DOCUMENT RESUME

ED 050 742 LI 002 801

AUTHOR	Hines, Theodore C.
TITLE	Vocabulary Control in Indexing the Literature of Librarianship and Information Science.
INSTITUTION	State Univ. of New York, Albany.
PUB DATE	68
NOTE	25p.; Paper prepared for the Conference on Bibliographic Control of Library Science Literature, State University of New York at Albany, April 19-20, 1968
EDRS PRICE	EDRS Price MF-$0.65 HC-$3.29
DESCRIPTORS	Citation Indexes, Classification, Conferences, Indexes (Locaters), *Indexing, Information Needs *Information Science, Library Research, *Library Science, Models, Permuted Indexes, Problems, Subject Index Terms, Thesauri, *Vocabulary Development
IDENTIFIERS	Bibliographic Control, *Library Science Literature

ABSTRACT

Problems in indexing library and information science literature occur because of the speed of introduction of new terms, the nature of class headings, and the uncertain terminology of the field. Vocabulary control requires control over the concepts selected (the depth of indexing), the form of expression of concepts and the syndetic apparatus of the index. The context in which vocabulary terms appear, subject and aspect, subject and class entry, other types of entries (author, title, series, etc.), depth of indexing, citation and keyword indexing, centralized and decentralized indexing, subject lists and thesauri, subject heading and classification are discussed. Indexing research ignores the codified record of past indexing experience including that of library subject heading work, which is the most carefully codified and tested, because of its use for a relatively shallow form of indexing. Ten general guidelines for planning indexing services for the literature are formulated, and aid to "Library Literature" for exploring new production methods and research, and expansion in staff, depth and scope, to produce a model index is proposed. (Related documents are LI 002 796 - 800 and LI 002 802 - 002 807). (AB)

VOCABULARY CONTROL IN INDEXING THE LITERATURE OF LIBRARIANSHIP AND INFORMATION SCIENCE

by

Theodore C. Hines
Associate Professor
School of Library Service
Columbia University

CONFERENCE ON THE BIBLIOGRAPHIC CONTROL OF LIBRARY SCIENCE LITERATURE

State University of New York at Albany
April 19-20, 1968

INTRODUCTION

In a sense, the title of this paper may be misleading. Except at the level of decisions about particular individual terms, there are no vocabulary control problems of indexing which are peculiar to the literature of librarianship and information science.

If librarianship were to be classified in accordance with the resemblance of its indexing vocabulary control problems to those of other disciplines, it quite clearly belongs to the soft or social sciences rather than to technology or to the hard or exact sciences.

The terminology of librarianship and information science is much more imprecise and shifting than that of science-technology. The concepts selected for indexing are, like those in the social sciences, likely to include a higher proportion of titles of works and names of persons and institutions than would be the case for the exact sciences, and a lower proportion of names of substances, procedures, and devices.

But for the purposes of this discussion, even this resemblance does not mean too much. While the proportions may differ, all types of vocabulary control problems occur in indexing the literature of librarianship. The field itself appears to be becoming, rather slowly and painfully, somewhat more of a science, at least in its more technical aspects. By its nature as a service profession, however, librarianship will inevitably remain social-science oriented as well. The field simply combines the vocabulary control problems to be found in almost all other subject areas or disciplines.

Further, the question of vocabulary control cannot, at least at this point in the development of the art, be profitably considered in any pristine isolation. The question of vocabulary control, in the broad sense, is the key question of subject indexing. The question of subject indexing is in turn the basic question of information science.

The intent of this paper, then, is not to provide answers to the question of vocabulary control in the indexing of our literature, but to indicate what some of the issues are, to try to clear up some current misunderstandings, to advance some tentative conclusions, and to suggest suitable areas for further exploration and research.

BASIC ELEMENTS

Vocabulary control may be said to involve three basic elements which it will be useful to keep particularly in mind as this discussion proceeds. These elements are:

1) Control over what concepts are to be selected, or definition of the scope of what constitutes indexable matter, and

2) Control over the form of expression of these concepts in the resulting index, together with

3) Control of the syndetic or cross-referencing apparatus of the index, together with appropriate scope notes and reverse cross-references and other appropriate indications of relationships among indexing terms.

It is the second of these elements, control over the form of expression of the concepts, which is most often assumed to be the topic when vocabulary control is discussed in the literature; yet these three are interdependent elements which cannot really be separated.

OTHER INFLUENCES

These elements are, or at least should be, influenced in turn by such factors as the size, physical form, and probable uses of the index in question. Only a small percentage of the problems of vocabulary control can profitably be decided upon in isolation from these factors.

CONTEXT

Among factors of this kind, one of the most important is often, it seems to me, ignored in discussions of vocabulary control in indexing, and should certainly be kept in mind here. This is the matter of context which is to appear with or under the indexing vocabulary terms which we employ.

Whether this context includes some things which sometimes are and sometimes are not considered as part of the indexing vocabulary itself, such as modifiers, subject subdivisions, or form divisions, is of obvious importance. It is important, too, in planning vocabulary control to know whether the references in the index appearing under the vocabulary terms are to be some form of reference number or locator which in itself conveys no meaning to the index user, or whether the reference is to be, say, a relatively full bibliographic reference in which the information about the author, title, and so on serve as further indexing modifiers or discriminators among the references listed under the particular expression of the indexing concept which is our entry or access point.

