Revision of British Standards BS5723 and BS6723 for brief progress report

Alan Gilchrist
Cura Consortium and TFPL Ltd.

Resumen

Las normas actuales de tesauros monolingües y multilingües llevan mucho tiempo necesitando una actualización. Esto se aplica tanto a los estándares internacionales ISO 2788 e ISO 5964 como a las respectivas normas nacionales que existen en varios países y a la norma norteamericana ANSI/NISO Z39.19. En el Reino Unido y en los Estados Unidos de América se está trabajando en revisar y extender los estándares, con espacial énfasis en la necesidad de interoperabilidad que impone en nuestro mundo actual la vasta red de comunicaciones electrónicas. En el Reino Unido se ha comenzado a trabajar con la Norma Británica, con la esperanza de liderar un estándar internacional que sirva a todos. Algunos de los temas que se están discutiendo todavía son el tratamiento del análisis de facetas, el estudio de otros tipos de vocabularios controlados como los esquemas de clasificación, las taxonomías y las ontologías, y el mapeo entre vocabularios.

Palabras clave: Tesauros. Vocabularios controlados. Formatos de intercambio. Interoperabilidad. Normas. Reino Unido.

Abstract

The current standards for monolingual and multilingual thesauri are long overdue for an update. This applies to the international standards ISO 2788 and ISO 5964, as well as the corresponding national standards in several countries and the American standard ANSI/NISO Z39.19. Work is now under way in the UK and in the USA to revise and extend the standards, with particular emphasis on interoperability needs in our world of vast electronic networks. Work in the UK is starting with the British standards, in the hope of leading on to one international standard to serve all. Some of the issues still under discussion include the treatment of facet analysis, coverage of additional types of controlled vocabulary such as classification schemes, taxonomies and ontologies, and mapping from one vocabulary to another.

Keywords: Thesauri. Controlled vocabularies. Interchange formats. Interoperability. Standards. United Kingdom.

Scire. 12 : 1 (en.-jun. 2006) 99-107. ISSN 1135-3761.

1. Are thesaurus standards still needed?

Since the 1960s, even before the renowned Cranfield experiments of 1967 (Cleverdon, 1967; Cleverdon, Mills and Keen, 1966) arguments have raged over the usefulness or otherwise of controlled vocabularies. The case has never been proved definitively one way or the other. At the same time, a recognition has become widespread that no one search method can answer all retrieval requirements. Indeed, a recent Forrester report (Brown, 2006) is advocating "Intelligent Content Services (ICS)", which turn out to be a collection of linguistic analysis, fact extraction, automated categorization, and taxonomies. In today's environment of very large networks of resources, the skilled information professional uses a range of techniques. Among these, controlled vocabularies are valued alongside others, and self-evidently, language still underlies retrieval.

The first international standard for monolingual thesauri was issued in 1974. At that time, their main use was in the support of postcoordinate indexing and retrieval systems applied to document collections or bibliographic databases. For many information professionals the only practicable alternative to a thesaurus was a classification scheme. And so the thesaurus, with normally its greater degree of specificity developed a strong following. After computer systems with full text search capability became widely available, however, the arguments against controlled vocabularies gained more followers. The cost of building and maintaining a thesaurus or a classification scheme was regarded as a strong disincentive.

Today's databases are typically immense compared with those of three decades ago. Full text searching is taken for granted, not just in discrete databases but across all the resources in an intranet or even the Internet. But intranets have brought particular frustration as users discover that, despite all the computer power, they cannot find items which they know to be present on the network. So the trend against controlled vocabularies is now being reversed, and many information professionals are turning to them for help. Standards to guide their compilation and use are still in demand.

A further incentive is the widespread expectation of a Semantic Web, enabled by "intelligent" software agents that can "understand" the metadata applied to networked resources. To facilitate this development, computers need to be able to interpret the terms in ontologies and other types of controlled vocabulary. Systems that can map from one vocabulary to another will be needed. The thesaurus standards, if updated to meet these new needs, may become more widely used than ever before.

2. Currently available standards for thesauri

- *ISO 2788-1986. Guide to establishment and development of monolingual thesauri* (which was based on the equivalent British Standard BS5723).

- *ISO 5964-1985. Guide to establishment and development of multilingual thesauri* (which was based on the equivalent British Standard BS6723).
- *ANSI/NISO Z39.19-1993. Guidelines for the construction, format, and management of monolingual thesauri.*

Although there are differences in style and emphasis, the three standards are broadly compatible with each other. The international standards have been adopted as national standards in several countries, including France, Germany, Canada and UK, among others.

