DOCUMENT RESUME

ED 116 680 IR 002 925

AUTHOR Painter, Ann F., Ed.
TITLE Classification: Theory and Practice.
INSTITUTION Drexel Univ., Philadelphia, Pa. Graduate School of Library Science.
PUB DATE Oct 74
NOTE 125p.
JOURNAL CIT Drexel Library Quarterly; v10 n4 Oct 74

EDRS PRICE MF-$0.76 HC-$5.70 Plus Postage
DESCRIPTORS *Classification; Cluster Grouping; Futures (of Society); Information Retrieval; *Library Automation; *Library Science; Library Technical Processes
IDENTIFIERS Dewey Decimal Classification; Library of Congress Classification; Universal Decimal Classification

ABSTRACT

In response to recent trends towards automated bibliographic control, this issue of "Drexel Library Quarterly" discusses present day bibliographic classification schemes and offers some insight into the future. This volume contains essays which: (1) define "classification"; (2) provide historical background; (3) examine the Dewey Decimal System, the Library of Congress Classification, and the Universal Decimal Classification; (4) discuss research and development of automated systems; and (5) make predictions for the future. (EMH)

Drexel Library Quarterly

A publication of the
Graduate School of Library Science
Drexel University

October 1974 Vol. 10 No. 4

Classification:
Theory and Practice

Ann F. Painter
Issue Editor

Publication Staff

Barbara P. Casini, Managing Editor
Vera Rocks, Circulation Manager
Meredith Gatschet, Designer
Joseph D'Auria, Printing Coordinator

Drexel Library Quarterly, Barbara P. Casini, managing editor, is published by Drexel University, Philadelphia, Pa. Issued in January, April, July and October.

Orders for subscriptions and single issues may be addressed to *Drexel Library Quarterly* at this address. Subscription rates: 1 year, $10.00; single copies, $3.00 each.

The *Quarterly* is indexed or abstracted in *Current Contents: Education, Current Index to Journals in Education, Information Science Abstracts, Library Literature, Library & Information Science Abstracts, PAIS Bulletin, Vertical File Index,* and *CALL.*

Library of Congress catalog card number: 65-9911. Microfilm edition prepared annually by University Microfilms, Inc., 313 N. First Street, Ann Arbor, Michigan.

Contents

Contents

Introduction

1

Classification and classification systems have formed the foundation of retrieval systems since man first began to record knowledge. Good histories and descriptions of classification schemes are few and far between, and usually little appears in them to explain the whys and wherefores. But obviously man has recognized the need to organize in order to retrieve.

Americans are not particularly classification-minded, as Mr. Stevenson will point out later. One of the great anomalies of American classification is the Library of Congress Classification, which has little to say for itself philosophically except that it works. One of my favorite quotes comes from Phyllis Richmond who once wrote of the Library of Congress Classification·

> In a discussion of classification research, the Library of Congress system does not fit any of the categories described. It is a pragmatic, functional system that is widely used with considerable consumer satisfaction. It is not logical, it is not scientifically or probabilistically built; it has little to dc with language or linguistics other than to provide the best classification of these subjects extant; in organization it sprawls in all directions; it violates all the postulates, principles and laws that are considered important in classification making; in some areas relationships are shown in hierarchies, but throughout most of the schedules nothing seems to be next to anything for any particular reason; yet it grows steadily without any serious sign of stress. Why does it work?[1]

Americans have been inclined to leave classification at just that, as long as it works, that is all that counts.

Introduction

Classification in the United States has developed quite uniquely. The so-called traditional schemes to which we are wed have been designed and used more as browsing tools or shelf organizers and hence tend to classify only generally. This has caused a one place on the shelf-one place in the scheme philosophy. The Europeans and Asians have used classification to organize concepts, rather specifically in indexes (classified catalogs), in order to retrieve information, not an "item". This difference in approach has indeed influenced every aspect of classification around the world.

Quite a few people have posed the question, especially now that automated bibliographic control is becoming a reality, Why bother? Let the computer do it. Classification is dead! There are other ways to access information. I leave this as a potential hypothesis—not yet researched or validated. Maybe because of my cataloger's inbred loyalty to classification as a self-evident truth, I assume classification is very much alive.

The purpose of this issue is to discuss classification today (primarily in the United States), with some insights into the directions of the future. There is no attempt to be comprehensive, thorough, exhaustive, etc. The authors have been asked merely to put some of their ideas and thoughts down. This is not a state-of-the-art, it is not a history, it is not a how-to-do-it manual for classifiers. One paper attempts to define classification and provides the theme of the issue. The historical paper intends to set the stage and indicate major trends. From a theoretical point of view both the traditional and the modern attitudes and characteristics toward classification are summarized. With the theoretical framework provided, the Dewey Decimal Classification, Library of Congress Classification and the Universal Decimal Classification are examined. And finally, there are two papers on the future—automatic classification and research. Admittedly this is a rather loose framework but it has allowed the reader an opportunity to see where American classification stands.

The working classifier may find the papers interesting and informative, perhaps reinforcing. The student may find them an introduction and summary on which to base further exploration. The researcher may not need really to dwell on them much at all. The papers are offered to the general librarian, not the specialist, with the hope that they may stimulate interest and improve awareness of the heart of the retrieval problem—classification.

Ann F. Painter, PhD
Professor
Graduate School of Library Science
Drexel University
Philadelphia, Pa.

Notes

[1] Phyllis A. Richmond. 'Some Aspects of Basic Research in Classification. *Library Resources & Technical Services* 4 (Spring 1960). 139-147

Classification: A Definition

4

Harris Shupak

The point of a definition is to provide a target for a concept, allowing the specification of its purpose, and in this specificity to label one or more variables of the universe. The task is formidable, especially as so many layers of meaning are implicated in the descriptors we use, with attendant expectations of their worth. With concepts becoming long-standing practices, this challenge is even greater, in this sense, I wonder if classification can be defined at all! With this warning given, however, I shall launch into my subject. Rather than give an initial definition and then attempt to prove why it is more or less true than other definitions, I shall illustrate various aspects of what I consider to be the practices of classification. Thus, my method should, with luck, back into the central problem of the paper.

A curious fact of our history on earth is the rise of stratified classes within human society, classes often based on the exclusive possession of differentiated skills deemed to be of significant value to these groups. Once these distinctions occurred, man no longer remained coequal with other men but, to paraphrase the words of a noted analyst, "some men became more equal than others." Stratified classes were early indications of man's ability to perceive distinctions and order his universe around them. Seen collectively, these stratifications are nothing more than the universe of his existence. Taken separately, they are the basis for hierarchical rankings and subdivision of that world.

Another example of what I may be allowed to call man's inherent process of artificially ordering the world he perceives, and one that illustrates another facet of the discussion, is the world of kinship systems in non-Western societies. In addition to the

Harris Shupak is Librarian at Camil Associates, Inc., Philadelphia, Pa.

forces of stratification, languages of kinship became formalized, precisely indicating the levels of relationship between branches of a family and its individual members. If an analogy can be made, it would be this: from a perceived difference (based on artificial criteria) between societal members, terms of address in kinship languages formalized these distinctions, giving notational relationship between these individuals. If some tribes locked these patterns in too rigidly, then at what cost would personal initiative have to suffer in order to break the deadlock of these expectations? Then again, with forces of diffusion and dispersion so widespread through this aged world, what changes were made to the set of these kinship orders? Could societies change the basis of kinship expectations without changing their stratified orders? This point, perhaps imaginary, is made to demonstrate the complexity of ordering and changing the universe of man's perceptions. Our classification, as a philosophy and a practice, stands in the same proportion of difficulty as these anthropological phenomena.

In this paper, I wish to illustrate classification as a process of naming and ordering this universe, but not merely an activity solely directed to some objective world of knowledge. The historical process of stratification-classification has been one of advancing knowledge as our understanding of natural and artificial orders has increased—to give new relationships a rightful and accurate place in the scheme of things. We have had to compare these changes to hierarchical orders previously constructed. In this way, changes in our classified orders came not in scattered bits, or bytes, as it were, but as alternatives to the hierarchy of established facts—and hence to knowledge itself

A carefully stratified order moves continuous time into separate epochs, thus, the extrapolation of time and circumstance was given definite boundaries Within each epoch, alterations could be observed in terms of that specific time period, with each one having its own level of development, contrasted with other ages, ordered and classified according to these distinctions This was our heritage of intellectual classification, changing in its sophistication as our accumulation of facts increased The order it created became the foundation for comprehending the universe of knowledge

Why do men have difficulty introducing radical change into their classified orders? Why has the existence of intuitive leaps been such an important device for accomplishing these changes? After hypothesizing and experimenting, sifting the studied relationships carefully, imagining through these concepts the new possible orders for them, after all this, how often do the gaps still exist? Man cannot consciously finish the job. The intuitive leap—that process of comprehension so slightly beyond the conscious world—accomplishes the extraordinary feat of interpolating these facts into the order that was not quite within reach of the thinker. With intuition, old classifications are destroyed and new ones created. These new classifications of phenomena, discovered in hard thinking and timely serendipity, have their own language, classes, their distinguishing characteristics from former classifications and their points of duplication. In some cases, the terms gaining access to these classifications are absolute—accurately part of the new order itself, unrelated, in its essentials, to previous classes and terms. In other cases, the terms of reference will be fuzzy, questionable, almost belonging to one or another class and to several different classifications. In gaining the best access to the hierarchy, how can one be sure of the accuracy of terms? In an absolute order there is no confusion. In separating a homogeneous world, however, how can man's order duplicate a natural order with the same degree of perfection? Therefore, no order can be absolute, it is only a temporarily derived stage for viewing the accumulation of facts to date. Yet, as a process special with man, is it not fascinating to recollect that our progress as a species was so fast, accurate, and unstoppable because we had gained control of such a power as classification?

If I have not explicitly defined anything yet, you may see that the difficulty rests with trying to pinpoint an activity so pervasive in man's growth and history. One of the large conflicts in thinking of our library classification, in contradistinction to the intellectual process I have described, is in deciding what classification really represents. When we invoke a Dewey Decimal number, are we seeing a pattern, a piece of the classified order, falling into our comprehension? Are we handling a representative from that order in the form of a document in which part of that classified order shall be revealed? Or, is it merely a place reserved for such a representative? The differences between these ideas can yield three definitions.

Historically, it may be said that classification is a process and an act of ordering and differentiating the universe, yet classification also mirrors philosophy, has its own terminology and is an organization of places to store things, be they ideas or documents. The mixture of these elements has caused a reign of confusion as to what purpose classification should serve. The Baconian influence on library classification has been well documented. Bacon's era extended library classification from the "art" of making philosophical charts of the universe (then at its apogee) to a conjunction, on a cosmic level, of collected bits of a developed process of mental classification. One respects Bacon for his ability to do this so beautifully, comprehensively and lastingly. Several hundred years later, through interminable changes, influences, and practices, the Baconian universe met Melvil Dewey, and there—in one of the more important historical events of man's history of classification—we find the first significant and lasting admixture of philosophical perception and the rather mundane practice of storing documents in libraries. In this encounter, a question was created that has not been solved. What is classification? Is it the practice of philosophical differentiation I have been describing or the art of accurately storing documents with a mind toward effective retrieval of related pieces of information?

Dewey lifted a sagging world of shelvers and card catalog makers, and gave librarians a chance to participate in the comprehension of rarity and beauty—the worlds of philosophy, the mirrors of man's universe, determined through these specialized perceptive and cognitive abilities Henceforth, when we hear about the classical debates of where to place certain documents in the Decimal Classification, it is not merely that a question of location is being argued. Indeed, one has the feeling that the interlocutors were questioning the inherent order of the universe itself. Why else would these practical storers of information give so much heat to the argument? Even though it was dealing with the documents of knowledge, the welding of philosophy into the Decimal Classification made it a process whereby the universe was divided into identifiable classes, further subdivided by these perceived differences, and given an appropriate notation for retrieving the documents of this order This may be a simplification, but I am illustrating classification, not giving a manual for the Dewey Decimal Classification (DDC). Later elaborations of this process found more enumerative schemes, in which this universe was explicitly developed for document retrieval and the collation of related materials The intent was to integrate materials as they were being stored

An ironic fact about the DDC is that it has managed to persist so long, undergoing numerous reordering of its scheme and yet remaining representative of our changing world, cohesive and contemporary. It is this duality of purpose that has given it so many problems. Perhaps it is being made to do too much? Perhaps it is only feasible for a limited point in time?

In beautiful opposition to the philosophy and practice of DDC, we have a modern science of classification which explores another extreme. It is called subject analysis, faceted classification, developing in its glory as a computerized operation. Beginning with a universe of discrete facts, ideas or subjects, it seeks initially to abolish formerly perceived classes, and substitute for the old method of finitely breaking down the universe, one which minutely orders these facts, bringing related aspects of documents together into classes which represent the cooccurrence of terms as analyzed in the documents themselves It is a process of building up the classification from these facts, or facets, without seeking to create a complete universe. With this method of classification, philosophy has been returned to philosophers. Computerized classification offers an opportunity to relate conceptually the documents of knowledge much more precisely. The classifications are not stable, but change frequently with the reordering of subjects in these documents. Indeed, one would wonder whether this is really classification, as it seems so antithetical to the progress of mental classification with which I began this paper.

These classifications are interesting in principle, but significant costs will have to be assumed to perfect their development as tools of classification. It is generally assumed the faceting can work for certain small classes of documents, but with general library collections they would be useless. In some experiments, classes created by computer algorithms had to be combined with classes from a traditional classification in order to reduce the amount of irrelevant materials retrieved. Even where computers can create such classifications, other practices must be appended—like in-depth indexing systems, independent notational systems for storing documents and the great interpretive involvement of librarians.

Are there any commonalities between these contrasted classifications that would allow us to define classification? Yes, one! In both cases classification must become an ordering scheme for locating these documents. At all costs, whatever scheme we invent or use, the classification must be able to perform efficiently at this central task—locating the document for the user Modern classificationists are attempting to take this qualification and to say. "If machines can do it faster, though not more comprehensively, and on an average perform as well, then our methods must be equally as valid. Not better, but at least as valid." For whatever slight cost advantage is given by this perceived equality, these classification systems will be developed for special collections, perhaps one day for general library use

It must be said again, therefore, that whatever the method, the finite dispersion of the universe or the monolithic creation of a world of discrete subjects, we still have to store and retrieve these documents. I think, now, that my definition of classification is apparent. it is the activity of storing documents for retrieval. No order is complete without a basis for distinguishing and differentiating the documents, but the locating and storing function—the notational device—must be independent of that scheme. I could conceive that, given the ability of successfully storing and locating documents, any scheme in the future might be adaptable to that purpose. So, in the end, I have offered a rather unstartling definition of classification. Is it reflective and worthy of my argument?

The problem is that I am forced to uphold it, but I do not fully believe it. Classification was initially described as a mental process of ordering the universe, and we have taken our library practice and reduced it to a mere act of storing and locating documents in a collection of materials. On one hand we can speak of classification in the highest sense—that which hierarchically orders the universe—allowing us to proceed from concept experience to concept experience, revising our categories, but building up our knowledge as a consistent attempted representation of the universal order. On the other, we merely speak of our library classification as a shuffling device, devoid of that presence which exists when the two are combined, as in the original encounter of the Dewey Decimal Classification and Baconian philosophy. Yet, our age has a new philosophy, and its herald, the computer, allows for quick, subtle and efficient manipulations of ideas, facts and subjects. If, therefore, my definition of classification is unsubstantive, I feel it must remain so. We have a different age upon us, and what have existed as inherent mental processes are changing, offering us new and different experiences and practices. The human process of classification and its substantiation in the library will mirror these influences. It is only a question of whether we will find the same kind of welding in library classification that will give us a new and unique opportunity to classify our documents and store them in the same mode, or some kind of dispirited shuffling system that is efficient, but lacking in human dynamism.

The Historical Context: Traditional Classification Since 1950

Gordon Stevenson

Introduction

Twenty-five years ago, when librarians in the United States spoke of classification, they were usually referring to two specific library classifications: the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LC). The habit of confusing the general idea of library classification with the possibilities and limitations of DDC and LC had been characteristic of United States librarians for generations. A clear distinction was not made between general principles of the nature, structure and uses of library classification and the application of these principles in specific systems. This approach to classification enshrined DDC and LC somewhere near the center of librarianship. Even today, our problem is not so much classification as what we think classification is and how we think about it. The way we thought about classification around 1950 was such as to give DDC and LC a legitimacy and permanency of the sort usually reserved for religious texts and sacred rituals. Unfortunately, this approach is still found to a great extent today; and though DDC and LC seem to be even more inextricably embedded in United States librarianship than ever, it is now necessary to identify these two systems as "traditional library classifications." They must also be identified as "general classifications," because they know no subject limitations.

DDC and LC are traditional in an historical sense because their roots are deep in the past, and in a practical sense because they are used by librarians today in essentially the same way they were used when they were introduced before and shortly after 1900. They are also traditional because of their structures.

Gordon Stevenson is Associate Professor, School of Library and Information Science, State University of New York, Albany.

They are internally structured with mutually exclusive, enumerated classes that are arrived at by a logical process of division that proceeds from broad concepts and disciplines to ever narrower and more specific subclasses. Since 1900, these systems have changed; but most changes have been quite superficial in terms of classificatory techniques. New classes have been introduced, finer subdivisions have been made, and old classes have been rearranged. But the traditional systems employ no basic structural or classificatory device that was not known before 1900.

In 1950, Jesse Shera critically evaluated the traditional classification schemes and the principles on which they are based.[1] In doing this, he succinctly defined their parameters and clarified the difference between traditional and nontraditional systems. In the meantime, we have learned a lot about classification and its theoretical and practical foundations. The past several decades have seen a more intense examination of the foundations of classification than any other period since the last quarter of the nineteenth century. The results of these investigations, experiments, philosophical speculations and theories have created an ever-widening gap between the traditional systems and the newer, modern systems. It is the purpose of the present review to consider the two traditional systems in their historical context and to comment on the idea of general, as opposed to special, library classification. With DDC and LC, we are dealing with two dinosaurs that one would have thought could not survive into the second half of the twentieth century. They would appear to be relics of the past, and their survival—indeed, their continuing vitality—raises important questions about the nature and uses of classification by librarians in the United States. It is the contention of the author that it is impossible to understand the condition of classification in the United States today or to speculate intelligently about its future without an historical perspective.

Classification Around 1950

General library classification as we knew it around 1950 was a product of decisions made around 1900. Expectations about the contributions of classification to subject control and access, ideas about the structure of classification systems, and general agreement about what constituted a proper subject catalog had long since been formalized and incorporated into the conventional wisdom of librarianship. A key historical event of almost unprecedented importance in the history of subject access was the rise of the dictionary catalog and the subsequent disappearance of the classified catalog from United States libraries. After that happened, the way we thought about classification and its uses changed fundamentally. By 1950, most librarians in the United States were not quite sure what a classified catalog was for or how it was different from an alphabetical subject heading catalog. Why this happened and its long-range impact on both classification and subject access are historical questions which have never been answered Added to this fundamental change in the use of classification was the phenomenal dispersal of DDC and later LC. All competing systems were swept aside and these two became such monumental edifices that they have never been seriously challenged in the United States. By the time Bliss published the final volume of his Bibliographic Classification in 1953,[2] hardly anyone in the United States took his work seriously. It is very possible that the Bibliographic Classification was a better classification than both DDC and LC, but it was published too late to have any practical impact in the United States.

As late as 1950, many, if not most, library schools in the United States taught all students the DDC system and saved LC for those hardy students who went on to take "advanced cataloging." It did not occur to anyone that there might be an alternate to DDC and LC. Most of what we knew about general principles, we learned from Berwick Sayers, the British classificationist and teacher,[3] but his work had some limitations It was not until the German translation of E. I. Šamurin's monumental history of classification was published in the late 1960s that we had access to a coherent survey of the great European systems in the full sweep of their historical evolution.[4] But by the time Šamurin's work was accessible in the West, few librarians in the United States were interested in the European systems. For most of us, classification in Europe was then, as it is now, a closed book. The reasons for this are also buried in our past. In the early decades of the American Library Association, there was a lively spirit of internationalism and an exchange of ideas about cataloging and classification. This ended in 1914 for reasons which are obvious and have nothing to do with librarianship. Since then, we have exported librarianship but have assumed that there is little worth importing. The new internationalism in descriptive cataloging that emerged from the Paris Conference of 1961 has not led to similar trends in subject cataloging or classification.[5]

The complete dominance of DDC and LC was due in a large measure to the long-range trend to centralized cataloging and national standardization. Improved subject control and access were not the only issues involved in this trend. Another was the rising costs of all technical services. When librarians thought it necessary to make a choice between DDC and LC, the overriding criteria that influenced their decisions were the economic consequences of the two systems. After 1950, the role of the library manager in making classification decisions increased. The spirit that animated change was made clear by Raimund E. Matthis when he said, "We must opt for the most workable tool at present available to carry forward the mundane but needful task of moving books and records from catalog department to shelves and catalog."[6]

With decades of the neglect of classification behind us, it was easy to accept without question the mystiques which began to surround DDC and LC. Of the two systems, we learned more about DDC than we did about LC. The massive size of the Library of Congress, its central role in national bibliographic control and its formidable staff of subject specialists gave it such an awesome authority that few librarians even considered subjecting the LC system to a serious, in-depth evaluation. Furthermore, the belief that we knew and understood the historical origins of the LC system was a myth. The extent to which LC is based on nineteenth-century European systems has not yet been documented. A reading of the works of Rudolph Focke casts doubts on much of what we think we know about LC's origins.[7] In developing a classification code, Focke drew up a series of rules which, when compared to LC, describe the foundations of that system quite precisely. It must say something about the LC system that Focke's code was written, not for shelving systems, but for the sort of classified book catalog common in German libraries around 1900. The historical implications of this are fascinating and we await a thorough study of LC's origins.

The DDC system, on the other hand, has been under almost constant critical scrutiny since its first edition in 1876. We are also reasonably well-informed as to its history. The fate of DDC after 1950 has been one of the strangest chapters in the history of modern librarianship. Today it is used by around 25,000 libraries throughout the world, and at least a third of recent editions have been sold outside of the United States. A complete French edition was published this year. The worldwide impact of DDC continues to pick up momentum. But in the United States, with the publication of the fifteenth edition in 1951, DDC was thought to be dead or dying. The fact that this edition, despite obvious limitations, began to bring DDC into the twentieth century was overlooked as librarians resisted changes which would require extensive reclassification. A decade and a half later, with the publication of the seventeenth edition, reactions in the United States were even more disasterous. We do not yet have a complete account of the extent of the erosion of DDC in the United States, but it appears to have been massive. In the mid 1970s, we are getting scattered reports of high school libraries changing from DDC to LC. Whether this change has been good or bad remains to be seen. However, it is ironical that DDC has been improved, but the changes necessary to make improvements have weakened its hold on librarianship in the United States. The use of DDC in the *British National Bibliography* (1950-) has been entirely beneficial and has helped to bring British classification experts into the editorial apparatus that guides the future of the system.

Reevaluation of Traditional Systems

While the library world at large went about working with the traditional systems, the complexity of the postwar world began to have an impact on the thinking of the more perceptive librarians in the United States and abroad. In the early 1950s, the most fundamental questions were raised about the very foundations of the traditional systems and about the validity of any general system of shelf classification. Criticism of library classification was nothing new, but never before had fundamental principles been so incisively examined. By 1950, Margaret Egan spoke of the "ferment over classification."[8] The impending impact of the computer, the dispersal of the ideas of Ranganathan and the tremendous increase in the production of scientific literature had an impact on how librarians thought about classification. There was a sense of urgency for the solution to problems of bibliographical control. Within this context, DDC and LC were examined and found to be grossly inadequate to deal with information in the modern world.