SUBJECT AND ASPECT

A few other basic distinctions may assist us on our way. I do not intend to deluge you with a group of specialized definitions, but for the purposes of this paper I would like to define a subject as the expression in words of the topic referred to, the text concept we are indexing. Only the subject proper is considered to fall under this definition, not the aspect of that subject which is treated in our text. For example, an article about the history of indexing is an article about the subject Indexing from its historical aspect, not about history. An article about circulation control in school libraries is an article about circulation control, and the aspect treated is that of this subject in a particular type of library. This distinction is easy to see in theoretical expression: it becomes a more difficult problem in practical indexing in some situations.

Please note that the fact that I have made this distinction does not mean to imply precluding entry under aspect rather than subject proper, but only that it is useful in indexing to know which is which, where possible, and to have an established policy for dealing with entry under subject or aspect.

SUBJECT AND CLASS ENTRY

A second distinction which will be useful later in this paper is one which I would like to make between subject and class entry. (Notice, please, the care with which I am avoiding the problem word specific.) Subject entry is entry directly under the subject represented in the text, or under a synonym or preferable form of expression of that subject. Class entries are those for which the indexer translates such an expression of the text subject into a larger containing class name - in which he decides to enter an article about the Enoch Pratt Free Library under Public Libraries, or even just Libraries.

Please notice that the use of class entry does not imply a classified arrangement of the resulting entries. The choice of entry procedure described does not necessarily imply faceted entries or chain indexing or alphabetico-classed entries like: Libraries - Public Libraries - United States - Baltimore, Maryland - Enoch Pratt Free Library.

Also, notice please, that I am not unaware of the philosophical truth that naming something actually constitutes classifying it, and that this is true even when the named class - a particular person or institution, say - has only a single member.

This is a concept which is not useful for our current purposes, whereas the distinction between subject as opposed to class entry is both useful and fairly common in indexing.

OTHER TYPES OF ENTRIES

Subjects - that is, the things or concepts discussed in the literature - are not, of course, the only desirable or useful indexing access points. In addition to class, or form, or geographical, or time groupings (for example, entry by such things as literary or physical form of the work, or geographical areas or time periods discussed) and similar entries, we may have entry by various other handles: by author, by title, by series, by sponsoring institution, and so on.

Expression of the first type of entries, once they have been selected for indexing, follows that for subject entries. In the case of the latter type, while we may feel that their choice is easy for the indexer, in practice there are many problems: cases, for instance, of multiple authorship, or of official subdivisions of larger corporate bodies where there is also a named individual author, and so on.

We may note briefly that, contrary to popular belief, these problems do not disappear when we can make multiple entry as opposed to being limited to a single main entry for them. Excessive multiple entry is not simply a problem of economy, it adds to the complexity of structure of the index for the user as well, and hence to other problems of vocabulary control.

FORM OF EXPRESSION OF THESE ENTRIES

For such forms of entry as author, title, and so on, we need not only to know what to choose as indexable matter, but also how to regularize the form of expression of what we have chosen, a means of vocabulary control. Notice particularly that all of these forms of non-subject access which are of particular importance in indexing the literature of a discipline may also appear in the literature - and do frequently appear in the literature - as subjects as well. For certain types of material, they may constitute the majority of indexable matter.

REGULARIZING THESE TYPES OF ENTRIES

Regularizing the form of expression of such entries is by no means a negligible problem, as librarianly discussion over rules for catalog entry shows in theory, and the difficulties posed in the Government-Wide Index in interfiling entries of this kind from different sources made under different rules vividly illustrate in practice.

With these complexities themselves I do not intend to deal, partly through cowardice and partly because the complexities are well known and have been discussed in detail for some generations - at least since Panizzi and probably since Tritheim. I remind you of them because it is the tendency of librarians to forget and of information scientists not to realize that these are not simple matters. It is not even just a question of how to identify author or title or series entries and how then to express them, but also the problem of whether such entries are or are not useful in a particular index or indexing situation - when, for example is a title non-distinctive?

MODIFICATIONS, SUBDIVISIONS, AND REFERENCES

I have briefly mentioned that indexes often include modifiers or subdivisions of topics, whether or not the access points are concepts in the text or not, and whether or not their expression is as subjects or as classes. Subdivisions are usually in some kind of regularized form, and a subdivision may include more than one reference. Modifications, on the other hand, are intended to individuate for each reference the aspect of the subject treated, and are only formalized to a limited extent, such as that of placing the most important word of the modifier at its beginning, except for prepositions or conjunctions.

The nature of the reference itself may also constitute a form of modifier or discriminatory context for the index user. We are all aware that some tools for vocabulary control - notably, for example, the Sears list[1] and the Library of Congress headings[2] - explicitly include an elaborate framework for entry subdivisions, both their nature and their form of expression.

Implicitly, these particular lists also include much more. They assume, in their structure and design, that a particular context and a particular form of expression - the unit card with all its tracings - will appear under the heading and its

subdivisions. I would like to point out that this actually constitutes a further extension of both topic subdivision and its expression, and that it is inherently part of the structure of these particular tools. The form of expression of all of these elements, and their order, constitute, subject to various exceptions requiring individual judgment or interpretation, the factors determining the arrangement of the entries and the structure of the resulting catalog.

DEFINITION OF INDEXABLE MATTER

We will further explore the use of these library vocabulary control listings later, but one other aspect of them should be noted at this point. It is assumed that they will be used to regubrize subjects chosen by a particular definition of the scope of indexable matter in the universe of material to be indexed, in this case, books to be subject headed. The definition of the scope of indexable matter for this purpose we probably owe to Cutter,[3] and Kaiser[4] commented on it in the same context of discussing the indexing of the literature of particular topics which concerns us today.

