Faceted classification in support of diversity; the role of concepts and terms in representing religion

Vanda Broughton
Emeritus Professor of Library & Information Studies
University College London

Abstract:

The paper examines the development of facet analysis as a methodology and the role it plays in building classifications and other knowledge organization tools. The use of categorical analysis in areas other than library and information science is also considered. The suitability of the faceted approach for humanities documentation is explored through a critical description of the FATKS project carried out at UCL. This research focused on building a conceptual model for the subject of religion together with a relational database and search and browse interfaces that would support some degree of automatic classification. The paper concludes with a discussion of the differences between the conceptual model and the vocabulary used to populate it, and how in the case of religion, the choice of terminology can create an apparent bias in the system.

Facet analysis is now widely agreed to be a reliable methodology for building controlled vocabularies in the form of classifications, thesauri and other subject retrieval tools, and Hjorland (2013: 1)has described it as 'probably the dominant theory of the late twentieth century'. Originally conceived by S. R. Ranganathan in the 1930s (Ranganathan, 1967), it was developed in a particularly British tradition by the Classification Research Group (1955; 1969) in the second half of the twentieth century. A faceted approach can be seen in a number of special schemes created by the CRG, in the second edition of Bliss's *Bibliographic classification* (Mills and Broughton, 1977-), and in several large scale thesauri such as *Art & Architecture Thesaurus* (1994) and *Thesaurofacet*. (Aitchison, 1970). There has been a resurgence of interest in faceted classification post 2000 (Broughton, 2006), and it is now seen in many different environments. All the general schemes of library classification presently show a more facet-like structure, the new International Standard ISO 25964 and the British Standard *BS 8723 Structured vocabularies for information retrieval,* on which it is based, acknowledge facet analysis to be a useful means of constructing a vocabulary, with many examples of faceted structures and arrangements, and the Library of Congress FAST Headings tool provides a faceted, or post-coordinate, version of LCSH. The inherent logic of the faceted system makes it hospitable to machine retrieval and some work has been done on representing faceted structures in web ontology languages (Miles and Bechhofer, 2009). There is also a substantial area of application in e-commerce where faceted search interfaces are now so common as to be almost the norm (Adkisson, 2005; La Barre, 2006).

The basis of the faceted approach is the idea that the concepts in any given subject area, or domain, can be analysed and individually assigned to a selection of theoretical categories, which are predominantly functional or linguistic in nature. The process of analysis generates a conceptual model of the domain, and a logical and systematic structure which is then used as a basis for the organization of documents. In Ranganathan's philosophy these categories are referred to as fundamental categories, and consist of *personality*, *energy*, *matter*, *space*, and *time*. Within a particular domain the fundamental categories are translated into *facets* specific to that domain; for example, in zoology, the *personality* category is represented by the *animals* or *organisms* facet, *energy* by the *physiology* or *bodily processe*s facet, *parts* by the *anatomy or organs and systems of the body* facet, and so on. When classifying or indexing documents, concepts are listed or combined in a predetermined and consistent order of their containing facets, known as citation order, or in Ranganathan's case, a facet formula.

Because Ranganathan's categories are relatively few in number his formulae for individual domains often contain what are termed rounds and levels. For example in Class I, Botany, the facet formula is I [P], [P2]: E where P (Personality) represents different genera and species of plants, and P2 (Second level Personality) are the parts and components of plants. This method of using second level personality to deal with the parts of an entity or system can also be found in History [P], [P2]: [E]: [T], and Political science [P], [P2]: [E]. The formulae can become very complex as in the case of Class J Agriculture J [P] [P2] [P3] : [E] [2P] : [2E], where additional formulae are provided for Soil J [P] [P2] [P3]: [1] [2P]: [2E], Propagation [P] [P2] [P3]: [3] [2P]: [2E], and Disease [P] [P2] [P3]: [4] [2P]: [2E].

In the United Kingdom the Classification Research Group elaborated Ranganathan's theory, and identified a larger number of categories, namely: *thing*, *kind*, *part*, *property*, *material*, *process*, *operation*, *agent*, *patient*, *product*, *by-product*, *space*, and *time*. This enabled more nuanced analysis, and the potential to distinguish between, for example, *processes* in plant husbandry (such as fertilisation, growth, disease, and fruiting), and *operations* (such as sowing, weeding, harvesting, and drying).

Categorical analysis beyond library science

The deployment of categories as part of an analytical technique is not restricted to library and information science. It is fairly common in various kinds of content analysis methodologies where text is involved, from about the mid-twentieth century onwards, notable examples being systems theory and grounded theory. The purpose is to build a model of the system or environment under consideration to better understand the various agencies and processes within it, and their interrelationships. The 'text' may be the aggregated terminology of the domain, obtained from vocabulary sources such as glossaries, dictionaries, or indexing languages, or it may be a corpus of actual texts such as articles, reports, narratives, or interview transcripts. Similarly, the nature of the categories is varied; sometimes, as with library and information science, a predetermined set of categories is used; in other situations categories are allowed to 'emerge' from the text as the analysis is carried out. Soft systems theory provides a good example of the former; for example, categories used in the soft systems analysis of business environments are often represented by the formula CATWOE (clients; actors; transformations; weltanschauung = world view; owner; environmental constraints) originally defined by Checkland (1989: 282). Even within information science, and among facet analysts, the choice of categories may be very diverse. Perry and Kent's (1956) *Semantic Code* uses *relationships*, *states*, *processes*, *substances*, and *objects* (although the last three are suggestive of *energy*, *matter*, and *personality*), and Gardin (1965) identifies *status*, *property* and *movement* among others in several systems designed for subjects as different as archaeology, mythology, and the Qur'an. Perhaps the most intriguing parallel example of categorical analysis is in Louis Guttman's (1959) research in social psychology which he calls facet analysis, and claims to have invented, apparently without any reference to the work of Ranganathan or others in the field of information work. Guttman's work is essentially mathematical in nature, applied to problems of statistical analysis, and it has been suggested that the link is through Ranganathan's mathematics background (Beghtol, 1998).

Facet analysis in alphabetical subject cataloguing

While the faceted approach was initially intended as a means of modelling classification systems (because its analytical basis provided a way of handling complex content of documents in a consistent and therefore predictable manner), quite early in its development it was also seen as relevant to subject cataloguing, that is document indexing, as well as document organization.