With a firm commitment to the fundamental role of classification in the organization of knowledge, Shera wrote a devastating critique of traditional classification.[9] Among the most enduring ideas from this influential essay is the proposition that traditional classifications are linear, and thus inadequate to deal with the many facets and multidimensional approaches of modern research. Inherent in much of the criticism that emerged around this time was the assumption that general traditional classification was inherently linear, but this was true only of our uses of specific traditional classifications and weaknesses in their structure. In any case, in the 1950s the world of knowledge seemed so immense, so unstable and so complicated that it was widely assumed that no general system would ever efficiently serve to provide subject access with any precision. So librarians learned to live with DDC and LC, and the locus of classification research was not to be found in librarianship, but in the information sciences.

Nontraditional Systems

The intense activity that has taken place in what may be broadly categorized as "nontraditional classification" can only be briefly noted here. The Universal Decimal Classification (UDC), which never lacked enthusiastic advocates outside of the United States, became the most widely-used special system UDC continues to move further and further away from its base in DDC (though both systems would clearly benefit if they were brought closer together again). Ranganathan became the most influential classification theorist of the twentieth century. Faceted classification became a practical reality, and hundreds of special faceted schemes were constructed. Strongly influenced by Ranganathan, the British Classification Research Group was founded in the early 1950s, and for the past twenty years has been struggling with the problem of finding a means of developing a new general classification system. Classification acquired a whole new vocabulary, with such terms as *links*, *rolls*, *planes*, *integrative levels*, *clumps*, and *isolates*, to mention only a few.

Throughout these years, theorists drew on widely scattered sources, such as systems theory, linguistics and psychology. If anything, we learned more about classification than we wanted to know. The optimism emerging from the rise of information science has turned to something bordering on despair Classification, we have learned, is not a physical thing consisting of schedules and indexes, it is a process that takes place in the human mind. But nonetheless never in the history of libraries have we known more about classification.

There is no problem in assessing the impact of these developments on traditional classification in the United States. The impact has been almost negligible. Somewhat cautiously, the DDC system has taken a few tentative steps towards the addition of concepts of faceted classification, though DDC is not and probably never will become a faceted classification. The LC system has not changed at all, and as more libraries adopt this system, the possibilities of change become more remote. The fact is that librarians who have adopted LC do not want it to change. The first law of classification dynamics is that the possibility of change decreases exponentially as more libraries adopt a given system. This law operates whether the changes might be good or bad in terms of the purpose of the system.

The high degree of standardization found in general library book collections does not extend to nonbook materials. Here, we find a great variety of local systems. The standardization of classification of sound recordings, for example, is not even on the horizon. Whether this is good or bad can only be answered subjectively if costs are discounted. One could argue that in organizing these local collections, librarians have a rare opportunity to use what they know about their collections, about the needs of their library users and about classification actually to construct good, working systems ideally suited to the functions and capabilities of their libraries. Such opportunities are not widely available to librarians who work with large book collections.

Summary and Conclusion

The development of applied library classification during the past quarter century has been captive to what went before. Many historical, economic, intellectual and emotional ties bound us to the misty past of the late nineteenth century. If some genius had devised a system better than DDC and LC, it is doubtful if the course of history would have been different. Even now, if we had a better system—and we could have a better system if we wanted one—it would probably not be taken seriously in the United States. If we have a problem, it is that we cannot even conceptualize a better system.

The future of classification in the United States will be determined by what it is that librarians want from a classification system. At the present time, they do not seem to want very much, and it may be that their modest expectations are well served. Our two systems do seem satisfactorily to serve the purpose of organizing materials on shelves. Or at least we are convinced that they are satisfactory for this purpose. But the conventional wisdom has not been subjected to any extensive and rigorous scientific research. The use of a shelf classification is a behavioral process. Something takes place in the mind of the user as he contemplates quantities of books on shelves. We know almost nothing about this process, and thus have no real way of evaluating the efficiency of either DDC or LC. We will probably continue to ignore this issue; but an issue we cannot ignore is the interrelationships of the computer, bibliographical access and classification.

More than anything else, our use of the computer will influence the future of traditional classification systems. The computer will either stabilize DDC and LC for many generations to come, or it will lead to the eventual abandonment of LC, a considerable reworking of DDC's notation, and possibly the development of a new general classification with extensive national ramifications. We have spent millions of dollars constructing networks and systems of bibliographical access based on computerized data bases. We have done this precipitously and with a single-mindedness of purpose that has failed to take into account the total implications of the enterprise. Not only has an inefficient and illogical system of subject headings been perpetuated on the MARC tapes, but each year thousands of titles are entered into this system tagged with LC class numbers which are almost completely useless in providing subject access through a computerized classified catalog. At the same time, in order to take advantage of the economic savings to be derived from networks and centralization, hundreds of libraries are switching from a system which shows some real potential for new modes of classified bibliographical access to a system with a nonhierarchical notation which is hopelessly antiquated for computerized retrieval systems.

Looking to the future, DDC has the capability of developing into a system that can exploit some of the potentials of the computer and at the same time provide a system of class numbers for shelving materials. The LC system, on the other hand, can probably continue to expand internally and provide a system of shelf numbers for the next fifty or more years. If this is what librarians want, and if it should come to pass, classification in the United States will, for all practical purposes, remain on the fringes of bibliographical access. Finally, in considering traditional classifications in both their broad historical context and in the complex world of today's libraries, one gets the uncomfortable feeling that we use these systems, not because they are the best or most efficient systems or even because we understand or like them very much, but simply because we are stuck with them.

Notes

[1] Jesse H. Shera, "Classification as the Basis of Bibliographic Organization," in *Bibliographic Organization, Papers Presented Before the Fifteenth Annual Conference of the Graduate Library School, July 24-29. 1950,* ed. by Jesse H. Shera and Margaret E. Egan (Chicago. University of Chicago Press, 1951), pp. 72-93.

[2] Henry E. Bliss, *A Bibliographic Classification* (New York. H. W Wilson Co., 1953).

[3] W. C. Berwick Sayers, *A Manual of Classification for Librarians and Bibliographers,* 3d ed. (London: Andre Deutsch, 1955).

[4] E. I. Šamurin, *Geschichte der bibliothekarisch-bibliographischen Klassifikation,* Bd. 2 (München-Pullach. Verlag Dokumentation. 1968)

[5] International Conference on Cataloguing Principles. *Report* (London International Federation of Library Associations, 1963).

[6] Raimund E. Matthis and Desmond Taylor, *Adopting the Library of Congress Classification System* (New York: R. R. Bowker, 1971).

[7] Rudolf Focke, "Allgemeine Theorie der Klassifikation und kurzer Entwurf einer Instruktion für den Realkatalog," in *Festschrift zur*

Begrüssung der sechsten Versammlung deutscher Bibliothekare in Posen, ed. by Rudolf Focke (Posen. Joseph Jolowicz, 1905), pp. 5-18.

[8] Margaret E. Egan, "Synthesis and Summary," in *Bibliographic Organization*, p. 257.

[9] Jesse H. Shera, "Classification as the Basis of Bibliographic Organization," pp. 72-93.

Traditional Classification: Characteristics, Uses and Problems

Josefa B. Abrera

Introduction

The "order of the sciences the order of things"[1] is the backbone of traditional classification schemes which have been used as instruments for bibliographic organization. On this premise emerged several concepts which have influenced the development of classification systems before the twentieth century. These concepts are:

1 the hierarchical order

2 the concept of classification for universal use

3 the enumerative system

The weaknesses and inadequacies of these schemes are attributed to the fact that their structures are derived from nineteenth-century principles of class logic rooted in the works of Plato and Aristotle. The history of the "grammar of classification" belongs to philosophy rather than librarianship. Despite the baffling contradictions which ensued regarding the "order of the sciences," it is advisable to look at some philosophical systems that have influenced bibliographic schemes, either directly or indirectly, to examine their deficiencies and determine why these nineteenth-century schemes are no longer adequate tools in the organization of materials and information in present day libraries. This will also provide students and practitioners with a conceptual framework that would assist them in understanding and synthesizing perspectives for classification schemes used in libraries.

Josefa B. Abrera is Assistant Professor, Graduate School of Library Studies, University of Hawaii.

Traditional Classification

Historical Prelude

Historical insight in the classification of the sciences* is provided by philosophers such as Plato, Aristotle, Bacon, Comte and Spencer. The table below is a schematic comparison of the different classical theories of the classification of the sciences according to the philosophers cited.

Table 1
Classical Theories of Classification of the Sciences

Plato (4th century B.C.)	Structure of the World of Forms	Collection and Division: Classify forms according to organized groups, as indivisible species, and in turn under genera
Aristotle (4th century B.C.)	Imitates nature	Doctrine of Predicables: Natural grouping of things according to structures and processes
Bacon (17th century)	Springs from one root and originates from the dominant faculties—Memory, Imagination, Reason	Tree System: Branches of a tree that meet in one stem
Comte (18th century)	Staircase Hierarchy: Morals Sociology Biology Physics Astronomy Mathematics	Law of Filiation: Decreasing generality to increasing complexity: complex dependent upon those that are simple
Spencer (19th century)	Abstract Sciences: modes under which we perceive Concrete Sciences: groups of sense impressions	Classification Hierarchy: Parallels Bacon's "tree system" and rejects Comte's "staircase hierarchy of knowledge"

*The term *sciences* is used in its unrestricted sense. It claims the whole range of phenomena, mental as well as physical—the entire universe is its field.[2]

Plato's Collection and Division

Plato's work on the theory of knowledge set forth the concept of knowledge as a priori and the deductive system of propositions dominated seventeenth-century thought and flourished in the nineteenth century. Plato advocated clear thinking, in terms of sharply defined abstract concepts. His theory of classification is reflected in his analysis of the structure of the world of forms.[3]

Aristotle's Predicables

The doctrine of predicables is the classification of conceptual relationships between a subject and its predicates. It is also referred to as Aristotle's doctrine of the categories—substance, quantity, quality, relations, place, time, position, state, action and affection. One recognizes from Aristotle's works some kind of overall classification of animals. Of course, there is the famous Tree of Porphyry which is a representation of the hierarchy of nature as Aristotle saw it.

Bacon's Intellectual Globe

Bacon in *Of Dignity and Advancement of Learning* has outlined a revolutionized classification of the sciences. In this work he reviewed the unchartered fields of knowledge and proposed a new classification of the sciences which is to supersede that of Aristotle. Bacon's plans for the advancement of learning included not only a reclassification of the sciences but also a reorganization of the divisions of human learning. Human learning emanates from the three dominant faculties of the understanding—memory, imagination and reason. This formed the basis of his analysis of knowledge.

"The divisions of knowledge," Bacon writes, "are not like several lines that meet in one angle, but are rather like the branches of a tree that meet in one stem."[4] Bacon's classification, particularly his analysis of history and sociology, has influenced the scheme of Spencer. The idea, common to Bacon and Spencer, is that the sciences spring from one root and branch off while Comte sees it as a "staircase hierarchy."*

*Comte asserts that for us to reach the supreme morals as soon as possible it is necessary that the study of each science is limited by the requirements of the one next above it.

Comte's Law of Filiation

In Comte's system of "positive philosophy," the Law of Filiation is associated with the Law of Classification. It determines the order of development by decreasing generality or by increasing the complexity of the phenomena—the mcre complex phenomena being dependent upon those that are simple. His "staircase theory of the hierarchy of knowledge,"[5] outlined in an elaborate scheme, is historically interesting but wanting from the standpoint of modern classification. However Comte is a link between Bacon and Spencer, for his writings on the Law of Classification of the sciences acted as a catalyst to Spencer's thoughts on the classification of the sciences.[6]

Spencer's Classification

Spencer's classification of the sciences parallels Bacon's concept of the sciences which is analogous to the "branches of a tree spreading out from a common root. ' He rejects the staircase arrangement of Comte's hierarchy. His classification combines the "tree" system of Bacon with Comte's exclusion of theology and metaphysics from the field of knowledge. It provides builders of classification schemes an excellent starting point.

In the preceding discussion of five philosophical systems, it is quite evident that the theory of classification is closely linked with the concept of the "universal order of things and ideas." The question thus arises. Is there such an order? If so, what is tne nature of this order? In the analysis of the processes involved in the classification and arrangement of things and ideas, one finds that the two processes complement each other, i.e., the former refers to the problem of sorting or grouping, whereas the latter addresses itself to the problems of unity, or the assembling of parts to form a whole.

The polemics that have gone on concern the problem of the order of ideas and things in structures, such as, order of complexity or order by class logic or order of power. Upon examination of the development of certain phenomena, one is bound to find that ideas reflect the evolutionary stages they go through *in time* from the simplest to the most complex.

Classification of the Sciences as a Model of Bibliographic Organization

An examination of library classification schemes of the prefaceted era would reveal a close analogy to the classification of the sciences which was advocated from the time of Plato to that of Spencer. The method used in constructing the schemes is deductive. Traditional classification begins with the assumption that classification is a process of division applied to a universe of knowledge. This universe is fragmented in stages by the application of various processes of division, namely:

1 Logical division
2 Physical division
3 Metaphysical division

One of the most fundamental divisions is the genus-species relationship. This is achieved by the classical method of logical division found in philosophical charts of learning, wherein all main classes spring from the traditional disciplines of knowledge. In physical division the parts of which an individual thing or aggregate is composed are distinguished—as in man· head, limbs, trunk, etc.; in a flower: sepal, petal, stamen, pistil, etc. In metaphysical division we distinguish a species in its genus and differentia. ın a substance, its different attributes, in a quality, its different variables or dimensions—thus, in man: animality and rationality; in sugar. color, texture, flavor, etc. Obviously metaphysical division can be carried out in thought alone In logical division, when the genus is concrete, its individual species can be exhibited in a museum case, likewise in physical division, the parts of an individual animal or plant may be separated physically, but in metaphysical division the parts cannot be displayed separately. e.g., taste or texture in salt can never be exhibited by itself alone.[7]

The deductive approach to such classification structure is based on the general assumption that the sum total of knowledge is arbitrarily divided into a number of main classes which are, in turn, subdivided into subclasses and so on down to a point where an *infima species* (irreducible unit) is arrived at.

The characteristic which dominates traditional classification schemes is the logical order of entities. This is accomplished by grouping entities according to the degree of likeness or similarity, then arranging them from complex to simple. Such a structure depicts a hierarchy of entities—the scheme order being that the genus and species follow a downward process until a unit in the hierarchy is irreducible (cf. Plato's method of Collection and Division) as opposed to the upward process (inductive approach) employed in modern classification schemes.

For our purposes we will use four general systems of book classification to illustrate the value and application of the classical concepts of classification as an instrument of bibliographic organization. These schemes are: Cutter Expansive (CE), Dewey Decimal Classification (DDC), Brown's Subject Classification (SC) and the Library of Congress Classification (LC), all of which manifest a close parallelism to philosophical classification systems characterized by a hierarchical structure, following the basic rules of the various processes of division.

Two of the schemes (DDC and LC), despite their uneven development and imperfections in their logical arrangement, notation and linear representation, are very much in use today. All four schemes are universal in range and scope. The schemes are hierarchical in nature, and in theory they follow the basic pattern of the inverted tree structure exhibited by taxonomic classification systems and are not based on "literary warrant."

In examining these schemes, one encounters combinations and variation of different principles proposed by individual philosophers who have formulated the concept of the classification of ideas. Except for LC (1901*), a product of team effort, the bibliographical systems produced during the nineteenth century were devised by individuals. DDC (1876), CE (1891) and SC (1906**).

The classification structure is manifested by the formation of classes and proceeds to sort out the subclasses or individual members of the class by enumerating the attributes or properties which differentiate one entity from the other. Thus the class

*LC's Class Z, Bibliography and Library Sciences, was completed in 1898.

**Work on Brown's Subject Classification began in the last decade of the eighteenth century.

Table 2
A Comparative Table of Main Classes Used in Four Book Classification Schemes[8]

DDC (1876)		CE (1891)		LC (1901)		SC (1906)	
0	General Works	A	Works of Reference	A	General Works	A	Generalia
1	Philosophy	B	Philosophy & Religion	B	Philosophy & Religion	B-D	Physical Science, Acoustics & Music
2	Religion	E	Biography	C-E	History	E-F	Biological Science
3	Sociology	F	History	G	Geography	G-H	Ethnology & Medicine
4	Philology	G	Geography & Travels	H	Social Sciences	I	Economic Biology
5	Natural Science	H	Social Sciences	J	Political Science	J-K	Philosophy & Religion
6	Useful Arts	L	Physical Sciences	K	Law	L	Social & Political Sciences
7	Fine Arts & Music	M	Natural History	L	Education	M	Language & Literature
8	Literature	Q	Medicine	M	Music	N	Literary Forms
9	History	R	Useful Arts	N	Fine Arts	O-W	History & Geography
		V	Recreative Arts, Music	P	Language & Literature	X	Biography
		W	Fine Arts	Q	Science		
		X	Language	R	Medicine		
		Y	Literature	S	Agriculture		
				T	Technology		
				U	Military Science		
				V	Naval Science		
				Z	Bibliography & Library Science		

Berries produces the species Cane fruits, Ribes, Huckleberries, etc., and under Cane fruits, the species Raspberries, Blackberries, Loganberries, etc. Hierarchies are thus created, based on successive application of characteristics. The Law of Likeness, which is the fundamental principle of the order of things, is employed. Ideas arranged according to likeness determine the order. In the main class concept, a generalia class is provided to accommodate materials treating a variety of subjects or subjects which are too general in nature to go to any other class. This class generally precedes the inclusive classes for the whole system, or it may be located within the subdivision of each class.

Division and subdivisions in these systems are arbitrary separations of closely related main classes. For example, one finds in CE, DDC and LC that the sciences are separated (e.g., Physical Sciences form a class apart from Technology, and Fine Arts from Useful Arts). On the other extreme, SC collocates Music with the Physical Sciences under Acoustics, thus stretching theory beyond practical considerations. Table 2 illustrates this characteristic.

For literary works, the arrangement is by national origin, genre and chronological sequence, except in SC which abandoned this concept in favor of four form divisions and alphabetically listing all individual authors, regardless of national origin or the period in which the work was written, under each literary form.

In the philosophical system of classification, only single concepts are included, e.g., *Forest exploitation, Forest utilization,* whereas bibliographical classification deals with compound concepts, e.g., *The effects of government regulations on forest exploitation and utilization.* Traditional classification attempted to cope with this problem by listing every possible concept that occurs, simple and compound, thus creating an enumerative schemes. Needham[9] states that the enumerative approach failed for the following reasons:

1 Enumeration can never be complete.

2 The theoretical application of class logic can be carried too far beyond practical reality, causing confusion because certain entities do not fall under any generic hierarchy.

3 Cross-classification occurs because there is an overlapping of

attributes in an array of classes and compound subjects are presented as if they were simple subsets of the preceding subject.

Traditional classification employs some form of notation, either pure or mixed. Notation is merely a coding device which displays the order of entities in a scheme and facilitates the mechanical arrangement of materials on shelves. The problems associated with notational requirements will be dealt with later.

Concomitant to the employment of notational devices in classification schemes is the introduction of synthesis and mnemonic features. Such features are exhibited in various ways, particularly:

1 Number-building devices which take the form of common tables for standard form divisions and geographic or area tables as illustrated in CE and DDC, and the categorical tables of recurring elements in SC. LC assigns each subject a set of standard subdivision and area tables.

2 Mnemonic features are introduced by means of auxiliary tables listing constantly recurring categories. Each of these categories is consistently denoted by the same notational symbol, thus enhancing the memory value of the notation.

In reviewing the principles of dividing knowledge set forth by classical philosophers, Plato's concept of structure of the world of forms and Comte's Law of Filiation have played an important role in the hierarchical features evident in traditional classification schemes. The ordering of concepts according to their degree of likeness, arranged in an evolutionary form from the simple to the more complex or vice versa, is the Baconian influence but probably more of the Aristotelian doctrine of predicables.

Problems

The originators and proponents of the traditional system in adopting the classical concepts worked out certain practical adjustments in the operational level of the schemes. These were made necessary because the implementation of the system called for a functional organization. More importantly, collections in libraries required the display of relationships not only by classifying similar items but by integrating those relationships that showed the effect of one class upon another. However, the classification in the traditional system excludes the latter. A problem in this approach surfaces. A philosophical system encompassing universal knowledge is inadequate as a model in devising a classification system which deals not only with complex concepts but also with the vehicles that transmit them.

The publication and dissemination of materials in a variety of subjects and physical formats are continuously increasing at an exponential rate. The inevitability of future discoveries and explorations renders the universe of knowledge continually changing in quantity. For effectiveness, a classification scheme derived deductively depends upon the invariability of the assumed sum total of knowledge. In effect such a scheme would require continuous revision and updating in order to keep abreast with the state of the sociology of knowledge. Thus a permanent complete scheme covering the whole field of knowledge is still an impossibility.

What needs to be understood is the fact that the deductive approach lacks the flexibility to accommodate new subjects whenever they occur *sans* revision. An enumeration of all subjects within a class or set of subclasses is nearly an impossibility. Furthermore, such enumeration is compounded by the problems associated with classification schemes that are universal in range and scope of their applicability. To produce a universal classification of knowledge, one must, theoretically. have all knowledge available. In reality only representative sam plings of the different branches of knowledge are covered in universal classification schemes. It is difficult enough to be certain that a set of subclasses completely covers the parent class and it is much more difficult to ascertain extant classes and predict what other classes may be added in the future.

The ever-recurring problem of synonymous terms and their standardization in a universal classification contributes to the problem in a scheme dependent upon the state of the theory of knowledge and its ramifications.

Schemes that are enumerative in nature have most of their compound subjects precoordinated. The tables of CE, DDC. SC and LC attempt to find a place for complex concepts that are likely to change within short periods of time or would vary from one document to another. The process of classification becomes, in most cases, nothing more than an exercise in approximation. For example, the tables of CE, DDC, SC and LC do not include compound concepts, such as, *The production of goats' milk cheese* or *The incidence of asthma in winter of 1962* or *Spring harvesting of wheat.*

The progress of knowledge has contributed to the instability of main classes because such categories tend to be names of collections of ideas that are very much colored by the theory and state of knowledge. An entirely different view of life and knowledge is expressed in classical schemes which give precedence to philosophy and philosophical writings, or the Russian scheme which gives more importance to Marxism or other socialist works.

Traditional classification schemes choose the major disciplines as their main classes. (See Table 2.) Aside from the fact that it is difficult to clearly draw the boundaries between main classes and determine the required number of main classes satisfactorily, there is the further disadvantage in using such disciplines as the *summum genus* in a scheme. It is often possible to show how the entities of one class vary until such entities begin to approximate the entities of another class. Then the suspicion is generated that there may be no fixed classes in nature and the once obvious differences observed in entities are all products of differing environments in which these entities are found and through which they have passed. A class organization of knowledge which includes concrete and empirical entities fails to be wholly adequate because it is incapable of organizing the varying characteristics that develop in entities in varying environment.

