Soergel

A GENERAL MODEL FOR INDEXING LANGUAGES: THE BASIS FOR COMPATIBILITY AND INTEGRATION

Dagobert Soergel School of
Library and Information Services

University of Maryland
College Park, Md. 20740

Paper presented at the International Symposium "Subject Retrieval in the Seventies New Directions" at the University of Maryland, School of Library and Information Services, May 14-15, 1971.

Summary

Classification theory is divided into two areas: analysis of conceptual structure and file organization, and the primacy of the first is stressed, A model for conceptual structure in terms of concept coordination and polyhierarchy is sketched, Some problems of file organization, namely post-coordination vs. pre-coordination and synthetic vs. enumerative schemes are discussed in relation to this model. A model for a classification scheme for different kinds of file organization is then proposed. The scheme would consist of a "core classification scheme" made up of elemental concepts and an "extended classification scheme" made up of combinations of elemental concepts. While the core scheme would be universal, extended schemes would be developed as needed in a specific application. This would make for flexibility while maintaining inter-system compatibility.

0 Introduction

The purpose of this paper is to give a perspective, not new results. It tries to put into perspective the problems of classification theory. These problems can be divided into two major areas: conceptual structure and file organization. It seems to this writer that classificationists have concentrated too exclusively on file organization and too often have looked on conceptual structure from the point of view of file organization and not as an area to be considered independently. This imposed many restrictions on the consideration of conceptual structure, and many aspects important for information retrieval have not been brought out. This might be one of the reasons why the results of classification theory have been neglected or sometimes have been reinvented in a rather amateurish manner in mechanized information retrieval systems where the restrictions imposed by file organization are by far less severe than in manual systems.

Contrary to this attitude we take the following position: the first, primary and basic task is to understand conceptual structure and its functions in the retrieval process. We say again that this task should be performed without any reference to the limitations imposed by particular kinds of file organization. File organization is the secondary, technical, almost ancillary task. File organization has to put into effect the insights gained from the analysis of conceptual structure for actual application in performing searches as far as is feasible with the equipment available in the particular system. It should be obvious that problems like pre-cordination and post-cordination, synthetic vs. enumerative schemes or alphabetical vs. classified order are problems of file organization. Whatever the file organization is, it should be based on the same conceptual structure. As we shall see later, this will increase considerably the effectiveness of information retrieval systems. Furthermore, this principle would serve to maintain compatibility between information retrieval systems with different kinds of file organization (e.g., a peek-a-boo file and a card catalogue).

1 Conceptual structure: concept coordination and hierarchy

1.1 Hierarchy

Due to schemes like UDC, DDC, and LC classification, misconceptions of hierarchy are widespread. Hierarchy is not a strait jacket in which the universe of knowledge has to fit somehow or other. On the contrary, a properly designed hierarchy is a device to assist in indexing documents and in performing searches for documents or other retrieval objects. Whenever a hierarchy sets constraint it is faulty; whenever it helps the indexer or searcher it serves its functions.

Based on this practical attitude to hierarchy we define hierarchical relationships as follows:

Concept A is broader than concept B, whenever the following holds:

In any search (most searches) for A all (most) items dealing with B should be found.

Given a set of concepts, the traditional approach to hierarchy building is to subdivide the set into mutually exclusive groups, subdivide in turn each of these groups into mutually exclusive sub-groups, and so on. The emphasis is on putting the concepts into some kind of orderly arrangement. If a concept does not fit naturally in that arrangement than it is forced somewhere. If, on the other hand, a concept would fit into different places it is more or less arbitrarily assigned to one of them: no concept is allowed to have more than one broader concept. This principle we call mono-hierarchy. It is quite obvious, especially in the light of our above definition, that this approach is very artificial and imposes many constraints. The modern approach is quite different. Each pair of concepts is analyzed whether or not the condition in the above definition holds. If yes, a hierarchical relationship is established. If no, no such relationship is established. While some concepts may end up with having only one broader concept, others might have two or more. Examples:

Constitution	broader concepts	Politics;
		Public law
Social psychology	broader concepts	Sociology
		Psychology
Railroad stations	broader concepts	Railroads,
		Stations, terminals

This we call poly-hierarchy. On the other hand, a concept may have no broader concept at all. These concepts on top of the hierarchy may be broad subject fields such as economics. But they may also be specific concepts which happen to have no broader concepts such as "Packaging" (no DDC-number for this concept as a whole) or "Weight and measures" (Wrongly placed under "380 Commerce" in DDC).

