saeed, H. and Chaudhury, A.S. (2002), "Using Dewey decimal classification scheme (DDC) for building taxonomies for knowledge organisation", *Journal of Documentation*, Vol. 58 No. 5, pp. 575-83.

Further reading

Smith, A.M. (2004), "An examination of PubMed's ability to disambiguate subject queries and journal title queries", *Journal of the Medical Library Association*, Vol. 92 No. 1, pp. 97-100, available at: www.pubmedcentral.gov/articlerender.fcgi?artid = 314110 (accessed 2 March).