Where main classes are used, the classification scheme must provide some rules for establishing order in the scheme. The problem of collocation presents another problem since there are no restrictions on proximities where schemes are essentially linear. Another problem relating to the lack of rules is the application of the rules of logical division. Logical division does not provide rules relating to compound subjects, nor does it give rules for the arrangement of classes in an order. In a theoretical scheme such rules may be deemed unnecessary, but when a scheme is used for bibliographic organization, such a

.32

rule is almost imperative. It is necessary to have a preferred order for arranging physical objects on the shelves or the entries in a catalog. There are further limitations to the use of logical division for bibliographic organization. The theory of division breaks down amidst the complexity and variety of concrete entities. Ideally when a genus is subdivided into species, whether once or through several stages, it is assumed that at each stage a number of definite species are included in that genus. For example, in the biological sciences such a division is clear-cut and definite. But in other classes, for the most part, it is not possible to expect entities to fall into the genus-species relationships which would fit into the perfect structure of a logical division. Neither would it be possible to completely exhaust the parent class or to enumerate all the individual members of the class that already exist or may be discovered. The most serious limitation to logical division is that it only deals with one kind of relationship, that of a thing and its kind, known in scientific jargon as *genus* and *species*. In library classification we are concerned with several other types of relationships, therefore it is necessary to apply other types of division in a classification scheme developed for use in libraries. The general relationship between genus and species is particularly not applicable to the organization of relationships involving spatial position. In addition, it is less possible to represent in class logic that part of the empirical sciences which deals with the continuous or discontinuous alteration of behavior of specific entities evolving from the changes in their environments.

Traditional schemes are further besieged by the problems of notational requirements. Schemes which were developed originally for the classification of ideas are used to classify ideas contained in a physical object. In arranging books in libraries. we are forced to consider their physical characteristics, thus requiring modification of any pure knowledge classification

The interpolation of a notation adds to the confusion. The notation, which serves as a location symbol, becomes a code representing the natural language. Notation introduces some undesirable inflexibility to a scheme and is "still-born" since it is incapable of growing. Notation is a necessity in a classification scheme since it displays and preserves the desired order of entities and acts as a shorthand code for the natural language. Since its function is to preserve order it is obvious that a notation must make provision to include new subjects as they arise

The quality needed for this aspect is referred to as *hospitality* To accommodate developing subjects, expansion is provided by reserving large blocks of unassigned numbers. Since there is no way of prejudging what subjects will develop either rapidly or slowly, faulty apportionment of notation arises. Many important classes are not developed sufficiently enough and there is very little room for expansion, or in some instances, one ends up with unwieldy, lengthy notation. On the other hand, subjects of very limited significance have comparatively large blocks of numbers assigned. This problem is closely tied to the fact that traditional classification schemes were constructed on the a priori basis of class division, with the exception of LC which was based on literary warrant. In most schemes one will find classes that are well developed and some classes that are not represented by any single publication, or if at all, very few indeed.

If knowledge were static, notation would not be a problem since it would be easy to add the notation to the scheme after it is completed. But since new subjects are created within classes or divisions, it is imperative that such new additions be located in their correct place within the scheme. If the notation is inflexible, then it will dictate the order, thus preventing its effective use. Notation does not improve the scheme *per se* but is a necessary evil in a working classification.

Adaptability to machine techniques requires that a scheme should have the facility to express generic relationships if hierarchical searching is required. Relationships among compound and composite subjects need to be made explicit through the use of notational symbols. In machine searching, it is necessary that concepts be associated consistently with one unique code. This process would be difficult and expensive to achieve in any traditional scheme based on main classes, since the notation which represents any particular concept keeps changing according to the class from which it is derived. On the other hand, the use of a notation that expresses hierarchical structure which is effective for machine storage and retrieval could be exploited so that the genus-species relationships could be displayed by lengthening or shortening the notation representing the concept. Thus the user would be able to broaden or narrow his search at the level of any particular element in a compound subject.

An enumerative scheme fed into a computer will not allow retrieval of a particular element. To be effective each class number needs a code representing only one subject and used consistently for that subject. Except for tables of standard subdivisions and area tables, this is not the case with traditional classification schemes created before the faceted approach was recognized.

Summary and Conclusion

This paper attempted to show that a classification scheme should not be evaluated on the basis of its completeness or "neatness" alone, but also on the extent to which it advances knowledge and achieves the purpose for which it was originally created.

Traditional classification schemes have proved inadequate as instruments of bibliographic organization in the face of the ever-expanding field of knowledge and of technological developments, particularly computers. The schemes are handicapped by limited recall capabilities due to the dispersal of related aspects of entities inherent in enumerative schemes and in one-dimensional linear classified arrangement. From them, however, certain fundamental principles, theories and concepts of the organization of knowledge have emerged which are crucial to the development of modern classification schemes. In 1970 Foskett wrote:

> In our technique for information control the time is ripe for the overthrow of existing paradigm, but we should not, at the same time, reject those aspects of it that can usefully contribute, for what we need now is not a blank slate, as was once thought, but a genuine synthesis.[10]

The problem of terms and their standardization in a classification for universal use can only be resolved by an overall acceptance of a single authority. Such standardization might be difficult to achieve since there is no such thing as an "all-around" view of the world. People's perception of reality is conditioned by the constraints of their cultural orientation. If we are seeking to accumulate a store of knowledge that may be employed in an eclectic fashion, we should strive to eliminate all vagueness. Furthermore, we usually believe that nature is not vague and it should follow that knowledge of nature should not be vague. In practice this vagueness cannot be eliminated. However, it can be reduced. The particular kind of organization that traditional classification schemes give to knowledge make it especially difficult to eliminate vagueness of connotation and denotation to any desired degree.

The rapid advancement of knowledge requires that schemes undergo frequent revisions and updating, even in areas where knowledge remains unchanged. Revisions will still be necessary but will take the form of extensions. In view of this, it is imperative that librarians become "independent classifiers." This means that librarians should have complete understanding of the principles, theories and concepts of classification so that they are in a position to amend, modify or revise any classification scheme within the normal limits of human error.

Although classification is a matter of picking out and conceptually grouping together certain entities of a heterogeneous field, it should be remembered that in the process its grouping of entities interrupts and disregards relationships between entities that fall into different classes and overemphasizes relationships between entities that fall into the same class. And this is especially true with schemes developed deductively. In effect we want a scheme that will reflect class organization and at the same time reflect cross-class relationship. A functional organization can preserve better than class organization specific variations in entities, and it would be foolhardy to sacrifice this advantage by allowing "likeness" to absorb or displace dissimilarities in classification. For knowledge may reflect a *knowledge* of a class of entities used as a justification for a particular classification and as an explanation for the fact that members of the same class behave differently. But it is not a knowledge organized exclusively by class relationships. Even if knowledge is about members of a single class, it contains references to entities belonging to other classes, thus it should be an organization in terms of relationships that cuts across classes. Futhermore such relationships should have the capability of being machine-manipulated and retrieved in a variety of ways at every accessible point.

A classification scheme must not arbitrarily group the materials of experience into few classes. There may be major classes, but there must also be numerous subclasses equipped witn cross-classification mechanisms.

Classification schemes that are closely associated with philosophical systems have a strong tendency to be regarded as either "natural" or "artificial," which is perhaps distortive of reality. For man to be fully satisfied with a classification system he needs to become aware of his own classifying activity and consciously to strive to control and master it. And this control and mastery are best exercised in a purposeful manipulation of classificatory concepts, with full awareness of the various ways in which complex entities could be classified and of the needs which any desired classification schemes must satisfy.

Notes

1 Ernest C. Richardson, *Classification, Theoretical and Practical*, 3d ed. (Hamden, Connecticut: The Shoe String Press, 1964), pp. 9-21.

2 Karl Pearson, *The Grammar of Science* (London. Walter Scott, 1892), pp. 29-30.

3 Francis M. Cornfield, *Plato's Theory of Knowledge, the Theatetus and the Sophist of Plato Translated with a Running* Commentary (London. Routledge & Kegan Paul, Ltd., 1964), pp. 268-273.

4 Quoted by Karl Pearson, op. cit., p. 445.

5 An elaborate discussion of Comte's staircase theory of the hierarchy of knowledge' is found in his *System of Positive Philosophy* (New York Burt Franklin [1966]), IV, pp. 161-165.

6 Pearson, op. cit., p. 448.

7 H. W. B. Joseph, *An Introduction to Logic*, 2d ed. rev (Oxford Clarendon Press, 1967), pp. 313-335.

8 W. C. Berwick Sayers, *A Manual of Classification for Librarians*. 4th ed., completely rev. and partly rewritten by Arthur Maltby (London Andre Deutsch, 1967).

9 C. D. Needham, *Organizing Knowledge in Libraries, an Introduction to Information Retrieval*, 2d rev. ed. (New York. Seminar Press, 1971). p 133.

10 D. J. Foskett, *Classification for General Index Language, a Review of Recent Research by the Classification Research Group* (London. The Library Association, 1970), p. 46.

Modern Classification: Characteristics, Uses and Problems

John H. Schneider

It must be clearly stated at the outset that this is not a review paper. Instead, I have taken this opportunity to present personal opinions regarding the role of certain types of classifications in our modern automated environment. Although there may be some statements the reader may not agree with, they will hopefully be offset by other concepts which do have merit and can be used by the reader to improve operational information systems or be incorporated into plans for new systems.

Classification As Used in this Paper

Since the word *classification* can be used in so many different ways, it is essential to indicate that in this paper *classification* refers to highly structured, hierarchical classifications found on the far right in Figure 1. This figure shows a spectrum of sources of index terms, concepts and/or notations used in various types of indexing and retrieval operations.

uncontrolled vocabularies → alphabetical subject authority lists → three level thesauri → word "Trees" → multilevel hierarchical classifications

Figure 1

In general, classifications are used to organize concepts (often expressed as single words or short phrases) in a logical, systematic fashion and to show relationships between concepts.

John H. Schneider is Scientific and Technical Information Officer, National Cancer Institute, Bethesda, Maryland.

They are created by grouping concepts that share similar characteristics, particularly those characteristics that are most significant for meeting the anticipated retrieval needs of a known user group.

The most useful classifications from a user standpoint are those that reduce the effort needed to retrieve precisely the needed items of information at exactly the level of detail required. Ideally, the classification should permit the user to select one or two categories containing all the required information instead of having to identify many separate concepts and link them together by a complex search strategy, often requiring several revisions, before the desired information is retrieved. The performance of a classification, then, is mainly a function of how well the developers of the classification have forseen the need of the users and grouped concepts together for users at the multiple levels of generality and detail likely to be needed by most users.

The preceding statement suggests that the performance of an information retrieval system can be improved by moving from left to right in Figure 1, which is arranged in order of increasing degree of organization and increasing delineation of the relationships between concepts. In actual fact, information systems that started with sources of concepts to the left of center in Figure 1 have generally been forced to move to the center or right-hand side of Figure 1 in order to improve performance This left-to-right shift has been well described by Lancaster.[1] who also points out that the thesauri are actually a rather extensive but covert or hidden classification.

When systems progress through the sequence shown in Figure 1, they seldom take the ultimate step which involves the development and use of a multilevel hierarchical classification and a hierarchical notation. It is because this type of classification has such tremendous potential for improving the performance of information retrieval systems. and because it is so seldom used, that I have chosen to emphasize it in this paper

What is a *Modern* Classification?

To be truly "modern" a classification must be (a) free of constraints associated with many existing "traditional" classifications so that it can be easily and frequently revised to keep it up-to-date, and (b) structured and used in a way that takes full advantage of the capabilities of computers and computer systems.

Full freedom from constraints is most nearly achieved when a "modern" classification is developed *de novo*. In other words, there is little chance of success in undertaking the thankless, very difficult task of trying to modernize a classification that is hopelessly out-of-date. The existence of highlevel committees who must approve each change, no matter how trivial or obvious, all but guarantees failure of any attempt to keep a classification in a fluid evolutionary state with frequent modification in response to changes in information being indexed.

If it is really necessary to have classifications which are used with only minor variation to arrange documents on shelves in hundreds, if not thousands, of libraries throughout the world, (and space does not permit me to do more than suggest that this may no longer be necessary), then it is clearly desirable to keep changes in the classification to a minimum, and modernization may actually be undesirable. This paper does not deal further with this difficult dilemma.

Instead, I am dealing with those situations where it is possible to take a fresh or a new look at a limited subject area and create new or extensively revised and open-ended classifications which conform as closely as possible to current thinking, current terminology and the present conceptual framework of those who work in the subject area. These modern classifications which cover selected subjects in considerable depth (rather than the near-universal and sometimes superficial coverage of some traditional classifications) are normally used in one rather centralized system by only a small group of indexers. The retrieved information is usually only a citation or citation plus abstract or some other document surrogate which does not need a unique classification number for determining its physical location on a shelf or in a file drawer. In these cases, the classification can be virtually constraint-free, which is the first criteria of a truly modern classification that will remain viable and useful for an extended period of time.

The movement away from the need for a one-to-one correspondence between a document and a notation derived from a classification also leads directly to the second criteria for a modern classification. In modern systems, *classifications should be structured and used as sources of multiple categories that are independently assigned to each document* and are manipulated through simple Boolean logic in computerized systems to retrieve only those documents assigned to any desired *combination* of categories. This use of multiple categories from a classification is exactly the same process as selecting multiple descriptors or keywords from a list of subject headings or a thesaurus. The difference, as described in more detail later, is that the categories from a classification can frequently be much more powerful descriptors than isolated keywords or phrases

This multiple assignment o' categories from a modern classification to each document is clearly different from the use of more traditional classifications to assign a single unique number to a document.

The Value of Modern Classifications

The problem of GIGO (garbage in, garbage out) in information systems can best be solved by thorough analysis and organization of the information as it is entered into the system. Assignment of categories from a modern classification to each entered data item specifically identifies that item by placing it with a group of other items having nearly identical characteristics. Because of its location within the hierarchy, this process also relates the new data to other data items already in the system As a result, the retrieval of very "clean" data is greatly facilitated.

Use of a modern classification to analyze and organize data at the time of input is particularly valuable in the "soft" areas of science, such as social science, political science, the humanities, and other subject areas where concepts are not closely linked to specific technical terms and are often described in wordy and imprecise phrases or in jargon that has meaning only to a small in-group.

In the physical sciences indexing problems arise more from the very large and increasing number of highly specific technical terms which must all be included in a search for information in a given area. Again, assignment of categories from a modern classification automatically places new input into small groups of closely related information items that can often be retrieved as a unit by specifying a single category number.

The problems mentioned above are compounded by the very large size of many data collections, the increasing specialization of both data and users and the rapid emergence of new subject areas resulting from the interconnection of two or more lines of research that were previously distinct entities.

Again, a constraint-free modern classification can effectively handle these problems, if it is kept up-to-date by frequent subdivision of categories to deal with new subspecialties and by the frequent addition of new categories to deal with new interdisciplinary topics. No matter how large the collection, the specificity of category descriptions, the detailed relationships built into the classifications, and the ability to specify the exact generic level needed for each search make it possible to retrieve data items in a very narrow subject specialty with *high precision*. When the categories are used in Boolean expressions, they become even more powerful precision devices.

At the same time. modern classifications give the user tight control of synonyms and near-synonyms as well as related concepts that can only be expressed by a string of words. This synonym control. along with the automatic grouping of closely related items by the assignment of categories from a modern classification, enables the user to achieve *high recall* of all relevant items, no matter how diverse the terms used to describe those data items.

In any information system there is an inverse relation between recall and precision. However, the preceding paragraphs state my conviction that modern classifications can be used to increase both recall and precision above the levels possible with other types of indexing tools. Experiments at the Smithsonian Science, Information Exchange [2] [3] have provided some experimental verification of this point.

Structure and Organization

Because computers can retrieve any desired combination of categories at the time of retrieval, it is no longer useful in mod ern automated systems to precoordinate categories from a classification at the time of indexing. Thus, traditional synthetic classifications, such as faceted classifications and colon classifications (to the extent that these terms suggest the need for synthesis or precoordination of different facets or elemental categories to form a unique notation at the time of indexing), are not truly modern classifications and are not well suited for use in modern automated systems.

Instead, to be fully effective, a modern classification should be enumerative rather than synthetic. That is, it should have a deep, multilevel, open-ended hierarchical structure which lists or enumerates all the unique concepts needed for indexing data at all levels of detail likely to be needed by the user. Whatever precoordination of words and phrases is desirable for identifying basic concepts should be built into each category as the classification is developed.

If this is done well, most of the concepts needed for indexing or retrieval will have a one-to-one correspondence with categories in the classification, and the need for post coordination will be greatly reduced. The resulting classification, structured along the lines just suggested, contains within the definition of each category a high level of "judicious precoordination" which Lancaster points out is useful for reducing the problem of noise in information systems. Ways to build precoordination into categories will be briefly outlined in a later section.

I have suggested elsewhere[4] that the acronym HICLASS be used to describe enumerative, "deep" or multilevel *HI*erarchical *CLASS*ifications with extensive precoordination built into the categories. The same paper describes how this type of classification was successfully used with no post coordination in a system for selective dissemination of information (SDI) based on single-hit matching between any *one* of several categories assigned to a user and any *one* of several categories assigned to a document.

Although the SDI system just referenced demonstrated that post coordination of categories from a HICLASS type of classification is not essential, a special type of post coordination would have improved the matching of users and documents About one-third of the documents which users rated as being of no significant interest would not have been matched with those users if answers to a few simple questions of the following type had been post coordinated with more substantive subject categories:

Does the information involve a human patient?

Does the experiment involve human tissue?

Was the experiment performed exclusively in laboratory equipment and not in a living animal (i.e., was it an *in vitro* or an *in vivo* experiment)?

Did the information involve newborn or very young animals?

In other subject areas, similar questions might cover geographic locations, ranges of years or other periods of time, anatomical sites, etc., if these aspects of the information were not already used as major categories. The questions themselves can take the form of a simple checklist which supplies special tags that can be checked off for both users and documents at the time of indexing and used as a form of post coordination at the time of retrieval.

In modern computer systems, answers to many of the types of questions just listed are best handled as a short string of bits in the computer record. Each bit in this bit string can be turned off or on, depen ng on the answer to a corresponding question or item in the checklist. Screening of these bits to make sure they match the user request is a simplified modern form of post coordination. A modern classification should be structured and organized to take full advantage of this ability to use the bit screening capability of computers.

Precoordination and Post Coordination

The best modern classifications probably fall somewhere between the faceted or colon classifications (which require extensive coordination of categories to define concepts needed for retrieval) and the enumerative, deeply-detailed, multilevel hierarchical classifications of the HICLASS type which have extensive precoordination built into the categories, and consequently need only minimum coordination of categories. It should be stressed again that coordination of categories in a modern system implies use of the computer only for *post* coordination at the time of retrieval, and not precoordination at the time of indexing.

A modern classification can be used to achieve the optimum balance between the amount of precoordination built into categories and the amount of post coordination required by system users. I have suggested elsewhere[5] that it is much better from a total information system viewpoint to tip the balance far toward the side of precoordination when the classification is developed. This is accomplished by having subject experts and potential users devote considerable time, effort and thought to building an enumerative classification with categories that fully describe each concept (with precoordination of all its components) and a deep multilevel hierarchical structure that clearly and accurately relates each concept to other categories in the classification. This operation (apart from revisions and updating, which all systems require) is performed only once by a very few experts.

In contrast, a number of existing systems now have hundreds of on-line users, often with very limited knowledge of the subject area, who make thousands (and for some systems, several hundred thousands) of searches each year. The development and use of a good modern classification is easily worth the effort, if it results in a saving of even a few minutes per search, even a slight increase in the recall or retrieval of useful information and a slight decrease in the "noise" or an increase in the precision of the retrieved information. These small savings by system users, multiplied by the number of users, the number of searches and the number of years of use, should more than offset the one-time cost and effort of building and using a good modern classification for structuring, analyzing and organizing the massive amounts of information in contemporary information systems.

Notation and the Index for Modern Systems

Two requirements for a modern classification system which have not yet been mentioned are. (a) a notation system that is maximally useful in computerized searching and (b) an extensive alphabetic index which serves as a 'lead in" vocabulary or "entry" vocabulary.

Since the index has no special features, it need not be discussed in detail other than to stress that it is essential for both users and indexers and that no classification is likely to be of much use without it. The index serves to lead "naive" users from a simple term in the alphabetic index to a sophisticated concept surrounded by related concepts in the hierarchical classification. In this way, the logical thought processes and relationships of categories built into the classification by subject experts are used indirectly in every retrieval operation, thereby upgrading the operation, no matter how inexperienced or how lacking in understanding of the subject the searcher may be. This results in a considerable upgrading of the content and usefulness of the retrieved data in most searches.

The word *notation* (sometimes called *class numbers*, which incorrectly implies that only numbers are used in the notation) refers to a string of characters used to uniquely identify each category in a classification. The notation makes it possible to use an unlimited number of words, phrases, synonyms, near-synonyms and variations in spelling or plurality precisely to describe the conceptual content of each category in the classification.

For the type of modern hierarchical classification I have advocated in previous paragraphs, the notation must also be hierarchical in order to reflect the structure and organization of the hierarchy. In other words, the notation for all subdivisions of a major category must be a meaningful, expressive notation rather than a meaningless string of characters. For example, notations 51.83, 51.832, 51.8345 and 51.83FT4 identify four categories which are all subdivisions of category 51.83, which in turn is a subdivision of category 51.8, which is a subdivision of the major category 51., etc. It is highly undesirable to use periods to set off each new number added to the notation (compare 51.83FT4 with 51.8.3.FT.4) since this adds extra and unnecessary characters and is more difficult to manipulate both manually and by the computer.

The advantage of this hierarchical notation is that the indexed items assigned to very specific categories (for example, 51 8345 or 51.83FT4) are clearly linked by the notation to every category that is more generic (51., 51.8, and 51 83). In this way, every character in the notation has meaning and reflects both subject content and precise relationships between categories in a very compact, concise way.

This type of notation permits users to select a notation at any desired generic level and let the computer identify and retrieve all subcategories of that category using only the notation as an instruction. No other automatic position, generic-to-specific posting referral process or mapping procedure is necessary for the automatic retrieval of all subcategories subsumed under the major category specified by the user.

Although the notation in a modern classification can contain letters or any other characters acceptable to the computer, it is best to use numbers, since even long strings of numbers (i.e., the 11 digit numbers required for a long distance telephone call) are relatively easy to memorize and manipulate for the few seconds needed to assign them to a document or enter them into a computer system. In fact, the first few numbers representing major categories can usually be recalled without any look-up process by those who use the classification regularly.

In contrast, strings of nonsense letters (mostly consonants) are very difficult to memorize and manipulate. The mnemonic approach sometimes used to build notations often results in a clumsy, complex, lengthy string of characters that are much more likely to introduce errors than simple strings of numbers.

A final comment on the notations is that they must be open-ended. Space must be left between major categories or groups of categories for the addition of new categories at a later time. Even more important, there must be no limit on the length of the notation. Although the classification should be organized in such a way as to keep the notation as short as possible, it must always be possible to add additional characters to the right of any notation to reflect new subdivisions. Any classification which sets a limit on the length of the notation has unnecessarily restricted growth and evolution of the classification, thereby creating increasingly difficult problems in keeping the classification up-to-date.

Procedure for Building

The desirability of developing new modern classifications rather than trying to modernize existing out-of-date classifications was stressed earlier. It therefore seems appropriate to suggest some useful guidelines for the building of such new classifications.

It is usually easy to outline the major categories for a given subject area. I would urge readers to select a topic of interest and see just how easy it is to develop a logical outline of major categories in just a few minutes. This is the only stage at which a tentative structure is based on preconceived ideas. After this point, the classification is developed and extensively modified only by creating new categories and arranging them in the best order for precise indexing of concepts obtained from input documents.

An extremely valuable procedure for organizing major categories in a subject area is to construct a two-dimensional matrix with different aspects or facets on each axis. For example, the field of radiation biology is best represented by a matrix of types of radiation vs. types of organisms, organs, cells and molecules being irradiated. This rapidly divides the field into many major categories that can then be subdivided as the need arises. The field of biochemistry logically falls into a matrix with major classes of compounds (proteins, amino acids, lipids, carbohydrates, nucleic acids) on one matrix and major analytical subdivisions (synthesis, chemical properties, physical properties, uptake and transport by the body, etc.) along the other axis. Much of biomedicine falls into a matrix with major disciplines (pathology, physiology, pharmacology, toxicology, clinical medicine) along one axis, and organ systems (lung. liver, stomach, skin, bone, etc.) along the other axis. Similar matrices can be constructed to cover large areas of information in most subjects.