Although there is a distinction to be made between a *concept* and the *term*(s) that are used to represent it (what Ranganathan calls the *idea plane* and the *verbal plane*) in practice it is hard to consider one without reference to the other, for a concept cannot be discussed without the use of its associated terms. A concept is often defined as an abstraction, an idea or general notion, whereas a term is a label for that idea, what may be designated as 'concept names' (Prasad and Guha, 2008: 501). The difference is readily demonstrated through the part which natural language plays. The same concept in different natural languages evidently has a variety of labels in, for example, English, German, Spanish or Russian. In a single language rich in synonyms, such as English, there may be many terms for the same concept, and in the case of homonyms, there may be several concepts related to the same term. Overall the association between a concept and its potential labels is very close, and it is easy to see how the same analytical methodology is relevant to both alphabetically and systematically organized systems. As Coates says of the role of subject cataloguing and classification in representing subject content (1960: 16), "the two disciplines only diverge at the subsequent phase in which the abstracted idea is reformulated by the subject cataloguer as a subject heading and by the classifier as a classification symbol. Both … provide answers to the question 'What is the subject of this document?' which are different in form alone."

The use of categories in subject indexing may be said to precede the work of Ranganathan on classification, since it was employed by Kaiser as early as 1911 to analyse and organize terms in subject headings in the field of commercial and industrial literature, using three categories: *concrete*, *place*, and *process*. Kaiser's systematic indexing was more than a pragmatic solution to complexity of content; it was both semantically and syntactically sophisticated, and well supported by a theoretical rationale (Dousa, 2013). His categories are easily interpreted as precursors of Ranganathan's fundamental categories of *personality*, *space*, and *energy.* A major contribution of Kaiser's work is the use of the categories to establish the relative importance of the subject heading terms, and hence the order of combination in the heading, what we now call citation order. For Kaiser a *concrete* would take precedence over a *process*, and would always be the entry word. Where a *place* was involved, there would be two entries, one of the form *concrete*, *place*, *process* (e.g. wine, France, export), and the other *place*, *concrete*, *process* (e.g. France, wine, export).

Other applications of facet theory to subject cataloguing are to be found in the work of Eric Coates, and, significantly, in the development of the British National Bibliography PRECIS indexing system at the British Library (Austin, 1984), which arose as a direct result of the CRG work in the 1970s. Coates (1960b) has written what is undoubtedly the authoritative UK study of subject catalogues, and one which employs categories as a means of formulating and ordering entries in the alphabetical subject catalogue. As in Kaiser's work, Coates uses the categories and their interrelationships as the basis on which the structure of a subject heading and the order of its components is determined, but his thinking is more specifically derived from Ranganathan's work on catalogue entry (Ranganathan 1945;1951). As he says in *Subject catalogues* (1960b: 44):

> From the practical point of view the order Personality, Matter, Energy, Space, Time is of the utmost importance, for this is the order in which the constituent parts of the Colon classification symbol are cited, and of course, it goes without saying that this order of citation determines where the subject is placed in the classified sequence. In his *Dictionary Catalogue Code* Ranganathan uses this facet formula as a basis for the construction of compound headings for the dictionary catalogue.

Hence, the relationship between the structure of the classmark, which controls the classified sequence, and the form of alphabetical subject entry in the catalogue is a close one, both underpinned by the use of fundamental categories. Writing during the very early days of the CRG, the categories Coates identifies (1960b: 57) differ slightly from the standard CRG list, but they

correspond closely in their general nature and scope: *thing*; *material*; *action*; *part*; *property*; *viewpoint*; *location.* To some extent they form a bridge, or transitional stage, between Ranganathan's PMEST and the later CRG set, with the addition of *part* and *property*, but the retention of the undifferentiated *action* (for *energy*). As with the rounds and levels of PMEST, it is sometimes necessary to specify *thing A* and *thing B*, and/or *action A* and *action B*. An interesting inclusion is that of *viewpoint*, which otherwise has only occurred in Otlet and LaFontaine's *Universal Decimal Classification* (1905-1907) where it is a generally applicable auxiliary table.

The use of facet analysis not only for the construction of classification schemes but also for alphabetical subject indexing, and slightly later as a basis for thesaurus construction (Aitchison, 1986), confirms it as a more general theory and methodology underpinning subject knowledge organization systems than might originally have been considered (Broughton, 2006). Much of the emphasis on the careful analysis of concepts, and the identification of relationships between them carries right through to digital tools and to ontology engineering.

Facet analysis in different disciplinary fields

Most of the original work of the CRG focused on scientific and technical subjects, and both the theory and the methodology are particularly well suited to the scientific domain. Much less attention has been paid to facet analysis in the humanities, although there are some especially interesting problems centred around the vocabulary of arts and humanities disciplines. These include the relationships between concepts in the classification and their lexical labels in the terminology per se, and the way in which equivalence can be determined in cross- and multi-cultural contexts. In Europe twentieth century research into categorical analysis paid more attention to the humanities, particularly in France as documented by de Grolier (1962: 61-89). Significant work included that of Gardin, initially for archaeology and later for iconography, and his conceptual analysis of texts such as the 'Mesopotamian tablets, the Koran, mythical tales (and more especially the myths of the Pueblo Indians)' (de Grolier, 1962: 84). A notable exception to the scientific focus of the original CRG work is Eric Coates' classification for the British Catalogue of Music (1960a), one of the earliest UK faceted schemes and the only example from a CRG member to feature a traditional humanities subject. However, like archaeology, music contains a substantial technical vocabulary, and it was this perhaps that made it more amenable to facet analysis.

Facet Analytical Theory in Managing Knowledge Structure for Humanities (FATKS)

The FATKS (Facet Analytical Theory in Managing Knowledge Structure for Humanities) project at UCL (https://www.ucl.ac.uk/fatks/) examined the theory and methodology involved in building a faceted classification intended for use as a tool for browsing and retrieval of humanities resources in an online environment (Broughton and Slavic, 2007). The concept behind the work was the belief that certain features of a faceted system which were particularly useful in supporting browsing and search in a subject context, were transferable to an online environment and could have distinct advantages in handling digital resources.

The work was occasioned in the first instance by the proposed merger of the two JISC (Joint Information Systems Committee) funded portals which at the time dealt with humanities resources in the United Kingdom, the Arts and Humanities Data Service (AHDS), which consisted of five independently managed data services, and the Humbul Humanities Hub (Broughton, 2002a; Broughton, 2002c). The millions of items held by AHDS were of very varied format and complex subject content; images were a significant element, but there were also sound recordings, film and

video, animation, and multi-media resources, as well as e-texts. As a consequence a major problem for the service was the difficulty of cross-collection searching.