Having introduced all hierarchical relationships useful for the search process one should of course try to bring the concepts in an orderly arrangement which expresses as many of the hierarchical relationships as possible. Hierarchical relationships not expressed by the arrangement have to be expressed by cross-references. We shall come back to this problem later.

1.2 Concept coordination

It is well known that by combination of concepts more compound concepts can be formed. The reverse of this process is to break down or factor compound concepts into less compound concepts. First of all the break down into semantic factors is useful for the detection of structural relationships between concepts as we shall see shortly. This is the aspect which interests us in this section. Second, semantic factoring may be used to achieve economy of the searching devices in mechanized retrieval systems (such as peek-a-boo systems or computerized systems).

A remark of caution is in order: We are not concerned with the linguistic decomposition of multi-word or compound terms, but with the semantic factoring of concepts according to their meaning. Thus, for example, "gross national product" is a multi-word term designating a concept the breakdown of which is not useful. On the other hand, the term "ship", simple from the linguistic point of view, designates a compound concept which may usefully be broken down into the semantic factors "Vehicles": "Water transport".

Of course, a multi-word term often designates a compound concept. In some cases, the conceptual semantic factoring goes along with the linguistic decoposition; for example: "Lead pipes" = "Lead" : "Pipes"

But, by no means does this apply in every case. To cite an extreme example: "White House" = "Agency" : "Chief executive": "United States". ("White House" here used in the sense used as "The White House announces…").

1.3 Interaction of Concept Coordination and hierarchy

In the early days of coordinate indexing it was suggested and it is still a widespread opinion that semantic factoring on the one hand and hierarchy on the other are opposite principles and that systems are either based on coordinate indexing or on hierarchical classification schemes. A simple example suffices to reveal the superficiality of this opinion: "Railroad stations" may be broken down in to "Railroads": "Stations, terminals". At the same time, "Railroads" and "Stations, terminals" are both concepts broader than "Railroad stations". This is a simple example showing the interaction of concept coordination and hierarchy with each other. In general, the following rules, familiar from the broadening or narrowing down a search request, hold:

Starting from a concept A:B:C one may <u>get broader concepts</u> by

(1) Dropping one of the components (dropping a restriction).

Example:

broader ↑

A B
Passenger transport: Stations, terminals

A B
Passenger transport: Stations, terminals: Local rail transit

(2) Broadening one of the components (weakening a restriction).

Example:

broader ↑

A B C^{1}
Passenger transport: Stations, terminals: Rail transport

A B C
Passenger transport: Stations, terminals: Local rail transit

If one weakens a restriction more and more, the restriction is finally dropped--(1) is a special case of (2),

Example:

broader ↑

A B C^{111}
Passenger transport: Stations, terminals: (Universal concept)

A B C^{11}
Passenger transport: Stations, terminals: Ground transport

A ^{1}B C
Passenger transport: Stations, terminals: Rail transport

A B C
Passenger transport: Stations, terminals: Local rail transit

The rules for forming narrower concepts are just the other way around. There is, however, a third possibility.

(3) There may be concepts narrower than A:B which cannot be derived by any of these methods. Example:

"Helicopter" is a narrower than "Vehicles: Air traffic" but cannot be derived by adding a meaningful semantic factor.