Each intersection of the two axes becomes a unique, distinct category in the classification, although sometimes a whole row or column from the matrix is used as a category. The notation should be synthesized from numbers that show how the major category was synthesized. For example. major disciplines can be assigned notations as follows: 52. for pathology, 53 for physiology, and 54. for pharmacology, etc. Organ systems can be assigned as follows. 43 for kidney, 52 for lung, 83 for skin, etc. Intersection of these two sets of numbers gives 52.43 for all kidney pathology, 53.52 for physiology of the lung, and 54.83 for pharmacologic agents that act on the skin. When combining two sets of numbers, the numbers representing the most open-ended and detailed aspect must always be placed to the right of numbers representing the broader aspects.

The important point in giving these examples is that use of matrices is the best way to build precoordination into the categories as the classification is developed. It is quite different from precoordinating or post coordinating individual categories after the classification is developed.

The next step after the initial set of major categories is created is to identify representative documents that fall in the selected area and assign concepts from those documents to categories in the classification. During this phase, it is necessary to add many new categories and subdivisions of existing categories to the classification—often at the rate of several new categories per document. In addition, existing categories must be shifted to new locations, deleted, or completely revised by using new words to reflect increased or decreased scope.

The structure of the classification must be extremely fluid and flexible at this stage. It should not be used for final indexing of any document until many hundreds of documents representing the whole subject area covered by the classification have been used to improve and flesh-out categories in the classification.

During this developmental period, the classification should match the conceptual organization, the way of thinking and the conceptual framework presented by authors of the documents and by representative users. Clearly this requires extensive input from subject experts and review by those who are actively working in the subject area.

The emphasis in this process of developing a new modern classification is on a very practical, pragmatic, empirical approach with as few rules or constraints on the structure or organization of the classification as possible. Whatever wording or organization of categories works best and seems most useful should be used. Any attempt to use vague, general or artificial concepts (i.e., personality, matter, energy, space, time, etc.) to organize or to create categories is of no value in developing a specific modern classification that accurately corresponds to the structure and organization of the subject area it will be used to index.

Even after the early pilot phase is completed and use of the classification for routine indexing begins, the same flexibility in revising the categories as soon as the need arises is required if the classification is to remain viable. The best person to make these changes is the individual who is indexing documents or formulating a search and sees the need to incorporate a new concept into the classification. *This should be done immediately, the first time such a need is identified.* The change must be made quickly and easily. It cannot wait for a decision by a committee.

An indexer or searcher should consult with others if there is any question about where to place the new concept in the hierarchy. However, if this is not possible, the concept should still be entered into the classification at once. Subsequent review of changes may shift the category at a later time, but in the meanwhile it is available for use the next time the same concept is encountered. The idea of postponing a needed change until review by some elite "authority" is completely unacceptable since the indexer or searcher is most likely to know how the new concept is described and related to the rest of the subject area by the author of the document or the individual requesting a search.

The process just described is designed to keep the classification modern and up-to-date. The focus must be on making the classification even more useful a year and ten years from now, rather than on whether the indexing of a few documents today or in the past becomes invalid because of a change in the classification.

In this connection, it is worthwhile mentioning that the type of modern classification advocated here has very little need for any manual reindexing of older documents when the classification is changed. Inserting a new concept or broadening the scope of an existing category has minimal effect on past indexing. Shifting a category to a new location with a new number can be followed by a corresponding change made automatically in the computer files. Subdividing a category means that documents indexed under any more general category can still be retrieved if the requester is willing to accept documents that could only be indexed at the more generic level (either because the category had not yet been subdivided or because the document covered so many aspects of a major category that it would have taken too much time to post it to all the subdivisions of that category). The inclusion of more generic categories reduces the precision of the retrieval, but is an excellent recall device which should be used for all searches unless the user specifies otherwise.

Three additional comments on building a classification may be useful. First, the categories should be organized and selected in such a way as to keep the notation as short as possible. Second, it is highly desirable to divide each category into only five or six major subcategories so that only one number or letter needs to be added to the notation for the major category. If this is impossible, then it is perfectly permissible to have up to 99 subdivisions and use two digits after the notation of the major category to represent each subdivision. Third, long lists of specific items (compounds, chemical elements, names of organisms, etc.) that need to be itemized under a major category are best arranged in alphabetic order with the first few characters of each word, followed by one or two numeric digits incorporated into the notation. In this way, the order of the notation reflects the alphabetical order in long lists of terms that all fall in the same class (antibiotics, bacteria, countries, names of individuals, etc.). Such alphabetic lists, imbedded in the classification, are much easier for both indexer or searcher to use than groupings of items into artificial subclasses.

In closing this section, I might mention that a computer system named AUTOCLASS has been designed for automated creation and updating of both the classification schedule and the alphabetic index that accompanies it.[6] Changes in categories result in corresponding changes in the alphabetic index, in cross-references to the changed categories and in cross-reference statements within the changed category, since all these linkages are recorded and used by the computer during each update step. Lists of changes made during the updating of categories and index terms or cross-references that need to be checked as a result of the changes are printed out during each update cycle. The existence of this type of automated system makes it much easier to build and update a classification than was previously possible.

Use in an Automated Environment

No matter what system is used for indexing and retrieval, it can probably be improved by using a modern classification in combination with the existing system. A good example of a "hybrid'' system would be the combination of free-text searchin- (to retrieve on any specific term in titles and/or text) plus categories from a classification (to permit retrieval of small groups of specific documents without having to specify every term needed to identify those groups).

Modern classifications used to supplement existing systems can be very simple, consisting of only a few dozen categories on a list that is checked off for each document entered. Or they can be much more extensive, approaching the type of deep multilevel hierarchy advocated in earlier sections.

Another use of modern classifications is for *automatic mapping* of words in free-text search systems. If all searchable words or terms are arranged in a deep multilevel hierarchy or word trees it is possible to use this classification as the front end to the search system. When the user enters a word, the computer can use the classification to identify all the terms and words subsumed under the selected word and include them in the search

This use of the computer to "expand" or "explode" or "map" a term can either be built into the computer as an automatic feature, or it can be an option which the user must ask for. Alternatively, the computer can display all the subsumed terms or categories that are narrower than the entered term (along with all the near-synonyms and related terms) and let the user choose those index terms or categories he wants to include in the search.

Modern classifications are also useful for computer-aided indexing at the time of input. The indexer enters a word or term or category into the system. and the computer displays the categories from the classification (or all the narrower terms from a classified thesaurus) that are equivalent, narrower than, or related to the entered term. This permits the indexer to select additional terms or categories from the displayed information by touching them with a light pen or some other touch-sensitive device.

To go even one step further, modern classifications are very useful for sophisticated automatic indexing or more precisely for automatic assignment of a document to classes or categories. If the alphabetic term list which is required as an entry vocabulary for the classification is very extensive. then every significant term present in the information being indexed can be looked up in that term list. The category numbers assigned to each term automatically place the term in its most logical location in the hierarchy of the classification. If a multimeaning term has several category numbers, then selection of the correct category is based on clues supplied by other categories already assigned to the document, particularly those categories identified by terms in sentences or paragraphs adjacent to the multimeaning term. This type of automatic classification or grouping into categories is based on content analysis that is supplied by the classification rather than on purely mathematical clustering algorithms.

Modern classifications are also an excellent mechanism for facilitating the exchange of information between two or more information systems, including systems located in many different countries. A modern classification can be developed to include a category for every concept (expressed by a variety of words, terms and phrases) in each of the independent systems. This new "central" or "common" classification forms the link between each of the other systems. The type of deep, multilevel, open-ended hierarchical modern classifications advocated in this paper are extremely useful for this modern application of classifications.

SDI systems which depend on precise matching of users with documents is still another area where modern classifications are of particular value, since categories can be subdivided to identify the specific interests of each user. The special uses of modern classifications described in the last few paragraphs do not in any way detract from the value of modern classifications to improve the retrieval performance of other types of information systems, as described in other sections.

Problems

Previous sections have stressed the many advantages and useful applications of modern classifications. Before concluding this paper, some of the disadvantages and problems must also be mentioned. These arise mostly from the need for subject experts with extensive background and experience who will undertake the development and oversee the continuous updating of the classification and its associated alphabetic index. The decisions that these experts must make regarding the words used to describe each category, the best subdivisions of each category and the most logical arrangement of categories are critical to the successful use of a modern classification. As a result of this need for expertise, the development and updating of a modern classification is more expensive and time-consuming than the creation and maintenance of other types of indexing tools.

However, the difference in time and cost may be more than offset by the fact that information indexed by the classification is better organized and analyzed and easier to retrieve by knowledgeable users than information indexed by most other methods. Since the number of users is usually many magnitudes larger than the number of experts who develop the classification and the total amount of retrieval time at multiple scattered locations is several magnitudes larger than the time spent on indexing, it is worth the extra effort to build and use the best possible indexing tool if it saves time and effort on the part of the users.

The size and complexity of enumerative hierarchical classifications with many *See Also* linkages between categories and links between the classification schedule and an alphabetic index to the categories present a major maintenance problem In the past this complexity has discouraged revision, and classifications have gradually become obsolete for lack of updating Development of modern automated systems for easier revision (such as the AUTOCLASS system mentioned previously) should significantly alleviate this problem. It must also be stressed that all indexing tools suffer if they are not continually updated and that this problem is not unique for classifications

Perhaps the biggest problem of all is whether any indexing tool is cost-effective. Free-text searching of any word in a title or abstract makes it possible to retrieve much useful information without any human indexing process. The extent to which this retrieval can be improved by using a modern classification or any other indexing tool, and whether this improvement is worth the added cost of the indexing, are questions that can only be answered by experiments which test retrieval performance of the various systems under carefully controlled conditions. Only a few results of these tests, such as the comparison of free-text indexing with the use of a modern classification mentioned earlier (see notes 2 and 3), have been published. They urgently need to be confirmed and extended by other researchers.

Conclusions

This paper has identified some attributes of a modern classification, and discussed why modern classifications are of value, how they should be structured and organized, how they lead to a useful balance between precoordination and post coordination, the type of notation and index needed for a modern classification, some guidelines for building a modern classification and some useful applications and disadvantages of using modern classifications. Stress has been placed on the use of enumerative multilevel hierarchical classifications and their advantages. It is hoped that readers will be stimulated by some of the ideas presented here to try to build such modern classifications for indexing information in various subject areas and to see for themselves how useful such classifications can be for achieving higher recall and higher precision than is possible with other types of indexing tools.

Notes

1 F Wilfred Lancaster, *Information Retrieval Systems Characteristics, Testing and Evaluation* (New York John Wiley & Sons, Inc . 1968), pp 32-38.

2 David F Hersey, Willis R Foster, Ernest W Stalder, and William T Carlson, Compaiison of On-Line Retrieval Using Free Text Words and Scientist Indexing, *Proceedings of the American Society for Information Science* 7 (October 1970) 265-267

3 David F Hersey, Willis R Foster, Ernest W Stalder, and William T Carlson, Free Text Word Retrieval and Scientist Indexing Performance Profile and Cost. *Journal of Documentation* 27 (September 1971) 167-183

4 John H Schneider, Selective Dissemination and Indexing of Scientific Information,' *Science* 173 (July 1971) 300-308

5 John H. Schneıder. The Case Agaınst Input Cost Reductıon and the Value of Hıerarchıcal Classıfıcatıons ın Automated Informatıon Systems. ın *Cost Reductıon for Specıal Lıbrarıes and Informatıon Centers,* ed. by Frank Slater (Washıngton, D C American Socıety for Informatıon Scıence, 1973). pp. 62-67

6 John H Schneıder. AUTOCLASS A Computer System for Facılıtatıng the Creatıon and Updatıng of Hıerarchıcal Classıfıcatıons To be presented at the Thırd Internatıonal Study Conference on Classıfıcatıon Research and publıshed ın the *Proceedıngs* (Avaılable from the author)

The Dewey Decimal and Library of Congress Classifications; an Overview*

Maurice F. Tauber and Hilda Feinberg

In the United States, two classifications are used primarily for the organization of materials in libraries. the Dewey Decimal Classification (DDC) and the Library of Congress Classification (LC). Each has inherent advantages and disadvantages for different types of libraries. Libraries have, in general, made their systems fit the needs of readers. However, as a rule, the closer the classification follows the order of the classification of knowledge, the more fully it serves the purpose of grouping together the books and ideas which are related. The basis of library classification by subject is the assumption that books on the same or related subjects will frequently be used together.[1] The classificationist attempts to develop a scheme which will "arrange books on the shelves in an order that will be recognizable as following some definite plan, will be in harmony with current studies, and will enable the finding of books together which have some likeness in a greater or less degree."[2]

Sayers has outlined the essentials of a library classification. What makes the value of one system as compared with another is its generalness of character, its order. the logical process of its subdivision, the quality of its terminology, and (at a later state) its practicality as shown in its notation and indexing.[3]

Maurice F. Tauber is Melvil Dewey Professor of Library Science at Columbia University.

Hilda Feinberg is Research Librarian at Revlon in New York City

*Some of the material in this article is adapted from *Technical Services in Libraries* by Maurice F. Tauber and Associates (New York. Columbia University Press, 1953). The article also contains material that will soon appear in a new edition of that work by Maurice F. Tauber and Hilda Feinberg.

The major disciplines should be represented, and they should be given space relative to their size, there should be flexibility to allow for extension of developing disciplines, reduction of contracting disciplines, and movement of disciplines or parts of disciplines from one section of the classification to another to express changing relationships.[4]

Dewey Decimal Classification

The most widely used scheme, and the oldest, is the Dewey Decimal Classification (DDC). It was devised in 1873 by Melvil Dewey for the Amherst College Library. First published in 1876, the arrangement of the classes was based to some extent on the classification scheme devised by W. T. Harris for the St. Louis Public School Library in 1870, which in turn was derived from Bacon's Chart of Learning.

As described by Mills, the importance of DDC lay in two significant advances it made over previous systems. 1) a notation was devised which exhibited great simplicity and flexibility, permitting a flexible shelf arrangement, 2) the comprehensive Relative Index, which showed those relative aspects of a subject which the systematic order scattered, solved to some extent what until then had been considered a serious drawback to systematic order.[5] The principle of relative location of books on the shelves was introduced, whereby the order of the books followed that of the classification scheme. This replaced the previously employed fixed location system of classifying books in libraries in which books were arranged according to size, accession number, or other considerations. Dewey's relative location meant that new titles could be inserted in their proper places alongside similar works already on the shelves without having to change the existing location symbols. This permitted continual moving of books from one shelf to another without destroying the logical order. The place of each book on the shelf was always the same in relation to the books on either side of it, although its actual position varied as books were added to the shelves, moved or withdrawn from the collection

Since the basic arrangement of DDC is systematic by conventional disciplines (history, literature, chemistry, etc.), and any given subject may be dealt with from various aspects and be classified in more than one discipline, it is the purpose of the

Relative Index to indicate all significant relationships between topics and show the relation of these topics to those in other areas, as well as their dispersion throughout the schedules.[6]

From 1876 to 1942, fourteen editions of DDC appeared. Since 1894 an abridged edition has been issued for small libraries and school libraries. At present, the eighteenth edition of the unabridged edition and the tenth edition of the abridged DDC are available. The abridged edition is useful for school and public libraries that do not predict growth larger than 20,000 titles.

The Dewey Classification may be described as an enumerative classification with provision of synthetic devices in some areas. As noted by Needham:

> Even schemes which are predominantly enumerative usually provide synthetic devices to cater for common form-division, space and time elements—for clearly all of these would otherwise have to be enumerated at more or less every division in the schedules. Any attempt to enumerate complex subjects is in practice found to be selective. it could never hope to encompass the unpredictable multiple relationships found in literature. Additionally—though this need not necessarily follow—it is likely that the enumeration will be unsystematic.[7]

Needham concludes that enumerative schemes are likely to have the following limitations in the schedules:[8]

1 Omission of some simple and complex subjects, duplication of others.

2 Conflicting principles underlying the placing of complex subjects.

As an example of the latter, *The harvesting of potatoes* may be found under *Potatoes, The harvesting of wheat* under *Harvesting*. Materials on the same subject may be found in two or more places.

Dewey incorporates numerous synthetic devices as may be represented by standard subdivisions, area tables, tables providing for subdivision of individual literatures, languages and other provisions. Recent editions indicate increasing use of synthetic elements, offering broader hospitality to complex subjects. DDC is becoming fuller in coverage and more capable of displaying complex topics.[9]

The notation symbols of DDC and other classifications are not necessarily connected logically with the principles upon which the formation and arrangement of the classes and their subdivisions are based, as they are added subsequently to the creation of the classification. The notation symbols are used to identify and shelve the books. The DDC notation is expansible; a new number may be created by the addition of another digit. The length of the notation to be used is determined by the individual library, taking into account its size, character and probable rate of growth. The small general library should find a brief notation of three to five digits satisfactory. The problem related to the complexity of long numbers resulting from attempts to gain greater specificity in close classification has resulted in a policy of segmentation of the notations to make them adaptable for libraries of varying size, for example: 258'.2'0922 may be segmented as 285, as 285.2, or may be used in its entirety at 285.20922. Since 1967, DDC numbers on Library of Congress cards have appeared in from one to three segments. The prime marks, which are not considered part of the notation, identify the varying levels at which notation is meaningful. Such segmentation makes it possible for a library to cut excessively long notations to more acceptable shorter numbers.

Criticism of Dewey Decimal Classification

Arrangement. In examining the arrangement of the classification, consideration should be given both to the order of the main classes and the order within the classes. No one type of arrangement is followed throughout the scheme—a number of arrangements, both natural and artificial, are employed. Yet, a logical process of division and subdivision of main classes is carried out in most instances. Dewey employs an arbitrary arrangement in some cases where an alphabetical arrangement would probably be more desirable from the point of view of the user, for example in class 546, an alphabetical arrangement for chemical elements would be preferred by the chemist. Alphabetic arrangements are provided in a few places in the schedules and auxiliary tables.

Another example of inconvenient arbitrary arrangement is the schedules for music, 780. The order of the numbers does not correspond to the importance of the subjects. The order of its classes has been criticized as separating the social sciences (300) from history (900), language (400) from literature (800);

and the separation of a particular science from its technology For example. the separation of chemistry from chemical technology invites criticism from some chemists using the classification.

The order of the main classes is not of crucial practical significance. particularly in larger libraries. In fact. the editor of the eighteenth edition stated:

> The primary basis for DDC arrangement and development of subjects is by discipline. as defined by the main and subordinate classes. while subject. strictly speaking. is secondary There is no one place for any subject in itself. a subject may appear in any or all of the disciplines . No other feature of the DDC is more basic than this that it scatters subjects by discipline.[10]

Of greater concern than the order of main classes is the rigidity resulting from a strict division by tens.[11] Accordingly many classes and subclasses are overcrowded. notably where the scheme fails to provide sufficiently for the interests and requirements of foreign. scientific or technical libraries In recent editions. the editors havè attempted to remove some examples of bias and to deemphasize the Western bias of the schedules As an example. the non-Christian faiths are developed in more detail in the 200 classification An additional provision recommends an option of using a letter or other symbol as an artificial digit to bring into prominence specific linguistic. ethnic or cultural approaches.

Other Criticisms Criticism has been expressed about a lack of foresight in relation to the growth and change in technical and scientific areas. Each edition attempts to update the scheme to keep pace with expanding knowledge through expansion of existing numbers. by the addition of more subdivisions. and through relocations of topics in the schedules This is done within the framework of the official policy which attempts to preserve the integrity of numbers. which means that vacated numbers cannot be used for new topics—at least until a time when such a change would be of relative insignificance to most users of the scheme Dewey realized that a classification which changed to a substantial extent with each new edition would not be acceptable to librarians Changes in each edition force the librarian to consider reclassification. requiring alteration of notations on catalog entries. reshelving. and refiling. entailing additional time and expense

A number of other objections to DDC may be cited. No Cuttering is supplied for DDC classification numbers on LC cards, a deficiency which requires time-consuming and costly shelflisting operations.[12] While DDC numbers have appeared on many LC cards since 1930, the number of titles classified by Dewey numbers has varied markedly since the service began—from a high of 99 percent of all cards prepared in the fiscal year 1933/34, to a low of 24 percent in 1965/66.[13] The DDC number on LC cards should be considered as only a suggested number which may no longer be valid if one is using a new edition of DDC. Libraries using DDC are obliged to perform expensive original classification for a substantial percentage of their titles.

DDC has been criticized as being too permissive. "This is a boon to custom cataloging or to local cataloging preference, but a Pandora's Box in centralized cataloging."[14] Examples are the classification of biography in 920 or the subject number, bibliography in 016 or the subject number, and extension of class numbers or building numbers beyond what is given on the LC card.[15]

The advantages that have been attributed to both the Dewey Classification and the Library of Congress Classification have been exhaustively recounted in the literature. Among the advantages of DDC are its up-to-dateness with successive revisions and its mnemonic features. Its notation is simple and comprehensive, but the length of notation used in many libraries presents a definite problem. It is adaptable for use in libraries of varying size and kinds. However, the Classification Committee, RTSD Cataloging and Classification Section, in 1964 recommended Dewey for libraries with general collections up to 200,000 volumes in size, and the Library of Congress system for those expected to be larger and for those small libraries with specialized collections.[16]

Use of the Dewey Decimal Classification

In 1960 it was reported that some 95 percent of public libraries, nearly 90 percent of college and university libraries and over 60 percent of special libraries in the United States used DDC. In Great Britain, over 500 libraries used it. It was reported at that time that it had been translated into some nine European languages, and into Chinese and Japanese.[17]

DDC has continued to be effective for most libraries for almost a century. It may be found in some form throughout the world. Of the sixteenth edition, 25 percent of the copies sold were to libraries outside the USA.[18] On the continent of Europe, the Universal Decimal Classification (UDC), a derivative of DDC, is used to a great extent, in spite of the fact that it uses long numbers and is subject to many changes.

Since 1934 the Decimal Classification Section of LC has periodically issued *Notes and Decisions on the Application of the Decimal Classification.* Since 1950, DDC has been used for the arrangement of the *British National Bibliography.*[19] Both the R. R. Bowker Company and the H. W. Wilson Company use the DDC in their bibliographic publications. Dewey numbers may be found in *Publishers' Weekly, American Book Publishing Record, Book Review Digest,* the Standard Catalog Series, the ALA *Booklist, New Serial Titles,* and in several national bibliographies. In addition, many commercial processing firms are prepared to classify by Dewey.[20] The DDC numbers on LC cards, on Wilson cards, and numbers derived from other published sources should not be accepted without further checking of local shelf lists and policies. Among other factors, one needs to know from which edition the numbers have been assigned

While many of the reasons for abandoning DDC lie within the classification itself, some of the contributing factors have been outlined by Maltby.[21] Failure of libraries to accept changes in succeeding editions of Dewey and tinkering with its numbers serve to lessen its practical use, absence of DDC and book numbers on purchased cards, as noted previously, increase the cost of classification, concern over the lengthening notation of DDC, routine recommendations to classify made by library surveyors, and a sincere conviction that another classification is better designed and more appropriate are all contributing factors.