Resources within the two portals were manually catalogued using Dublin Core, developed and managed by the Dublin Core Metadata Initiative (DCMI), and the de facto standard for digital resource cataloguing. Dublin Core, however, like other cataloguing standards, has no recommended authority for subject metadata and so within the portals several different vocabularies were being used. At the time of the research, any attempt at integrated search had been abandoned in favour of separate searching of the individual data services, even though this meant that searchers failed to retrieve much of the relevant material' (Broughton and Slavic, 2007: 728). It was felt that a single system would be more appropriate for the merged service, and ideally one which would offer more effective cross-collection searching and retrieval.

By 2000 the potential for faceted approaches to information retrieval in managed electronic environments had already been established (Broughton 2001; Duncan, 1989; Gödert, 1987; Gödert, 1991; Ingwersen and Wormell, 1992). Work on the faceted approach in the 1990s saw the development of applications such as 'view-based' and 'facet space' systems that, within a Windows environment, allowed the simultaneous display of multiple facets using cascaded-menus and interactive windows as an aid to search formulation and retrieval (Pollitt et al., 1996; Allen, 1995a; Allen, 1995b). The discussion had also begun with respect to the unmanaged world of the Internet (Broughton and Lane, 2000; 2004; Ellis and Vasconcelos, 1999; 2000).

Some of the perceived advantages of a faceted scheme over conventional classifications included:

- a scheme's logic and predictability and the regularity of the structure
- capacity for representing the subjects of semantically complex resources through combination of concepts
- clear rules (system syntax) for that combination
- flexibility in combination that allows for multiple paths to resource discovery and retrieval.

All of the above also enhance a system's suitability for use in a digital context, especially in terms of machine comprehension and manipulation of the system, and the possibility of some automatic processing of the materials for subject cataloguing and indexing. Although it was not at that time a particular concern, as the World Wide Web expanded it became clear that large scale human cataloguing of online resources was not a viable proposition, and research in the following years was also concerned with automatic metadata generation as an alternative.

Building the FATKS prototype classification

The immediate task was to build the prototype classification and to test its effectiveness as a classification and indexing tool. The prototype named FAT-HUM consisted of three distinct semantic components. Firstly, two disciplines, religion and the fine arts, were chosen as examples of humanities disciplines, partly because of recent work on their facet structure. The vocabularies for these fields would be developed in some detail and used as a test bed for the classification work. Secondly, to provide for any broader subject coverage that might be required they were located within the framework of the Broad System of Ordering (https://www.ucl.ac.uk/fatks/bso/about.htm), a general knowledge organization structure containing about 6,800 concepts. BSO had been originally developed by Eric Coates in 1978 as a general search and tagging tool and a switching language specifically for wide scale environments, such as international affiliations of libraries and documentation centres. Finally, frequently

occurring concepts such as place, form, and ethnicity were catered for by the use of the auxiliary tables of the Universal Decimal Classification (UDC), themselves also subject to recent revision (Broughton, 1998; Broughton, 2002b) as part of a general project to improve the facet structure of UDC (McIlwaine and Williamson, 1994). All of these elements assumed a faceted approach as the natural underlying principle of their construction, and together represented a coherent interlocking structure, as shown in Figure 1.

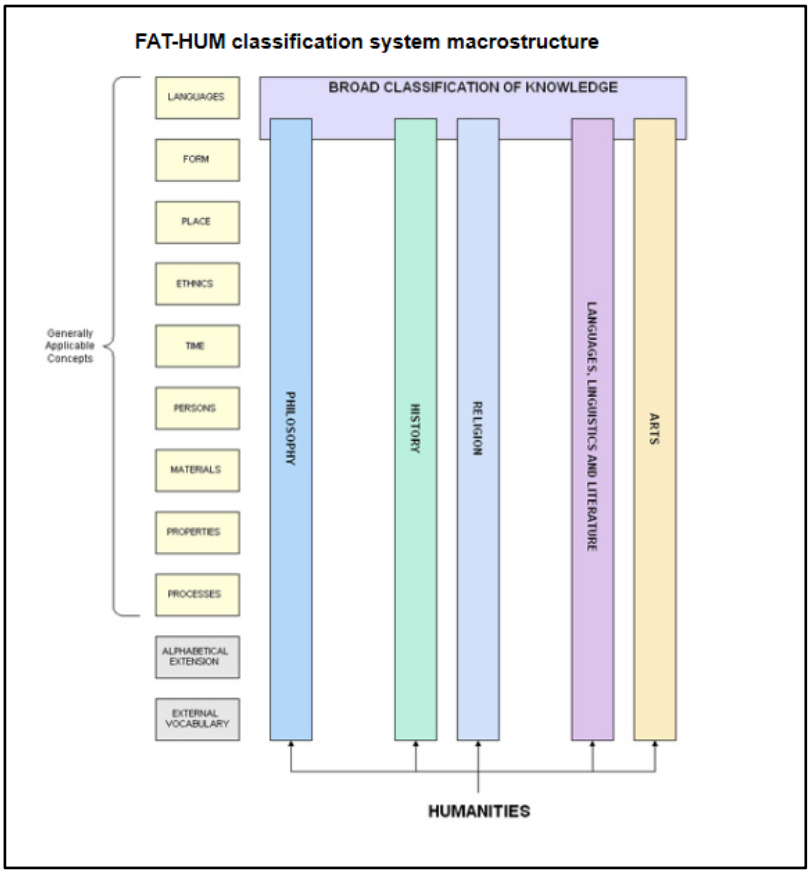

Figure 1. FAT-HUM classification macrostructure

Of the core constituents the principal demonstrator was the religion vocabulary. This work built on the recent revisions of religion in the *Bibliographic classification Second edition* (BC2) and the *Universal decimal classification* (UDC). In both of those cases one major objective of the exercise was to alleviate the bias towards Christianity shown by most western Classification schemes through a more equitable provision of notation and detail in the vocabulary. The treatment of religion in major classifications had long been the subject of complaint from librarians, with Dewey usually regarded as the worst offender (Zins and Santos, 2011: 881). The revision of Class P of Bliss's *Bibliographic Classification* (Mills and Broughton, 1977b), and of Class 2 of the UDC (Broughton, 2000), by applying a faceted approach to the terminology of religion, created classifications in which all of the major religions were regarded as equal from a classificatory perspective. A further consequence was the use of a generic template for subdivision, and the provision of a standard pattern for the detail under each faith.

The FAT-HUM vocabularies for religion adopted this general methodology, but benefited considerably from the experience gained with BC2 and UDC. They were more cleanly and systematically structured than the original BC2 schedules, being influenced greatly by the underlying database structure of the UDC which brought to the fore some of the constraints imposed by

machine management of the terminology (Slavic and Cordeiro, 2002; 2004). The development of the schedules for individual religions in UDC was also a work in progress and our thinking had similarly progressed alongside the revision work.