The most recent editions of both of the international standards were written well before the era of the personal computer. Although it is seven years younger, the American standard too is very much oriented to the world of print. Plainly it is time that the worldwide community of thesaurus users had more up-to-date guidance. Work is now under way, on both sides of the Atlantic, to revise the standards and bring them into the 21st century.

3. Which aspects of the standards need revision and/or extension?

The principles of thesaurus use and construction have not changed fundamentally since they were established, even before the first national or international standard, in the Rules and Conventions of the Thesaurus of Engineering and Scientific Terms [4]. But the context around them has changed and this needs to be reflected. As a simple example, guidance on thesaurus displays should now provide for display on screen as well as on printed pages.

The current international standards provide no guidance at all on software to support the task of thesaurus construction. Good software products provide many functions to make the editorial work more efficient, and avoid errors that infringe the principles of thesaurus construction. The availability of a specification of minimum functionality should encourage the development of reliable products.

The current standards cater for vocabularies used by information professionals trained in the arts of indexing and searching. Today, however, few users have the luxury of a librarian to help them. Tools are needed that will support untrained end-users, many of whom have little patience with the perceived complexity of a controlled vocabulary.

Perhaps the greatest deficiency in the current standards concerns interoperability, that is to say the ability of one system to work in harmony with others. Instead of applying one thesaurus to one discrete database, users now want to search across a multiplicity of resources, which may have been indexed with different thesauri or classification schemes, or no controlled vocabulary at all. Technology is already providing some very welcome innovative approaches to this challenge, but

much more could be done if the standards were to provide for mapping between vocabularies.

Interoperability also requires standardized formats and protocols to support the exchange of controlled vocabulary data between computers, not just at the stage of searching, but also to support indexing, the construction of the vocabularies, and the sharing of networked services.

In summary, what we need is a new or revised standard to make good all the above deficiencies and in general to support today's needs. Since the information community works across national boundaries, it should be international.

4. What is being done?

An international standard has to be backed by an international committee, with members from all the ISO (International Organization for Standardization) member countries which express an interest. But the procedures for forming such a committee are lengthy, and the communication overhead tends to slow down the work of drafting. Each time the existing standards have come up for review over the last 15 years, the easiest option has simply been to confirm them without amendment. Finally in 2000, members of BSI (British Standards Institution) committee IDT/2/2 decided to pick up the challenge, at least in part. A Working Group was formed to develop a new British Standard, which would cover all the aspects mentioned and supersede the existing BS 5723 and BS 6723. As mentioned above these are identical to ISO 2788 and ISO 5964 respectively. The new British Standard, BS 8723, would then be offered to the international community, in the hope that it would form the basis of an international standard. It is hoped that this indirect procedure may avoid some of the delays.

Meanwhile in the USA, NISO (National Information Standards Organization), APA (American Psychological Association), ASI (American Society of Indexers) and ALCTS (Association for Library Collections and Technical Services) held a workshop in 1999 to investigate the feasibility and desirability of developing a standard for electronic thesauri. The workshop concluded that a new standard should be developed, to supplement rather than replace Z39.19. Its prime concern would be with interoperability, and it should include other types of controlled vocabulary as well as thesauri. A committee was formed and a consultant appointed. The draft version has been circulated for consultation, has been approved, and is now being prepared for publication, under the title *ANSI/NISO Z39.19-200X. Guidelines for the construction, format and management of monolingual controlled vocabularies*.

A third initiative has been the formation of a Working Group in IFLA (International Federation of Library Associations) to consider the guidelines for multilingual thesauri.

Members of these three separate initiatives have maintained informal communication, so that the outcomes may be aligned as closely as possible.

5. Plans and progress for BS 8723

The new standard *Structured vocabularies for information retrieval* will have the following scope and structure:

- Part 1 sets out definitions and other matters common to all information retrieval applications of structured vocabularies.
- Part 2 deals with thesauri, covering all the scope in the existing BS 5723 (= ISO 2788) plus additional guidance on electronic functions of thesauri and thesaurus management software.
- Part 3 covers other types of structured vocabulary (such as classification schemes, search thesauri, subject headings lists, taxonomies and ontologies).
- Part 4 covers interoperability between vocabularies, addressing situations where one thesaurus or classification scheme, etc., has to be mapped to another. Multilingual thesauri are treated as a special case of such mapping. Thus the whole scope of BS 6723 (= ISO 5964) is included, within a much wider frame of reference.
- Part 5 sets out the protocols and formats needed for exchange of vocabulary data.

