Kevin Harris
Polytechnic of North London
School of Librarianship and Information Studies

Part-Controlled Vocabulary for Literature Studies

Harris, K.: **Part-controlled vocabulary for literature studies.**
Int. Classif. 13 (1986) No. 3, p. 133–136, 18 refs.

Part-controlled vocabulary involves the supplementation of a limited controlled vocabulary with freely chosen index terms which are not subject to any authority procedures. Such an indexing device is probably widely used in various kinds of information service, but has not been formally recognized as a design option for librarians. The concept and its implications are explored, and its potential for literature studies, where it appears to be highly suitable, is considered. (Author)

1. Introduction

Using Shakespeare studies for examples, Heiner Schnelling (1) has proposed the use of "a structured set of fixed indexing terms, which will have to be complemented with free terms" in an attempt to overcome certain terminological problems in alphabetical subject catalogues. Elsewhere, James Anderson (2), in describing the Modern Language Association's classification and indexing system has hinted at an apparently similar approach, with the use of "uncontrolled 'identifiers'", although in practice thesaurus procedures operate to ensure full vocabulary control in today's MLA *International Bibliography* (3). Research into the potential of *The Dickens House Classification* (4) for adaptation to special literature collections has indicated that systems which employ such part-controlled vocabularies could be well suited to literature studies, as these two instances already imply. It is the purpose of this paper to outline the concept of part-controlled vocabularies and briefly to consider their potential for the systematization of literature studies.

One of the best-known thinkers and writers on vocabulary control, F.W. Lancaster, has discussed the idea of the "hybrid" vocabulary, which he describes as "one that combines a controlled vocabulary with natural language":

Typically, in this situation, a relatively broad controlled vocabulary of perhaps several hundred terms provides a kind of overall superstructure for the system. Documents are indexed by one or more of these broad descriptors and also by natural-language terms extracted from title or text or both. The natural-language words allow specificity in searching. The broad controlled terms provide for generic search and give context to the natural-language terms . . . The joint use of a limited controlled vocabulary and an uncontrolled natural-language vocabulary offers powerful retrieval capabilities and has been shown to operate very effectively in a number of applications. (5)

However diverse the applications of this kind of retrieval device, it remains true to say that there is little in the literature of librarianship which helps us explore its possibilities. It seems quite certain that the device is widely used, perhaps often unconsciously, as a pragmatic "finger-in-the-dyke" option – controlled vocabularies being expensive to maintain – in many kinds of libraries. But as a first-choice, planned option it is unlikely to have been considered, and librarians are offered little guidance based on professional practical experience.

Part-controlled vocabularies as here described should not be confused with other kinds of "hybrid" vocabulary, such as that used by the British National Bibliography (a combination of *Dewey Decimal Classification* and the BNB *Supplementary schedules*) prior to the introduction of PRECIS, which are in effect combinations of two or more related controlled vocabularies. They should also be distinguished from what Lancaster describes as a "postcontrolled vocabulary", which "might best be developed empirically, from strategies used by searchers . . ." (6): such a device is clearly likely to consist more of specific terms than a fundamental, broad classification.

One very reasonable explanation for the fact that little work has been done on hybrid or part-controlled vocabularies might have to do with the lack of demand for them in scientific and technical subject areas. For such subjects, which often for economic reasons are more likely to be the testing ground for theoretical developments, tend to be represented by vocabularies which are readily susceptible to analysis and logical division. They are also subjects where traditions of high specificity and completeness of search have been established; computerized databases calling for pure controlled vocabularies or free-text searching are commonplace; and where financial support for, and commitment to, information services and systems is relatively easily found. As a result, such factors as uncertainty of vocabulary, lack of emphasis on precision in searching, and lack of funding, which characterize information services in the humanities, have not influenced general information retrieval theory or practice to any great degree. Factors like this, however, may well prove to be influential in stimulating the use of vocabularies which are partly controlled, where before there was none. And in many special collections in the humanities, subject access is so undeveloped as to be virtually non-existent, so that the economic part-controlled vocabulary option may be highly attractive.

In practice, then, the system would be based on a fundamental classification which could be derived by adaptation and amendment of existing tools. It is envisaged that in certain areas of literature studies, particularly for author-based collections, *The Dickens House Classification* might serve as a useful foundation or framework: elsewhere, in other areas, one would hope that the indexing system underlying the MLA *International Bibliography* (7) would prove a valuable model. The MLA Thesaurus is not available in print, although there are plans to publish sections of it (8).

At the indexing stage, the classification provided would be used as far as it goes: thereafter, where greater

specificity is called for, the indexer would supply terms which are not subject to authority procedures. These terms would be selected on the basis of the indexer's familiarity with the material in the collection and the users' requirements, as well as on the language used in the document.