The concepts in religion were initially divided into ten groups, using the standard CRG type categories which had been employed in the revision of BC2 and subsequently in UDC. Other work, especially in the fine and creative arts, suggests that additional categories are needed to fit the somewhat different conceptual structure of these fields; for example *form*, *style* and *genre* are usually necessary for the organization of material in disciplines which are imaginative or representational. Nevertheless, despite their non-scientific nature, the religion concepts were found to map to the categories reasonably well (Figure 2). Some imagination was needed to interpret some of the more abstract concepts, such as 'schism' or 'mission', but this was only on a par with other subjects, and seemed to be more related to the understanding of the categories than to differences in disciplinary approach.

Category	Facet in religion
Thing (entity)	Religions and faiths; specific named religions and religious movements e.g. Hinduism
Theory and philosophy	Abstract concepts relating to religion e.g. nature of God
Parts	Structures, organizations and institutions within a faith, administrative divisions e.g. religious orders, charitable organizations
Properties	Attributes of religions e.g. monotheistic
Processes	Action concept that occur internally or without particular external agents e.g. schism
Operations	Action concepts carried out by agents, normally adherents e.g. worship, social work, missions
Patients	Recipients of operations e.g. the young, refugees
Agents	Persons who carry out operations e.g. ministers, social workers; also the means to perform operations, such as buildings, equipment, etc. e.g. mosques, prayer wheels
Time	Periods e.g. mediaeval, nineteenth century
Space	Places e.g. Europe, Asia Minor, Japan

Figure 2. Fundamental categories applied to religion

What might be termed a 'pseudo-category', *theory and philosophy* was added, as is common in CRG type faceted schemes, although these highly abstract terms might as easily have been accommodated within *properties* which they mostly consist of. Ultimately, in the UDC revision of religion, an additional facet of *evidences* was inserted to contain concepts such as sacred texts and other foundational reasons for belief. This was the only non-standard category, but it is close in nature to *agents*, which it sits next to in the sequence, and it might have been regarded as a subset of that category as it is in this prototype. This process of analysis demonstrates quite nicely that the method is an art rather than a science, particularly when it comes to concepts with more fluid meanings. The standard categories were completed by space and time, which in the FAT-HUM case were derived from the UDC auxiliary tables for those concepts.

Within in a facet the vocabulary is further organized into *arrays* or *sub-facets*, of terms which share a particular *characteristic* or *principle of division*. It is often the case that the defining characteristic or principle is derived from elsewhere in the vocabulary, as in Figure 3 below:

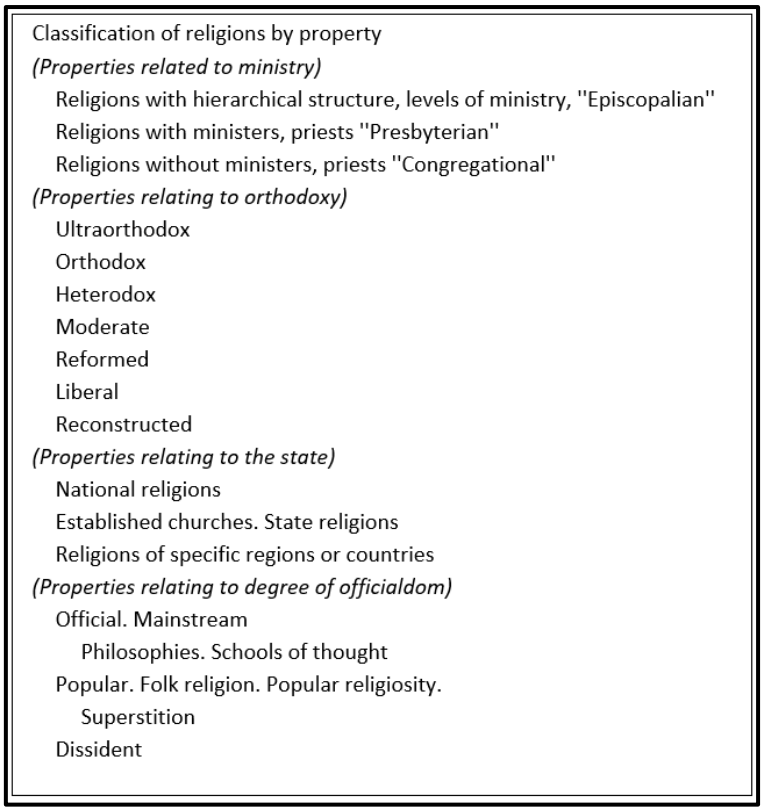

Figure 3. Organization of facets into arrays

At this stage equivalence relationships are acknowledged by bringing together synonyms or near synonyms at the same location. As a final element in the structure, hierarchical relationships are identified. This is normally a straightforward process in faceted scheme construction, since the concepts within a facet are all of the same categorical nature, and the only possible relationships between them are those of hierarchy, either subordination, super-ordination or co-ordination. The conventions of classification are that these hierarchical relationships are shown in the visual display by indentation (Figure 4).

Figure 4. Hierarchical structure in FAT-HUM classification

In all knowledge organization systems a primary means of making structure, status and relationships evident is the use of markup or encoding. In a classification scheme a system of encoding is already in place, namely the notation; this has the immediate function of identifying the individual concept or class, and maintaining the linear order or sequence, but other relationships and properties can be displayed through the judicious use of the notation. Notation used in the source vocabularies was quite diverse, and not necessarily aligned with the project objectives. Broughton and Slavic (2007: 736-7) note that:

> In BC2, from which the domain vocabularies were drawn, the notation is of a relatively unusual type, being ordinal, non-expressive, and retroactive in terms of synthesis, automatically imposing citation order if the rules for classmark building are carefully followed. While this provides an easy, elegant, and painless way to maintain order in the linear environment of the library shelf, it is not helpful for machine management. The UDC notation, which is expressive of hierarchy and uses a large number of symbols as facet indicators, was a better model for a notation for the prototype. It was decided that the notation for each concept would clearly indicate the subject area to which it belongs, the facet it comes from (facet indicator), and its hierarchical position within the facet.

Consequently a completely fresh system of notational coding was devised which incorporated all of these features (Figure 5).