Of the above, drafts of Parts 1 and 2 are completed, have been issued for public consultation, have been approved and are now in the press. Parts 3 and 4 are currently under development and final drafts are scheduled for completion this year. Part 5 will be written after all the other parts are completed, since the content is necessarily derived from them.

This has been, and remains, an ambitious project (and now that the ANSI/NISO standard has appeared it seems even more so, as the scope of the British Standard is wider). A number of issues have arisen which have implications for continuing work on the standard.

6. Facet analysis

Facet analysis is traditionally considered in the context of classification schemes, and is not an easy technique to explain. It has an important role in thesaurus construction too, and is hardly mentioned in the current standards. One of the challenges in drafting Part 2 of BS 8723, which deals specifically with thesauri, has been to cover facet analysis briefly but sufficiently.

7. Other types of vocabulary

Part 3 raises bigger issues. The intention here is to cover other types of structured vocabulary that are increasingly being applied to information retrieval in one way or another. The application area where the need is most obvious is for the directories of websites on the Internet or intranets, where we can already see hundreds and thousands of website administrators struggling to provide browse access to their resources. The principles of classification are fundamental to their efforts, but large schemes such as the Dewey decimal classification or the Universal Decimal Classification (UDC) are not usually appropriate for the context. The term *taxonomy* is often used for such directories, and they often combine elements of classification with some thesaural features. Other than that, there is little commonality of approach. The challenge for BS 8723 is to identify what features of taxonomies could usefully be "standardized", and provide guidance that will be intelligible to the very mixed community of potential users.

For the other types of vocabulary proposed for Part 3, the question arises of why to include them at all. There are already dozens of good textbooks on classification schemes, for example, and the big schemes such as Dewey and the UDC have become standards in themselves. It would seem presumptuous to write a new standard now, and consensus might be difficult except at the level of broad principles. It is useful, though, to show the complementary nature of thesauri and classification schemes, being alternative ways of arranging and presenting a list of defined concepts and their relationships.

Another reason for including classification schemes is to begin to tackle the interoperability issues. Large collections of resources have already been classified with Dewey, LCC (Library of Congress Classification) or the UDC, and/or indexed with LCSH (Library of Congress Subject Headings), MeSH (Medical Subject Headings), the AAT (Art & Architecture Thesaurus) or AGROVOC, to name but a few. The world of information users wants to be able to access any combination of these resources, with any of the controlled vocabularies. Tools are needed for "cross-walking" or mapping between the vocabularies. Teams of researchers are already busy developing such mapping schemes, often by automated means. To inform their work they could benefit from guidance on how to map terms, codes and concepts. The guidance needs to be based on descriptions of the elements of the vocabularies (concepts, terms, class codes, notation, captions, relationships, etc.), sufficiently clear that a programmer with no training in classification can follow the rules and set up a mapping table that works. It follows that BS 8723 Part 3 needs to describe the vocabulary elements and what they are for. In that way it lays the groundwork for Part 4, which will cover the mapping process proper.

That argument still leaves open the question of how much more to say about the other types of vocabulary, for example how to build and maintain them. Prob-

ably we will not have a firm answer to that question until Part 3 is issued as a draft for comment. On the basis of a well-developed draft, we hope the community of information managers will provide good feedback on what is or is not wanted. Meantime, the safest answer may be to include the minimum.

8. Mapping between vocabularies

For Part 4, the aim is to provide guidance in situations where one vocabulary must interoperate with another. The simplest case is that of a multilingual thesaurus, in which each of the vocabularies represents a different natural language, each preferred term has an equivalent in each language, and all the vocabularies have identical structure. Greater complexity arises where two thesauri have scopes and structures that overlap, but are not identical. In this case, some mappings may be between terms that are not equivalent —perhaps one is broader than the other; perhaps the two have overlapping meanings. The situation becomes even harder to manage when a mix of thesauri, classification schemes, taxonomies and ontologies are to interoperate, and where a single term in one scheme may have to be represented by a combination of terms in another. This is not just a hypothetical situation —it is the real challenge that will have to be overcome to make the vaunted Semantic Web a reality.

