By JOHN ALBERT SANFORD and FREDERICK R. THERIAULT

Problems in the Application of Uniterm Coordinate Indexing

Dr. Sanford is deputy chief, Technical Library Division, National Security Agency, and Dr. Theriault is chief of the documentation branch.

THE LIBRARY of the National Security Agency has completed the organizational and experimental work necessary for the creation of a large-scale uniterm coordinate index. Production is now on a routine basis. Over 70,000 documents have been cataloged. This report is written at this time to make our experience available to other librarians who may be considering the use of this system.

We wish we could answer all the questions that have been raised about coordinate indexing in the literature. Many earnest librarians with very considerable professional experience have been deeply troubled by its potential pitfalls. Perhaps we have been very lucky. Perhaps the pitfalls will vanish in any other large-scale test. We do not know. We can report only that our system works. We do not know of any other means to gain such tight control of large masses of documents so economically and rapidly.

There have been problems, however; and some of them were formidable. Our version of the uniterm system of coordinate indexing is certainly not the last word in desirable development. It contains some of the whimsical invention and much of the rough-and-ready crudities of Henry Ford's Model "T" automobile. But, like the Model "T," it runs. As will be apparent in the account that follows, we have had to introduce many adaptations and changes of the system as originally outlined in the literature. Since we were pioneering, matching our wits against the new system day by day has been challenging. We hope that in solving our own problems we have made a contribution to the developing science of documentation.

For any reader unfamiliar with the uniterm system of coordinate indexing, the scheme is conceived as follows: The ideas presented in the title of a document, *plus additional ideas embodied in the text* if the title is not sufficiently descriptive, are broken up into separate words, dubbed "uniterms." The document is assigned an arbitrary number. A 5″ x 8″ master index card is prepared for each uniterm, and the number assigned to a document is registered on all of the uniterm cards that describe that document. Thus, the uniterm card bearing the heading CATALOG will have inscribed on it the document number of every report having anything of moment to do with catalogs or cataloging. A document that is a catalog of spare parts for automobile windshield wipers will have its number recorded on each of the following cards: CATALOG, AUTOMOBILE, WINDSHIELD, WIPER, PARTS. To find this document, the above cards are compared. Wherever the identical number appears on two or more cards, that number represents a document wherein the ideas intersect, i.e., coordinate.

We thought our first problem was the creation of a list of these uniterms for our subject matter. Here, however, our experience recommends three immedi-

ate departures from the system as proposed by Documentation, Inc.

1. Let the documents themselves generate their own uniterms. Catalog 1,000 documents. They will produce about 1,000 uniterms. Weed this list carefully, combining synonyms. With this core, catalog another 1,000 reports, using the "approved" basic list where possible, and then repeat the weeding operation. Once some 8,000 uniterms have been chosen, the rate of addition falls off very rapidly, even in highly varied subject matter. The curve begins to grow nearly flat when between 40,000 and 50,000 documents have been cataloged. The useful limits of a uniterm vocabulary are so soon reached that above 10,000 terms only highly specific items, such as new trade names and equipment designations, need to be added.

2. Forget all about "free" and "bound" terms as set forth in the literature of the subject. "Bound" terms almost inevitably free themselves sooner or later, and the intermediate step serves only to make extra work. Multiple words, however, should be used for exact description of concepts, wherever the idea expressed is a unit.

3. From the start, use *see* and *see also* references on the headings of uniterm cards, in accordance with standard library practice. We have found no other satisfactory solution for problems of near synonyms, for synonyms-in-some-meanings of words, and for all the other perplexities born of the fact that uniterm coordinate indexing uses the living fabric of language for its base.

Our next problem was the discovery that we needed to develop satellite indexes around our coordinate index. Here the needs of libraries will no doubt differ, but we soon found that in the coordinate index we were building a heavy-duty machine unsuitable for light work. We decided to employ traditional library methods for all types of document reference questions they served best, and to turn to the coordinate index where traditional methods broke down when laden with the peculiar burdens which documents engender. The combination of old and new methods has turned out to be an unexpectedly harmonious arrangement.

Problems of work flow came next. Our system as it finally evolved represents the solution of a series of problems in practical operation and hence may be of interest. Because the approach lends itself so readily to rapid processing, attention paid to "time and motion" pays large dividends in total production. Our basic requirements perhaps differ little from those of a great many libraries:

A collection of at least 200,000 technical reports and other documents needed improved information control. They were scattered throughout the organization in several dozen informal collections. A good many individual office files also needed to be surveyed. Each collection, small and large, had been organized according to a scheme chosen by its compilers. No professionally built catalogs existed. Large numbers of duplicate copies of reports were known to be wasting file space among various collections. The total number of reports to be processed probably approached one million.