590 J 14	Social behaviour
590 J 141	Self-emptying. Kenosis
590 J 142	Food and diet
590 J 1424	Food laws
590 J 14245	Dietary requirements. Dietary limitations
590 J 14247	Abstinence. Fasting. Prohibition
590 J 1425	Rules concerning specific foods and drinks
590 J 143	Personal hygiene and appearance. Personal conduct
590 J 1433	Cleanliness. Washing
590 J 1435	Clothing. Headgear
590 J 1437	Cutting, wearing, arrangement of the hair

Figure 5. Notational coding in FAT-HUM classification

In this example the notation '590' indicates the broad subject class (i.e. religion) within the total structure, 'J' denotes the facet status of the concept (i.e. religious activities, *operations*), and the following number relates to the individual concept or class, its position within the linear sequence, and within the hierarchy. So 590 J 14247 Abstinence. Fasting is shown to be following 'Dietary requirements' and preceding 'Rules concerning specific foods', and also to be three levels of hierarchy down from 590 J 14 'Social behaviour'. From this data it is possible to infer the broader term (or containing class) and any narrower terms (or subordinate classes). Although this feature of coding was not exploited in the FATKS project, it was subsequently used in the BC2 work to automatically generate a thesaurus format from the encoded data for the classification structure (Aitchison, 2004; Broughton, 2008).

The notation was also central to the representation of complex subject content, since it supported combination in a manner in which the various constituents of the complex subject remain evident.

590**A** Theory and philosophy of religion
590**A3** The Holy. The sacred. The supernatural. Object(s) of religion/worship

59033 Hinduism
59033A3 The Holy. Brahma. Absolute being
5906 Judaism
5906A3 Kedushah. The Holy. The Sacred

Figure 6. Combination between facets in the FAT-HUM prototype

Here the concept of the Holy has been taken from the general facet of philosophy and added to specific named religions (Figure 6). Complicated notations can be built up when facets are combined from the core humanities vocabularies with auxiliary table concepts (Figure 7).

Notation:	Description:	Facet:
57071	Eastern churches	590 Religions and Faiths
59071224	Autonomous Orthodox churches	590 Religions and Faiths
(D52)	Japan	(D) Common auxiliary of Place
59071224(D52)	Orthodox church in Japan	

Notation:	Description:	Facet:
5904	Buddhism	Religion and faiths
J448	Divination. Augury. Soothsaying. Oracles	Religious activities. Practice. Subfacet: Ceremonies
(K01)	Persons as agents, doers, practitioners	Common auxiliary of Persons
5904J448(K01)	Soothsayers	

Figure 7. Complex subject representation in the FAT-HUM prototype

Hence, in the prototype tool the notation works together with the internal rules of the system, or system syntax, to support the internal conceptual structure. The notation also bears the burden of synthesis of compound classmarks through the imposition of citation order based on facet status.

The second stage of the project was to build a relational database to hold the classification data and act as an editorial tool for the management of the prototype classification by its owners; this also involved the major task of populating it with the data.

The third element was to facilitate access to the data through the development of a search interface for end users (Figure 8). In addition to simply acting as a repository, the database was therefore also able to support a facet search and browsing function that included, for example, the facility to expand and collapse hierarchies. Because of the way in which the notation was representative of the internal conceptual structure of the system, the prototype was able to manage combinations of concepts, applying citation order of elements in line with the system syntax and building valid classmarks; used in this way it could potentially enable computer assisted indexing. Strictly speaking notation so employed can support a considerable level of automatic indexing, searching (through query formulation and modification) and retrieval without being evident to the end user; hence it is a viable means of supporting discovery in a more intuitive style without the end user needing to learn or understand the syntactic rules.

Figure 8. Browsing interface in the FAT-HUM prototype

The development of the prototype classification demonstrated the feasibility of building a system that translates the conceptual approach of facet analysis into a manageable data structure that can support all the semantic and syntactic features of a fully faceted vocabulary.

At the end of the project it was felt that three important objectives had been achieved:

- the validity of facet analysis for humanities work had been demonstrated in the creation of the conceptual model
- this was reflected in the structure of the relational database built as an editorial tool to hold the classification data. The complexity of the conceptual model could be replicated there, as could the system syntax
- the search and browse interface provided a verbal means of access to the content of the prototype.

The second and third objectives together formed a basis for potential non-human classification of resources in religion, an important issue as unmanaged information continues to grow at an unprecedented rate.

The prototype performed well, but our emphasis then was predominantly on the mechanics of the exercise and the usability of the methodology rather than the intellectual analysis of the domain vocabulary as such. Synthesis of classmarks for complex subject was seen to work effectively, the subjects were fully represented and the notation complied with the system rules. However, the whole was subject to a lack of clarity in the semantics of the built notations.

Representing complex content through structure and terminology

The matter of factoring (now referred to as splitting) has long been a concern in the verbal indexing world. Standards for thesaurus construction routinely address the question of how one should represent a concept that is either syntactically or semantically composite. There are numerous examples provided in the literature and in the standards themselves, but those below serve to illustrate the problem. Syntactic factoring, or splitting, is considered where the entry terms, or concepts, consist of multiple words in a phrase, e.g.:

Horse racing
Brain surgery
Mothering Sunday
Information retrieval
Blue movies
Cats' eyes

Various rules and conventions apply to whether a term such as 'Horse racing' is a valid entry term, or if it should be created post-co-ordinately, that is at the time of search, from the two simple terms 'horse' and 'racing'. In such a case the term 'horse racing' would not feature in the indexing language nor would it be used in a heading or as a descriptor. Exceptions to this rule include where the term is a well established one (Information retrieval), where the compound has a meaning beyond its constituents (Mothering Sunday), or where the constituents standing alone have a different meaning to that in the compound (Blue movies, Cats' eyes).

Semantic factoring is more complicated, and is now not routinely covered in the standards. Nevertheless it is a concern for compilers of knowledge organization systems, since it involves a key relationship between a concept and the lexical labels used to describe it. Semantic factoring refers to a situation which is the inverse of syntactic factoring, and where a compound concept is represented by a single term. For example, the term 'gingivitis' means 'inflammation (of the) gums', but neither of its semantic components are identifiable in the term itself. In a thesaurus the term 'gingivitis' would be included, but in a faceted classification it might very well not be, on the assumption that the cataloguer or classifier would understand its meaning, and create a classmark as needed based on the notation for these two elements.