Library of Congress Classification

Second in usage to the Dewey Decimal Classification in this country is the Library of Congress Classification (LC), an enumerative scheme which was originally developed from an incomplete expansive classification founded by Charles Amni Cutter. The LC classification was designed to be a pragmatic and expansive scheme for the holdings of the Library of Congress and what it might be expected to add in the future. The individual subject schemes of the classification were independently created by a number of subject specialists for each discipline, who worked individually and in groups, and have been published by the U.S. Government Printing Office since 1901. Each main class is published separately, and though developed separately, each represents a unified part of an overall scheme for the organization of library materials. Since very few libraries expect to grow to the size of the Library of Congress, the scheme represents adequately in most cases the needs of the majority of libraries in this country Because of the extremely large holdings of the Library of Congress, and its status as a depository for copyright works, the schedules are generally comprehensive and provide to a large extent for the scholarly works likely to be held by academic, research and large public libraries.

Expansion of the classification is governed by literary warrant, depending upon the acquisition of new materials by the Library of Congress. Thus, the development of the classification is directly affected by the acquisitions policy of the library.

The original designers of the scheme provided for expansion by leaving gaps at places which were predicted to be appropriate in the future. Such predictions of the advancement of knowledge are impossible to foresee, thus the accuracy of the placement of the gaps will of necessity be approximate. There are five single-letter classes that have not been used, and many double-letter combinations available for future expansion.

The individual schedules are kept current by 1) *LC Classification–Additions and Changes* (Quarterly) published by the Library of Congress, 2) the addition of supplementary pages of *Additions and Changes* to reprinted editions of individual schedules, and 3) publication of new editions of the individual classes when appropriate. The Gale Research Company offers a compilation, *L.C. Classification Schedules. Additions and Changes through 1970*, and *Additions and Changes for 1971-72.*[22]

Criticism of the Library of Congress Classification

Arrangement. The classification schedules have been built up continuously as material requiring new subdivisions and revisions in the existing schedules has been added to the Library of Congress collection. While there is general uniformity of structure and format throughout the schedules, the classes, divisions and subdivisions have been developed and revised to meet the needs and use made of the large collection. Consequently, no strict uniformity among individual schedules in regard to subdivision for form, geographic areas or periods is evident. Subjects are followed by subject subdivisions, progressing from general to specific as far as possible. The schedules frequently provide for an alphabetical order of subdivision for subclassification employing topical Cutter numbers to represent individual topics, rather than classified subdivisions.

Many classes are equipped with special tables and directions for subdividing the classes more minutely. These tables are peculiar to the one subject to which they apply and can seldom be used to subdivide other topics, thus they exhibit minimal mnemonic characteristics. While no facets are common to the whole scheme, the separate classes are provided with some synthetic devices to varying degrees. Classes H and P have synthetic capabilities, class Q does not.[23]

The order of the main classes, although somewhat arbitrary, is based on the major traditional disciplines. The quality of detail varies from one part of the scheme to another. The LC is comprehensive, but not universal. As might be expected, a general classification scheme designed for a library identified as the congressional library of the country places emphasis on political and social sciences and on history While providing for these areas in depth, LC offers as well, a comprehensive treatment of language and literature The Library of Congress makes available through its printed cards, book catalogs and MARC tapes, classification numbers for the major subjects likely to be represented in general libraries of all sizes For subjects like law, medicine, science and technology, many libraries with extensive holdings have had to use special schemes. The Library of Congress does not assume responsibility for comprehensive collecting in such special fields, and thus cannot provide as detailed a classification as might be needed by specialist libraries The National Library of Medicine has its own classification,[24] and this has been adopted by many other medical libraries

Other Criticisms. As the result of a broad survey, the basic satisfaction of librarians with LC has been confirmed.[25] Among complaints were a lack of a general index to all of the schedules, the fact that many parts of the schedules lack adequately detailed instructions, the difficulty of keeping track of changes in classification in the present format (a loose-leaf or index-card method of publication is preferable, ideally, new revised schedules incorporating the changes should be printed more frequently), the failure to supply author numbers in the literature schedules, since most academic libraries do not use PZ 3 and PZ 4, and the lack of a manual of instruction for applying the scheme.

By far the most frequent reason given for not accepting LC without change was that the library did not use the PZ fiction class, but instead classified such titles in the various national literatures. As noted by Bead, perhaps no decision has produced more comment than the grouping of all fiction in English in PZ 3 and PZ 4. This material includes not only American and English fiction, but also foreign fiction translated into English [26] As indicated by Bead, the original purpose of classing all fiction in English in PZ 3 was no doubt to bring together at the Library of Congress a special collection of fiction, arranged alphabetically by author, which a reader could easily use for browsing without first consulting the catalog.

Some progress in meeting some of the above objections has been accomplished. Two general indexes to LC were announced in 1974. *An Index to the Library of Congress Classification, with Entries for Special Expansions in Medicine, Law, Canadian and Nonbook Materials*, Canadian Library Association,[27] and *Combined Indexes to the Library of Congress Classification Schedules*, edited by Nancy B Olson, U.S Historical Documents Institute [28] As mentioned previously, Gale offers *L.C. Classification Schedules. Additions and Changes* through 1972. Considerab'e interest exists to create a general manual of instruction. While no manual has yet been issued by the Library of Congress, *The Use of the Library of Congress Classification* by Schimmelpfeng and Cook[29] and *A Guide to the Library of Congress Classification* by Immroth[30] offer some degree of assistance. In regard to fiction, Library of Congress cards for fiction in English now include, in addition to the usual LC number, another number for the nationality and period of the author.

Use of the Library of Congress Classification

In whole or in part, the scheme is being used increasingly, particularly in academic libraries.[31] [32]Hoage in 1961 located 256 libraries using the LC system.[33] Richard Angell indicated that between 800 and 1000 libraries were using LC in 1964.[34] He predicted that in the ensuing eight years this growth would double. The growth has been marked, ana is related not only to libraries changing from another classification, but also to new developing libraries and to departmental libraries of universities, particularly science and technology collections. In addition to academic libraries, there have been a number of public libraries as well as state, historical and special libraries which have adopted LC. In regard to the size of libraries using LC, there is evidence that this is not a significant factor. Small as well as large libraries find the classification appropriate for their collections. Over the years, there has been a tendency to regard LC as a system only for arranging materials for large research libraries. In recent years it has become clear that the system is suitable for all types of libraries even, in some cases, school libraries. Some foreign libraries, particularly those involved with governmental responsibilities and general research, have accepted LC as an effective arrangement of materials for their use and services. There is no question about the ability of adults to use books arranged by LC. There was some apprehension about children and young people not being able to locate materials classified by this system. This has been disproven by the experiences of both the St Paul and the Buffalo public libraries

Reclassification[35]

The pressures of growth, expanding knowledge and publication, and rising costs have made evident the inadequacies of outmoded classification systems. In the late nineteenth and early twentieth centuries, libraries generally changed from local systems to the Dewey Classification. Beginning in the 1920s the trend in reclassification shifted toward the introduction of the Library of Congress Classification as the advantages of that system for large research libraries became apparent In recent years, a significant development has been the indication that LC is suitable for smaller libraries too.

A survey conducted in 1966 revealed that 85 percent of those libraries approached which had shifted to LC had formerly used Dewey or "Modified Dewey."[36] The survey indicated that 59 percent preferred to reclassify the entire collection, 41 percent did not. Usually the overriding reason for adopting partial reclassification is economic—recataloging projects are expensive Although a librarian may be cognizant of the difficulties and costs involved in programs of reclassification, he may institute the project on the basis of two assumptions. 1) that the use of a classification such as that of the Library of Congress (for most changes have been directed to LC) achieves a grouping of the books in the collection that is of greater educational significance and shows the users the currently accepted relationships among the branches of knowledge more effectively than did the system that is being replaced, and 2) that the adoption of a new classification, which involves abandoning a system that has been found expensive to handle technically, will in the long run be an efficient administrative device. These assumptions are based on the testimonies of librarians who have grappled with the problem and on the results of general surveys.

The reasons for reclassification include economic considerations, problems relating to the system in use, the desire to improve services for the user, and reasons relating to administrative factors. DDC is the most abandoned library classification, and LC the most frequently adopted.[37]

In the Report of the Classification Committee, RTSD Cataloging and Classification Section, 1964, on the type of classification available to new academic libraries, it was stated that

> LC has the advantage of not being logical in exposition, as a rule, and while it is practically impossible to memorize, it is easy to expand without upsetting existing classified books The advantage of a non-logical classification is apparent in dealing with rapidly advancing subjects, as the sciences, where a major change in thought can throw out a whole branch in a previous arrangement of knowledge LC can interpolate where DC must compromise.
>
> Dewey has to be expanded through further breakdown, sub-classification or re-naming and reassigning classes LC can be expanded by interpolation because the whole system does not have to be logical but can, to a considerable degree, grow like Topsy without regard to its environment It has been possible to abridge Dewey, but not LC[38]

In regard to notation the report added that the mixed notation of LC is more complex than the pure notation of DDC. However, the LC numbers on the average are shorter than DDC numbers. Dewey's notation is positional, each position represents a classification level. LC notation is ordinal. Each class has a number of its own, not necessarily related to preceding or following classes. LC is much broader and more comprehensive than DDC.[39]

The Library of Congress Classification is supported by the substantial resources of the world's largest library operation. 'Any reasonably comprehensive classification system developed and maintained by the considerable means of a federally supported agency, that is, the Library of Congress, is the logical classification system for general library use.''[40] The Library of Congress card service is backed by some of the best trained professionals to be obtained. Through its MARC program, The National Program for Acquisitions and Cataloging, its Cataloging-in-Publication Program, its book catalogs and its cards, it represents a true cooperative and centralized operation. The program of centralized or shared cataloging on an international basis brings to the library a greatly increased inflow of material which will, in effect, increase cataloging production and will, in turn, be responsible for a substantial increase in the establishment of new numbers.[41] The principal advantages of using LC are economy in cataloging, speed in processing and the benefits to be realized from tying into a large centralized cataloging operation.[42]

The advantage of conversion to LC lies primarily in accepting the classification numbers as they appear on the cards, otherwise the economy is not fully realized Unnecessary checking and verification of data on the cards should not be performed except in cases of obvious errors Changes in LC call numbers and other variations results in a situation where the library cannot take advantage of the Library of Congress services. A large number of libraries do not, or are not able to, take full advantage of centralized cataloging We cannot expect the program of cooperation and centralized cataloging and classification to be any more than empty words unless catalogers stop thinking of all kinds of reasons for not taking advantage of it.

It has been determined that there are fewer changes in LC class numbers than in DDC, and that Library of Congress cards give

LC class numbers plus LC Cutter numbers on 85 percent of the cards.[43] DDC numbers appear on Library of Congress cards for about 35 percent of titles for which cards have been printed, including titles in all languages. However, as indicated by Benjamin Custer, over 95 percent of cards sold are for English language titles, and an analysis of these cards received by a sampling of orders in this country indicated that approximately 80 percent contained Dewey numbers.[44] On this basis, it may be estimated that, for the type of materials collected by an undergraduate library in the United States, close to 80 percent of the Library of Congress cards may be expected to contain numbers representing different editions of Dewey, provided that the Library of Congress continues to assign Dewey numbers at the same level as were assigned at the time that the analysis was made.

Conclusions

Recent developments in the application of computers to libraries, and the planning and establishing of networks of all kinds (national, regional, state and others) must force classification reevaluation. "The fact that centralization of bibliographic processing through automation is closer to reality than at any time in the past century is a strong impetus against continuing one's provincial ways. Under these conditions it will be come imperative that large libraries consider how they may be assimilated into a national network. [45] Shell has expressed the opinion that LC can be programmed to do all that we have required of an enumerative scheme up to the present, so that effective electronic searching, printouts of lists of materials for any segment of LC, book catalogs, inventory control, etc., can all be done with the aid of computers. "Future demands for more sophisticated searches may have to be met by the application of a new language which will be used for certain types of in-put and information retrieval, but not for the organization of books on the shelves or in the card catalogs."[46] Hines has shown that both Dewey and LC notations may be manipulated by computers. He encountered no problems in programming, arranging or finding of items in either DDC or LC.[47] DDC does not offer the advantage of a purely numeric notation when the complete call number is considered.

Angell has indicated two factors which may influence the LC classification in the future. First, the transfer of the library's bibliographic records to computer operation will render the shelf arrangement less important. It is envisaged that consoles may replace conventional catalogs as now used, providing the facility for browsing which is presently offered by open stacks. Second, there is the need to economize on space in anticipation of future accessions. Should arrangement by size be utilized as a space saver, the need for classification as a means of shelf arrangement no longer exists.[48] Progress towards automation cannot be expected to be rapid in most libraries, and may not be possible in the foreseeable future in others. There will, therefore, be a need to have the schedules maintained according to standards of today.

Matthis and Taylor note that any perfect system is a dead system, and a classification system based on a total view of knowledge is preposterously presumptuous.

> Essentially the argument has now moved beyond theoretical discussion of the "best" classification system and settled upon the real issue—the promise and prospect of centralized cataloging and classification. No one classification system will ever solve all of the problems, but the practice of "rugged individualism" in cataloging no longer makes sense and should no longer be tolerated.[49]

Notes

1 Leo B. LaMontagne, *American Library Classification with Special Reference to the Library of Congress* (Hamden, Conn. Shoe String Press, 1961), p. 6.

2 Thelma Eaton, *Classification in Theory and Practice. A Collection of Papers* (Champaign, Ill., Illini Union Bookstore, 1957), p 23

3 W. C Berwick Sayers, *Canons of Classification* (London Grafton, 1915), p 14

4 Christopher D Needham, "Dewey Decimal Classification," in *Reclassification, Rationale and Problems. Proceedings of a Conference*

on Reclassification Held at the Center of Adult Education. University of Maryland, College Park, April, 1968, ed. by Jean M. Perreault (College Park. School of Library and Information Services, University of Maryland, 1968), pp. 9-10.

5 J. Mills, *A Modern Outline of Library Classification* (London Chapman & Hall, 1967), p. 57.

6 Theodore Samore, *Problems in Library Classification, Dewey 17 and Conversion,* Library and Information Science Studies, No. 1 (Milwaukee. The University of Wisconsin, 1968, published in cooperation with R. R. Bowker), p. 20.

7 Christopher D. Needham, op. cit , pp. 12-13

8 Ibid., p 13.

9 Jean M. Perreault. 'Afterword,'' in *Reclassification, Rationale and Problems,* p. 187.

10 Benjamin A. Custer, 'Introduction, in Melvil Dewey, *Dewey Decimal Classification and Relative Index,* 18th ed , vol 1 (New York. Lake Placid Club, Forest Press, 1971), pp 17-18.

11 Melvil Dewey. Decimal Classification Beginnings,' *Library Journal* 45 (1920): 151-154.

12 Raimund E. Matthis and Desmond Taylor. *Adopting the Library of Congress Classification System, A Manual of Methods and Techniques for Application or Conversion* (New York R. R. Bowker, 1971), p.2.

13 John McKinlay, ''More on DC Numbers on LC Cards. Quantity and Quality,'' *Library Resources & Technical Services* 14 (Fall 1970). 518.

14 ''Statement on Types of Classification Available to New Academic Libraries,'' Report of the Classification Committee, RTSD Cataloging and Classification Section, May 15, 1964, *Library Resources & Technical Services* 9 (Winter 1965)· 107.

15 Ibid , p. 108

16 Ibid., p 106

17 J. Mills, op cit., p 58

18 Theodore Samore. op cit , p 3

19 J. Mills, op. cit , p 58.

20 Barbara Westby. "Commercial Processing Firms, a Directory." *Library Resources & Technical Services* 13 (Spring 1969) 209-286

21 Arthur Maltby, ed., *Classification in the 1970's, a Discussion of Development and Prospects for the Major Schemes* (Hamden, Conn Linnet Books and Clive Bingley, 1972), p. 102.

22 Gale Research Company, *L.C. Classification Schedules. Additions and Changes through 1970* (Detroit. Gale Research Company), *Additions and Changes for 1971-72.*

23 J. Mills, op. cit., p. 97.

24 U.S. National Library of Medicine, *Classification,* 3d ed (Bethesda. Md.: U.S. National Library of Medicine. 1964).

25 Maurice F. Tauber, "Review of the Use of the Library of Congress Classification," in *The Use of the Library of Congress Classification,* ed by Richard H. Schimmelpfeng and C. Donald Cook (Chicago American Library Association, 1968), pp. 1-17.

26 Charles C. Bead, "The Library of Congress Classification Development, Characteristics, and Structure," in *The Use of the Library of Congress Classification,* p. 24.

27 Canadian Library Association, *An Index to the Library of Congress Classification, with Entries for Special Expansions in Medicine, Law, Canadian and Nonbook Materials,* Preliminary ed., ed by J. McRee Elrod, Judy Inouye and Ann Craig Turner (Ottawa, 1974).

28 *Combined Indexes to the Library of Congress Classification Schedules,* ed. by Nancy B. Olson, 1st ed , 15 vols. (Washington' U. S Historical Documents Institute, 1974).

29 Richard H. Schimmelpfeng and C. Donald Cook, eds., *The Use of the Library of Congress Classification.*

30 John Phillip Immroth, *A Guide to the Library of Congress Classification,* 2d ed. (Littleton. Colo . Libraries Unlimited, 1971).

31 Edward G. Holley, "The Trend to L.C Thoughts on Changing Academic Library Classification Schemes,' in *Library Lectures, March 1965-May 1966,* ed by Sue B. Von Bodungen (Baton Rouge. Louisiana Louisiana State University Library, 1967), pp. 29-46. H. F. McGaw, comp., "Academic Libraries Using the LC Classification System," *College & Research Libraries* 27 (January 1966): 31-36.

32 Maurice F Tauber, ' Review of the Use of the Library of Congress Classification "

33 Annette L. Hoage. "The Library of Congress Classification in the United States. A Survey of Opinions and Practices. with Attention to Problems of Structure and Application" (D.L.S. dissertation, School of Library Service, Columbia University. 1961).

34 Richard S. Angell. "On the Future of the Library of Congress Classification," in *Classification Research. Proceedings of the Second International Study Conference, Elsinore, Denmark, Sept. 14-18, 1964*, ed. by Pauline Atherton (Copenhagen. Munksgaard, 1965), pp. 101-112.

35 Maurice F. Tauber, Subject Cataloging and Classification Approach the Crossroads," *College & Research Libraries* 3 (March 1942). 153-154, ' Reclassification and Recataloging in College and University Libraries. Reasons and Evaluation. ' *Library Quarterly* 12 (October 1942). 827-845, Reclassification and Recataloging of Materials in College and University Libraries," in *The Acquisition and Cataloging of Books*, ed. by W. M. Randall (Chicago: University of Chicago Press, 1940), pp 187-219, Reclassification of Special Collections in College and University Libraries Using the Library of Congress Classification,' *Special Libraries* 35 (April 1944). 111-115, ' Reclassification and Recataloging, in *Technical Services in Libraries* (New York Columbia University Press, 1954), Chapter XIII.

36 Maurice F. Tauber. "Review of the Use of the Library of Congress Classification," p 5

37 Elton E. Shell, 'A Rationale for Using the Library of Congress System in Reclassification, in *Reclassification, Rationale and Problems*, p. 31.

38 'Statement on Types of Classification Available to New Academic Libraries," pp. 105-106.

39 Ibid.

40 Raimund E Matthis and Desmond Taylor, op cit . p 2

41 Charles C Bead, op cit . p 28

42 Raimund E. Matthis and Desmond Taylor, op cit . p 9

43 Statement on Types of Classification Available to New Academic Libraries," p 107.

44 Benjamin A Custer, Letter, *Library Resources & Technical Services* 9 (1965) 212.

45 Edward G. Holley. 'The Trend to L C . p 36

46 Elton E. Shell, op. cit., p. 52.

47 T. C. Hines, "Computer Manipulation of Classification Notations," *Journal of Documentation* 23 (September 1967): 216-233.

48 Richard S. Angell, op. cit.

49 Raimund E. Matthis and Desmond Taylor, op cit., p. 3.

UDC: Present and Potential

Hans Wellisch

The Universal Decimal Classification (UDC) was last presented in detail to an American public in the first volume of the Rutgers series on Systems for the Intellectual Organization of Information[1] by Jack Mills who gave a detailed and closely reasoned overview of the theoretical foundations of the system The difficult problems and intricate discussions reported in this book and the esoteric language used by Jean Perreault in his collection of theoretical essays[2] may have estranged the last remaining adherents of the UDC in this country rather than endearing the system to them. The erroneous impression was created that the UDC was a system fit only for philosophers and strange European classificationists but not a practical tool for information retrieval. This view was reinforced by the popular but fallacious notion that classification systems as such were outmoded, and in fact dead as doornails, as far as scientific and technical information and its retrieval were concerned since computers would do the job better and faster if only their memories could be made large enough. Added to these artificially generated obstacles to the promotion of the UDC as a viable retrieval tool was the complete lack of a usable English edition of the scheme, because no comprehensive schedules have been published in English since an abridged edition in 1961 which is, of course, now completely out-of-date.

Present Applications

The scheme, however, is alive and well in most parts of the world, including the American continent (in particular in the Latin American countries), with the sole exception of the United States where it is still looked upon as an oddity rather than a

Hans Wellisch is Visiting Lecturer, College of Library and Information Services, University of Maryland, and a member of the Central Classification Committee, International Federation of Documentation.

viable retrieval tool despite valiant efforts of some American information scientists to demonstrate not only its usefulness but also its applicability to computerized stores of information. The pioneers in this field were Malcolm Rigby,[3] Robert Freeman [4] and Pauline Atherton,[5] who were followed by T. W. Caless and others;[6] quite recently, important work has been done in Canada by M.A. Mercier and his collaborators[7 8] in the construction of a computerized retrieval system for water resources information in project Environment Canada, known as WATDOC, which is based on the UDC, resulting in a concordance between the classification schedules and the *Water Resources Thesaurus* of the U.S. Department of the Interior. Other noteworthy practical and large-scale computerized applications of the UDC have been made in Germany,[9] and the U.K.,[10] Denmark[11] and Switzerland,[12 13] these and other computer applications were summarized at two international seminars devoted to the topic,[14,15] and in a report by Rigby.[16]

The growing interest in the construction of both general and specialized thesauri in the latter half of the 1960s led to another fallacious idea, namely the superiority of verbal retrieval tools which purportedly could keep pace with changing terminology much better than relatively rigid hierarchical classification schemes. Two pilot projects were undertaken to test the validity of this proposition when UDC was compared with the Engineers' Joint Council's *TEST* thesaurus[17] and the *MeSH* subject heading list used by *Index Medicus*.[18] Results showed conclusively that the UDC could handle almost all concepts listed in the two thesauri, while conversely the thesauri showed some serious lacunae in their coverage, on the other hand, these projects also revealed some structural faults in UDC which, although their nature had been known for a long time, were put in sharp perspective by these comparative tests. The overall conclusion drawn from these experiments was that, far from being competing retrieval tools, classification systems of a faceted or semi-faceted nature (such as the UDC) complement verbal retrieval tools of the thesaurus type, and vice versa; moreover, a sound classification scheme is, in fact, indispensable for the successful construction of a thesaurus.

Indeed, universal classification schemes are now needed more than ever, be it as backup systems in situations where detailed information is retrieved by specially devised schemes (verbal or classificatory) but where marginal subjects have to be handled by a general scheme, or be it as "switching devices" between two or more different retrieval systems that are used simultaneously.[19] These aspects as well as many others relating to tho theory and practice of the UDC have recently been treated in an excellent and exhaustive book by A.C. Foskett[20] to which the reader is referred, since it would be presumptuous for anybody to try and paraphrase the wealth of material brought together there in easily readable and thought-provoking form. Foskett's book also deals extensively with the present shortcomings of the system, both on the conceptual and managerial level, and his proposals for improvements in both respects will hopefully have a profound influence on any future developments concerning the UDC.