What is indexable matter in assigning a subject heading to a book is the subject or subjects of the book taken as a whole entity, not those subjects which may be discussed at various points within it. A book which is about library cataloging may contain the best information in the library about subject headings, but by the library definition subject headings as treated in this book do not constitute indexable matter.

The principle, of course, may be and is extended to smaller units than books - to journal articles, reports, book chapters, and so on - but it is always applied to the bibliographic unit selected as a whole, and not to the information contained within that unit.

I may be unduly belaboring something which is already quite obvious to an audience which consists mainly of librarians, but in the broader field of indexing misunderstanding of this point appears to me to have caused both confusion in developing indexing systems and a significant failure to take maximum advantage of library experience in vocabulary control. Let me, even in a paper which already threatens to become overlong, develop this point. It is important to what is to follow.

DEPTH OF INDEXING

Library cataloging practice, especially in relation to vocabulary control, has not really been very seriously considered by modern indexing theorists and those in information science, although it seems to me that it has much to contribute. This may be because the library concept of indexable matter leads to "shallow" rather than "deep" indexing. Although there is no really satisfactory method of counting what constitutes an "entry" in an index on a uniform basis, there sometimes seems to me to be a further mistaken idea held by those who make this criticism that a multiplicity of index access points, regardless of their nature, makes an index "deep" and therefore somehow good. This is not the point I wish to make.

Library subject cataloging does indeed have a concept of indexable matter - the subject or subjects of the bibliographic unit as a whole - which produces few entries indeed as compared with the number required for intensive indexing of the literature of a discipline. We need only to think, for example, of the subject indexes to Chemical Abstracts, where new chemical information constitutes what concepts must be chosen for indexing[5], to see that this is the case.

FORM OF EXPRESSION

Defining indexable matter, however, as we noted at the beginning of this paper, constitutes only the first of several aspects of vocabulary control, the selection of concepts or things for indexing. The second, as you will remember, is the control then of the form of expression of that indexable matter, the area usually thought of as vocabulary control proper as exemplified by the use of heading lists or thesauri.

It is in this area that the library experience has been very great indeed. Building upon Cutter's[6] all too briefly expressed basic outline, we have now nearly a hundred years of experience in this particular art, codified in innumerable lists, with learned commentary by such experts as Haykin,[7] Metcalf,[8] Frick,[9] and, among my own immediate colleagues, by Tauber,[10] Frarey,[11] and Lilley.[12] The question of depth of indexing should not be allowed to prevent proper use of this experience in meeting indexing problems.

THE IDEAL INDEX

An ideal index would have its scope clearly defined in a number of ways. The interested user would be able to determine what sources were covered, what concepts or types of concepts had been selected as indexable matter, what the form of expression of the concepts would be, how this expression would be modified or subdivided, precisely what context would accompany the resulting entries, and exactly what the arrangement of the entries was. All of these things would be definable in terms so rigorous that, barring minor clerical slips, one indexer using the same criteria should be able exactly to replicate what another indexer had created as an index to the same material. Insofar as this were possible, we might be able to call indexing a science rather than an art.

In point of fact, as a number of studies show, conventional indexing falls far short of this ideal, even when the test is to have the same indexer re-index the same material under the same conditions but after a lapse of time. A significant number of the differences which arise seem to be due to problems of vocabulary control. At least one study, that by Dr. Ann Painter,[13] would indicate that librarians applying library-style cataloging rules are somewhat more consistent in the entries they produce than are other types of indexers.

The extent to which this is due to the definition of indexable matter, or to the types of controls of form of expression involved is not discussed as such in her study, but it might be supposed that both have a role to play.

EXISTING INDEXES MEETING "IDEAL" CRITERIA

There are, of course, existing indexes which come close, very close indeed, to realizing most of the criteria of exact definition which I have given as necessary for truly scientific indexing, and for which a second indexing would produce an index almost if not exactly indentical with that produced the first time.

The indexes which approach this ideal of rigor in the definition of the procedures followed are indexes of the keyword in-or-out-of-context type, or indexes following the citation indexing principle.

As a means for indexing the literature of a discipline so as to permit reasonable retrospective search, these tools turn out to be rather mediocre indexes if you add one additional criterion to the above: that the index should also offer, within

a reasonably usable compass, at least as good access to the material indexed as those indexes which we subjectively recognize as "good" manual indexes.[14]

In many ways, of course, the actual examples of these methods of indexing are neither as pure nor as simple as they appear to be on first acquaintance.

CITATION INDEXES

A citation index restricts the concept of indexable matter very narrowly indeed, to the works cited by the works indexed. In practice, however, repetitions of citations given in a single article are suppressed, and judicious elimination of some citations not (by subjective judgment) of value to the index user might both lower the bulk of the resulting index and improve its usefulness. Citations must be regularized in form to permit their arrangement and merging in a citation index, forms of author names must be determined, and even these steps are not simple, nor readily to be done without the exercise of human judgment on each individual item, or without what amounts to vocabulary control lists of acceptable abbreviations and journal names.

TITLE KEYWORD INDEXES

There have been a number of studies, perhaps the best-known of which is that by Montgomery and Swanson, which when read superficially appear to indicate that keywords chosen from titles correspond closely to human indexing of the same material. The Montgomery and Swanson[15] study found that 85.8% of titles in the Index Medicus contained either the index term used or its "synonym." A replication of the study done at Columbia[16] indicates that it included as synonyms many very broad classes to which the index term belonged and vice-versa, and that if synonyms were more conventionally defined only fifty-odd percent of the titles contained the indexing term or a synonym for it. This figure agrees more closely with other studies of the same kind than are less often cited.[17] In addition, of course, synonymy is one of the major problems causing scatter in indexing, which vocabulary control systems are intended to minimize.