One of the original perceived advantages of faceted classifications was the economy in scheduling since there was no need for the constant repetition of concepts in pre-co-ordinated class names. Published volumes of these early schemes (including the Colon Classification itself) were invariably slim, providing only the key concepts in the relevant domain. These represented a kind of conceptual skeleton which would become populated only in the classified catalogue or on the shelves as examples of combination of concepts were added as a result of cataloguing documents. The consequence was that the classification itself, while it was conceptually sound, lacked many terms of a semantically compound nature. This severely compromised the classification from a search perspective, since it failed to include large numbers of what would be sought terms. The revision of BC2 resolved the difficulty by expanding the schedules to include pre-synthesised classmarks (what UDC would label as 'examples of combination') built according to the system rules. This practice had several objectives: to ensure that sought terms were represented in the vocabulary; as a corollary to ensure their appearance in the alphabetical index to the schedules; and to provide the cataloguer with examples of correct application of the rules for combination. Incidentally, the Dewey Decimal Classification also follows this latter custom, but with the primary purpose of demonstrating number building rather than to facilitate the display of these semantic compounds.

This example (Figure 9) from Bliss Class H, Health Sciences (Mills & Broughton, 1980: 229) shows how this translates into the classification display (note that not every class is included):

HWU	Liver
HWU BJ	(Physiology)
HWU FG	(Investigation)
HWU FKW B	Scintigraphy
HWU H	(Pathology)
	(Enlargement)
HWU JK	Hepatomegaly
HWU L	(Inflammation). Hepatitis
HWU	(Necrosis)
HWU MAT	Acute yellow atrophy, parenchymatous hepatitis
HWU MDB	(Degeneration)
HWU MDR	(Metaplasia)
HWU MDT	Hepatic cirrhosis, cirrhosis of liver

Figure 9. Schedule expansion in BC2 Health sciences

Without the expansion of the schedule to include these pre-built classes terms such as 'scintigraphy', 'hepatomegaly', 'hepatitis' and 'cirrhosis' would not feature in the system, and would not be easily discoverable by its users.

The role of terminology in the religion classification

Although such a large technical vocabulary is not common to all disciplines it does seem that where there is a cultural dimension to the discipline, as is the case with humanities fields, a similar phenomenon could be observed, and this is certainly true of religion. But in focusing on the structural and technical aspects of the FATKS prototype, we had not been able to consider the semantic problems of the terminology. Ongoing revision work on UDC had begun to grapple with this, and the development of the Religion schedules in that scheme had provided some customised expansions for individual faiths, which used the specific vocabulary of the faith to label the classes, rather than a generic name. A generic, or neutral, special auxiliary was provided which could be combined with any faith to create a faith specific classification; a number of these were created editorially and published as special expansions into which were imported the terms associated with concepts in that particular belief system. This achieved the same aims as the BC2 medicine class, namely making a much larger and more 'technical' vocabulary visible and accessible.

2–144.2	Names of god(s)
2–23	Sacred books. Scriptures. Religious texts
2–24	Specific texts. Named texts and books
2–282.5	Prayer books. Books of prayers
2–442.45	Dietary requirements. Dietary limitations
2–523.4	Centres of worship (religious significance)
26	Judaism
26–24	Tanakh. The Hebrew Bible
26–442.45	Kasruth. Kosher regulations

26–523.4	Synagogue. Beth knesset
27	Christianity. Christian churches and denominations
27–523.4	Church buildings (religious significance)
273.4	Anglican church
273.4–282.5	Book of Common Prayer
28	Islam
28–24	The Quran
28–442.45	Halal. Dietary requirements. Dietary limitations
28–523.42	Mosques

There are several interesting features of this phenomenon. Firstly, the presence of a natural language dimension to the vocabulary means that each primary facet may have very many unique terms associated with it. Secondly, while these may be broadly equivalent between one faith and another it is likely that because of cultural differences they do not map exactly to the generic concept used as the basis of combination. An understanding of prayer in Judaism may be very different from that in Buddhism, whereas the nature and composition of an oxygen atom in sulphur dioxide is not different from that in nitrogen dioxide, nor does it have a different name. Clearly in the case of specific named religious entities there is an inevitable mismatch. For example:

Judaism + sacred text = Hebrew Bible
Hinduism + sacred text = Vedas

where Bible and Vedas are conceptually comparable, but not conceptually, nor linguistically, equivalent. One can well argue that, in the same discipline, the nature of God in Buddhism is very different from the nature of God in Islam, just as, more obviously, practices and rites are different. In practice it is possible that these differences do not matter very much, and that it is useful for the benefit of the data structure that they be regarded as equivalent since that rationalises the data structure. Using the same notational coding enables cross-searching to retrieve material that is comparable in function if theologically distinct, for example, all scriptures, or all marriage rites.

Terminology and religious bias

One advantage of the customised classifications which had not been fully appreciated until more recently is the way in which they resolved some of the problems of perceived bias in classification schemes. While a faceted classification should in principle provide an even handed and unbiased model for knowledge organization in culturally diverse fields, the choice of vocabulary within the classification is also a significant factor in avoiding cultural bias. Even where there is equal notational provision, and a uniform pattern of classification for every culture, bias may still be implicit in the choice of terms and the unconscious dominance they reflect.

The precise use of language in culturally sensitive domains has the power not only to misrepresent, but also to disadvantage and to offend. In the last few years there has been much research into the way in which minority groups have been marginalized through the use or misuse of language. Communities which have been affected in this way include those characterised by gender, race, sexual orientation and political status, although surprisingly there has been little published about apparent discrimination on the grounds of religious belief, despite widespread complaints about

Christian dominance in the large library classification schemes. It is clear that very often the bias occurs in the choice of language as much as the structure of the scheme and the disposition of the notation, and it is important to have this in mind when constructing knowledge organization systems. A study of religious terms used in some large vocabularies adopted by automatic indexing tools (Broughton, 2019) shows some very regrettable choices, with at best a strong lean towards the terminology of Christianity, and at worst some archaic expressions displaying an arrogance towards non-Christian faiths, such as 'Hindoo' or 'Mohammedanism'. Figure 10 shows some examples of entry terms in WordNet which are common to most religions but which are associated with strongly Christian language; while the concepts are essentially neutral the accompanying text is not.

Source term	Synonyms
altar	Communion table, Lord's table
baptism	a Christian sacrament signifying spiritual cleansing
bless	make the sign of the Cross over someone
festival	religious festival, church festival
monk	Brother, Carthusian, Trappist, Cistercian
preaching	an address of a religious nature usually delivered during a church service
scripture	Bible, Christian Bible, Holy Writ, Word (the sacred writings of the Christian religions)
service	church service, prayer meeting, chapel service, vesper
sin	mark of Cain

Figure 10. Some entry terms for religion in WordNet https://wordnet.princeton.edu/

Some recent work (Broughton and Lomas, 2020) has also looked more closely at the nature of religious vocabulary and how the selection of terms can create, or alleviate, bias in the representation of communities in a spectrum of information sectors. While Broughton (2019) had examined the situation in artificial intelligence applications, Broughton and Lomas (2020) analysed practice in both the library and archival domains. Broughton and Lomas's work shows that there is a correspondence between the language used in a knowledge organization system and the standard model of interreligious attitudes. A scheme which has minimal provision for religions as a whole, and a very limited number of terms that are not Christian in origin may be regarded as exclusivist in its view of the value of different faiths, whereas another that has more equitable treatment with a fuller vocabulary indicative of a variety of religions is inclusivist in approach. A classification scheme that apparently treats all the religions equally with equivalent allocation of notation and a rich mixture of terms and concepts from the whole spectrum of world faiths may be said to be pluralist. A pluralist view should be regarded as desirable in a modern knowledge organization system intended for use internationally which may be considered a given in a digital environment. Looking at progression towards the goal of pluralism it is seen that both bibliographic and archival practices are moving towards this position.