Although the present situation, as indicated above, is not at all as gloomy as it is sometimes painted by people who have but scanty knowledge of the UDC, it would be foolish to deny that the system is in urgent need of revision and reform. Its framework, still largely cast in the mold of the Dewey Decimal Classification (DDC) of which it originally formed an extension, suffers from overcrowding in classes 5 and 6, and unhelpful dislocations of closely related subjects, as in the notorious case of theoretical chemistry (54) being separated from chemical technology (66), and other such incongruencies and even outright follies. The high degree of detail and specialization, once the hallmark and pride of UDC, now threatens to suffocate the system, too many additions and "refinements" were often made by people with little insight into the workings of a general classification scheme and interested only in promoting their own special field, so that the UDC is now weighed down by unreasonably long and complicated notations and a growing redundancy of concepts listed in many different parts of the scheme. leading to complexity and ambiguity in application and to consequent retrieval failures.

Added to this are serious shortcomings in the management of the system which have been pointed out by Wellisch[21] and Foskett[22] but have been dealt with so far on a patchwork basis, if at all. Foskett also made the proposal to transfer the responsibility for the English edition of the UDC to the newly-formed British Library which might consider the adoption of the scheme for the classification of its open-shelf reference collection.

While both existing and imaginary shortcomings of the UDC were formerly pointed out mostly by individuals (UDC users and developers as well as outside critics), the last few years have seen more concentrated efforts at constructive criticism and revision, backed by institutions and by the UDC s own governing body, the Central Classification Committee of the International Federation of Documentation (FID) Some of the more important of these proposals for renewal and reform will now be briefly presented, particularly since some of them are of such recent date that they are not yet covered by Foskett's book

Reform or Revolution?

Two trends are now clearly discernible one is concerned with the upkeep of the present framework of the system in more or less unmodified form, making only routine amendments and extensions while at the same time using new techniques of presentation and indexing The other trend is towards a more revolutionary shake-up of the whole system with the aim of creating a new universal classification scheme or, as some of the proponents of this school of thought have suggested, a New UDC or NUDC. On the face of it, it may seem that there is a basic contradiction here, and that the simultaneous pursuit of such divergent aims can only lead to a dissipation of already scarce resources in manpower and money which would be better spent in concentrating on either one of these trends Although such a danger no doubt exists there is some justification for proceeding along both lines simultaneously

The UDC in its present form is still the most widely used system of classification for information retrieval, despite its many faults and a certain lack of enthusiasm displayed even by its defenders and users Almost everybody agrees that the system is in great need of a thorough overhaul but the design and construction of a completely new scheme is a major undertaking that must necessarily take many years until it can be presented to the world, and will then take a few more years to be tried in actual retrieval situations so as to debug the system for use at least during the remaining decades of this century and perhaps beyond Meanwhile, the existing framework must be kept in a viable state, (a) by taking into consideration new developments in all branches of knowledge and (b) by putting at the disposal of users, present and potential, the tools that make it possible to utilize the system

Objective a is met in part by the traditional method of piecemeal revision of existing sections, and although this is at present a tedious and cumbersome process, some measures have already been adopted to rid the system of its hyperdemocratic procedure, which in the past often delayed urgently needed revisions for months and even years As a result, several hundred amendments and innovations in dozens of important subject fields have been introduced during the past few years, giving the lie to the often-heard argument that the UDC cannot cope with new developments in science and technology or in the social sciences (just the latter having undergone almost complete revision and updating which is still not complete but has already resulted in a much improved class 3) Another part of the revision procedure is the impending reallocation of class 4 (emptied more than ten years ago in order to accommodate new subjects and overcrowded sections of classes 5 and 6, but as yet not reoccupied). Various proposals were submitted and discussed, and at present it seems that the following allocation of subjects has the best chance to be approved as a new class 4 schedule

4 Man and his natural environment Material resources Science and technology in general

41 Man as an individual Medical sciences, anthropology, psychology.

42 General biology, botany, zoology

43 Agricultural sciences. Plants and animals

44 Animal biology and husbandry (cf 43 for plants and crops only).

45 Mineral resources. Mining and mineral dressing

46 Materials. Testing, sampling, etc

47 Handling and transport of materials and persons

48 Management business, household, etc

Objective b is currently being met by a large number of new and revised full, medium-sized or abridged editions of the tables in more than 20 languages (which, incidentally, now form the largest existing general multilanguage thesaurus of terms, where each language is linked to any of the others through the relevant UDC number, although much remains to be done in order to reconcile vocabularies and sometimes even the interpretation of certain UDC numbers in different languages, cultures and political regimes) The most important among these editions which will hopefully be forthcoming within the next couple of years is a new English-language edition

English Basic Medium Edition

As mentioned above, the use of the UDC in the English-speaking world has been seriously hampered by the lack of a usable English abridged or medium-sized edition. Plans are now being made to bring out a revised and updated Basic Medium Edition (BME) in English by the Central Classification Committee in collaboration with the British Standards Institution (the body responsible for all English-language editions of the UDC). The publication of this edition will be the UDC's contribution to the Melvil Dewey Centenary in 1976. It will serve two basic purposes. (a) it will put at the disposal of UDC users the long awaited comprehensive tables in English which could be used for most information retrieval work except where very fine detail of classing is needed (for which almost complete full tables are now available in English). (b) it will serve as the master file for the creation of other medium-sized editions in various languages and will be constantly kept up-to-date in the editorial offices by mechanized equipment At present, a committee is trying to determine the degree of abridgment from the full tables for every major subject field so as to assure a balanced presentation in the forthcoming BME, since critique had been levelled at the somewhat uneven allocation of detail in previous medium-sized editions that were published in German and French

Index in Thesaurus Form

It is now generally recognized that a well-constructed thesaurus, using the standard relational devices of USE, BT, NT and RT, is a more flexible aid to the classifier than the conventional type of relative alphabetical index, and certainly much better than the mechanically produced one-line indexes of the German editions which are more in the nature of concordances A pilot project, undertaken by a group of Belgian experts, resulted in the construction of a thesaurus-type alphabet cal index to part of class 33, economics, although this is a notoriously difficult field for any kind of index because of the vague and constantly shifting terminology of the discipline, the results are very encouraging and much superior to the type of relative index used up to now in the English and French editions of the UDC

Subject-Field Reference Code

Turning now from these projects which are still geared to the existing framework of the UDC to the more ambitious plans for future remodeling of the whole system, at least three approaches have been made, and some tentative outlines have already been published for discussion.

The worldwide information and documentation network inaugurated by Unesco under the name of UNISIST recognized in its first report the necessity for an internationally applicable classification system for recorded knowledge by means of a Broad System of Ordering (BSO) and came to the conclusion that the UDC would be suitable for this purpose, although it might have to be substantially changed and updated:

> The use of the Universal Decimal Classification in particular . has been advocated. Its further potential has yet to be realized, and both a continuing programme to strengthen UDC and further studies and experiments to test its applicability to retrieval systems are desirable.[23]

From the outset it was clear that BSO would not be as elaborate as UDC but would rather be a much more general system serving primarily two functions (a) as a tool for broad indication of subject fields and disciplines, (b) as a switching code for other retrieval systems (classification systems, including the present UDC, as well as verbal indexing tools, such as subject heading lists and thesauri) which could thus achieve a minimal measure of mutual compatibility while still catering to the specialized needs of experts and practitioners in a particular subject field The original name of the project was later changed to Standard Reference Code (SRC), sometimes also referred to as a roof code (the *R* in SRC then standing for *roof*)

Initially there were some misgivings on the part of UDC experts ~~and~~ users that the new SRC, for the development of which FID had assumed responsibility, would virtually be the end of the UDC without necessarily resulting in a better or more useful tool (the properties, scope and actual application of the SRC as yet being in the realm of speculation), while the proponents of the SRC were apprehensive lest the more traditional ideas inherent in UDC would exercise an undue restraining influence on SRC This conflict was resolved by the formation of an independent group of experts in FID who will deal only with the development of SRC (although some members of the group and its coordinator, Dr. I. Dahlberg, are also UDC experts active in the formulation of more or less radical redevelopment programs[24] for UDC and conversant with the actual problems of document retrieval systems).

One of the first actions of the group was to give the initials SRC again a somewhat changed meaning as "Subject-field Reference Code," and to state that it would serve as

> A tool for interconnection of information systems, services and centers using diverse (often incompatible) indexing/retrieval languages.
>
> A tool for tagging (i.e shallow indexing) of subject fields and sub-fields
>
> A referral tool for identification and location of all kinds of information sources, centres and services.[25]

In early 1974, some 90 top-level subject fields had been identified, and a more detailed list of a second and third level breakdown will be elaborated and discussed at meetings during 1974, to be submitted for final approval at the forthcoming Third International Conference on Classification Research in Bombay in January 1975 At present, only a very rough tentative outline exists, so that it is impossible to assess the value of the system for its stated objectives as compared with other universal systems (including the UDC which gave the impetus to the whole enterprise) and to judge whether the international scientific community will be persuaded to use it, and if so, for what purposes, since the system is expressly intended not for the retrieval of individual documents from any specific store, but only as a kind of identification system for the location of blocks of information (whatever that may be) and whole collections Whether the considerable effort expended on the construction of the SRC scheme will be justified by these rather limited and somewhat nebulous goals remains to be seen

UDC as a Universal Faceted Classification

An elaborate plan for a thorough reform and revision of the UDC was submitted by A F Schmidt, head of the Classification

Committee at the German Standards Institution, who is responsible for the German UDC edition, in collaboration with J H de Wijn, who is in charge of the Dutch UDC edition.[26] To those familiar with the principles of UDC it might come as a surprise that a reform proposal should in fact only confirm what has actually been the inherent nature of the UDC since its beginning, namely its basically faceted structure (devised long before the term *facets* had been coined by Ranganathan, who conceived of the idea after having studied the structural features of the UDC).

But while it is true that UDC has always displayed facets in the form of its General auxiliaries and has indicated them by various nonnumerical symbols serving as facet indicators, this principle is countermanded innumerable times in the schedules themselves where the age-old method of simple decimal subdivision and enumeration, basically inherited from DDC, is used where the application of existing facets would not only be more logical (in terms of the structure of the system) but would also result in better and simpler retrieval Allow me to give just two examples. Anything connected with a country can and should be expressed by the geography facet (an elaboration of DDC's geographical subdivision device), e g where (73) is U S A , 63(73) is U S agriculture, 72(73) is U S architecture, etc , but the geography and history of a country are still main numbers, viz , 917 3 and 973, exactly as in DDC This means that in inverted files where documents can be grouped by the geography facet to give (73)63, (73)72 etc , the documents on the U S A are dispersed to at least three different and noncontiguous places (since the unfortunate interpolation of biography, 92, between geography and history has also been retained in UDC) Another example is the classing of persons for which a quite detailed and generally applicable auxiliary schedule, -05, exists, thus we find 52-05, astronomers, 62-05, engineers, 681 11-05, watchmakers, etc But for some unexplained reasons, there are also many direct subdivisions for persons, such as 262 14, clergymen (359 8 military chaplains, is separate!), and 78 07, musicians, using a different special auxiliary, 07, to do the job for which the general auxiliary, -05 was devised

The Schmidt-de Wijn proposal is intended to put an end to these incongruities and to put the UDC on a truly faceted basis without any exceptions thus cutting back substantially on the ever-growing numerical subdivisions which have become an impenetrable undergrowth stifling the sound trees of the system. Their plan provides for three levels

1 Superstructure

2 Direct subdivisions

3 Recurrent subdivisions

The superstructure would consist of about 70 to 80 superclasses with a two-figure notation (which might coincide with the upper level of SRC). followed by classes with a three-figure notation and then four-figure subclasses, all structured by the present principle of decimal subdivision. Within any of these three levels, further subdivision would be possible by distinctive notational devices and the appendage of either recurrent or special subdivisions (corresponding roughly to the present general and special auxiliaries). Finally, a relatively large but still manageable number of 'Recurrent subdivisions would be developed, e.g., concepts and features that cut across all disciplines and are more or less applicable to all or most of them Again, this principle is not basically new, but would now be applied much more consistently, freeing the main numbers from all unnecessary ballast and harmful duplication (as in the example of civil and military clergymen, neither of which should be enumerated at all but only indicated by a suitable person facet in the religion and military administration schedules respectively)

The following general facets or recurrent subdivisions are proposed

General features (e g abstract concepts time, relation, size quantity, quality, criterion, experience, etc)
Processes
Actions
Methods
Energy and power
Objects (materials, persons as individuals and as groups products, documents)
Languages
Philosophies
Cultures
Cosmic and geographic units

Schmidt also proposes some changes in the use of symbols as signposts and relational indicators, and considers that a UDC restructured along these lines would be even more amenable to computerized information retrieval than the existing form which, as already pointed out, has proved to be computer-compatible when relatively small adjustments were made, or even where the existing tables were used without any change. The authors of this proposal hope that their plan would provide not only the "roof" for SRC but also the "pillars" to bear it, i.e., the necessary substructure of more detailed indication of document content needed in actual retrieval situations which no doubt are of more importance to individual researchers than an internationally standardized code for blocks of information.

NUDC

A similar approach to the restructuring of UDC, a New UDC or NUDC, is taken by the Czech classificationists D Simandl and L. Kofnovec,[27] who also take the need for a SRC as their starting point but proceed to develop a methodology for the construction of a revised UDC rather than proposing a new structure as such. Their approach is based on the relative importance of subjects as indicated by the volume of literature generated in various fields (based on an analysis of abstracts in the Soviet abstracting journal *Referativnij Žurnal* and in other abstracting and indexing services), it results in the following rough percentage breakdown of the fields of knowledge

Technology	35%	
Chemical engineering		10%
Electrical and mechanical engineering		10%
Natural sciences	30%	
Chemistry and physics		10%
Earth sciences		5%
Medicine and agriculture	15%	
Social sciences and humanities	15%	
Others	5%	

The authors also provide a more elaborate table in which the 80 or so superclasses of a possible SRC are assigned two-figure notations and which shows a suggested breakdown of these main subject fields which is substantially different both from the one to which we have become accustomed in the

10-main-class framework common to DDC and UDC and from the one suggested by the SRC committee. This means that the biggest difficulty in the design of a new scheme seems to be the lack of consensus among various groups of experts on what constitutes a "super class" or main discipline It remains to be seen whether the Simandl-Kofnovec methodology can be fruitfully amalgamated with the Schmidt-de Wijn structural approach, and whether the final product of these endeavors will lead to a universal retrieval system which can cope better with rapid changes and innovations through inherent reliance on basic building blocks rather than on ad hoc additions and amendments to an inherently rigid and outmoded framework

Conclusions

This necessarily brief survey of the most recent developments in the complete revision or reform of the UDC shows clearly that there is still quite some life in the old tree whose roots go back more than one hundred years if we trace its ancestry to Dewey s scheme, first conceived in 1873 and published in 1876 It is also evident that the UDC is still a truly international scheme, with people in many countries contributing to its further development. These are by no means the isolated efforts of starry-eyed idealists, but constr'ictive attempts made by experts and backed by national and international organizations

Now that the initial euphoria of the early 1960s concerning the use of computers in information retrieval has evaporated, and the subsequent infatuation with slightly spruced-up subject heading lists under the grandiose name of *thesauri* has been replaced by more sober assessments of the requirements for construction and utilization of these and other retrieval tools—all of which rely in the last analysis on classificatory principles—there is indeed room for a new appraisal of existing classification systems and their restructuring in the light of both theoretical and practical insights gained over the last quarter of a century The UDC, so often declared to be dead (especially by those who did not know the purportedly deceased in life) will probably have to play an important role in these future developments towards a truly universal and international retrieval code, even though the phoenix to arise out of the ashes may have little outward resemblance to its venerable predecessors

Notes

1 Jack Mills, *The Universal Decimal Classification* (New Brunswick, N J Graduate School of Library Service, Rutgers, The State University, 1964).

2 Jean M Perreault, *Towards a Theory for UDC* (London Bingley, 1969)

3 Malcolm Rigby, *Mechanization of the UDC Final Report on Pilot Project to Further Explore Possibilities for Mechanization of UDC Schedules* (Washington American Meteorological Society, 1964) PB 166 412

4 Robert R Freeman, The Management of a Classification Scheme Modern Approaches Exemplified by the UDC Project of the American Institute of Physics, *Journal of Documentation* 23 (1967) 304-320

5 Pauline Atherton and Robert R Freeman, *Final Report of the Research Project for the Evaluation of the UDC as the Indexing Language for a Mechanized Reference Retrieval System* (New York American Institute of Physics, 1968) AIP/UDC-9 Also published in FID/CR Report No 9 (see note 14 below)

6 Thomas W Caless et al , *Strategies for Manipulating Universal Decimal Classification Relationships for Computer Retrieval* (Washington Biological Sciences Communication Project 1970)

7 G A Cooke, D M Heaps, and M Mercier The Study of UDC and Other Indexing Languages through Computer Manipulation of Machine-readable Data Bases, in *International Symposium UDC in Relation to Other Indexing Languages, Herceg Novi Yugoslavia 1971 Proceedings* (Beograd Yugoslav Centre for Technical and Scientific Documentation, 1972)

8 Helen E McCuaig and M Mercier *A UDC Water Thesaurus Concordance Development and Use* (Ottawa Environment Canada 1973)

9 J Webber "Die *Documentatio geographica* ein Beispiel fur die Anwendung der DK in der Geographie in *Seminar on UDC and Mechanized Information Systems, Frankfurt 1970 Proceedings* (Copenhagen Danish Centre for Documentation 1971), pp 21-41 FID CR Report No 11

10 F H Ayres C F Cayless and J A German Some Applications of Mechanization in a Large Special Library *Journal of Documentation* 23 (1967) 34-44

11 B. Barnholdt, A Computer-based System for Production of a UDC-classed Library Catalogue at the Technological University Library of Denmark," *Libri* 18 (no. 3-4, 1969): 191-196.

12 B. Studeli, Innerbetriebliche Informationsverteilung und -speicherung mit Hilfe von EDV-Anlagen auf der Basis der Internationalen Dezimalklassifikation," *DK-Mitteilungen* 15 (no 1, 1970) 1-4

13 Maurice W. Downey, "Data Collection and Transcription in the Cataloguing Section, Report on a Pilot Project in the ETHZ Library, Zurich," *Libri* 22 (no 1, 1972) 58-76

14 *Seminar on UDC in a Mechanized Retrieval System, Copenhagen, 1968, Proceedings* (Copenhagen Danish Centre for Documentation, 1969). FID/CR Report No 9.

15 *Seminar on UDC and Mechanized Information Systems, Frankfurt, 1970, Proceedings* (Copenhagen Danish Centre for Documentation, 1971). FID/CR Report No 11

16 Malcolm Rigby, *Computers and the UDC. A Decade of Progress, 1960-1970* (Rockville, Md National Oceanic and Atmospheric Administration, Scientific Information and Documentation Division, 1970)

17 Hans Wellisch, A Concordance Between UDC and *Thesaurus of Engineering and Scientific Terms (TEST),* Results of a Pilot Project, ' in *International Symposium UDC in Relation to Other Indexing Languages, Herceg Novi, Yugoslavia, 1971, Proceedings* (See note 7 above)

18 Einar Ohman and C Olivecrona, Some Notational Hierarchic and Syntactic Problems in Connection with Concordances Between UDC and Thesauri, in *International Symposium UDC in Relation to Other Indexing Languages, Herceg Novi, Yugoslavia, 1971 Proceedings* (See note 7 above)

19 Geoffrey A Lloyd, The Universal Decimal Classification As an International Switching Language, in *Subject Retrieval in the Seventies Proceedings of an International Symposium 1971* (Westport Conn Greenwood Press and School of Library and Information Services, University of Maryland, 1972), pp 116-125

20 Anthony C Foskett, *The Universal Decimal Classification The History, Present Status and Future Prospects of a Large General Classification Scheme* (London Bingley 1973)

21 Hans Wellisch, Organisatorische Neuordnung des DK-Systems *Nachrichten fur Dokumentation* 22 (1971) 55-63

22 Anthony C. Foskett, op. cit., pp 64-68.

23 *UNISIST. Study Report on the Feasibility of a World Science Information System* (Paris: Unesco, 1971).

24 Ingetraut Dahlberg, Possibilities for a New Universal Decimal Classification," *Journal of Documentation* 27 (1971) 18-36. Originally published in German in *Nachrichten fur Dokumentation* 21 (1970) 143-151

25 Andre van der Laan and Jan H. de Wijn, "UDC Revision and SRC Project. Relations and Feedback, *Unesco Bulletin for Libraries* 28 (no 1, 1974), 2-9.

26 Adolf-Friedrich Schmidt and Jan H. de Wijn, Some Possibilities for a New Reformed' UDC (Suitable for Extension of the Standard Reference Code)," *DK-Mitteilungen* 16 (no. 5, 1972)· 19-21

27 Ladislav Kofnovec and Dušan Simandl, "An Objective Approach to Building Up a General Classification Scheme,' *DK-Mitteilungen* 17 (no 2, 1973) 5-7.

Automatic Classification: Directions of Recent Research

Zandra Moberg

As the pervasive computer technology has come more and more to dominate many aspects of library science, directions of change in librarianship have been largely determined by this technology. "Library science" has been broadened to "library and information science," symbolizing the unlikely union of computer scientists and electrical engineers, on one hand, and humanistic librarians on the other on common professional ground.

In classification this synthesis has been manifested in extensive investigations into computerized classification. At first these new systems, alluding as they do to *clumps* and *passes* and *thesauri* and *algorithms*, seem to bear little resemblance to the subdivision-of-knowledge approach encountered in philosophy and in traditional library science classification. It is gratifying for librarians to note, nevertheless, that W. C. Berwick Sayer's classic definition, which specifies only that the arrangement of items be *useful*,[1] pinpoints the *raison d'être* of all automatic classification attempts, and is, even with the further refinement that things be assembled in an order of likeness,[2] comprehensive enough to include them all

Justification of Research

Impetus for research and development in automatic classification has grown from very practical considerations Little research is undertaken in any field from purely theoretical interest—someone must be willing to pay for it, and this implies that the results should have potential value in application Automatic classification is needed for at least two important

Zandra Moberg is Assistant to the Director of the Material Aids Program American Friends Service Committee Philadelphia, Pennsylvania

reasons. The first reason is utility to the user. Precision and recall from automatic indexing alone leaves a great deal to be desired and classification can be used to expand the queries and thereby to retrieve more relevant documents.

An even more compelling justification for automatic document classification in obtaining financial backing for research has been economics. searching only one section of a classified document collection. especially a very large one, requires dramatically less computer time as well as human time. The goal of research should be to produce effective, practicable classification by computer rather than manually, given the geometrically increasing volume of information to be processed.

Automatic Classification Arrangements

It should be emphasized that "automatic classification" may refer to two distinct kinds of arrangements. it may be a scheme for the grouping of index terms or it may mean classifying the documents themselves Investigators in the sixties worked on systems of one kind or the other. The names of Cleverdon, Salton, Lesk, Needham. and Sparck Jones are cited repeatedly for work on the former. significant contributions to research on the latter were made by Borko, Doyle, Rocchio and Dattola. A current trend seems to be to coordinate the two kinds of classification into one system

Another basic dichotomy in classification, which has become a dominating issue posed by the use of computers, is that of semantic classes versus statistical mathematical classes. Computers manipulate and store symbols quantitatively, but classes need to have appropriate names for communication of meaning between people [3] Acceptance of this premise seemed to mean that classes must have semantic unity, which pointed the way to experiments in linguistic analysis

Semantic Classification

Language analysis by computer, with a view to assignment of linguistic symbols to documents as content and class identifiers, has been one of the projects of Gerard Salton's SMART system SMART the most sophisticated and elaborate information storage and retrieval system in the United States, was designed at Harvard between 1961 and 1964. Operating at Harvard and Cornell, SMART has been supported by the National Science Foundation.[4] Under the aegis of computer scientist Salton, SMART is an experimental automatic system which serves as a testing ground for many ideas in information storage and retrieval, language analysis being conspicuously among them, in the sixties.