For example, a limited vocabulary might be established for a collection of works on English poetry, and might include a general class for 'literary devices'. A study of 'typology' (as used by seventeenth century religious poets) would presumably belong here – together with works on related concepts such as allegory and symbolism – and could be indexed without a class or authority record being provided for the term. Resulting index entries such as

allegory – Metaphysical poets
typology – George Herbert

draw attention to the lack of cross-referencing which this system implies: nonetheless, as discussed below, it may be questioned whether the maintenance of a reference structure at this level is justified.

2. Discussion

A number of points arise from a consideration of such an indexing procedure.

2.1 Characteristics of literature studies

Literature studies is generally regarded as being a comparatively static field, its terminology changing relatively slowly. This is no longer the case. The expansion of inter-cultural and inter-disciplinary approaches, together with a remarkable increase in the attention paid to critical theory over the last 15 years or so, have reshaped the field so that, in certain characteristics, it contrasts strongly with its previous history. Philosophical, linguistic and ideological influences predominate, and aesthetic evaluation is in retreat, in a truly international re-examination of the theoretical base (9). Literature studies (unlike other humanities subjects perhaps, such as history) appears consistently to be showing some of the characteristics of scientific and social science fields in the way that new, predominantly more specific concepts are constantly being introduced.

There is a significant difference, however: in scientific fields, important new contributions can lead to major, even radical reassessment of the fundamental structure of the field or fields concerned. Literature studies is more cumulative, less oriented towards the concept of "falsifiable statements"[1], although of course subject to trends and fashions. What this means is that the fundamental structure of the field is relatively consistent. Like a relief map which also shows roads and railways, its basic arrangement remains unchanged, though new roads and structures spring up frequently. The advantages offered by part-controlled vocabularies in literature studies are clear: the basic controlled vocabulary can be expected to serve for quite some time, and not need major revision every five years as might be the case in, say, biotechnology. The significant area of change in terminology will be in the addition of new, specific concepts.

2.2 Characteristics of the vocabulary

Part-controlled vocabulary accentuates the distinction, to which attention is drawn by Robert Fugmann, between general and individual concepts.

It is typical of individual concepts *that they are expressed solely* by lexical expressions, *e.g. proper names.* (10)

General concepts, on the other hand, are characterized by "a multiplicity of expressions", typically of a "non-lexical mode". Fugmann points out that "it is inherent in non-lexical expressions that they cannot be looked up in case of demand, because they defy effective alphanumeric arrangement" (10). It is clear that for such language there is an important role for vocabulary control.

Fugmann's analysis may be compared with Stephen Wiberley's study of the language of the humanities, in which are distinguished singular proper terms, enumerable proper terms, general proper terms, and common terms.

General proper and common terms . . . are very imprecise: their definitions are often characterized by change over time or a wide range of meaning, and their referents frequently include a diversity of subjects or objects (11).

One of the distinguishing characteristics of the language of literature studies is the high incidence of singular proper terms. These are not class terms and are thus much harder to fit into any hierarchical structure. They are clearly strong candidates for the free part of the vocabulary . As Fugmann says,

Where individual concepts are concerned, it is advisable in most cases to represent them by natural language terms (12).

However, a further and perhaps most significant characteristic of the language of the field is not catered for by these studies: that is, its *range* of specificity, from such general concepts as "style" or "characterization" to individual concepts such as "*Paradise lost*" or "Leopold Bloom". Heavy use is made of the terminology at each end of the scale. Hence the desirability of a controlled indexing language, supported by the use of free terms for at least the most precise concepts.

2.3 Subject knowledge of indexers

The necessity for the indexer to show particular familiarity with the language and requirements of the field is an acknowledgement and re-affirmation of the traditional close relationship between search and research in literature studies. Typically, staff in special literature collections are selected with greater emphasis on their subject knowledge than on information retrieval expertise: there are understandable reasons for this, which have to do with the lack of demand for refined information systems, the conceptual complexities of the field, and the traditions of bibliographic "rummaging". A strategy incorporating a part-controlled vocabulary might be regarded as a opportunistic choice which would help to maximize the skills and knowledge of available staff.

2.4 The demands on indexers

The risk of "classification by attraction" increases as the specificity of the vocabulary decreases. A number of writers[2] have commented on the phenomenon of classi-

fication by attraction, which is partly a function of incomplete schedules. The limitation to a part-controlled vocabulary takes from the indexer the comfort of being able to begin indexing at the most specific level: the indexer is obliged to think in terms of generic classes to which a specific concept might belong, particularly in this field where polyhierarchies abound. There may be a temptation to use established terms rather than more appropriate free terms. Thus there is an extra emphasis on the need for the initial subject analysis to be sound *before* the free terms are added and an index entry is made.