The following diagrams show hierarchical structures generated by these rules. In summary, we may say:

Semantic factoring or concept coordination on the one hand and hierarchy on the other are not opposite and mutual exclusive principles. On the contrary, they interact with each other.

This finishes our study on conceptual structure. We can now go on to problems of file organization.

2 File Organization for Retrieval

As we said before, problems such as pre-coordination vs. post-coordination, enumerative vs. synthetic schemes and alphabetical vs. classified arrangement are problems of file organization. We shall look at these problems from a somewhat different point of view, which will lead us to somewhat different and more refined distinctions.

2.0 The problem defined

The problem to be solved by file organization may simply be stated as follows: Documents deal with compound concepts, made up of many components and called document delineation. Example: Title: The Glideway system, a high-speed ground transportation system in the Northeastern corridor of the United States.

Components of compound concept (document delineation):

Traffic networks; Traffic modeling and simulation; Traffic ways; Stations, terminals; Vehicles; Propulsion of vehicles; Rail transport; High-speed transport; Schedules; Passenger traffic; United States.
Searches are also for compound concepts, called search request formulation, but these search concepts are usually made up of fewer components. Examples:

(1) Stations, terminals for rail transport.

Components of compound concept (search request formulation):

Stations, terminals: Rail transport

(2) A regional network for passenger transport in the Northeast of the U.S.

Components of compound concept (Search request formulation):

Transportation network: Passenger transport: United States

The problem is then to retrieve those documents the delineation of which is equal to or narrower than the search request formulation. (In the case of non-inclusive searches, one wishes to retrieve only those documents with a document delineation equal to the search request formulation.) We shall discuss several possibilities to solve this problem, that is, several possibilities of file organization. We shall confine ourselves in this discussion to inverted files. We shall call "entry-concepts" those concepts under which entries are made in the inverted file, By entry we mean any identification leading somebody looking under the concept to the document; an entry may be a document number, or a document description such as a catalogue card, or a document itself, as on the shelves .

2.1 Principal solutions: post-coordination vs. pre-coordination - a quantitative view

2.1.1 The most important parameter in characterizing file organization in our context is the degree of compoundness of the entry concepts. This is a "quantitative" version of the dichotomy post-coordination vs. pre-coordination. (a) At the one end of the scale we have files where the entry concepts are elemental or at least of a very low degree of compoundness. The usual application of peek-a-boo cards would be a concrete example. The number of entry concepts is comparatively small in these files. A document delineation (the compound concept assigned to a document) is made up listing many of the elemental or nearly elemental concepts (see figure 5); an entry is made under each of these elemental concepts (multiple entry with numerous entries). In the same way, a search request formulation is made up as a combination of elemental or nearly elemental concepts, which may easily be found in the comparatively small list of entry concepts. A search for this combination is then made; this type of file is useful only in the case where it is feasible, from a mechanical point of view, to search for combinations of entry concepts, such as in the case of peek-a-boo files or computerized files. All documents that have delineations equal to or narrower than the search request formulation due to combination are retrieved. By "narrower due to combination" we mean A:B:C being narrower than A:B. As we have seen, this is to be distinguished from A:B' being narrower than A:B due to the fact that B' is narrower than B, If, in a peek-a-boo file, documents on A:B' are to be retrieved, too, generic posting from B's to B has to be introduced. (See fig. 6) this was the one end of the scale, post-coordination.

(b) On the other end of the scale we have files which use very compound entry concepts such as files where documents are arranged on shelves by subject. In this case, we have a huge number of entry concepts. A document delineation is made up of one very compound concept (see figure 5); only one entry is made (single entry). In preparing a search one has to find in a first step among the huge number of very compound entry concepts the one which is to equal to the search request formulation and, for an inclusive search, in addition all those which are narrower than the search request formulation. In a second step, one can then retrieve the documents entered under these concepts. We shall come back shortly to the important problem of how to find the appropriate compound entry concepts. This was the other end of the scale, extreme pre-coordination.