Conclusion

Facet analysis is a sound and reliable methodology for constructing classification schemes and knowledge organization tools of all kinds. It has been shown over time to provide a strong theoretical basis that is transferable to a wide range of disciplines, and that all these domains are susceptible to analysis including those in the humanities. The FATKS project demonstrated how a well constructed prototype can accommodate both manual subject cataloguing and some degree of

automatic indexing of resources in religion. However, most of the work in the FATKS exercise was concerned with the data structure of the classification, and with equalising the conceptual framework for a variety of religions so that a balanced view was presented. Subsequently it was seen that the choice of terminology was as important as the conceptual structure in fairly representing subjects where there is considerable cultural variation in one or more facets.

In a world of increasing diversity it is important that these concerns are acknowledged when building organization and retrieval tools, and that our theory is used to best purpose in dealing with them.

Note

Figures 1-8 are adapted from documentation on the FATKS website https://www.ucl.ac.uk/fatks/

References

Adkisson, H. P. (2005) *Web design practices: use of faceted classification.* www.webdesignpractices.com/navigation/facets.html.

Aitchison, J. (1970) 'The Thesaurofacet: a multipurpose retrieval language tool', *Journal of Documentation* **26**(3), 187-203.

Aitchison, J. (1986) 'A classification as a source for a thesaurus: The Bibliographic Classification of H. E. Bliss as a source of thesaurus terms and structure', *Journal of Documentation* **42**(3), 160-181.

Aitchison, J. (2004) 'Thesauri from BC2: problems and possibilities revealed in an experimental thesaurus derived from the Bliss Music schedule', *Bliss Classification Bulletin* **46**, 20-26.

Allen, R. B. (1995a) 'Two digital library interfaces that exploit hierarchical structure', in *Proceedings of DAGS95, Electronic Publishing in the Information Superhighway, Boston, Massachusetts, May 30 - June 2,* pp.134-41.

Allen, R.B. (1995b) 'Retrieval from facet spaces', *Electronic Publishing*, **8(**2/3), 247-58.

Art and architecture thesaurus (1994). New York: Oxford University Press. Available online at: http://www.getty.edu/research/tools/vocabularies/aat/.

Austin, D. (1984) *PRECIS: a manual of concept analysis and subject indexing,* 2nd edn. London: The British Library Bibliographic Services Division.

Beghtol, C. (1995) 'Facets as interdisciplinary undiscovered public knowledge: S. R. Ranganathan in India and L. Guttman in Israel', *Journal of Documentation* **51**(3), 194-224.

Broughton, V. (1998) 'The development of a common auxiliary schedule of property: a preliminary survey and proposal for its development', *Extensions and Corrections to the UDC* **20**, 37-42.

Broughton, V. (2000) 'A new classification for religion', *International Cataloging and Bibliographic Control* 2000 **4**, 2-4.

Broughton, V. (2001) 'Faceted classification as a basis for knowledge organization in a digital environment; the Bliss Bibliographic Classification and the creation of multi-dimensional knowledge structures', *New Review of Hypermedia and Multimedia* **7**, 67-102.

Broughton, V. (2002a) 'Facet analytical theory as a basis for a knowledge organization tool in a subject portal', in M. J. Lopez-Huertas and F. J. Munoz-Fernandez (Eds.), *Challenges in knowledge representation and organization for the 21 st century. Integration of knowledge across boundaries. Proceedings of the Seventh international conference of the International Society for Knowledge Organization, Granada, Spain*, *10-13 July 2002, Advances in Knowledge Organization* **8**, Wurzburg: Ergon, pp. 135-41. Also available at: http://www.ucl.ac.uk/fatks/paper2.htm.

Broughton, V. (2002b) 'A new common auxiliary for relations, processes and operations', *Extensions and Corrections to the UDC* **24**, 29-35.

Broughton, V. (2002c) 'Organizing a national humanities portal; a model for the classification and subject management of digital resources', *Information Research Watch International* **June,** 2-4

Broughton, V. (2006) 'The need for faceted classification as the basis of all information retrieval." *Aslib proceedings*', **58**(2), 49-72.

Broughton, V. (2008) 'A faceted classification as the basis of a faceted terminology: conversion of a classified structure to thesaurus format in the Bliss Bibliographic Classification, 2nd Edition', *Axiomathes* **18**(2), 193-210.

Broughton V. (2010) 'Concepts and terms in the faceted classification: the case of UDC', *Knowledge Organization* **37**, 270-79.

Broughton, V. (2019) 'The respective roles of intellectual creativity and automation in representing diversity: human and machine generated bias', *Knowledge Organization* **46**(8), 596-606.

Broughton, V. and Lane, H. (2000) 'Classification schemes revisited; applications to web indexing and searching', *Journal of Internet Cataloging* **2**(3/4), 143-55.

Broughton, V. and Lane, H. (2004) 'The Bliss Bibliographic Classification in action: moving from a special to a universal faceted classification via a digital platform', in I. C. McIlwaine (Ed.), *Knowledge organization and the global information society. Proceedings of the Eighth international conference of the International Society for Knowledge Organization, University College London, 13-16 July 2004, Advances in Knowledge Organization* **9**, Wurtzburg: Ergon, pp. 73-78.

Broughton, V. and Lomas, E. (2020) 'Philosophical foundations for the organization of religious knowledge: irreconcilable diversity or a unity of purpose?', *Knowledge Organization* **47**(3) in press.

Broughton V. and Slavic, A. (2007) 'Building a faceted classification for the humanities: principles and procedure', *Journal of Documentation* **35**(5), 727-54.

Checkland, P.B. (1989) 'Soft systems methodology', *Human Systems Management* **8**(4), 273-89.

Classification Research Group (1955) 'The need for a faceted classification as the basis for all methods of information retrieval', *Library Association Record* **57**(7), 262-8.

