Subject indexing in the Web age

Vanda Broughton

*An interview to Vanda Broughton**

by Laura Ballestra**

BLISS Classification has a very long history, starting in 1940 for the first version through the new edition system realized by the BLISS Classification Association in 1967. Since 1977 you are the editor, with J. Mills, of BC2. What has been the evolution of BC2 during the last 30 years and what are the actual perspectives?

BC2 has changed quite a lot during the period since the publication of the Introduction and Auxiliary Schedules in 1977. Initially there was some tension between the desire to implement the theory which had been developed during the previous twenty years, and the need to meet the expectations of users that the scheme would not change too much. The original plans for the revision envisaged a fairly broad classification, comparable in size to Dewey, in which the major changes would consist of modernising the vocabulary and rationalising the citation order. As it became clearer that few libraries would convert from BC1 to BC2 (most of the BC2 libraries are new users of the classification) the revision has become more radical with more substantial vocabularies and more provision for variations in the citation order. Such terminologies meet better the needs of special collections, and are more suitable for the complex nature of digital information.

Could you tell us something about the BC2 team? How do you work?

The methodology of revision has also changed somewhat over time. Until the 1990s, the original revision fund and subsequent research funding and sponsorship allowed for the employment of a full-time research worker (myself) in addition to Jack Mills. This is not to say that we did all the work. Members of the Bliss Classification Association (BCA), and particularly, the Classification Research Group (CRG) also developed schedules. This is reflected in several classes authored by people other than the main editors. Even those schedules worked on primarily by Jack Mills and myself were commented on by a wide range of interested parties: all were taken to the CRG for discussion, and the draft schedules were circulated to quite a large group of BCA users and others who provided valuable commentary and feedback. This continues to be the pattern of working even now that we're dependent on voluntary contributions. The contribution of the Cambridge College libraries (several of which adopted BC2 in the 1990s) has been particularly valuable and a number of new draft schedules have been developed within the University. In the near future we hope to make the draft schedules available on the new BCA website in the expectation that a wider community will be able to respond to them. To supplement the intellectual work on the classification, the physical production of the schedules is controlled by a suite of computer programs which generate the finished version, and the alphabetical index. Recently some new programs have been commissioned which allow us to produce a compatible thesaurus, and we hope that this will appeal to a wider variety of users, particularly those indexing digital resources.

Do you think BC2 can be translated into other language?

In theory, a classification is the easiest type of controlled vocabulary to translate into another

* Lecturer at School of Library, Archives and Information Studies, University College London.
She's one of the most important experts on faceted classification. Since 1977 she's the editor, with J. Mills, of BC2.
** Librarian at the Università Carlo Cattaneo (LIUC), Castellanza.

An Italian version of this article has been published in "Biblioteche oggi", 26 (2008), 2.

natural language since the concepts are represented by the notation. This means that the class headings or captions need not be particularly concise since they are not used in indexing, and any amount of text defining or explaining the class can be part of the class description. In practice we find that the situation is not quite so simple, for there may not be an exact correspondence between the conceptual structures of subjects in different languages; for example, English has no word which means slugs + snails, as is the case in German and Dutch, so the hierarchy for molluscs is different in these languages (despite their being very closely related). Therefore some of the problems that affect multilingual thesauri also affect classifications, but generally speaking the conceptual nature of the classification avoids other problems associated with indexing languages. So, yes, there seems no reason why BC2 shouldn't be translated, as indeed Dewey and UDC have been, into many different languages.