(c) In the middle of the scale we have files using moderately compound entry concepts, such as in subject heading catalogues. The number of entry concepts is large, but not as large as in (b). A document delineation is made up of a few subject headings (see figure 5); an entry is made for each of these (multiple entry with a few entries). In preparing a search, one has first to find the appropriate subject headings from among the large number of subject headings; this poses similar, if less severe, problems as finding the very compound entry concepts in (b). In a second step, one can then scan the entries under one of those subject headings to retrieve the pertinent documents. If it is mechanically feasible, one might also search immediately for a combination of subject headings.

2.1.2 Remark: We have linked in this discussion the degree of compoundness of the entry concepts and the number of entries made for a document in spite of the fact that these two parameters are in principle independent from each other. The linkage set forth here holds if one starts from the requirement that the document delineation be of the same degree of precision with every type of file organization. In actual systems the degree of compoundness goes up less than the number of entries goes down. As a result, the delineations of the documents become less precise (see figure 5).

2.1.3 After this digression we come back to what is the basic problem of this paper. We have seen that in systems using compound entry concepts the problem arises of finding the appropriate entry concepts for indexing or searching among the large number of entry concepts. It follows, that a mechanism for the retrieval of the appropriate compound entry concepts has to be provided. We could call this mechanism a secondary or auxilary information retrieval system. Preempting the next section we may state already that it is here that the considerations of section 1 on conceptual structure come into play and are applied to "conventional" systems.

We could, for example, express the very compound entry concepts of a shelving classification by elemental concepts. To make this more concrete: we could write up a catalog card for each compound entry concept. The compound concept would serve as "title". We could then write down the elemental concepts, which in combination make up the compound.

Once this is done and the search request formulation is also expressed by elemental concepts, there is actually no substantial difference between retrieving documents the delineation of which is made up of elemental concepts, and retrieving compound entry concepts equal to or narrower then the search request formulation. The following illustration should clarify this point further:

In a file of newspaper clippings, the clippings are the documents; they are arranged in folders according to themes which are very compound concepts; that means, we have a shelving classification, the themes being the entry concepts, and we could set up a secondary information retrieval system to retrieve these themes. We could for example make up a catalog card for each theme, as discussed above. But we could also look at this file in another way: We could look at each folder as being a document, and at the theme of the folder as the delineation of that document in terms of elemental concepts. In this view, our catalog cards would stand for documents, the elemental concepts serving as indexing terms; our IR-system would become a primary IR-system, retrieving documents (namely the folders) and not a secondary IR-system retrieving entry concepts.

In the majority of Systems using Compound entry concepts presently in use, with the notable exception of faceted classification, the auxiliary information retrieval system is rather weak compare the remarks on LCSH in 2.2.1 and on LCC and DDC in 2.2.2).

We may summarize these considerations as follows: Searching consists of two steps:

Step 1: Find the appropriate entry concepts to be used in the formulation of the search request.

Step 2: Retrieve documents by combination of the entry concepts found in step 1.

The"work load" of searching for the appropriate compound concepts may be distributed between the two steps. In a peek-a-boo file, entry concepts are elemental concepts, therefore no combination searching necessary in step 1, and combination of many entry concepts in step 2. In shelving classification, the entry concepts are very compound (ideally as compound as document descriptions); therefore there is combination searching involving many components in step 1, retrieving the appropriate entry concept (or concepts, in the case of inclusive searchers), but there is no combination searching in step 2. Systems in between use moderately compound entry concepts so that both steps involve combination searching, with less components in each step.

We have already mentioned that the problems of file organization are much more difficult in systems using compound entry concepts than in systems using elemental or nearly elemental entry concepts, such as peek-a-boo systems. The rest of this paper concentrates on problems of systems using compound entry concepts (pre-coordinate systems). We first deal with the question how retrieval mechanisms for compound entry concepts can be designed. We then go on to the problems of selection of entry concepts and of their arrangement in a file.