None of the studies known to me which compare title keywords with more conventional subject indexing employing vocabulary control consider the problem of subarrangement of entries under keywords, although there has been considerable research by Tukey[18] and by Kollin[19] in seeking rigorous means of producing

a reasonable subarrangement within keyword-from-title indexes. Kollin has also sought to deal with synonymy.

For practically all keyword-in-context indexes, the subarrangement is accidental, in that it is based on the word following the keyword itself. This at least means that if the concept is expressed by a multiple-word term, such terms fall together in the index. Not even this much is true of keyword-out-of-context indexing.

For machine purposes, most existing keyword-out-of-context indexes are subarranged by an accession or other number not logically useful to the user. This is not a necessary restriction, of course, but by definition indexes of this kind do not provide subject subdivision or modifiers, or even subarrangement by author or title.

It is certainly true that we still lack anything resembling rigorous means of evaluating indexes or indexing methods, despite the substantial contributions to the literature since the Cleverdon studies.[20] It does seem to me, however, that studies indicating inconsistency in human indexing are not by any means an adequate argument for abandoning attempts to secure consistency in entry expression, nor an adequate argument for achieving consistency in the choice of indexable matter by restricting that choice to words - not even concepts - which happen to appear in titles.

For the moment, it seems to me that the most convincing arguments against title keyword indexing as the means for providing an index for retrospective searching of the literature of a discipline are subjective and circumstantial. They are nonetheless quite convincing, as I think anyone who tries to consult cumulations of B.A.S.I.C., the index to Biological Abstracts, will find if he tries to check such subjects as Blood, or Rat, or Rats. Since there is even less correspondence between keywords from titles and concept subject indexing in the social sciences than in the physical sciences or life sciences, the method certainly does not seem promising for indexing of the literature of librarianship.

Real indexes based on keywords taken from titles are often enriched (added to) where the title is not expressive, have forms of words in titles altered on input by human beings, hyphenate where Webster would not in order to make subjects expressed in more than one word arrange as subjects rather than as isolated words, and involve other deviations from the pure definition of their scope and expression. It is probably quite safe to say that these changes made to improve the form of expression of the subject are generally in the direction of similar decisions earlier made by those compiling manual indexes and catalogs, and recorded in such tools as the library sub-

ject heading lists. Such deviations require human judgment and add to the cost as well as the quality of the resulting index. I know of no studies of amount of alteration in existing indexes of this kind, or of the add-on cost of various amounts of change.

Essentially, too, insofar as they remain rigorously defined, title keyword indexes shift off onto author and/or editor both the choice of indexable matter and the form of its expression. Even when e authors, as was the case with at least one set of the pr. edings of the American Documentation Institute (now the American Society for Information Science), are warned in advance that this type of indexing will be done, are information specialists or scientists themselves, and are highly motivated to bring their papers to the attention of their colleagues through the index, the results are neither suitable for use without considerable post-editing, nor very satisfactory even for a small index even when this post-editing has been done - at least, as compared with our subjectively "good" manual index with anything approaching the same number of access points. For indexes of this kind we still retain for an indexer, too, the problems of regularization of author names, titles, and so forth.

STICHWORT ENTRY AND ROTATIONAL INDEXING

We learn from history, Spengler tells us, that man learns nothing from history. Stichwort indexing - the idea of getting a kind of subject indexing by entering under the most important subject word in the title, and inverting the title to provide the necessary context - is a very old idea, going back at least to the late 15th century, which is still practiced today. Indeed, it is often practiced in the indexes issued by very learned scientific journals whose pages urge improvement in indexing techniques for the scientific literature. A great deal of the indexing in Poole[21] was essentially Stichwort. Crestadoro[22] even suggested a technique he called rotational indexing - or making Stichwort entry on all of the (manually determined) "important" words in the title. Such techniques were not successful, primarily for lack of adequate vocabulary control.

AUTHOR AND EDITOR INDEXING

It would, of course, be nice if authors and editors supplied titles with everything in them regularized and which expressed what the text was about, even though this would still not supply useful indexing handles for all of the index-

able matter which might be required and economic for a particular index. But the idea is less enchanting than it at first appears, at least for other than use in relatively small listings intended for current awareness purposes.

To make such a system a success, we would be confronted with the problem of teaching all authors, or at least all editors, how to be indexers. It is hard enough just to teach indexers this, despite the fact that they are quite properly more strongly motivated toward indexing than we have any right to expect authors or editors to be, and know more about the structure and possible uses of the particular index into which their entries must fit. For that matter, they also know that theirs is a particular index with a particular set of users. As we have seen, keyword-from-title indexes, however regularized, cumulate badly, and are capable of growing only to a certain size without becoming unmanageable to use for lack of meaningful subarrangement. This is a matter of vocabulary control in the provision of subheadings or modifiers with a role in determining overall index structure as well as a vocabulary control problem in subject expression.

THESAURI FOR AUTHOR OR EDITOR USE

Related to the idea of having editors or authors provide "proper" titles for keyword indexing is another currently popular suggestion. This is that we provide editors, or authors, or somebody, with lists of regularized words or terms, or thesauri, or subject heading lists, and have decentralized production of indexing terms which will appear with the article or report and may subsequently be centrally filed, or filed by the user, thus producing instant indexes.