We could say that BC2 is a fully faceted classification scheme; which kind of problems did you find in defining the structure of so different subject fields?
The standard methodology used in BC2 is, perhaps surprisingly, transferable and quite reliable for a huge range of subjects. It has to be the case for a universal classification scheme that the citation order and general structural principles are applied consistently otherwise too many unpredictable and potentially conflicting situations would be created, so to some extent the specific needs of individual subjects have to be considered as a secondary to the overall structure of the classification. In practice, the number of alternatives in BC2 allows for the demands of different kinds of collections and users. We always start by using the normal range of CRG categories (thing, kind, part, material, property, process, operation, product, by-product, agent, space, and time) and applying standard citation order, but often this needs some modification in individual subjects. Almost invariably this is quite clear from looking at the literature of the subject and taking subject specialist advice. What is unavoidable is the need to become very closely acquainted with the terminology of the subject, and this is one of the most time-consuming aspects of the revision work, especially in technical subjects where the relationships between terms may not be at all evident at first glance. This has been very much the case with Chemistry, and was also so for Mathematics, a particularly abstract and difficult subject. It is also quite often necessary to use some "new" categories, not included in the standard list; form and genre in the creative arts are familiar examples. Whatever the subject, at the heart of facet analysis is a very detailed and precise examination of the terminology, and a very clear specification of the relationships between terms.

How many libraries are adopting BC2?
There are currently about 50 libraries using BLISS, and most of these are BC2 users. New libraries continue to join, but a major potential use of BC2 must be as a more general indexing tool (as opposed to a system for physical organization of resources), which is why we have begun to explore the thesaurus option.

You are an expert of both thesauri and classification scheme. What is the relationship between thesauri and classification?
At one time the thesaurus and the classification scheme would have been regarded as quite different kinds of tool, but the work of Jean Aitchison on the *Thesaurofacet* and subsequent faceted thesauri showed how facet analysis principles could be used to construct both types of vocabulary. In fact, using facet analysis to create a systematic structure is a very helpful method of identifying the various relationships between terms that will be needed in a thesaurus. After *Thesaurofacet* it became quite usual to publish a thesaurus in two parts: a systematic, or classified, structure, and an alphabetical list of terms with thesaural cross-references. Jean's methodology has been followed up in current work on the BC2 thesaurus, where we've been able to use a faceted vocabulary as the basis of the finished classification, the alphabetical index to the classification, and a thesaurus. This is achieved by encoding the source vocabulary in such a way that these three kinds of display can be generated by computer programs almost totally automatically. This shows, I think, that the different displays are just variant ways of presenting the same terminology. What one does find is that labels such as thesaurus, taxonomy, ontology, and so on, are now used interchangeably, and that non-LIS specialists in particular don't see the need to make these distinctions between indexing tools.

In England many websites are adopting thesauri as control vocabularies for indexing web pages. Do you think that thesaurus is a good application for semantic metadata of websites?
I think a thesaurus is a really excellent way to assign semantic metadata, since the structure of the thesaurus support a more sophisticated use of the metadata. For instance, the thesaural cross-references provide an easy way to navigate the collection of resources, and to explore related material by broadening or narrowing the search, or searching on related terms. There is also potential for using the systematic display of the thesaurus as a browsing structure, particularly with digital resources, where hypertext linking is ideally suited to the purpose. The controlled nature of the thesaurus vocabulary should also improve retrieval in a general way, and it can be used behind the interface to formulate or modify user queries without the end-user needing to be aware of its presence. Interestingly, several recent studies of automatic metadata generation have shown that machine extraction of terms, which are then mapped to a controlled vocabulary such as a thesaurus, provides an inexpensive way of assigning good quality semantic metadata without human intervention.

The Biblioteca Nazionale Centrale di Firenze is adopting a faceted thesaurus in developing italian subject headings Nuovo soggettario*. What do you think about the possibility of using a general thesaurus as vocabulary source for subject headings?*
When I teach students to construct a thesaurus, we always finish the exercise by using it to index some of the documents from which we've gathered the terms at the beginning. This is followed up by showing how these index descriptions can be used to make subject headings if a standard citation order is applied to the thesaurus terms or descriptors. This works very well, and although the resulting headings are not quite the same as pre-coordinated headings such as LCSH, they are very regular in structure and, I think, easier for users to understand. The Library of Congress FAST project, which aims to rationalise and simplify the LCSH headings for use in a digital context, uses headings with a similar sort of structure, which suggests that they too are thinking along the same lines.