It should also be noted that there are certain disadvantages in using too broad a classification, as far as shelf arrangement and classified catalogues are concerned, as summarized in the introduction to the BNB *Supplementary schedules*:

Lack of specificity is on the whole a less serious hindrance to the use of a classified catalogue than poor order, but if a great quantity of material has to be placed merely 'at the nearest general head' it becomes impossible to supply a precise direction from the subject index. Also, if the classification has been partially developed, classing 'at the nearest general head' can itself produce illogical order. (15)

With computerized systems, of course, this problem need only apply to shelf arrangement.

2.5 Updating

It will be necessary regularly to extend the controlled vocabulary in a small proportion of cases, or more substantialy over longer periods: in other words, the basic list should not be regarded as "closed". Examples of the kinds of instance which might justify an update or a special entry would be –

(a) concepts newly introduced into the field which are likely to become the subject of increased critical attention;

(b) ambiguous terms and common synonyms or near-synonyms.

One approach, having prepared a classification to a broad level only, might be to treat the most specific listed terms as terms in array, and then to introduce a rule stating that additions to the controlled list should be made for terms in array (*ie* coordinate terms) but not for more specific sub-classes.

Continued editorial control over the language of a given field is obviously important for retrieval, and for this reason the MLA Thesaurus in particular would be welcomed. The introduction of a large number of specific terms into the language of the field need not be threatening, if the experience of INSPEC, which indexes material in certain scientific and technical fields, is any indication[3]. INSPEC indexes about 4.500 documents per week, using a controlled vocabulary (the INSPEC Thesaurus) and free terms: an average of about 12 free terms is introduced per document, and there is a very low level of duplication among them. In spite of this, when the INSPEC Thesaurus is revised every two years, only a very small proportion of the free terms is absorbed. A large number of these terms prove to be redundant or just variations on the controlled terms, rather than 'new' concepts.

There appears not to be anything to suggest that this kind of approach in literature studies would be significantly different in its effect or in its demands.

2.6 Hybrid vocabnlaries and full text

The trend towards computerization of databases, whether online or not, appears to be the catalyst for thinking on hybrid vocabularies, largely because of the advent of full-text databases: –

. . . some information services that offer free-text searching of their databases arrange for some degree of enhancement, with human indexers assigning additional terms and instituting vocabulary controls. In effect they are hybridizing the two processes, incurring all the problems and reaping the rewards of both. (16)

Such systems, however, are distinctly different because they are full-text and do not depend upon document surrogates. This means that searching may be first and foremost free-text, with the vocabulary control very much an afterthought, an add-on which is designed to compensate for inadequacies. Nonetheless, the idea is obviously very similar: it springs from the same need to address the recall and precision dilemma without either getting over-involved in technicalities or sacrificing quality and ease of use – a "pragmatic compromise", in the words of Ernest Perez. Perez outlines the use of

. . . a small, controlled vocabulary of general headings in a separate record field to use as enrichment to fulltext. These necessarily broad terms will not be sufficiently precise in themselves, but when used in Boolean logical combinations with each other and with full-text terms, they can deliver the broad concept retrieval which is the weakness of straight, free-text searching . . . The specificity and precision power of the full-text remains while the general recall ability of controlled vocabulary is approached at reasonable or acceptable cost. (17)

Where free searching of full-text is not an option (for example, where document surrogates are used to comprise a catalogue) the compromise may be approached from exactly the opposite direction. A controlled vocabulary is provided, but it is not or cannot be expected to offer adequate specificity and precision, so that the enhancement device of natural language terms is adopted.

There is clear justification, in literature studies at least, for the emphasis on the control at the broader level, since erring on the side of recall rather than precision is a well-known characteristic of searching in this field, for the very reason that suitably refined systems have been few and far between.

Thus we can expect to see retrieval systems based on broad controlled vocabularies and, assuming a computerized catalogue, the provision of access to specific concepts by a form of free-text searching, within the title field and the keyword field.

2.7 The extent of the controlled vocabnlary

Clearly, the size of the basic controlled vocabulary calls for careful consideration. If it is not sufficiently large or developed it could be no advantage whatever, leading to failed searches, confusion for indexers, and a reversion to uncontrolled indexing. If the controlled vocabulary is too large, the advantages of speed and ease of indexing may have been forfeited. Ideally, partcontrolled vocab-

ulary appears to be most suitable in providing one or two freely chosen specific terms in the indexing of documents at the journal article level. Most existing humanities subject access systems are wholly unable to deal with the specificity of material at this level, but the concept of a part-controlled vocabulary represents a relatively trouble-free and inexpensive means of improving them. Of course, where a collection is allowed to grow virtually unchecked, then some regular but straightforward development of the vocabulary will be necessary anyway. The Shakespeare Data Bank, to which Schnelling refers, is likely to have to adopt some form of hybrid vocabulary, if it has not done so already: the basic list of 120 headings will obviously be very far from adequate (18).