The task was to weed out the duplicates, select items from the remaining originals which were worth keeping, and to create an index for them without assembling a definitive central collection. A central index was desired but a central file could not be contemplated: among the wealth of duplicates, too many items were unique and were required for frequent reference use in the departments then holding them.

Our organization is built on four teams of three members each, with a support group of twelve people located at the central cataloging department. Three teams operate in the "field," visit-

ing any desired component of the agency's organization and cataloging its documents. The fourth, a "home" team, operates in the central department in association with the support group, and is responsible solely for cataloging newly issued reports. Each team has a "leader" who is responsible for its entire, independent operation, including public relations with the people whose files he is organizing. He is also the uniterm cataloger for his team. He is assisted by a descriptive cataloger and by a clerk.

At the beginning of the operation performed on any file drawer of documents, the clerk of the team copies only the titles in informal lists. Once daily he returns to check these against the authority title file in the central catalog department. Duplicates of documents already cataloged are noted. Upon returning to his team, the clerk rubber-stamps these items "Duplicate Copy." Henceforward, these may be destroyed with confidence when no longer needed locally. The remainder are stamped "Record Copy" with a space provided for registering a permanent index number. These originals provide the team with the material for the ensuing day's work.

Because desk space is usually limited in the office being visited, each team is restricted to one typewriter, normally operated by the descriptive cataloger. The descriptive cataloging is performed directly on fanfolds. Because of the total needs of the system, the process is simple and swift. We record (a) title, (b) corporate author, (c) personal author, (d) series number, (e) contract number, if any, (f) collation. No tracings, subject headings, or other time-consuming notations are required. They are cared for, using simple short cuts, elsewhere in the system.

Document and fanfold are then passed to the team leader, who verifies the accuracy of the descriptive cataloging. He scans, studies, or dissects the document as its importance or difficulty seems to require, and writes out in longhand in a space provided on the fanfold all the uniterms he believes the document requires for indexing "in depth." This means that he attempts to record *all* of the subjects concerning which this document could conceivably be useful. Always, if the document concerns some subordinate topic—a part of a larger machine, a step in a process—the next larger concept is set down as the first uniterm. Then come all the words that answer the classic reporter's questions: "Who?" "What?" "Why?" "When?" "Where?" "How?" Then, uniterms to cover any ideas given special treatment in the document or which are important in the conclusions. The team leader is not afraid to scribble out a long list. He knows (a) that the ensuing processes in indexing these terms into the system are so economical of time that it is desirable to err on the side of over-completeness, and (b) that on any week's work his lists of uniterms will average nine terms per document.

Having finished his list, he examines it critically. The best test we have found for good uniterming is this: Do the words, read consecutively, come close to forming a complete and intelligible sentence? If so, no essential has probably been omitted. Next question: Do the terms cover all the ideas for which this document could be wanted? Here, of course, the human factor enters heavily —the cataloger's knowledge, background, and plain brain power. We know of no other system, however, where overly-careful and too-detailed indexing can be so cheerfully applauded by top management. It is certainly true that perfectly satisfactory indexing can be performed by catalogers with much inferior technical subject background than is required in any taxonomic system of classification.

The team leader's final chore is to assign a permanent accession number to the document and to record it on both

document and fanfold. He chooses this number from a block of "open" numbers currently assigned for his use. When the document is refiled by his clerk, his part of the operation is completed.

The routine in the central office employs copies of the fanfolds for various needs. The original and one carbon go first to the uniterm control officer, who must approve all new terms, adjust cross references, and eliminate useless synonyms. The original then goes to the clerk who types the Multilith stencils, and finally to the desk where the biweekly document accessions list is prepared. The carbon is routed to the posting clerks. The second carbon is filed immediately in the title authority file; the third in the accession-number file until replaced by the permanent printed card.

For economy, stencils are cut with a micro-elite typewriter on commercially available Multilith mats of narrow gauge having perforated sprocket edges which prevent slipping, since the typewriter is equipped (at very small cost) with sprocket guides above the platen. When these sprocket edges are torn off along the perforation, the stencil is the correct width to print 3″ x 5″ cards on long sheets. The press will accommodate two of these masters side by side so that press time is reduced to half. The stencils are pre-printed, again for economy, with whatever legends are standard for this library's cards. The finished sheets of printed cards, being completely uniform in register, can be machine cut, ready for filing. Satellite files are maintained by title, corporate author, personal author, series number, and contract number.