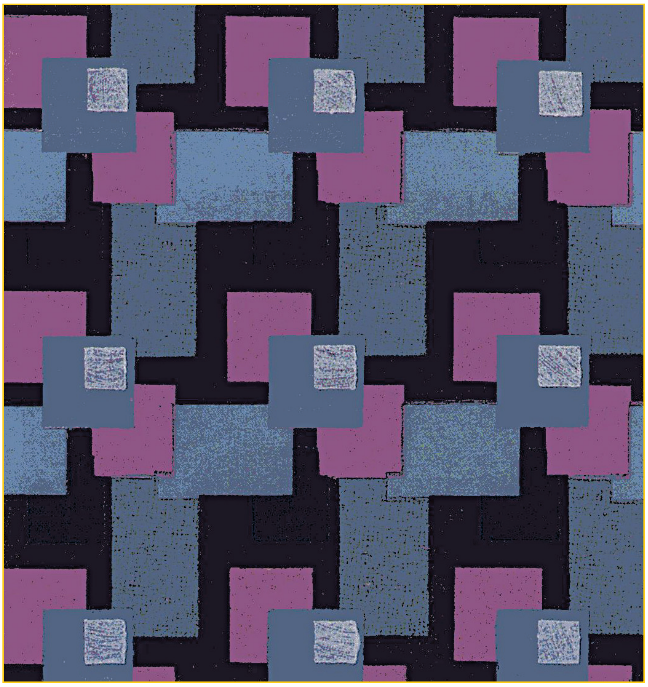
Maintaining a thesaurus or adopting a classification scheme is of course a cost. Do you think that in the Internet actual trend, with web 2.0., taxonomies and ontologies, there is really more attention to the problem of indexing information?
Yes, I think there is currently much greater interest in a properly structured approach to web retrieval, and that many computer scientists are now far more aware of the theoretical issues related to good information architecture, rather than assuming that very rapid machine processing can by itself solve all the problems. I would have said ten years ago that computer specialists had a great deal to learn from LIS specialists, but I think that they already have begun that process, and that there is now much greater awareness of techniques like facet analysis, and of the different ways to structure vocabularies to improve organization and retrieval. There is also, of course, a great deal of research into finding ways of automatically generating thesauri and ontologies and so on, but the results are nearly all much less sophisticated than intellectually built tools, and not nearly so effective for indexing and retrieval.

How do you think librarians can present their knowledge about indexing to computer people for expanding the use of classifications and thesauri in the Web?
As I've said in the previous answer, I believe that computer people *are* now more aware of LIS theory than in the past, but we still must continue to research and speak and write about what we know, so that the body of knowledge is available to all. I think that part of the reason why classification theory was so slowly taken up by information technologists was that there was very little dissemination of the research work, and that it took place within a quite limited group. That is certainly true of the work of the CRG, which is represented only in a very small number of publications, and which took a long time to come to the attention of a wider community. While researchers may now be more familiar with LIS theory, there is still a need for librarians to speak up for themselves and their skills in the workplace, where technical staff may regard them as ignorant about information retrieval. It's essential that we maintain a bridge between intellectual, practical and technical aspects of cataloguing and indexing, and ensure that all these aspects are kept in balance. It's much more likely that young LIS professionals today will have acquired technical skills as part of their professional education, but we must make sure

that we don't try to avoid these more 'difficult' aspects of practice, and allow ourselves to be pushed aside. If we understand information technology better, then we can really engage with the computer scientists and speak to them in their own language.

You saw many technological changes during last 30 years. What is, from an indexing point of view, the main change that you can describe?
I suppose the biggest change that has occurred is the move away from physical collections towards digital and hybrid libraries. This must have affected nearly every area of professional practice, and indexing is no exception. When I first qualified, and went to work as a researcher in the 1970s, there was no Internet, no digital materials, no electronic catalogues, and just a very few online indexing services such as MedLine, the index to medical literature. Computers were just beginning to be introduced to libraries, and as library school students we were taught about what they *might* be able to do, with hardly any real examples of computing in practice. The problems for indexers and classifiers were largely related to the physical organization of collections, and to the way in which complex subject content could be reduced to linear order, not just on the library shelves, but in card catalogues, and in printed bibliographies and periodical indexes. The development of facet analysis provided a very neat way to represent subject information in a logical and predictable manner, and it greatly aided both indexing and retrieval, because, however complicated the subject, the indexer knew the rules for synthesising the classmarks from the constituent elements of the subject, and could predict where in the sequence the document would be located.