While a number of journals, particularly engineering journals, have begun to include entries from a list of this kind with the articles when they are published, I am so far aware of only one which uses these entries in the index published by the journal itself. At least one publisher of technical journals has included the entries with articles in some of its journals, but does not use them either in the indexes published for the journals or in the extensive in-house indexing of its journals. As far as I know, no index covering a group of journals uses the index entries thus produced, and I know of no special libraries using them. The approach does not seem to have caught on, and this seems likely to be for a combination of reasons involving vocabulary control: lack of definition of indexable matter in using lists, possible unsuitability of the lists, and lack of an indexing structure into which entries can be fitted.

Decentralized indexing today, then, does not seem to prove notably more successful - perhaps not as successful - than decentralized indexing efforts in the past, of which Poole's index is perhaps the outstanding example. In Poole's case, of course, the index was centrally edited, and he used experienced catalogers and indexers, but lacked a list for vocabulary control.[23]

SUBJECT LISTS AND THESAURI

Current thesauri or vocabulary control lists seem to have an interesting difference from library subject headings and information file listings which I have not seen previously discussed in the literature. Most of them, despite provisions for updating and for the addition of new terms are, in comparison with the library lists, actually classifications arranged in alphabetical order.

Let me see if I can clarify this difference. In using a library classification scheme, despite its synthetic aspects, what we are basically dealing with are rows of pre-established pigeon holes, and our task is to place our item in the most appropriate one. When we use a library heading list, our approach is first to determine the subject of the work and then to use the list to regularize its expression or, where experience has shown this to be necessary or desirable, to classify the work in some way instead. But if the subject has not been given in the list and we are not told by analogy with the list to deviate from our general instruction to enter under the subject, we create our own heading in the spirit of the list, add it to the list, and go on with our work.

It is true that, for some types of entries and by some subject headers, this procedure is ignorantly more breached than observed. But in most cases, the principle is clearly followed: a book about Man o' War is not entered under Racehorses, nor a book about Mt. Washington under Mountains, nor even under Mountains - U.S., despite the fact that neither Man o' War nor Mt. Washington appear in the lists.

VOCABULARY CONTROL OR VOCABULARY LIMITATION?

For various reasons - mostly, I suspect, technological in inception and only secondarily intellectualized - most thesauri would have the indexer place each concept he chooses for indexing under an existing heading in the list - the thesaurus of the American Petroleum Institute is the only exception to this known to me.

Placing each concept under an existing heading is an act of classification in that, as is the case with library classification schemes, it means that the indexer must choose an existing pigeonhole from an array before him; the only difference is that the arrangement of the array, and of the entries in the resulting index, is alphabetical rather than classified.

This leads to an important point about the nature of library subject heading lists for vocabulary control, a point which carries over to published indexes such as the Wilson indexes which follow similar principles for vocabulary control. It is clear that, at least for very large classes of subjects, their presence on the list is implicit, even if they are not actually printed in the list or have not been previously used in the index or catalog.

Subjects of this kind include the names of persons and institutions, geographical names, particular species of birds or animals or fishes, kinds of games - the list is literally endless, even if we do not include types of subjects which are sometimes (mistakenly, in my view) entered only under broader class headings by some catalogers and indexers: names of computer languages, names of particular chemical compounds, or names of particular devices or machines. This is sometimes justified in library practice by stating that direct entry should be made down to the level of the species, but that varieties should be given class entry.[24] This may be easy to see in biology, but becomes more difficult and less readily justifiable in other subjects.

Note that in the library lists, or for the bulk of the indexing done in the fashion of, say, the Wilson indexes, control of the form of expression of concepts not previously indexed or entered in vocabulary control lists is done either by rule (as in the case of using library entry rules for names of persons and institutions, for example), or by analogy with previous indexing decisions recorded in the list or index, or by using another index or list or reference work as authority (Chemical Abstracts for names of compounds, for example, or a particular gazetteer for place names).

In the latter case, the authorities chosen become in fact extensions of the heading list or thesaurus, effectively extending the list for vocabulary control purposes by literally millions of items without swelling its bulk. Where individual decisions must be made for subjects, these may be added to the list at the time of first need and become authority for the form of expression of future occurrences of the same concept when it again appears as indexable matter. All entries are, of course, added to the index itself, where they may serve as authority just as if they were in a separate listing.

Notice that this method of vocabulary control certainly does not preclude classed entry of the concept in the index in place of or in addition to (as usually seems preferable to me as it did to Kaiser[25]) subject entry proper, nor does it prevent the use of the same list for the control of the vocabulary of expression of the classed concept.

Class entry is not usual in library catalogs, with certain well-established exceptions, such as alphabetico-classed entry for historical topics. To some extent the lack of classed entry is made up for by the classed shelflist and classed array of entries on the shelves. This is not true of indexes to classified abstracting services, since the classification in this case is usually broad rather than narrow and is, in any case, not cumulated. It serves instead to provide groupings which are readily scanned for current awareness purposes.

In indexing, classed entries or classed arrays are frequently more desirable either to augment subject indexing or to group some types of entries for special purposes, generally, based on common needs of users of the particular index or indexing service.

Perhaps because of the concurrent use of shelf classification, library heading lists (although they include a syndetic apparatus which serves some of the same purposes) do not include the kind of classification of the headings themselves found in some thesauri expressed as "broader terms" and "narrower terms" or "generic for" and "specific to". Cutter felt that a classified listing of library headings would be very useful, but was too difficult and expensive to maintain. The experience of current thesaurus builders may be helpful in answering these questions. It will be interesting to note, too, how these listings will deal with topics whose class relationships, at least for an index of broad coverage, are not, as Haykin puts it, "obvious or common:" ink, for example, or the White House.