The "posting" operation, as the process of recording document numbers on uniterm cards is called, caused real trouble. Here lay the most formidable problem we encountered in the application of coordinate indexing. The process seems simple enough, but once it is begun difficulties multiply. Each card must be pulled, recorded upon, and refiled. The work is boring and fatiguing. Errors are easy to make and difficult to detect. Workers get in each other's way. While posting is going on, any reference use of the index usually means that one or the other operation must stop. Posting was hopelessly slow in relation to the smoothness and speed displayed in all other steps. It is not an exaggeration to say that this bottleneck threatened the collapse of our entire system.

The solution proved to be a simple one. We installed an IBM punch in the catalog department and equipped it with two standard "programs." A document number punched (and verified) on the first card is automatically reproduced on all ensuing cards until the operator wishes to change it. In changing from one document number to the next higher one, the operator touches only the final digit keys. One typist working two hours a day can keep up with the punching from all fanfolds generated by all four teams working at full production. At the end of each week the accumulated IBM cards are dropped into a machine sorter.

Now the posting operation is a quite different matter. Our coordinate index is housed in the familiar library "Kardex" type file. Beginning with the first one required, the poster withdraws one tray at a time, disturbing reference workers and other posters no more than does any other catalog department worker when she removes a drawer from the main card catalog in any library to file a new card. On the uniterm card for AUTOMOBILE this clerk posts the number for the document on windshield wipers, *and all others concerned with automobiles that the library has cataloged that week.* The IBM sorting machine has even placed all the automobile entries in correct numerical order. The posting operation is swift and highly accurate.

We had our fingers crossed concerning the reaction of our library users toward

the coordinate index, but we soon discovered that our misgivings were groundless. Unless this agency's employees are miraculously different from the general public, no one else needs to worry either. It is true, however, that the most enthusiastic response has been from our engineers and others with training in some academic discipline. Use of the system numbers several hundred questions each week.

On the premise that our customers could not care less whether the answer to their reference question came from a book, a serial, or a document, we placed the coordinate index and its satellite catalogs right beside the library card catalog. All who will may use them. Habitual library customers almost invariably prefer to consult the coordinate index unaided after their first five-minute indoctrination course in the system. Reference personnel are available, of course, to help any newcomer, or anyone else with a problem. We think it is sound public relations to offer to help everyone. Everyone, including the reference staff, is taught to think of the coordinate index as his heavy artillery. Where author, title, or serial number of a document is known, the satellite catalogs provide quick reference.

Much has been printed speculating on the amount of "noise" a large installation of coordinate indexing would produce; that is, the number of false coordinations of the man-bites-dog variety which would interfere with effective research work. The gloomy predictions have not been borne out by our year of operational use of the system. Now, it may well be that there are subject fields in which "false hits" would embarrass the reference librarian. We can only report that, in our library, a little care and forethought in the catalog department has kept the number of false coordinations so small, in our subject matter, that the annoyance is negligible. We have found that: (a) the more specific the subject field we are cataloging the tighter is the information control gained; (b) the more specific the uniterming the fewer are the false hits created; (c) skillful uniterming is a logical fractionating process, not a mere slicing of a document's title into separate words —this is especially true in the exact sciences; (d) wherever the man-bites-dog difficulty can be foreseen by a cataloger, the addition of a simple delta sign after the index word will signal the reference user which one is the correct reading, e.g., GERMANYΔ—IMPORTS—FRANCE means *only* German imports from France, not French imports from Germany. We have used these "delta flags" freely for words which cause us trouble. Their total number, however, has remained small.

There are, no doubt, more problems which we shall encounter as time goes on, but these are all the difficulties which we have met so far. A potential problem, that of an unwieldy pile up of numbers on "popular" uniterms, was solved just as it arose with us by a timely paper from the Office of Basic Instrumentation of the U. S. National Bureau of Standards.[1] On the question of "browsability" of the system we refer the reader to the excellent discussion of "browsability and suggestability" in the same paper. In their conclusions we heartily concur.

We have discovered no completely valid method to test the reliability, or the percentage of completeness of retrieval of information, of our index. We can testify that to date it has never failed to produce any "known" document. The expressions of pleasure we receive concerning the quality of our reference service leads us to conclude that the percentage of retrieval is high, perhaps even very high.

[1] William Wildhack, and others, "Documentation in Instrumentation," *American Documentation*, V (1954), 223-37. This article contains a useful bibliography on documentation experiments reported abroad. We employ standard uniterm cards for all except the most heavily used terms.