2.2 Retrieval Mechanisms for entry concepts

In this section we are concerned with the retrieval of compound entry concepts in terms of their conceptual components (as specified for example in a search request). We are not at all concerned with alphabetical indexes where a compound concept may be found under the term used to designate it.

2.2 1 The first possibility for such a retrieval mechanism is to represent the poly-hierarchical structure formed by all the concepts in a linear arrangement with hierarchical cross-references[1] If one chooses classified order, many hierarchical relationships can be expressed by the arrangement alone and cross-references are needed for the remaining ones only. If one chooses alphabetical order, all hierarchical relationships have to be expressed by cross-references. In principle it is not necessary for this purpose that the compound concepts be expressed by

[1] that means, Broader Term- and Narrower Term-cross references; these may be complemented by Related Term-cross references, which are also useful for retrieving the appropriate entry concepts. In LC subject headings, all these are lumped together as see also- cross references.

semantic factors, as long as all hierarchical relationships are known. However, the task becomes much easier if one expresses the compound concepts by semantic factors, since the derivation of hierarchical relationships, the determination of the arrangement and the introduction of cross-references can then be done much more systematically and can even be automated (See fig. lb). (As to the arrangement, compare section 2.3(2), where this question is dealt with in detail).

Someone looking for an entry concept appropriate for his indexing or searching purposes will enter the list at a broader concept which he knows, He will then go down, in; the classified arrangement as well as following the cross-references, until he finds the appropriate entry concept.

We illustrate the process in a system where compound entry concepts are expressed by semantic factors. Someone has expressed his search concept by A:B:C:E. He enters the list at any of the components, say A. There he will look through the narrower concepts, either listed at the same place or indicated through cross-references. He will either find the entry concept he is looking for or he will find a broader concept, say A:B:E. In the latter case, he looks through the narrower concepts given for A:B:E, and there he finds A:B:E:C (which, of course, is the same as A:B:C:E, the system using a different order than the searcher). Starting from A:B:E:C he will also find all narrower entry concepts, either listed immediately or indicated by cross-references.

If there are many entry concepts with more than two components, this is a very cumbersome and ineffective method. In general, if the number of entry concepts is Large, cross-references do not provide a convenient means for retrieving entry concepts, as anybody following the cross-references in LC Subject Headings can confirm.

2.2,2 The second possibility is to establish an actual information retrieval system for entry concepts. In such a system one would express the search question by a combination of concepts contained in a "core classification scheme "consisting of elemental or nearly elemental concepts. One would then retrieve all entry concepts (subject headings, LC class numbers) which are equal to or narrower than the search request formulation. Such a system could be peek-a-boo system (if the entry concepts are numbered serially), an edge-notched card file or a computerized system. The most likely possibility, however, would be a printed index of the combinatorial type. Foskett's rotated index is such an index. It shows every entry concept under each of its single components. The same purpose is achieved by a KWIC-index, indexing strings of terms or strings of notational symbols. More convenient but also of much larger size would be an index showing each entry concept under each pair of components. Even further goes the SLIC index, which shows a compound entry concept under each combination of components. The PRECIS-system could also be used for producing such an index. The chain index is another example. However, the chain index rather confuses the matter by being two things at once: an index to entry concepts in terms of their constituents as well as an alphabetical index.(1) It would be much clearer and probably much more useful, too, to separate these two functions and to provide a chain index in which constituents are expressed by their notation, and an alphabetical index to the schedules.

An index constructed according to one of these methods would make the use of, for example, the Library of Congress subject headings much easier both in indexing and in searching (Comp. fig.7)

Two further remarks are in order:

1. Combinatorial indexes usually are designed in such a way that it is easy to retrieve those entry concepts which are equal to the search request formulation or narrower than the search request formulation due to combination (see 2.1), that is, for the search request formulation A:B the narrower entry concept A:B:C is found easily. The problem of is narrower than B, retrieving also the narrower concept A:B', where B' is not at all or not as well solved (in peek-a-boo systems this problem may be solved by generic posting, as we have seen in 2.lJ compare fig.6). The searcher has therefore to be careful while using combinatorial indexes.