The language analysis experiments on SMART employ manual intervention in the processing of documents in that the thesauri are constructed manually. This means that judgments on which terms are to be classed together are made for SMART by people, not by machines, and that, strictly speaking, this is not automatic classification. However, it is not always logical or practical to separate indexing from classification, Salton's findings with the manual thesauri and Saiton's thinking have influenced directions pursued by subsequent researchers, and theoretical approaches outlined by him will be seen to have been carried out by others working on automatic classification.

The thesaurus entries in the SMART system constitute the classes of a keyword classification which was constructed by subject experts and committees of subject experts. The synonym thesaurus, or dictionary, builders determined which words and phrases would be important content identifiers for a given subject area, they were required to come up with all possible words or word combinations, so that the computer could be programmed to recognize them. There were separate dictionaries for different fields. There was also a suffix dictionary, listing approximately 200 English suffixes, a statistical phrase dictionary, a syntactic phrase dictionary, the usual negative thesaurus, word stem dictionaries, and concept hierarchies constructed specifically for different fields. These dictionaries were used in extensive experiments in semantic analysis employing several hundred automatic content analysis methods.[5]

To give an idea of the complexity of some of the procedures, in one—the syntactic phrase dictionary—each syntactic phrase entry consists of a specification of component concepts, syntactic indications, and syntactic relations permitted between concepts, all indicated by numbers. There are four possible basic kinds of syntactic indications, each being divided into twenty syntactic types. Syntactic dependency types are expressed in the form of syntactic dependency "trees," vertical displacement along a given path of the tree denoting syntactic dependency. Parts of speech are also prescribed.[6] Despite the enormous amount of work that must have gone into this thesaurus and associated algorithms, syntactic phrase analysis is not mentioned in the detailed evaluation studies and may be presumed to have been abandoned.

The dictionaries entailed great expenditure personally, as well as monetarily, it would seem:

> . . . the task of constructing a subject dictionary . . . is one which demands many skills, including a great deal of persistence and tenacity. . . . a committee is often appointed to thrash out controversial questions which frequently ends by satisfying no one. . . . any saving which might result from automatic search and retrieval methodology might be promptly lost through the elaborate preparations required to build dictionaries.[7]

One can infer the *Sturm und Drang* on the thesaurus committees. Salton and Lesk concluded that automatic or semiautomatic dictionary construction is imperative, above all, "to eliminate the human element"![8] Furthermore, in the exhaustive retrieval evaluation studies the performance of language analysis to characterize documents using these thesauri is not as effective as expected,[9] with the exception of the regular synonym dictionary. The numerous dictionaries, the hundreds of methods and countless runs on SMART for language analysis are expensive. Although the experiments are designed to meet the first practical need stated above—utility to the user by provision of more effective retrieval tools—Salton recognizes that ultimately cost is a consideration of overriding importance.[10]

Salton's current thinking on language analysis is that linguistics does not have much of a role to play in information retrieval,[11] having established once and for all, it would seem, that linguistic analysis by computer to characterize documents is not a fruitful avenue of investigation.

If automatic syntactic and semantic analysis has proven a dead end for classification, it is probably due to the nature of language itself and not because of inadequacies in anybody's algorithm. Salton and Lesk state it rather strongly: " . . . no human intermediaries exist who could resolve some of the ambiguities inherent in the natural language itself, or some of the inconsistencies introduced into written texts by the authors."[12] They leave it unclear whether or not they meant at that time to suggest that they thought machines ought to be able to resolve those ambiguities better than the subjective human intermediaries. It is an accepted premise of linguistics that the set of words in a given language is infinite and that the possible combinations of symbols are similarly infinite,[13] which presents a formidable challenge even to the most sophisticated software-hardware combinations.

Mathematical Keyword Classification

Salton and Lesk have turned from dictionary construction to the consideration of automatic methods of mathematical keyword classification. Their term-document matrix association procedure is based on the work of Doyle and others, they suggest that term association process be applied to the matrix to achieve thesaurus groups. In 1966, Salton and Lesk stated prophetically that "it would be nice if it were possible to give some generally applicable algorithm for constructing hierarchical subject arrangements,"[14] and went on to outline possible methods of automatic and semiautomatic methods of hierarchy formation based on keyword co-occurrence and resulting in a classification *tree*—a technique later developed and applied in the system of the Moore School at the University of Pennsylvania.

While Salton's experiments assumed that membership in a common class means that words must be related semantically, Karen Sparck Jones and others at the Computer Laboratory at the University of Cambridge have bypassed the foregoing difficulties with language by creating a purely statistical keyword classification system. "We cannot ask direct questions about the meanings of words if we are using automatic techniques," states Sparck Jones flatly in her monograph on automatic classification, but in a mathematical keyword classification it does not matter, for if two words always co-occur in a given set of documents they are necessarily able to be substituted for one another in retrieval, and it makes no difference whether they are semantically or conceptually related or not.[15]

Acknowledging previous work on statistical association in classes by Doyle, Cleverdon, Borko, Needham, Salton and others.

Sparck Jones has devised a series of controlled experiments to test the effectiveness of selected automatic keyword classification methods. Funded by the British government, the project has attempted to describe logically systematic comparisons of the recall and precision measures achieved by these methods.

Any mathematical classification using a matrix is a two-stage process, involving, first, the construction of a similarity matrix for object pairs based on co-occurrence of terms in documents. The class-finding procedure is then applied to this matrix to identify groups of similar terms. Four different routines for constructing the matrix and four group-finding procedures have been compared by Sparck Jones: strings, stars, cliques and clumps, so named for the kinds of links between elements of the classes. Runs were made using different combinations of each of these variables, comparing different values of a given parameter in a base environment with a view to drawing conclusions about the nature of a "good" classification.[16]

Sparck Jones has found, first, that automatically obtained term classification does give better retrieval than unclassified terms alone, which means that automatic keyword classifications are worth constructing and that the means of constructing them on a large scale is readily available by computer. Her results also show that the choice of similarity definition on the matrix does not affect retrieval performance very much and that the choice of class finding procedure also does not matter very much—strings, stars, cliques and clumps all did about the same. One constant did become apparent. Whichever of the four class definitions was used, restricting the class to a very strongly connected set of elements gave noticeably better retrieval performance.[17]

Another striking finding was that restricting the vocabulary of the classes by excluding the very frequent terms, treating them as classes unto themselves, promoted higher retrieval performance. And Sparck Jones found, contrary to expectation, that higher recall was accompanied by higher precision values, attributable to the fact that the classes were not mutually exclusive, and the context of one class or another defined more sharply the homographic words.[18] Her thorough evaluation included external comparisons in which she found her best results to be roughly comparable to Salton's with his best manual regular thesaurus.[19]

Discussing the necessity of updating a collection to take into account new documents, since addition of each new one could create new patterns of co-occurrences of terms, Sparck Jones has acknowledged the pervasive problem of control of very large collections and the last sentence of her monograph points out the direction being taken by current investigations. the use of automatic keyword classification in conjunction with automatic partitioning of a collection into units (document classification) for searching.[20]

In later research on term classification, Sparck Jones found the villain in low retrieval performance sometimes to be the supply of terms themselves, between which strong term connections could not be formed because relevant documents had not been separated from nonrelevant ones.[21] This provides strong support for the notion that automatic keyword classification will work best if the documents are classified or grouped by likeness and that the two kinds of automatic classification are indeed complementary.

Clustering Techniques

Extensive work was done on SMART on automatic mathematical document classification in the late sixties, namely the clustering techniques worked out by Rocchio and Dattola with a view to shortening search time.[22] Groups, or clusters, were created by correlating documents on a matrix according to keyword co-occurrence. One representative document description, called the centroid vector, was generated to represent all the documents in a given cluster, being a ranked list of the most frequently occurring index terms in the cluster. Queries were matched initially against the centroid vectors of the group and then against selected pertinent documents within the group. Standard precision-recall evaluations were carried out and investigations on questions pertaining to optimum cluster size, amount of overlap between groups, query clustering, and how many documents to allow unclustered were performed. The SMART investigators were able to conclude that cluster searching appears to offer large savings in search time, at no substantial loss in recall and precision, for all searches not requiring either a very high recall performance or a very high precision.[23] Generally, more clusters with fewer documents in each gave better precision and recall.[24]

A new wrinkle in current investigations into automatic classification is automatic hierarchy construction. The document clustering methods are potentially hierarchical, and Salton suggested at one time a multilevel search procedure by grouping the centroid vectors themselves into broader an: broader groups for expanding the search.[25] This proposed multilevel procedure is based on a principle to be worked out in others' algorithms, that the breadth or generality of a subject class is a function of how widely and how frequently member index terms occur. Salton's idea was depicted as wider and wider circles in a document space to indicate the levels.[26]

Text Organizing System

Production of a total system for processing data bases incorporating advanced techniques in information storage and retrieval into one practicable package has been the goal of the Text Organizing System at the Moore School, University of Pennsylvania, which was put together during the past ten years for the U.S. Office of Naval Research by a group including Prywes, Lefkowitz, Litofsky, Köymen and others. The work is still in process. In contrast to the testing ground function of SMART or Sparck Jones's program package of an orderly series of controlled tests with concurrent evaluation studies, the Text Organizing System is the product of applied research, intended for use with specialized or private data bases.[27]

The unifying component of this system is a classification algorithm, called CLASFY, contributed by Litofsky,[28] which successively subdivides items to create a hierarchy or classification tree, based on occurrence of index terms assigned to the items until document groups or keyword groups of the desired size are obtained. In the first step of the process candidate index terms are extracted from the text automatically. These are then selected manually for eventual use in the classification algorithm assisted by printouts from the computer of word frequencies and similarly spelled words which are used as guides for reducing the term vocabulary. The resulting directory of index terms determines which terms shall represent the documents in which they are contained. The classification algorithm is then applied to the documents represented by the index terms, successively subdividing the collection hierarchically, to produce a reordered data base, arranged in an order reflecting similarity of groups. Documents adjudged to be similar are allocated to common "cells" of approximately equal size. Moving up the tree, index terms showing content of document groups beneath them are indicated at the nodes of the tree, the terminal nodes being the cells, the actual location sites of the documents.[29]

The classes at these nodes on the tree are denoted by numbers, and document descriptions are built by combining, in order, the numbers of the successive nodes under which the document is grouped to produce the document's canonical classification number.[30] As in Salton and Lesk's manually constructed concept hierarchy tree, each successively higher node is assigned one more digit, the number farthest to the left representing the most general (i.e., frequent) class, with numbers representing the more specific (less frequent) classes toward the right. The class, or node, numbering system is thus the numerical notation of a synthetic classification scheme, the notation being built up from a hierarchy of node numbers representing mathematical classes derived from frequency statistics of index terms. In Salton and Lesk's hierarchy, it will be remembered, the classification numbers symbolize words and concepts; in the Text Organizing System, the synthesized classification number classifies a document.

CLASFY is a three-step algorithm which partitions the collection into more groups each time it is applied, each level in the tree resulting from another application of the subdivision process. In the first pass just the keywords of the collection or subgroup are partitioned; the next two passes assign documents on the basis of matching similar keywords in them to one of the groups. The algorithm continues to be reapplied until groups of like documents reach specified optimum size.[31] The more unique (to the collection) terms a document contains, the farther down the tree it will be.

The tree describing the entire collection is made available to the user in two directories. the key-to-node directory which lists index terms along with all the documentanonical classification numbers to which they have been assigned; and the node-to-key directory lists all the node numbers with assigned keywords. Facsimile tables of these directories are manually searched in printout or microfilm form, a first step in retrieval.[32]

The final step of the integrated text processing in T. O. S. is the creation of a keyword classification by means of the same subdivision algorithm. The keyword vocabulary is successively subdivided into mutually exclusive sets of keys on the basis of classification numbers assigned to each key in the document classification process—the more documents a term appears in, the higher up in the tree it will be. The product of the mathematical keyword classification is an Affinity Dictionary, which is intended to be used both to expand a search by identifying interchangeable terms, and to consolidate terms in updating the system.[33]

Like Sparck Jones's thesaurus, the Affinity Dictionary is an automatically derived keyword classification, the difference being in the mathematical techniques. Any comparison of retrieval success is not possible because the keyword classification component of the Text Organizing System can not be isolated out from the total system, and because no recall precision values are available on it as yet. It is of interest to note that the divisive, hierarchical T. O. S. algorithm achieves one condition noted above which Sparck Jones's tests demonstrated to be an important factor for success: that frequent terms should be classes unto themselves and that less frequent terms should be grouped. This is exactly what CLASFY does with keywords.

In the absence of retrieval success figures, which are still being worked on at the Moore School,[34] the authors use a rather curious criterion for evaluating their document classification scheme. "... the quality of a classification system is measured by how well it minimizes the average number of keys, per cell,"[35] i.e., the fewer keywords characterizing a class of documents, the more alike they must be.

Hoyle's Integrated System

Meanwhile, W. G. Hoyle at the National Research Council of Canada has recently developed an integrated indexing and classification system similar in many respects to the Text Organizing System. The automatic indexing procedure is the basis for automatic generation of a classification scheme. documents are assigned to categories on the basis of keyword occurrence, as in CLASFY. Hoyle's procedure does not involve reapplications of the divisive algorithm to create a hierarchy but he does suggest the possibility of "super categories" of keywords by this method. A specialized thesaurus, analogous to the Affinity Dictionary, is also a last step of the process. Hoyle's method, however, does not partition the collection into mutually exclusive groups, but produces rather an "ordering of relevance." Further, the ordering employs weighing of terms, which the Test Organizing System does not. Comparing his document and keyword classification system to a manual one using the same material, Hoyle found "reasonable resemblance."[36]

Van Rijsbergen's Hierarchical Clustering

Another avenue to reducing computer time in retrieval by scanning only a subset of a classified collection is hierarchical clustering, called cluster-based retrieval. Research on this approach has been reported recently by van Rijsbergen at Cambridge, with "helpful comments and criticism" by Karen Sparck Jones. Document classification follows logically from Sparck Jones's findings on the importance of the collection properties, this method arranges the collection in a hierarchic system of clusters. The clusters are obtained by a single-link cluster method applied to a dissimilarity matrix to generate a stratified hierarchy of clusters.[37]

In the Text Organizing System, it will be remembered, the documents are located only at the tips of the branches of the tree, and Salton's suggestion for multilevel search with clusters created hierarchy by using the centroid vectors or keyword groups. Hierarchic clustering is innovative for arranging the documents themselves hierarchically. Evaluated for retrieval effectiveness, this technique was reported cautiously by van Rijsbergen as being "quite competitive."[38]

The Mathematical Theory of Hierarchy

It may be observed that several of the automatic classifications described have included or hinted at hierarchy formation. These mathematically obtained hierarchies are formed on a different basis from hierarchies heretofore encountered in library science. *Hierarchy* has had a common meaning of a division of knowledge into progressively narrower classes which necessarily bear a generic-specific relationship to one another. The new automatic hierarchies are based on the occurrence and co-occurrence of words in documents. Words occurring less frequently, in fewer documents, determine a lower position in the hierarchies, more frequent terms appearing in a wider range of documents are indicators of a higher point, thus *generic* is replaced by *more frequent* and *specific* comes to mean *less frequent*. Salton propounded the new mathematical theory of hierarchy in 1966:

> . . . there seems to be some relationship between the frequency of occurrence in a given collection and its place in the hierarchy. More specifically, those concepts which exhibit the highest frequency of occurrence in a given document collection, and which by this very fact appear to be reasonably common, should be placed on a higher level than those concepts whose frequency of occurrence is lower.[39]

It is again gratifying for librarians to note that the profession has not rested on theory at variance with a reality being transformed by automation. At the Elsinore Conference on Classification Research in 1964, classification was defined as "any method creating relations, generic or otherwise, between individual semantic units, regardless of the degree of hierarchy contained in the systems and whether . . . with traditional or more or less mechanized document searching,"[40] a detailed but flexible definition which will accomodate any ilk of hierarchic or nonhierarchic classification based on word frequency and distribution figures.

Conclusions

In all of the above-reported research, in which methods have been sought to generalize classification from indexing, there has been implicit recognition of the inextricable relatedness of indexing and classification as two facets of one process. No automatic system should be expected to retrieve satisfactorily from a system of document description in which the classification bears no relation to the index terms as is sometimes the case in the Library of Congress system. This emphasis on coordination of the two is a less obvious theoretical contribution of automatic classification research.

Research will continue to be carried out and supported if for no other reason than the economic one of reducing computer time on large collections. Richmond's prognosis for automatic classification in the most recent *Annual Review of Information Science and Technology* is dim because "its utility as a satisfactory means of document representation has not been demonstrated except for small, homogeneous collections in well-defined subjects of narrow scope"[41] (probably an allusion to Sparck Jones's findings). This statement can,apply to keyword classifications only, document classification aims at the creation of just such "small homogeneous collections in well-defined subjects" within the larger collections. And within such homogeneous document subgroups, retrieval success with keyword classification can be expected to be optimal. What seems to be called for at this time in management of large collections is coordination of keyword classification and document grouping, which is what the Text Organizing System of the Moore School and others have been attempting to do. And, since the ultimate basis for existence of information systems is, to go back to Sayers, utility to the user, interest will not be expected to wane in the refinement of systems to approach the elusive goal of high precision and recall.

Notes

[1] W. C Berwick Sayers. *Manual of Classification* (London Grafton, 1944). p. 1.

[2] Ibid., p. 4

[3] P. Baxendale, "Content Analysis Specification and Control," in *Annual Review of Information Science and Technology*, vol 1, ed. by Carlos Cuadra (New York: Wiley, 1966). p. 91.

[4] Gerard Salton, ed., *The SMART Retrieval System Experiments in Automatic Document Processing* (Englewood Cliffs. Prentice Hall. 1971). vii-x.

[5] Gerard Salton and M E. Lesk. 'Information Analysis and Dictionary Construction," in *The SMART Retrieval System*, pp 119-132.

[6] Ibid., pp. 128-130.

7 Ibid., pp. 132-133.

8 Ibid., p. 133.

9 Gerard Salton and M. E. Lesk. "Computer Evaluation of Indexing and Text Processing," in *The SMART Retrieval System*, p. 178.

10 Ibid., p. 146.

11 Gerard Salton, personal note.

12 Gerard Salton and M. E. Lesk. "Computer Evaluation of Indexing and Text Processing," p. 166.

13 Joseph Greenberg, *Anthropological Linguistics* (New York. Random House, 1968), pp. 75-76.

14 Gerard Salton and M. E. Lesk. "Information Analysis and Dictionary Construction," in *The SMART Retrieval System*, p. 132.

15 Karen Sparck Jones, *Automatic Keyword Classification for Information Retrieval* (Connecticut: Archon, 1971), pp. 9-10.

16 Ibid., pp. 45-69.

17 Ibid., p. 195.

18 Ibid., p. 13.

19 Ibid., p. 239.

20 Ibid., p. 242.

21 Karen Sparck Jones, "Collection Properties Influencing Automatic Term Classification Performance. *Information Storage and Retrieval* 9 (no. 9, September 1973): 510-513.

22 S. Worona. "Query Clustering in a Large Document Space. in *The SMART Retrieval System*, p. 299

23 Gerard Salton, "Cluster Search Strategies and Optimization of Retrieval Effectiveness," in *The SMART Retrieval System*, pp 238-239

24 Robert T. Grauer and Michael Messier, "An Evaluation of Rocchio s Clustering Algorithm," in *The SMART Retrieval System*, p 257

25 Gerard Salton. "Cluster Search Strategies and the Optimization of Retrieval Effectiveness, in *The SMART Retrieval System*, pp 225-226

26 Ibid., p. 226.

27 Noah Prywes, Allen Lang and Susan Zagorsky, "*A-Posteriori* Indexing Classification and Retrieval of Textual Data," *Information Storage and Retrieval* 10 (no. 1, January 1974): 15.

28 Barry Litofsky, "Utility of Automatic Classification Systems for Information Storage and Retrieval" (PhD dissertation, University of Pennsylvania, 1969), p. 99.

29 Kemal Köymen, "TOS. A Text Organizing System (PhD dissertation, University of Pennsylvania, 1974), pp. 161-167.

30 Ibid., p. 153.

31 Ibid., pp. 161-175.

32 Ibid., p. 190.

33 Ibid., p. 17.

34 Ibid., p 217.

35 Ibid., p. 185.

36 W. G. Hoyle, "Automatic Indexing and Generation of Classification Systems by Algorithm," *Information Storage and Retrieval* 9 (no. 4, April 1973): 234-241.

37 C. J. Van Rijsbergen, Further Experiments with Hierarchic Clustering in Document Retrieval, *Information Storage and Retrieval* 10 (no. 1, January 1974), p. 3.

38 Ibid., p. 13.

39 Gerard Salton, "Information Analysis and Dictionary Construction," in *The SMART Retrieval System*, p. 132.

40 P. Baxendale, "Content Analysis Specification and Control," in *Annual Review of Information Science and Technology*, vol. 1, ed. by Carlos Cuadra, p. 89.

41 Phyllis Richmond, "Document Description and Representation," in *Annual Review of Information Science and Technology*, vol. 7, ed. by Carlos Cuadra (Washington, D.C.. American Society for Information Science, 1972), p. 88.

The Future of Classification*

Phyllis A. Richmond

About one hundred years ago, the British philospher, William Jevons, dismissed classification as a "logical absurdity". His view has not been an unusual one. Some of the more interesting adventures in information science, such as Mortimer Taube's Uniterms, stemmed from an irremediable loss of faith in classification as a way or organizing knowledge. Furthermore, Kurt Goedel's now famous proof pulled the rug out from under systems based solely on deductive logic, so that one was left with the necessity of organizing knowledge on some other basis.

In effecting change to a new basis, two areas were obviously wide open for investigation. The first was words—index terms, descriptors—the path taken by Taube and others. The second area was inductive logic—classification systems built from the ground up primarily. The most notable of these have been the faceted classification systems, but they are not the only ones. Maps, graphs, patterns, statistics and probability theory have been invoked either to show relationships or to find them.

The area of words has had considerable attention for some time. The great weakness of systems based on words turned out to be the very richness of language. Part of the problem was the willingness of creative minds to employ old words with new meanings—a practice Robert Fairthorne deplored as words "used in public with private meanings."[1] The effort to escape classification by means of verbiage alone was not an unqualified success, nor was it a total failure. The successful development of thesauri during the past fifteen years has shown

Phyllis A. Richmond is Professor of Library Science, Case Western Reserve University.

*Delivered at the Institute on Classification in the 1970s, Faculty of Library Science, University of Toronto, 25 May 1974.

that a controlled vocabulary can be quite useful in a homogeneous subject field, even though it has minimal classification features—minimal meaning to about seven levels.[2] Subject headings, descriptors, index terms, unit terms and many other kinds of terms are still very much with us.

Methods of linguistics, particularly computational linguistics, are still being investigated as a means of describing knowledge.[3] Machine translation is currently in abeyance though it may not remain so. Some methodologies originally developed for syntactic analysis, content analysis and similar processes are viable, as is Gerard Salton's ultrarefined little SMART system at Cornell.[4]

The problems of definition of words remain. Relationships between concepts for which words stand and the very sources and uses of words themselves still need much more research. Interest in relationships has led from word lists to thesauri to attempts at mapping by means of directed graphs and other similar devices.[5] The mapping has reintroduced a factor of classification into the subject analysis process, just as *see also* cross-references inserted a measure of classification into subject heading compilations. A less semantically-oriented type of relationship study has produced little classification systems by means of some of the techniques of applied mathematics. William Goffman's "indirect method" is an example of this.[6] Although it has been used deliberately for word classification in only one publication so far,[7] it has some interesting possibilities.[8] In sum, one may say that in the area of words, as an alternative to classification, the trend has led right back to classification.