Classification Research Group (1969) *Classification and information control: Papers representing the work of the Classification Research Group during 1960-1968*. London: Library Association.

Coates, E. J. (1960a) *British Catalogue of Music classification.* London: British National Bibliography,

Coates, E. J. (1960b) *Subject catalogues: headings and structure.* London: Library Association.

Coates, E. J. (1973) 'Some properties of relationships in the structure of indexing languages', *Journal of Documentation* **29**(4), 390-404.

Coates, E., Lloyd, G. and Simandl, S. (Eds.) (1978) *Broad system of ordering: schedule and index.* The Hague: International Federation for Documentation (FID). Available online at www.ucl.ac.uk/fatks/bso.

de Grolier, E. (1962) *A study of general categories applicable to classification and coding in documentation.* Paris: Unesco.

Dousa, T. M. (2013) *Julius Otto Kaiser and his method of systematic indexing: an early indexing system in its historical context.* Unpublished PhD thesis. University of Illinois. Available at: http://hdl.handle.net/2142/46755.

Duncan, E. (1989) 'A faceted approach to hypertext', in R. McAleese (Ed.), *Hypertext: theory into practice.* London: Intellect, pp. 157-63.

Ellis, D. and Vasconcelos, A. (1999) 'Ranganathan and the Net : using facet analysis to search and organize the World Wide Web', *Aslib Proceedings* **51**(1), 3-10.

Ellis, D. and Vasconcelos, A. (2000) 'The relevance of facet analysis for world wide web subject organization and searching', in A. R. Thomas and J. R. Shearer (Eds.), *Internet searching and*

indexing: the subject approach. Binghamton, NY: Haworth, pp. 97-114. (Also published as *Journal of Internet Cataloging* **2**(3/4).

FATKS (Facet analytical theory in knowledge structures) Project details and documentation at www.ucl.ac.uk/fatks/.

Gardin, J. C. (1965) SYNTOL. New Brunswick: Rutgers State University.

Gödert, W. (1987) 'Klassificationssysteme und Online-Katalog. Classification systems and the on-line catalogue', *Zeitschrift for Bibliothekswesen un Bibliographie* **34**(3), 185-95.

Gödert, W. (1991) 'Facet classification in online retrieval', *International Classification* **2**, 98-109.

Guttman, L. (1959) 'A structural theory for intergroup beliefs and action', *American Sociological Review* **24**(3), 318-28.

Hjørland, B. (2013) 'Facet analysis: the logical approach to knowledge organization', *Information Processing and Management* **49**(2), 545-57.

Ingwersen, P. and Wormell, I. (1992) 'Ranganathan in the perspective of advanced information retrieval', *Libri* **42**, 184-201.

ISO 25964 International standard for thesauri and interoperability with other vocabularies (2011). NISO website. www.niso.org

Kaiser J. (1911) *Systematic indexing.* London : Pitman.

La Barre, K. (2006) *The use of faceted analytico-synthetic theory in the practice of website construction and design.* Unpublished PhD dissertation. Indiana University. Available at: https://netfiles.uiuc.edu/klabarre/www/LaBarre_FAST.pdf.

McIlwaine, I. C. and Williamson, N. J. (1994) 'A feasibility study on the restructuring of the Universal Decimal Classification into a full-faceted classification system', in H. Albrechtsen and S. Oernager (Eds.), *Proceedings of the Third International Society for Knowledge Organization (ISKO) Conference: Knowledge organization and quality management, Copenhagen, Denmark, 20-24 Jun 94*. Frankfurt/Main: INDEKS Verlag, pp. 406-13

Miles, A. and Bechhofer, S. (Eds.) (2009) *SKOS Simple knowledge organization system: reference. W3C Recommendation 18 August 2009.* Available at: http://www.w3.org/TR/2009/REC-skos-reference-20090818/.

Mills, J. and Broughton, V. (1977a) *Bliss bibliographic classification*, 2nd edn. London: Butterworths.

Mills, J. and Broughton, V. (1977b) *Bliss bibliographic classification,* 2nd edn. *Class P Religion, occult, morals and ethics.* London: Butterworths.

Mills, J. and Broughton, V. (1980) *Bliss bibliographic classification,* 2nd edn. *Class H Anthropology, human biology, health sciences.* London: Butterworths.

Otlet, P. and LaFontaine, H. (1905-1907) *Manuel du repertoire bibliographique universel.* Brussels: International Institute of Bibliography.

Perry, J. W., Kent, A. and Berry, M.M. (1956) *Machine literature searching.* New York: Interscience.

Pollitt, A. S., Smith, M. P. and Braekevelt, P. (1996) 'View-based searching systems: a new paradigm for information retrieval based on faceted classification and indexing using mutually constraining knowledge-based rules', in C. Johnson and M. Dunlop (Eds.), *Information retrieval and human computer interaction. Proceedings of the joint workshop of the information retrieval and human computer interaction specialist groups of the British Computer Society.* GIST Technical Report G96-2, Glasgow University. Glasgow: Glasgow University, pp. 73-77.

Prasad, A.R.D. and Guha, N. (2008) 'Concept naming vs. concept categorisation: a faceted approach to semantic annotation', *Online Information Review* **32**(4), 500-10.

Ranganathan, S. R. (1945) *Dictionary catalogue code.* Madras: Thompson; London: Grafton.

Ranganathan, S. R. (1951) *Classified catalogue code.* Madras: London.

Ranganathan, S. R. (1967) *Prolegomena to library classification,* 3rd ed. New York: Asia Publishing House.

Slavic, A. and Cordeiro, I. (2002) 'Data models for knowledge organization tools: evolution and perspectives", in M. J. Lopez-Huertas and F. J. Munoz-Fernandez (Eds.), *Challenges in knowledge representation and organization for the 21 st century. Integration of knowledge across boundaries. Proceedings of the Seventh international conference of the International Society for Knowledge Organization, Granada, Spain*, *10-13 July 2002. Advances in Knowledge Organization* **8** Würzburg: Ergon, pp. 127-34.

Slavic, A. and Cordeiro, I. (2004) 'Core requirements for automation of analytico-synthetic classifications', in I. C. McIlwaine (Ed.), *Knowledge organization and the global information society. Proceedings of the Eighth international conference of the International Society for Knowledge Organization, University College London, 13-16 July 2004, Advances in Knowledge Organization* **9** Wurtzburg: Ergon, pp. 187-92.

Zins, Chaim and Santos P. L. V. A. C. (2011) 'Mapping the knowledge covered by library classification systems', *Journal of the American Society for Information Science and Technology* **62**, 877-901.