[1] We may note, parenthetically, that this remark applies to some degree to the "relative" alphabetical index to DDC, and even, if still less, to the alphabetical indexes for the LC schedules.)

2. Some of the considerations of this section apply also to combinatorial indexes used in primary information retrieval *Systems.*

2.3 Selection and arrangement of entry concepts

In systems using compound entry concepts there is the problem what entry concepts to include and also the problem how to arrange the entry concepts in the file (catalog, shelves). Both problems are usually discussed under the heading "enumerative vs. synthetic schemes". In the following, we give a refined analysis of these problems. We introduce three aspects according to which classification schemes should

be analyzed.

(1) The first aspect iS concerned with the problem: how are the Compound concepts designated Are the

y designated by their own, independent symbol (possibility (la)) or are they designated by a chain of constituent symbols, each constituent symbol presenting one of the conceptual components(possibility (lb)) . Examples for possibility (la) are the LC classification (the independent symbols being LC class numbers) and subject headings, (the independent symbols being natural language terms). (We may remark that not too seldom subject headings are made up of a string of constituent symbols, especially if standardized subheadings are used.)

An example for the possibility (lb) is, of course, faceted classification.

Remark: It is possible to have independent symbols for the compound entry concepts and still express them by semantic factors. For possibility (lb) it is obviously necessary to express the compound entry concepts by semantic factors.

(2) Sequence of entry concepts.

We first remark that once a mechanism for the retrieval of compound entry concepts as described in section 2.2 has been established, the sequence of the entry concepts is less significant. We could even number them serially as they arise. This would then be a system in the category (la) above (independent symbol for compound entry concepts). Usually,however; in such systems one of the following procedures is applied:

(2al) The entry concepts are arranged according to the alphabetical sequence of the terms chosen to designate the entry concepts (this is of course the case of subject headings).

(2a2) Or the concepts are arranged according to independent notational symbols chosen to designate the concepts. The notational symbols usually lead to some kind of classified order. There is plenty of discretion and arbitrary decision-making in the arrangement. For example, if a subdivision by country is used in different places, a different sequence of countries can be chosen in each instance.

(2bl) In systems where the entry concepts are designated by strings of constituent symbols, the place of an entry concept is completely determined by the string. This makes sure, for example, that at every place where a subdivision by countries is used the countries appear in the same sequence. But there may still be considerable or complete discretion as to the sequence of constituent symbols in the string ("citation order"). See the example given in figure 1b .

(2b2) With all the procedures for sequencing discussed up to now it is necessary to look up an entry concept in a listing in order to determine the symbol used for its designation. Provided every new entry concept is allowed in the system (see below) this is avoided in schemes that prescribe a citation order completely in every instance, such a faceted classification.

Remarks:

1. The designer of a system which uses independent notational symbols is free to adhere to the restrictions put forward in possibility (2bl) or (2b2)(faceted classification) in constructing his sequence of entry concepts.

2. The constituent symbols used in (2bl) and (2b2) may be either terms to be arranged alphabetically or notational symbols.

(3) The third aspect is the degree of ease with. which new entry concepts may be introduced. Is the introduction of new entry concepts forbidden at all or are there well set procedures by which they have to be approved? What is the time needed to introduce a new entry concept?