The area of inductive logic was partly the foundation of faceted classification, but here again it was not all clear sailing. The problem of relationships turned up again, and Jason Farradane has been emphasizing this factor for over twenty years.[9] Derek Austin, who was engaged to work full time on a New General Classification for the Classification Research Group in London, found the relationship factor so highly significant that, while he did not complete the New General Classification, he made use of the relationship-in-classification idea to the extent of developing it in his PRECIS system.[10] PRECIS terms, among other things, carry enough of their context with them to be a miniature, multiple-entry classified index to a given title.

Continuation of Present Work

The idea of a New General Classification has not been given up in London. Regular meetings are still being held and the problems of classification still retain their interest.[11] Classification is by no means a dead issue.

For the immediate future, one may expect continuation of present work. This includes the Depth Classification schedules being produced at the Documentation Research and Training Centre in Bangalore, an ongoing effort that already has produced well over fifteen schedules on subjects mostly, but not entirely, scientific and technical.[12] Statistical and probabilistic methodology will certainly continue to be applied in the attempt to make automatic classification systems. Karen Sparck Jones, in particular, has been persevering in this direction.[13] Modern mathematics has areas, notably in topology, which may be applicable to classification. Some attempts have been made along the line of physical models for three-dimensional classification in an attempt to improve the visualizing process begun in mapping class relationships via directed graphs and the like.[14] At the moment, this work is primarily a teaching tool.

Looking toward the future, UNISIST, with the cooperation of the International Federation for Documentation, has set up a Working Group to prepare the background for and to begin to develop a Subject-field Reference Code (SRC). This code is designed to produce a Broad System of Ordering, which is "a mechanism for shallow indexing, whose goal is to locate and transfer large blocks of information, rather than specific documents or data, between different discipline and mission-oriented systems, using, eventually, different natural languages."[15] This broad classification scheme is to be universal in scope, flexible enough to keep up with changes in the fields of science and technology, easily updated, simple in structure so that it can be adopted and adapted inexpensively, and usable for both manual and computerized systems. These noble goals are not entirely new, but it will be very interesting to see what transpires. One has the impression of *déjà vu*, but hope does spring eternal and within the given parameters it may prove possible to produce such a scheme.

Meanwhile the traditional forms of classification will continue. expansions of the Universal Decimal Classification, phoenix schedules in Dewey and new editions of the Library of Congress system. Various suggestions have been made to get more mileage out of these various schemes by using a variety of techniques, such as merging the subject headings, class descriptive terms and index terms for the Library of Congress and Dewey systems, with rotation and permutation of the individual words involved, in order to offer the user more access points.[16] John Immroth has made a study of the Library of Congress system in this respect and has developed a means for chain-indexing some of the classification.[17] His method appea:s to work better with those parts, notably in literature, which have some degree of hierarchy in them.

Influence of New Factors on the Scene

In PRECIS, in automatic keyword classification, in the SMART system and others, the computer has been introduced as a convenient tool. This gadget promises to become as commonplace as the telephone. Currently, there is a hand-held calculating machine which can be programmed with cassettes, thus giving the user the option of staying in his office and getting as much or more computer power than was available with early machines of the first generation. The computer has already been applied to classification in a project covering the whole middle edition of the Universal Decimal Classification (in English).[18] The computer can be used as a sorting device with the notation of almost any classification system. It is used for printing and maintaining the Library of Congress subject heading system. It is the mainstay of all automatic classification attempts and a good many indexing ones as well.

For comparative classification studies, it is helpful because it can dredge up more materials for study in an hour than one could get in a year by manual means, assuming machine-readable text, and one can work with total data instead of samples. The intellectual aspects of comparative classification, however, are not amenable to mere computer manipulation be cause of the varying theoretical foundations of the more common classification schemes.

The computer can do a lot more than just count. Even a brief look at the usages to which it has been put in the humanities (other thari for concordance-making) will indicate the variety of methods that have been used to deal with masses of historical data, literary text, archaeological findings, architectural rendering possibilities, and so on. The cleverness with which the computer has been assimilated into the methodology of the research scholar is impressive. One can, for instance, use a quantitative history approach to reveal unsuspected occurrences for which the answer must then be sought by more conventional means.[19] As machine-readable data bases, such as MARC, become available and access to such bases via console spreads, it will be possible to replace one classification with another automatically, though the results may not be very satisfactory because of the variation in systems and also because a one-to-one correspondence between classes is present less than half the time.

It is much more likely that a totally new classification might be adopted if a satisfactory one comes along. This is probably one thing that keeps the Classification Research Group going. The Bliss Bibliographic Classification and the Colon Classification were hardly touched because they arrived on the scene when the three major systems were already entrenched. In big libraries, total reclassification, as a rule, has only been done under dire necessity, as, for example, at Cornell in 1947 when the old homemade system had virtually collapsed.

With centralized processing and computerized distribution, the adoption of a new scheme, while still a major problem, would not be an unthinkable one. Therefore, classification research and the search for a new general system are no longer the knowledge-for-knowledge's-sake undertakings that they have appeared to be for the last fifty years.

In addition to the computer, there is another major factor on the scene which eventually may affect classification. This is the concept of the wired city. The combination of cable television, the telephone and the computer can bring into any subscriber s home a variety of services. Some of the projected configurations sound like Big Brother, especially as there will be changes in who controls access to information, not to mention who has access to information and when. On the other hand, the conveniences of the wired city would be great time and energy savers. Classification might well be involved in the information retrieval aspects of such a system. One can conceive of browsing in a classified catalog when one dialed up the local library. When specific information was sought, data could be displayed on the TV screen and the reference interview could be conducted over the line, so to speak, instead of over the counter.

A third new factor on the scene is the concept of the information utility. "A utility can be defined as a system providing a relatively undifferentiated but tangible service to a mass consumer group and with use charges in accordance with a pricing structure designed for load levelling."[20] Normally one would not consider the library in this category. However, a step in that direction has been taken with the formation of networks among libraries in order to share the cost of delivering bibliographic material to cooperating institutions. The middleman rather than the user pays the cost. Information centers and commercial services which serve the paying customer are much closer, but they are still independent agents, not a utility. If the whole lot were thrown in together, including the agencies that produced the information-product, and the user paid according to his needs, one would have something closer to a utility.

Perhaps the librarians nearest this concept are those who argue that it would make sense to contract out for all or some of the data services based on machine-readable data bases rather than to try to maintain them in the library where they would only be partially used at great overhead cost. A resident bibliographer would act as agent between the would-be user and the vendor to ensure that the best possible connections were made.[21] The user would ultimately pay the selected vendor. Veaner has suggested, for example, that cataloging services might be "purchased totally from a vendor or obtained from *his* resident staff, much as computer centers buy specialized expertise through the 'resident s.e.' (systems engineer)." [22] In view of how much cataloging still has to be done locally at present, this seems less likely than the purchase of access to on-line reference tools. In either case, however, the technical process of actually calling up the data is going to require classification and indexing, particularly if the query is on a subject. "I need to know how to synthesize rubrene." "What is available on the etiology of multiple sclerosis?" "I want to consider the degree to which modern poets may have been influenced by current scientific views on plate tectonics. Please give me all the review articles you can find on this theory." If the distribution of information is to follow the pattern of the distribution of electricity or gas, the price will have to be lower than at present and the quality of the end product higher. The idea of an information utility will be an interesting thing to watch. The analogy may not be as applicable as it sounds.

Unsolved Problems

There are still at least four major unsolved problems in classification. The first is the problem of continuous updating. Every classification system is out-of-date the minute it goes to the press. If it goes out at noon, by one o'clock additions and changes have already begun to come in. Existing classification systems are kept up-to-date by corrections and additions made either at regular intervals, as with the Library of Congress system, at irregular intervals by international cooperation, as with the Universal Decimal Classification, or mainly by editions, as with the Dewey Decimal and Colon Classifications. The continuous process at regular intervals is probably the most satisfactory for the user, but even here it is easy to fall behind current knowledge.

The second problem is virtually unsolved. How does one represent objective reality adequately? Any system delineated on the pages of a book leaves a great deal to be desired. Subjects may be scattered through a multiplicity of disciplines. Hierarchies with only two dimensions are unrealistic. Connections or splits or mergers among subjects may be hard to show. The trend toward interdepartmental cooperation tends to wipe out specific boundaries in some places and raise them in others. Complex systems, for example, exist everywhere, but the study of them excludes most of those in the humanities.

The third problem in classification is that we do not yet have an organizing philosophic basis for current thought in the late twentieth century. The philosophy may be here but unrecognized, or it may be in process but has not yet emerged publicly. As yet we cannot reorganize our body of knowledge according to principles more realistic for its content. This seems like a minor matter, but it is not. Each age has its own way of looking at the universe and its own body of knowledge and belief follows that insight. No new major classification has come forth in such terms for the last half of the century. Perhaps part of the difficulty the Classification Research Group has had in producing their projected New General Classification is that they have no philosophic system to hang it on. Thus they are generalizing particulars and dealing with methodologies when what they really need is a broad organizational pattern from a suitable philosophic system.

The final major problem is that of how to develop a completely open-ended system with infinite hospitality in array, chain and concept capture. In one way, this is a part of the first problem, that of continuous updating, but it is also a technological difficulty, a philosophic c ne and a notational problem. The solution must be partly a creative one. It seems more likely to come in a flash of creativity or even by chance. But since chance favors the prepared mind, one must still explore all possibilities while awaiting insight.

A Few Research Problems

Some of the possibilities to prepare the ground for advances in classification will be listed here. They seem relatively mundane compared with the problems given above. In the process of their undertaking, they should suggest other research, which in turn will lead to more, until ultimately we will have a better basis for classification-making than exists today. The topics are as follows:

1 *Frequency distribution study of classification number/subject heading correlation with words of titles in nonliterary works.* Is title page classification the rule or the exception?

2 *Frequency distribution study of the coincidence of classification number and first subject heading.* Is this common practice or wishful thinking? One should get a Zipf-Bradford curve if the former is actually the case.

3 *Study of built-in ambiguity in a classification system.* What differences can be demonstrated in classifiers' interpretations of a given classification scheme? Can variations be eliminated so that a user can count on getting everything *specifically* on the same topic in the same class?

4 *Study of variations among systems of classification.* Instead of critical analysis with a view to abstracting "the best," what could be taken from each for augmenting the record, thus

allowing users a variety of entrance points? (This is in part taking the opposite approach from that in problem number 3.)

5 *Study of using variant classifications for different subjects or materials in the same collection.* Is one grand scheme for all as effective as an eclectic system where the principles for classifying each subject would be suited to that subject or to the kind of material? One can think of good reasons for classifying government documents, serials, literary works, certain audiovisual materials and possibly scientific literature differently from the rest of a library's collection.

6 *Frequency distribution study of the effect of classification unity or scatter caused by cataloging series as separates vs. analyzed sets.* Can regularities be discovered which would suggest practical means of deciding between the two methods when an item is first received?

7 *Investigation of depth classification at the chapter and section heading level in monographs.* Under what circumstances is the effort worth the results? How do these additions increase accessibility? Should the subject index be added?

8 *Study of cut-off levels in classification.* Can one produce subject bibliographies evaluated critically so that the user can employ classification to get the level of sophistication he requires? (This is only partly a classification problem. It is also a problem of Dr. Koh's "data quality control."[23])

9 *Cross-classification of data bases in machine-readable form.* What linkage should be devised between machine-readable data bases so that the user could progress from information retrieval to bibliographic retrieval to document retrieval from a single console? (This does not mean sticking a classification number on every word!)

10 *Classification from machine-readable text.* Can a method similar to *content analysis* be used in conjunction with classification principles to derive an automatic classification of any text? This assumes some kind of semantic relationships will be discovered and identified during the content analysis process, as opposed to the mathematical kind of relationships sought by Sparck Jones.

11 *Application of a mathematical means, such as that of Goffman, to the terminology of class descriptions as a means of finding relationships among classes.* Can such a means be used to pull together classes scattered among different subjects?

12 *Study of the confusion factors in automatic classification. metaphor, allusion, synonymy, analogy.* Is automatic classification possible where these confusion factors are used to express new ideas? Since humans take in new knowledge by fitting it into existing patterns and can classify by an "inductive leap,"[24]can a means be devised to do this automatically?

Conclusion.

Currently it looks as if classification has taken a new lease on life. Twenty years ago a colleague called it "a grand intellectual exercise" with the implication that it did not have much value beyond that. Now, with extension of the range of immediate access to information, tremendous increase in the sheer bulk of information to be communicated, recognition of the interdependence of subject matter in a great many disciplines, technological capabilities beyond the wildest dream of twenty years ago and emphasis on quick and effective communication, classification is becoming more and more the entry point of choice. A part of this is due to the demonstrated weakness of reliance on terminology alone. Nevertheless, it is rather obvious that classification without indexing is just as impossible as indexing without classification. The two go hand-in-hand. In conclusion, one may iterate that past is prologue. New opportunities call for tradition-shattering creativity, with a promise of at least the *possibility* of widespread usage of results.

Notes

[1] Robert A. Fairthorne, "Content Analysis, Specifications and Control, in *Annual Review of Information Science and Technology*, vol. 4 (Chicago: Encyclopedia Britannica, 1969), pp. 98-99.

[2] A good example is *Thesaurus of Engineering and Scientific Terms* (New York: Engineers Joint Council, 1967).

[3] Cf. Donald E. Walker, "Automated Language Processing,' in *Annual Review of Information Science and Technology*, vol. 8 (Washington. American Society for Information Science, 1973), pp. 69-119.

4 Gerard Salton, *The SMART Retrieval System. Experiments in Automatic Document Processing*, Prentice-Hall Series in Automatic Computation (Englewood Cliffs, N.J.: Prentice-Hall, 1971).

5 L. Rolling, "The Role of Graphic Display of Concept Relationships in Indexing and Retrieval Vocabularies," in *Classification Research, Proceedings of the Second International Study Conference Held at Hotel Prins Hamlet, Elsinore, Denmark, 14th-18th September 1964*, ed. by Pauline Atherton (Copenhagen: Munksgaard, 1965), pp. 295-320, European Atomic Energy Community—Euratom, *Euratom-Thesaurus. Keywords Used within Euratom's Nuclear Energy Documentation Project* (Brussels: Center for Information and Documentation CID, 1964), Lauren Doyle, "Semantic Road Maps for Literature Searchers," *Journal of the Association for Computing Machinery* 8 (no. 4, October 1961). 553-578; "Indexing and Abstracting by Association," *American Documentation* 13 (no. 4, October 1962). 378-390, "Some Compromises between Word Grouping and Document Grouping," in *Statistical Association Methods for Mechanized Documentation. Symposium Proceedings, Washington, 1964*, ed. by Mary Elizabeth Stevens et al., National Bureau of Standards Miscellaneous Publication 269 (Washington, D.C.: Superintendent of Documents, 1965), pp. 15-24, "Is Automatic Classification a Reasonable Application of Statistical Analysis of Text? ' *Journal of the Association for Computing Machinery* 12 (no. 4, October 1965): 473-489.

6 William Goffman, "An Indirect Method of Information Retrieval," *Information Storage and Retrieval* 4 (no. 4, December 1968). 361-373.

7 Jane R. Moore, "On Interrelationships of the Sciences and Technology as Expressed by a Categorized List of Journals and Modified by a Classification System, ' *Journal of the American Society for Information Science* 24 (no. 5, September-October 1973): 359-367.

8 Phyllis A. Richmond. "Synthetic Classification." To be published in the Ranganathan Memorial Volume.

9 Jason E. L. Farradane, "A Scientific Theory of Classification and Indexing and its Practical Applications," *Journal of Documentation* 6 (no. 2, June 1950). 83-99, "A Scientific Theory of Classification and Indexing. Further Considerations," *Journal of Documentation* 8 (no. 2, June 1952). 73-92, "Fundamental Needs and New Fallacies in Classification, in *The Sayers Memorial Volume, Essays in Librarianship in Memory of William Charles Berwick Sayers*, ed. by D. J. Foskett and B. I. Palmer (London. Library Association, 1961), pp. 120-135, "Analysis and Organization of Knowledge for Retrieval," *Aslib Proceedings* 22 (no. 12, December 1970): 607-615.

10 Derek Austin, "Precis Indexing," *Information Scientist* 5 (no. 3, Sep-

tember 1971): 95-113; "Two Steps Forward . . . " in B. I. Palmer, *Itself an Education*, 2d ed. (London. Library Association, 1971); pp. 69-111, "A Conceptual Approach to the Organization of Machine-held Files for Subject Retrieval," in Ottawa University, Faculty of Philosophy, *Conference on the Conceptual Basis of the Classification of Knowledge, Ottawa, Ontario, Canada, 3 October 1971*. Preprint (in press), "Classification and Subject Indexing at the British National Bibliography," *Canadian Library Journal* 30 (March-April 1973). 122-130.

11 "Classification Research Group. Bulletin No. 10," *Journal of Documentation* 29 (no. 1, March 1973): 51-71.

12 Cf. *Library Science with a Slant to Documentation*, Vol. 1, no. 1-3, Vol. 2, no. 1-2; Vol. 3, no. 1, 3-4; Vol. 4, no. 2-3; Vol. 7, no. 1, 3-4; Vol. 9, no. 2-3; Vol. 10, no. 1 (1964-1973).

13 Karen Sparck Jones, *Automatic Keyword Classification for Information Retrieval* (Hamden, Conn.. Archon Books, 1971), "Collection Properties Influencing Automatic Term Classification Performance," *Information Storage and Retrieval* 9 (no. 3, September 1973). 499-513.

14 Phyllis A. Richmond and Nancy Williamson, "Three Dimensional Physical Models in Classification." Paper for Third International Study Conference on Classification to be held in Bombay, January 1975.

15 "UNISIST Seeks Broad Classification Scheme," *Information Part 1. News. Sources. Profiles* 6 (no. 2, February 1974): 41.

16 Phyllis A. Richmond, "LC and Dewey. Their Relevance to Modern Information Needs," in *Proçeedings of the ALA Preconference on Subject Analysis of Library Materials, Atlantic City, New Jersey, June 19, 1969* (in press), "A Reconsideration of Enumerative Classification for Current Needs," *Ciencia da Informacão* (Instituto Brasileiro de Bibliografia e Documencão), in press.

17 John Phillip Immroth, *Analysis of Vocabulary Control in Library of Congress Classification and Subject Headings* (Littleton, Colorado. Libraries Unlimited, 1971).

18 Robert R. Freeman and Pauline Atherton, *Final Report of the Research Project for the Evaluation of the UDC as the Indexing Language for a Mechanized Reference Retrieval System*, UDC Project, Report No. AIP/UDC 9 (New York: American Institute of Physics, 1968).

19 Roy E. Schreiber, "Studies in the Court of Wards, 1624-1635—The Computer as an Aid," *Computer Studies in the Humanities and Verbal Behavior* 1 (no. 1, January 1968): 36-42.

20 Allen Veaner, "Institutional Political and Fiscal Factors in the Development of Library Automation, 1967-71," *Journal of Library Automation* 7 (no. 1, March 1974): 22.

21 Richard De Gennaro, "Providing Bibliographic Services from Machine-readable Data Bases—The Library's Role," *Journal of Library Automation* 6 (no. 4, December 1973): 215-222.

22 Allen Veaner, op. cit., p. 23.

23 Hesung C. Koh, "HABS. A Research Tool for Social Science and Area Studies," *Behavior Science Notes* 8 (no. 2, 1973). 172-173, 191.

24 William Whewell, *Novum Organum Renovatus, Being the Second Part of the Philosophy of the Inductive Sciences*, 3d ed. (London. John W. Parker, 1858), Book II, pp. 88, 114. For usage in classification context, see Phyllis A. Richmond, "Contribution toward a New Generalized Theory of Classification," in *Classification Research, Proceedings of the Second International Study Conference*, pp. 47-50, J. Farradane, J. M. Russell, and P. A. Yates-Mercer, "Problems in Information Retrieval. Logical Jumps in the Expression of Information," *Information Storage and Retrieval* 9 (no. 2, February 1973): 65-77.

Contributors

Josefa B. Abrera, an Indiana University, Graduate Library School, PhD ("Bibliographic and Information Control of a Small/Medium Public Library"), has been teaching at the University of Hawaii since 1970. Her areas of teaching include cataloging, classification, management and library automation. She is active in the American Library Association, particularly in the Resources and Technical Services Division, and also in the American Society for Information Science. She brings to this issue a sound understanding of the scope and nature of traditional classification from both the theoretical and practical views.

Hilda Feinberg is presently Research Librarian at Revlon in New York City. She has taught at Queens College and Pratt Institute, and in 1972 received her PhD from Columbia ("An Analysis of Automatic Derivative Indexing").

Zandra Moberg is presently a part-time student in the Graduate School of Library Science, Drexel University. She has some background in linguistics and has done independent study in the area of automatic classification. She has shown an ability to take an area which is difficult for the traditional classifier to understand and translate it into the comprehensible.

Ann F. Painter is presently Professor at the Graduate School of Library Science at Drexel University. She has worked in classification research and application for several years, in addition to her teaching responsibilities. She is active in both the American Library Association and the American Society for Information Science in classification research and interest groups and has served on two American National Standaros Institute Z39 committees concerned with classification-indexing standards.

She has published a book of readings, several articles and some specialized classification schemes.

Phyllis A. Richmond is another figure well-known in library classification circles. She is one of the few people in the United States who is really "into" classification research. She has taught at both Syracuse University and (presently) Case Western Reserve University in library education. She has worked diligently in the American Library Association and the American Society for Information Science with those most interested in classification research, as well as on an international level with conferences on the subject. Her publications are numerous and well worth a special literature search and review.

John H. Schneider has been active in classification research for a number of years. He was a founding member of the Special Interest Group/Classification Research of the American Society for Information Science. He has participated in their programs both locally (Washington, D.C.) and nationally. As his paper indicates, he plans to take part in the Third International Conference on Classification Research in India. He is no stranger to the principles and applications of modern classifications.

Harris Shupak is presently the librarian at Camil Associates, Inc., in Philadelphia. His interest in classification developed last year as a student at the Graduate School of Library Science, Drexel University. Several of his explorations into both the theory and the practice have been stimulating and thought-provoking. His essay here offers the same promise, it sets the pattern for the issue.

Gordon Stevenson has been teaching at the State University of New York at Albany since 1970. He received his PhD in Library Science from the Graduate Library School, Indiana University, Bloomington, Indiana, his topic, "The Classified Catalogs of German University Libraries.' He has written articles for several periodicals, including *Library Quarterly,* and is professionally active in the American Library Association on the Catalog Code Revision Committee of the Resources and Technical Services Division.

Maurice F. Tauber, presently Melvil Dewey Professor of Library Service, Columbia University, cannot be introduced in a few lines. His name has become synonymous with cataloging, classification and technical services. Teacher, researcher, consultant and author in and on classifi:ation for a number of years, he has provided the profession with stimulus and expertise in most of the states of the United States and abroad. In this era of change from DDC to LC, he and his collaborator, Hilda Feinberg, bring a sound summary of the problems and issues of the struggles between the two classifications.

Hans Wellisch is a visiting lecturer at the University of Maryland. He has been involved in classification research for some time and like many non-U.S. educated librarians has a great appreciation of the Universal Decimal Classification. In 1971, he coordinated the International Symposium on Subject Retrieval in the Seventies and later coedited the proceedings.