The Sears list did and the Library of Congress list does include, however, classification numbers from Dewey and Library of Congress classifications respectively. These were intended primarily as scope notes or suggestions for classification, however, not to provide a classification of the headings. In the case of the Library of Congress list, they were intended to be included when and only when, the class was co-extensive with the heading, but this has been done only inconsistently.

In the thesauri, it does not seem to me to be clear exactly how this classed apparatus, as opposed to the syndetic apparatus carried over from library schemes, is intended to be used: that is, whether it is intended for the indexer - and if so, in what way - or for the user of indexes based on the thesaurus. Where

indexes are maintained in machine-readable form, and where entries are automatically also posted to the next upward step in the hierarchy on the machine-readable record (though not, for reasons of bulk, included in the printed indexes), it will be interesting to see the extent and nature of use of this feature.

In a sense, too, these classifications, in the thesauri, serve, in the list though not in the index, as a kind of upward cross-reference of the type usually avoided in indexes. If users actually do employ the thesaurus as an aid in searching, as appears to be the intent of some of these lists, and as might be made necessary by the very extensive use of class entries, we may be able to test the effect of upward cross-references. While upward cross-references have been advocated by some cataloging experts[26] they seem to propose selective and judicious, rather than overall, use of such a feature.

A separate heading list or thesaurus is not, of course, required to achieve exactly the type of vocabulary control associated with listings. Heading lists grew from the common indexing practice of achieving regualrization of expression and guidance in the choice of indexable matter by consulting the previous indexing used in the same index or indexing service, or by consulting other indexes upon which their own may be modeled, as catalogers frequently consult the Wilson indexes for form of expression for new subjects. Provided additional apparatus such as a record of reverse cross references and scope notes is provided, within or outside the index, the result is the same as with a separate subject authority listing, though sometimes less convenient to use. The use of the index itself, as guidance to interpretations of the list, for example, is usually an essential aid to the indexer even when a thesaurus or list is maintained.

The separate vocabulary control listing is desirable for convenience, as an aid to starting a new index, as a device to standardize form of expression across indexes (even indexes with entirely different interpretations of indexable matter), as an aid in creating local entries which will fit with those issued by a centralized service, and as a place to record decisions or control apparatus (reverse cross-references, scope notes, etc.) which would swell the bulk of the index itself without aiding the user. A separate authority list, too, is easier to edit and serves as an easier-to-use record of chances in entry form.

If we seek to make a general assessment , however premature this may be, of the new thesauri, and to judge them on the basis of their success in use, it becomes evident that at least some of those most discussed have never been used on any significant scale. Those which seem to me to be the most successful

are those which have been based, in their form of expression of terms and the choice of terms to be included, upon actual indexing practice and experience as well as upon the useful advice of subject experts.

Of the larger and more ambitious listings, nearly all of the most successful have been based on substantial library or indexing practice or substantially refined after actual use: Medical Subject Headings (MeSH), the Bureau of Ships Thesaurus, the American Petroleum Institute Thesaurus and the National Aeronautics and Space Administration Thesaurus. Some of these, particularly the NASA Thesaurus, seem to have gone beyond the library lists in important respects: the provision of a classed listing or a listing by broad categories; of a separate list of subjects to be used; of indications of the broader and narrower class relationships of topics separate from the syndetic apparatus proper; of permuted listings of the terms and of different and clearer control terminology ("use for', and 'refer from', for instance - although the latter was formerly used in the Sears list).

In other respects, they seem still to lack important features to be found in the library listings, or their related apparatus. There are few clear, or at any rate published, explanations of the way in which indexable matter is to be selected. The discussions of this in Haykin and in Sears are certainly not completely unambiguous, any more than the definition used by Chemical Abstracts indexers, but they do seem reasonably functional. Perhaps the most outstanding lack in the thesauri is that of adequate provision for subject subdivision. I may be jumping to conclusions too early, but it is probably safe to say that the indiscriminate use of roles and links is dead; that their effective use in future indexes will be selective, more infrequent than has previously been advocated, and designed for machine use, not publication.

It is evident that much research needs to be done before vocabulary control in indexing can become anything resembling an exact science. Since indexing vocabularies are inevitably linguistic in nature - and this applies even to classification schemes - and since the material to be indexed is expressed in language, it seems impossible, in the same sense that machine translation is impossible, that there will ever be an exact and rigorous means for carrying out the total task. This makes it all the more important to develop exact and rigorous methods which contribute to useful indexing in whatever areas this is possible, so as to limit those areas in which separate judgments for each item are required.

Much of current indexing research which has to do with vocabulary control seems to have been done ab ovo, without regard to the codified record of the information we have gained empiri-

cally through years of experience. Because this record has been more carefully codified, and tested over a longer period of time, the library experience in subject heading work is probably that which has the most to contribute. It seems to have been largely ignored because it is primarily used for a relatively shallow form of indexing in the definition of the scope of indexable matter.

TENTATIVE CONCLUSIONS

It would seem useful, then, to offer a group of tentative conclusions about vocabulary control based largely on that experience, but considered and presented in the context of indexing the literature of librarianship.

1) Previous indexing, provided an adequate record of syndetics is maintained, can constitute as rigorous a control of vocabulary as any listing especially designed for that purpose, although it may be less convenient to use.

2) Limitation of the size of vocabularies may be required for technical reasons. If this is the case, however, the reasons for the limitation, and its nature, must be clear to the indexer and to the interested and concerned user.