But the idea was more far-reaching than this, because the structure it gave to data also proved highly appropriate to a non-linear context, a thing which first became evident with the appearance of bibliographic databases and the first electronic catalogues (just a specific example of a bibliographic database) in the 1980s. It was probably then that people began to think about data structure, simply because it had to be understandable by machines, and therefore more consistent and regular. Nowadays, with the advent of the Internet the problems of linear ordering take second place, because there is no actual "collection" to arrange, although there is still a need to think about presentation of resources and knowledge organization tools continue to be relevant in that context. I suppose that another significant feature is that information no longer necessarily comes in the convenient forms of thirty years ago such as books and journal articles, so that what we're attaching our index labels to can be very much more diffuse, and, of course, it is more susceptible to change because of the ease of editing and updating online resources. As a result, resources tend to be less stable and indeed different versions can disappear altogether in a way that print resources never did. In addition to these changes in the form of information, subjects are more liable to re-interpretation and definition, and new subjects come into being, not only through research and the creation of new knowledge, but also through cross-disciplinary and inter-disciplinary study and the adoption of methodologies from other disciplines. This diffuse and elusive nature of information means that current indexing tools need to be particularly flexible and responsive if they are to be effective.



I think it is very gratifying that indexing theory, particularly facet analytical theory, shows itself still to hold good in the modern information milieu and that it has proved highly adaptable to this much more demanding situation. This is mainly because the emphasis is on the *methodology* of analysis, the identification of roles and categories, and the determination of relationships between

terms and concepts, rather than on the creation of fixed structures of classification as was the case for enumerative classifications. This precise and rigorous analysis is of course essential to machine handling of data, and this is undoubtedly what makes it attractive to non-LIS professionals.
I think it is important to remember that the principles of indexing and of constructing vocabularies remain true, and although we are in a very different environment, the fundamental theory that was developed in the mid twentieth century continues to provide us with a very sound basis for managing information.

Vanda Broughton's bibliography

Essential thesaurus construction, London, Facet, 2006.
Essential classification, London, Facet, 2004.
(with Jack Mills) *BLISS Bibliographic Classification 2nd edition Class W Fine arts*, Munich, Saur, 2006.
A faceted classification as the basis of a faceted terminology, "Axiomathes", 2008 [DOI 10.1007/s10516-007-9027-7]
Henry Evelyn Bliss; the other immortal or a prophet without honour?, "Journal of librarianship and information science", 40 (2008), 1, p. 43-58 [DOI 10.1177/0961000607086620]
(with Aida Slavic) *Building a faceted classification for the humanities: principles and procedures*, "Journal of documentation", 63 (2007), 5, p. 727-754, <http://dlist.sir.arizona.edu/1976/01/Broughton_Slavic_jdoc2007_preprint.pdf>.
Classification and subject organization and retrieval, in Bowman (ed.), *British librarianship and information work 2001–2005*, Aldershot, Ashgate, 2007, p. 467-488.
Classification and subject organization and retrieval, in Bowman (ed.), *British librarianship and information work 1991–2000*, Aldershot, Ashgate, 2006, p. 494-516
The need for a faceted classification as the basis of all methods of information retrieval, "Aslib proceedings", 58 (2006), 1/2, p. 49-72.
Meccano, molecules and knowledge organization: the continuing contribution of S. R. Ranganathan, ISKO UK event *Ranganathan revisited: facets for the future*, University College London, 5 November, 2007, <http://www.iskouk.org/presentations/vanda_broughton.pdf>.
Facet analysis as a fundamental theory for structuring subject organization tools, NKOS workshop of the European Conference on Digital Libraries, Budapest 17-21 September, 2007, <http://www.comp.glam.ac.uk/pages/research/hypermedia/nkos/nkos2007/presentations/BROUGHTON3.PPT>.