What are the criteria for approval for a new entry concept? A criterion might be for example literary warrant, that is one might require that the number of entries made under the new entry concept is expected to exceed a certain number. This problem is related to the problem of multiple entry. If multiple entry is allowed then one may always use two or more less compound entry concepts to make up the delineation of the document instead of introducing a new entry concept. (Note, however, that a compound entry concept available in the system should always take precedence over a combination of less compound concepts.) In this case one should use literary warrant and/or "search warrant" as a criterion. We shall come back to this problem in the following section. If multiple entry is not allowed, such as in LC classification and in those applications of faceted classification Where *a* policy decision for single entry has been made, the situation is more difficult. In the case of LC classification one can either admit that documents are forced into an entry or one has to update the schedule in very short intervals. In the case of faceted classification the indexer is allowed to form new entry concepts as he deems necessary and a procedure has to be established to update the index to the entry concepts accordingly. Note, however, that a scheme of the LC type that allows for the inclusion of very specific entry concepts and for "immediate updating" and that provides a mechanism for the retrieval of entry concepts as described in section 2.2 is operationally equivalent to a faceted classification scheme.

2.4 A unified classification scheme for different kinds of file organization

From the perspectives developed in this paper there emerges a practical proposal for the design of a classification scheme to be used in connection with different kinds of file organization. One starts from a "core classification scheme" consisting of elemental or only moderately compound concepts. These concepts are called (core) descriptors, and they are represented by an independent symbol, such as a notation or a term. The core classification scheme is presented as a linear arrangement with cross references. In a faceted classification, the schedules would be the core classification scheme. Starting from the core classification scheme, entry concepts are formed.

In a peek-a-boo or other post-coordinate system, only descriptors are used as entry concepts.

In a card catalog or similar systems, the descriptors themselves may be used as entry concepts, too. But further entry concepts are formed by combination of descriptors as it becomes necessary during the development of the catalogue. In the beginning documents dealing with A:B will be entered under both A and B. If it turns out that there are a lot of search requests for A:B or a lot of documents dealing with A:B, then A:B is introduced as an entry concept, and documents dealing with A:B are entered only there. This reduces both the number of entries and the effort necessary for searching (in searching for A:B it is no longer necessary to scan all the cards entered under A or all the cards entered under B). A document dealing with A:B:C is entered under A:B and under C (or under B:C, if this is an entry concept), A general rule may be formulated as follows.

(Compound concepts denoted by lower case letters): Let d be the delineation of a document. An entry for the document is made under every entry concept x with d narrower than x unless d is also narrower than entry concept y and y is narrower than x.

In a shelving system, entry concepts are formed be combination of descriptors as required by the single entry rule.

In the case of card catalogs and shelving systems, an index to the entry concepts is prepared as has been described in section 2.2. This index also tells, for example, a searcher looking for B that he should also look under A:B.

The core classification scheme together with the additional entry concepts may be called an "extended classification scheme".

A few additional remarks are in order at this point.

1. On multiple entry vs. entry under compound concepts.

Take the above example of documents on A:B. In one case, they are entered under both A and B. Searcher 1, searching for A, is lucky because he has all entries together at one place in the catalog. The same is true for searcher 2, searching for B, Searcher 3, searching for A :B, however, is disadvantaged because he has to scan all the entries under A (or all entries under B) to find those on A:B. If the compound entry concept A:B is created and arranged after A, searcher l, searching for A, is still lucky. Searcher 3. searching for A:B, is now lucky, too. Searcher 2, searching for B, however, is now disadvantaged because he has to follow a cross-reference to another place of the catalog. Giving up the advantage of having fewer entries, we could help searcher 2 by arranging the new compound entry concept at a second place as B:A and making entries for this second place, too. For searcher 3, searching for A:B, this would also be convenient, because he now could enter the file either a A or at B. Speaking in terms of the model sketched in section 1, this means: the compound concept A:B is arranged under each of its broader concepts. Note that such a system provides a retrieval mechanism for compound entry concepts, as described in section 2.2, right in the file itself. If we come to more compound concepts having more components, the size of the file increases very rapidly if one uses this procedure. One must then select some particularly useful places where to put a given compound concept among all the possible ones. This is in essence the purpose of the PRECIS system. For each document a delineation is prepared as a combination of descriptors, structured according to special rules. The document delineation is included into the system as an entry concept (if it was not already included before). This entry concept appears at different places in the arrangement, and an entry for the document is made under each of them, preferably giving the document number. In the application of the PRECIS system in the British National Bibliography, we encounter a little peculiarity which might be confusing: the index to BNB is an index to documents. However, the full document descriptions are listed in BNB by DC class numbers, and the class number is the only means to look up a document description. Therefore, in the BNB index, class numbers are given instead of document numbers. This should not detract from the fact that the index is an index to documents and not an index to class numbers of the type discussed in section 2.2.