3) While it is useful to have expert assistance in defining the scope of particular terms, or in suggesting terms for inclusion in a list, this can constitute only a strengthening of, not a substitute for, building a vocabulary from material like that to be indexed, fitting the terminology into a particular indexing structure.

4) Vocabulary requirements, particularly the questions of class entry, of class and subject or aspect subdivision or modification, and the nature of the reference, vary widely with the estimated size of the index, whether it is designed to be cumulated or not, and whether or not it is to be periodically closed and post-edited or not.

5) Experience would appear to show that, at least in the current state of the art, indexing for a discipline like librarianship to provide for both current and retrospective search cannot achieve sufficient vocabulary control to be as useful as the best of existing tools by means of citation indexing or title keyword indexing, although the former may be a useful adjunct for certain types of searches in depth, and the latter may be a useful current awareness tool, particularly in indexes of relatively small size. It would seem premature to invest in either before we have adequate investment in current, on going indexes.

6) While post-coordination of indexing terms may be a very valuable supplement to published conventional subject indexes, their vocabulary requirements appear more and more to be the same as those of more traditional forms of indexes.

7) Author, editor, or other source indexing, other than by encouragement of the use of more explicit titles and greater bibliographic regularization for use in title keyword indexes for current awareness purposes, does not appear to offer a practical solution to the problem of producing an index to the literature of a discipline.

8) Lacking a far greater knowledge of the indexing process than we have at present, decentralized indexing, with or without a thesaurus control device, for use in smaller journal or report indexes and then later centralized cumulation into an index to the literature of a discipline, does not appear practicable. The opposite procedure - centralized indexing for an index to the literature of a discipline, producing index entries which can also be used, for example, to provide indexes for individual journals, appears more likely to be practicable at present, although it would require careful design and considerable experimentation.

9) Large indexes designed for cumulation and retrospective search are not practicable without some form of subject subdivisions or modifiers to produce useful subarrangement of the material.

10) It seems possible to make a number of specific statements about the actual form of expression of indexing terms:

a) Entry terms should be in the form of expression most likely to be known to the user, with reference from other forms: ASLIB rather than Association of Special Libraries and Information Bureaux; COBOL rather than Common Business Oriented Language.

b) In general, entry forms should be in the plural rather than in the singular, and word-by-word filing should be used. While some very successful large indexes use the singular and letter-by-letter, they all require more filing modifications than might otherwise be necessary. The singular is often used because it seems to produce, expecially with letter-by-letter filing within the subject proper, more classed groupings than entry under the plural. The plural form, however, is not only a more natural way of expressing a topic or subject (a work refers to computers, not to computer), but permits useful discrimination in those cases where the singular connotes the general and the plural particular aspects of it (Engraving, Engravings).

c) Homographs should be qualified by an expression indicating what is meant; placing the expression in parentheses is a widespread means of doing this, and might be accepted as standard.
d) Many concepts cannot be expressed except as phrases, and should be so expressed.
e) Headings composed of adjective and noun should not usually be inverted, particularly if the purpose of the inversion is only to achieve a classed arrangement of the entries.
f) Subdivisions of or within a subject intended to constitute conceptual and, therefore, arranging breaks, should be clearly indicated by punctuation, typography, or spacing.
g) Subjects may be subdivided by facets, aspects, generalized classed groupings (applicable to a range of similar subjects), ad hoc classed grouping, or by tailored modifiers, like those in Chemical Abstracts subject indexes. Subdivision should take into account the number of entries likely under a topic, display or lack of it; and the usefulness of the division methods chosen, as well as problems of cumulation. It might be pointed out here that the library lists are quite sophisticated, providing subdivisions which may be used with any reading subdivisions printed once which may be used with any of certain classes of heading, and divide like instructions. This is far more sophisticated than such devices as roles, obligatory sub-faceting in some proposed faceting schemes, and is an area which deserves and requires further research.

Again, these observations must appear commonplace to an audience of librarians, but many of us seem attracted to more sophisticated systems, or machine-based systems, where other aspects of indexing may be better performed than in conventional indexing, but where we are accepting advice from less experienced or qualified people in questions of vocabulary control, often, again, erroneously believing that the advice reflects equipment requirements.

It seems unquestionably true that newer techniques, primarily the use of machine readable copy, computer-based production of indexes, and the use of computer facilities for special-purpose searches and bibliographies can lower unit costs, improve speed of production of indexes, and provide for greater flexibility in the indexing process.

In seeking to extend control over the literature of librarianship and documentation, however, we should keep in mind that we need a flexible tool for retrospective searching, and that more rapid indexing for the major existing indexing service would either enable concurrent production of a current awareness

service at a reasonable cost, or substitute for it.

Further indexing vocabulary analysis would be desirable for library literature, unquestionably, but it would seem sound to base this on Library Literature, the only substantial index in the field, and the only index with an established, comparatively sophisticated vocabulary control, and with the only list of headings used for substantial amounts of literature in the field. The major vocabulary problems now seem to be those of speed of introduction of new terms, the nature of class headings, and the uncertain terminology of the field.

In the interest of increasing the access to our literature, then, it would seem most reasonable to build upon existing strength particularly when we consider the present high quality level achieved with such a small staff. I would propose aid to the Wilson Company in exploring new production methods, and for research in indexing vocabularies procedures, as well as urging support for expanding the staff, depth, and scope of their index to be Library and Information Science Literature, through subsidy if necessary, to make it a model index.