Guidance on mappings is not straightforward to write. Firstly, although the developers of knowledge structures already have a good intuitive feel for when and where it is a good idea to use mappings, there is little written material available setting out the principles. Secondly, conventions have not yet been established for how to represent the cross-vocabulary mappings, distinguishing them from the relationships internal to one vocabulary. Thirdly there are variables arising from the context of where the mappings will be applied —at the point of indexing, or of searching, or perhaps some other operation. And fourthly, whether or not the vocabularies are structurally equivalent affects the advisability of mapping. All these factors add to the difficulty of articulating clear, concise guidance. Again, the first edition of this part of the standard may have to be limited to the most essential points, pending feedback from the community of users.

9. Conclusion

Our expectation in the UK is that our project will be just the start of a wider effort. It is important to compare our results with those of NISO. Feedback from the information community will be vital, to shape further development of the standard. In view of the eventual need for an international standard, the Working Group has already been inviting informal inputs as widely as possible. Awareness of the concerns of other language communities, cultures and perspective needs to be built in at an early stage.

It should be stressed here that it is only by historical accident that these publications appear as standards; and, as such, they differ markedly from normal standards. And while they are guidelines rather than standards, nor are they textbooks. In essence they are attempts to discover best practice through a process of consensus building, made far harder by the existence of different communities of information professionals working in this area.

Despite the unfriendly ring of the word "interoperability", the hope is that a better understanding of the principles and practice of building linked vocabularies will assist the development of information-sharing applications. The knowledge organization community already has the experience to guide such projects. Let us try to capture some of the expertise in a standard so it can be made more widely available.

10. Postscript

This year 2006 has seen slow but effective progress on the preparation of BS8723. It has to be remembered that the four members of the Working Group are working on this revision in addition to their every day jobs, and work on the standard is conducted largely by email (with multi-coloured tracking of the manuscripts!) and with whole day meetings every three months.

Part 1: Definitions, symbols and abbreviations and *Part 2: Thesauri*, which were in the press at the time of writing of the paper presented at Ibersid 2005, are now formally published. *Part 3*: *Vocabularies other than thesauri* and *Part 4: Interoperability between vocabularies* are very nearly complete, but have not yet been sent to the British Standards Institution (BSI) for two reasons. The first is that there is still some internal discussion on how best to deal with "authority files" in Part 3. Where such types of vocabulary exist, they are usually less complicated than subject authority files, but their proper and adequate treatment in the Standard has been surprisingly difficult. The second reason for the delay is much more significant and involves a complex and continuing debate with a panel of experts on the difficult and evolving issues surrounding the purpose and content of Part 5, which is attempting to deal with the protocols and formats needed for the transfer of vocabulary data between computer systems. It would be premature to give an account of the debate so far, which has generated a huge and continuing number of email correspondence between the members of the Working Group and the panel of experts. Suffice it to say that the attempt is being made to take account of ongoing development work in the W3C community, as well as coalface applications of such schemes as the different versions of MARC, ADL (Alexandria Digital Library protocol), ZThes, etc., and the use of XML in a variety of versions for their coding. These discussions may have some minor implications for Parts 3 and 4; but it is the plan, depending on a satisfactory outcome of the discussions on Part 5, to issue the three remaining parts to BSI for official publication as consultation documents.

Meanwhile approaches have been made to a number of countries regarding the possibility of BS8723 being used as the basis for an international standard. The rules issued by the International Standards Organisation (ISO) in Geneva state that support from at least five countries supporting the proposal and having members on the relevant ISO Committee. So far, positive replies have been received from Spain, Denmark and France and approaches have been made to Canada and Germany.

11. Acknowledgements

This paper is an edited and updated version of an earlier version presented to a meeting of ISKO in London. I am grateful to my colleagues on the BSI Working Group —Stella Dextre Clarke, Ron Davies and Leonard Will.

References

Brown, Matt; Ramos, Laura (2005). Searching for a better search. Cambridge, Mass.: Forrester Research, 2005.

Cleverdon, C. W. (1967). The Cranfield tests on index language devices. // Aslib Proceedings. 19 (June 1967) 173-192.

Cleverdon, C. W.; Mills, J.; Keen, E. M. (1966). Factors determining the performance of indexing systems. Cranfield: College of Aeronautics, 1966.

Thesaurus of engineering and scientific terms (TEST) (1967). New York: Engineers Joint Council and US Department of Defense, 1967.