2.8 Part-controlled vocabulary as an economic option

The economic factor should be recognized for the influence that it undoubtedly has. Literature studies is a perfect example of a field of knowledge where it is generally misguided to theorize on an "ideal" system without including the availability of funds as a major consideration. And this factor applies not only to the acquisition of hardware but also to considerations of staff time spent on cataloguing, indexing, maintaining the controlled vocabulary, and bibliographic searching. Calculations made at the Dickens House Museum, London, as part of a proposal for computerization in 1983, indicated that the maintenance (at the required standard) of a card catalogue would take approximately four times as much staff time as would be the case using a computer: this is explained, of course, by the depth of indexing which was called for and the need for innumerable analytical entries. Because part-controlled vocabulary is an economic option which promises reasonably efficient retrieval without convoluted complexity it is likely to fit well into plans for organizing knowledge in special literature collections.

3. Conclusion

The concept of part-controlled vocabulary, involving procedures which are probably fairly widely if haphazardly practised, needs to be more clearly recognized as a design option. Its use in literature studies in particular looks promising.

It seems possible that the principle might well be extended to various kinds of small special library, and even to certain areas of public librarianship. One interesting area of investigation, for example, would be in those newspaper and cuttings libraries which do not as yet depend heavily upon full-text databases: this is a field where for much of the time the imposing size of the vocabulary necessary for meaningful control has prohibited development.

Now that computerization is becoming a realistic option in special literature collections (in the UK, for example, the Bronte Society, the Dickens House Museum, and the Keats House all expect to begin automating their catalogues and procedures during 1986–87) some recognition of the potential advantages of a planned part-controlled vocabulary is surely overdue. It would also be invaluable to have more accounts, like that by Schnelling, detailing particular applications.

Notes:

1 The concept is Karl Popper's: see, for example, chapter 3, "The criterion of demarcation between what is and what is not science" in Bryan Magee, Popper. Glasgow: Fontana 1973.
2 See for example (13) and (14).
3 For information provided in this paragraph I am indebted to Mrs. Gill Wheeler at INSPEC, Hitchin, Hertfordshire.

Refereuces

(1) Schnelling, H.: Pattern indexing: an attempt at combining standardized and free indexing. In: Int. Classif. 11 (1984) No. 3, p. 128–132.
(2) Anderson, J.D.: Contextual indexing and faceted classification for databases in the humanities. In: Information choices and policies: proceedings of the American Society for Information Science, 42nd Annual Meeting; 1979. White Plains, NY: Knowledge Industry 1979, p. 194–201.
(3) Anderson, J.D.: Essential decisions in indexing systems design. In: Indexing specialized formats and subjects; ed. by H. Feinberg. Metuchen, NJ; London: Scarecrow 1983, p. 9.
(4) Harris, K.: The Dickens House classification. London: Polytechnic of North London 1986.
(5) Lancaster, F.W.: Information retrieval systems: characteristics, testing and evaluation. 2nd.ed. New York: Wiley 1979, p. 287–8.
(6) Lancaster, F.W.: Trends in subject indexing from 1957 to 2000. In: New trends in documentation and information: proceedings of the 39th FID Congress. London: Aslib 1980, p. 229.
(7) 1984 MLA International bibliography of books and articles on the modern languages and literatures. New York: Modern Language Association 1985.
(8) Mackesy, Eileen M.: Personal communication.
(9) Makinen, Merja: Personal communication.
(10) Fugmann, R.: The complementarity of natural and indexing languages. In: Int. Classif. 9 (1982) No. 3, p. 141.
(11) Wiberley, St.E.: Subject access in the humanities and the precision of the humanist's vocabulary. In: Library Quarterly 53 (1983) No. 4, p. 430.
(12) Fugmann, R.: On the practice of indexing and its theoretical foundations. In: Int. Classif. 7 (1980) No. 1, p. 17.
(13) Needham, C.D.: Organising knowledge in libraries. London: Deutsch 1971, p. 141–142.
(14) Foskett, A.C.: The subject approach to information. 3rd.ed. London: Bingley 1977, p. 284.
(15) Supplementary Classification Schedules: prepared to augment the Dewey Decimal Classification for use in the BNB. London: Council of the British National Bibliography 1963, p. 2.
(16) Rothman, J.: Is indexing obsolete? Keyword indexing and free-text searching. In: Indexing specialized formats and subjects (op. cit. 3, p. 24.)
(17) Perez, E.: Text enhancement: controlled vocabulary vs. free text. In: Special Libraries, 73 (1982) No. 3, p. 190–1.
(18) Marder, L.: The Shakespeare Data Bank: a modest beginning. In: Shakespeare Newsletter, 34 (1984) No. 2, p. 1.

Kevin Harris, British Library Research Fellow School of Librarianship & Information Studies. Polytechnic of North London Ladbroke House, Highbury Grove, London N5 2AD, England