2. The entry concepts for a classified catalog using multiple entry can of course be formed using a faceted classification scheme. Each entry concept is then designated by a string of constituent notations. A document is indexed by as many entry concepts as necessary, the appropriate notational strings are put on the catalog card, and the card is filed at the appropriate places.

The important point in this proposal is that different institutions using the same core classification scheme could extend it in different ways, adapted to their specific needs, but still maintain compatibility between their systems. Even nonessential features of the core classification scheme (for example, the sequence of main classes or facets, respectively) could be changed without destroying compatibility on a conceptual level. (There may be some practical difficulties arising from the use of different notations in both systems. But these can easily be resolved by the application of computers.) Existing schemes, such as the Library of Congress classification scheme, could be made compatible by expressing the entry concepts in terms of the core classification scheme. A properly designed core classification scheme could thus take the role of this old dream, a universal classification. This is made possible by concentrating on the basics of conceptual structure and leaving aside details of arrangement and file organization on which agreement cannot be achieved and is not even always desirable.

The approach developed in this paper has basic implications for the design of thesauri, in particular for the design of a universal accumulative thesaurus, into which we cannot enter here.

Acknowledgement

I wish to thank Thomas D. Wilson for his thorough criticism and comments which lead both to a correction of errors and a clarification of formulations.

References

Austin, Derek
Prospects for a new general classification.
J. Librarianship, vol. 1, no. 3, 1969, p. 149-169.
Similiar in the whole idea, but far more specific as to the kind of classification. See esp. section "Implications for the Future" for section 2.4.

Cordonnier, G.
Metalanguage pour les traductions d'intercommunications entre homines et son adaptation dans le domaine des machines pour recherches documentaires. Kent, A., ed. Information retrieval and machine translation, vol. 2. New York, Interscience, 1961, p. 1091-1137. For whole paper, esp. section 2.1.3.

Foskett, D. J.
Classification and indexing in the social sciences.
London, Butterworth, 1963, 190 p.
p. 165 and following for section 2.2.2, rotated index

Jonker, F.
A descriptive continuum: a "generalized" theory of indexing, ICSI Proceedings, vol. 2,1953, p. 1291-1311. For section 2.1.1

Libbey, Miles A.
The use of second order descriptors for document retrieval.
Amer. Doc., vol. 18, no. 1, 1967 , p. 10-20.
For section 2.1.3 (same principle, but we do not agree to the implementation suggested in this article)

Needham, R. M.
Research on information retrieval, classification and grouping 1957-61.
Cambridge, CLRU, 1961, 177 p.
p. 11 for section 1.1, definition of hierarchical relationships (our definition is developed from Needham's definition)

Scheele, Martin
Thesaurus - Baustein jeder Fachdokumentation Nachr. Dok., vol. 1, 1964, p. 1-4. For section 2.1.3

Sharp, John R.
The SLIC Index.
Amer. Doc. vol. 17, no. 1, 1966, p. 41-44.
For section 2.2.2

Soergel, Dagobert
Mathematical analysis of documentation systems.
An attempt to a theory of classification and search request formulation.
In: Information Storage and Retrieval, vol. 3, no. 3, 1967, p. 129-73.
For section 1

Soergel, Dagobert
Outlines of an algorithm for the analysis and
comparison of classification systems
Freiburg i. Br., Selbstverlag, 1965, 30 p.
appendices * mimeo *
App. 2, section A(l) for section 1.1.3
(quoted as source only)

Whole report relevant for section 3.