NOTES

1. Sears List of Subject Headings. 9th ed. Edited by Barbara Marietta Westby, with Suggestions for the Beginner in Subject Heading Work by Bertha Margaret Frick. New York, H.W. Wilson, 1965.

2. U.S. Library of Congress. Subject Headings Used in the Dictionary Catalogs of the Library of Congress. 7th ed. Washington, 1966.

3. Cutter, Charles Ammi. Rules for a Dictionary Catalog. 4th ed., rewritten. Washington, GPO, 1904. (U.S. Bureau of Education. Special Report on Public Libraries, pt. 2) pp. 23, 66-67.

4. Kaiser, J. Systematic Indexing. London, Pitman, 1911. (The Card System Series, v.2) paragraph 74.

5. Bernier, Charles L. "Indexing and Thesauri." Special Libraries 59(2), February 1968. p.100. Bernier also notes emphasis and extensive review as other criteria for selection of indexable matter.

6. Cutter, op.cit., pp.66-77.

7. Haykin, David Judson. Subject headings; a Practical Guide. Washington, GPO, 1951. This may serve as a manual for the Library of Congress Subject Headings, although there are differences from Library of Congress practice.

8. Metcalfe, John. Alphabetical Subject Indication of Information. New Brunswick, N.J., Graduate School of Library Service, Rutgers- the State University, 1965. (Rutgers Series on Systems for the Intellectual Organization of Information, v.3) See also his: Information Indexing and Subject Cataloging. New York, Scarecrow Press, 1957. Also: Subject Classifying and Indexing of Libraries and Literature. New York, Scarecrow Press, 1959. Metcalfe is contenticus, interesting, and basically sound.

9. Frick, Bertha Margaret. "Suggestions for the Beginner in Subject Heading Work" in Sears List (see note 1), pp.14-29.

10. Tauber, Maurice F., ed. Subject Analysis of Library Materials. New York, School of Library Service, Columbia University, 1953, is still very valuable. Se also, for example, Chapter X of his Technical Services in Libraries. New York, Columbia University Press, 1953. pp.150-176.

A very considerable amount of research in this area by Dr. Tauber and his students and associates has been done over the past 15 years, and is continuing.

11. Frarey, Carlyle J. Subject Headings. New Brunswick, N.J., Graduate School of Library Servi..., Rutgers- the State University, 1960. (The State of the Library Art, v.1,pt.2) See also Tauber, Subject Analysis. pp. 147-166.

12. Lilley, Oliver L. Terminology, Form, Specificity and the Syndetic Structure of Subject Headings for English Literature. New York, School of Library Service, Columbia University, 1959. 2 v. Unpublished doctoral dissertation. Lilley has published a considerable number of important papers in this area.

13. Painter, Ann F. Analysis of Duplication and Consistency of Subject Indexing Involved in Report Handling at the Office of Technical Services. Contract CC-4753. Washington, D.C., Office of Technical Services, March 1963.

14. I am indebted to John O'Connor for this useful, deceptively simple, and most valuable benchmark. See his Mechanized Indexing; Some General Remarks and Some Small-scale Empirical Results. Philadelphia, Institute for Cooperative Research, University of Pennsylvania, (1960?). Information Systems Branch, Office of Naval Research Contract No. Nonr 551(35).

15. Montgomery, Christine and Don R. Swanson. "Machinelike Indexing by People." American Documentation 13(4), October 1962. pp.359-366. The precision of this figure (85.8%) is interesting.

16. Miller, Doris. Index Medicus; Feasibility of Subject Indexing by Computer. New York, School of Library Service, Columbia University, April 1967. Unpublished paper. Miss Miller examined only 582 titles, and found that only 214 or 37% of these titles had either exact replication, plural/singular difference (16 titles) or dictionary synonyms (27 titles) of the headings.

17. O'Connor, John. "Correlation of Indexing Headings and Title Words in Three Medical Indexing Systems." American Documentation 15(2), April 1964. pp.96-104. O'Connor's correlations were 19-45% for one service, 40-88% for aother, and 13-39% for the third. See also Kraft, Donald H. "Comparison of Keyword-in Context (KWIC) Indexing of Titles with a Subject Heading Classification System." American Documentation 15(1), January 1964. pp.48-52.

Kraft found that 64.4% "of the title entries contained as keywords one or more of the ... subject headings under which they were indexed." p.48. By including all terms which might conceivably by regularized by rule and/or extensive dictionary lookup, Miss Miller got a top figure of about 75% of titles in her sample, thus neatly bracketing Kraft. Her lower figure corresponds closely with O'Connor's.

18. Tukey, John W. Unpublished paper presented at the American Society for Information Science annual convention, New York, 1967.

19. Kollin, Richard C. Personal communications. The results of this work may be seen, however, in Pandex, a recently issued multidisciplinary index. New York, Pandex, Inc., 1967- .

20. Rus, Alan M. "Evaluation of Information Systems and Services." In Cuadra, Carlos A., ed. Anual Review of Information Science and Technology. vol.2. N.Y., Interscience, 1967. pp.63- 86.

21. This was actually a series of indexes. See Winchell for a description. The series indexed well over 1/2 million articles. Poole felt the need for a heading list, and probably was responsible for the first American Library Association list. See Poole, W.F. "Plan of the New Poole's Index." Library Journal 3, May 1879. pp.48-57.

22. Crestadoro, A. Art of Making Catalogues of Libraries. London, 1859.

23. Poole, op.cit. p.49.

24. Haykin, op.cit. p.11.

25. Kaiser, loc.cit.

26. See Frarey, Subject Headings, pp.39-46.