Automated Techniques in Comprehensive Indexing

By SISTER M. CLAUDIA, I.H.M.

Marygrove College

IN the words of its preface, the *New Catholic Encyclopedia,* being published by McGraw-Hill in 1967, "proposes to meet the need for an authoritative work of reference for the English-speaking world. It is not a revision of the *Catholic Encyclopedia* of 1907–1914, but a completely new work, abreast of the present state of knowledge and reflecting the outlook and interests of the second half of the twentieth century."

"In addition to providing full information on the doctrine, organization, and history of the Catholic Church through the close of Vatican II, the *Encyclopedia* includes within its scope the persons, institutions, religions, philosophies, scientific developments, and movements that have affected Catholicism in the past or are of particular concern at present."

There is a total of about 17,000 separate articles in the *Encyclopedia,* written by some 4,800 scholars, each qualified in his field. Every article appears over the name of its author. The contributors are men and women, Catholic and non-Catholic, from all parts of the world.

The *Index* volume covers the approximately 15 million words of the *Encyclopedia,* more than 7,000 functional illustrations, and about 300 maps. It also includes the sigla, or list of abbreviations, used in the bibliographies and the complete list of contributors with full identification.

Preparation for the index began in the fall of 1963. The editor spent the first weeks visiting the editorial offices of the standard encyclopedias to see what methods were being used in editing encyclopedia indexes. These included the Britannica, Compton, and World Book in Chicago, and Collier's, Americana, and Grolier in New York. The tour also gave an opportunity for interviews with John Rothman of the *New York Times,* Robert Kingery of the

The author is Librarian of Marygrove College, Detroit, Mich., and the 1965–67 president of the Catholic Library Association. From 1963 to 1966 she was on leave from Marygrove to serve as index editor for the *New Catholic Encyclopedia* and worked in Washington, D.C. Sister Claudia bases her paper on one presented—with the benefit of visual aids—before the Society of American Archivists in Atlanta, Ga., on Oct. 7, 1966.

New York Public Library, Edwin Colburn of the H. W. Wilson Co., and Giuseppe Sergio Martini of the Library of the United Nations. The initial objective of the interviews was to obtain an answer to the vital question: Was any form of automation being used for indexing? The answer in every instance was the same: automation was not yet being used in any form, but there was a possibility that it would be in the near future.

Concurrently with the interviews the staff had been conferring with the Executive Committee of the *New Catholic Encyclopedia* to determine what kind of index they had in mind. After a presentation of various types, the committee agreed that the index should be similar to that of the old *Catholic Encyclopedia* but improved in format. It was also decided that it would omit the cataloging of items under general headings since such listings, while helpful, are not strictly index material.

The McGraw-Hill time schedule called for indexing from page proof within an 8-month period, the first volume and a half of page proof to be delivered to the index department in May 1965 and the last in December of the same year. It was the staff's considered opinion that they could not compile the kind of index indicated within the time schedule set by the publisher unless they used some form of mechanized aid and started indexing from galleys instead of waiting for page proof. After extensive reading, study, and consultation, the editor and her staff recommended, therefore, that the NCE enter into a contract with Documentation Inc. (DOC INC) of Bethesda, Md., for services in aid of the compilation of the index. The proposed contract would make it possible for the indexers to use printer's galleys instead of page proofs; the correlation of galley and page proof references would be accomplished by a computer program based on a series of simple conversion processes and additional input.

The presentation of the plan to the NCE Executive Committee brought enthusiastic support from the majority of the members, but it took 8 months to clear the way for a contract with DOC INC. By that time the pilot study that had at first been suggested was no longer necessary; DOC INC, in investigating the feasibility of the proposal, had really made one. The contract was signed on July 31, 1964, and work began in August of the same year. DOC INC agreed to share the cost of programing, systems design, and project management.

While waiting for the contract to be negotiated, the index staff tried to anticipate work that could be done ahead of time. A process file of article titles to be included in the encyclopedia was typed

up in duplicate. The indexers selected all the slips for biographies (of which there were approximately 8,000), took them to the Library of Congress day by day for a period of several months, and checked them in the official catalog to establish entry forms and to note cross-references traced. This undertaking more than paid off later when many problems arose because of the NCE's numerous medieval names.

The second task was to draw up a filing policy, based in the main on the Library of Congress filing rules. It was not obvious at the time, but this was probably the best first job the staff could have done because the "sort" was basic to the computer program designed for the project. Many additional aids were compiled at this time: an alphabetical list of popes with dates (most lists are chronological); a list of all bishops and archbishops, past and present, of the United States; patterns for topics that would probably run into multiple headings; and lists of forms to avoid.

At this time, too, the staff indexed about 200,000 words to try to find out what the NCE's special problems would be, and they compiled a tentative style manual based on this experience. The manual was modified three times in the course of the project in order to take care of new problems encountered in seeing the work through. All editors' corrected galleys were processed through the index department so that any changes could be incorporated on the index copy. Much could be said of the training of the index team—but that is another paper.

Once the contract had been signed the indexers worked closely with personnel from DOC INC. Coding was the first step in the preparation of copy for keypunching. The code worked out made use of plus signs, slashes, and brackets. The plus signs enclosed numbers to insure proper sort; the slashes indicated material to be disregarded in arranging entries; and the brackets signaled reference to the "lookup" table, which indicated how abbreviations or numbers should be filed. All coding was automatically removed before final printing. A checklist of diacritical marks used in the encyclopedia also resulted in a table of word marks to be used in indicating by code the marks to appear in the final printing.

The Master Title List (MTL) of articles to be included in the encyclopedia was numbered by machine and coded for entries that seemed likely to give trouble. The diacritical marks were highlighted by a superimposed red check; the St.'s, the Mc's and the Mt.'s were bracketed; and the parts to be disregarded in medieval names and those of sovereigns were slashed. The marked copy was then xeroxed and sent to DOC INC for keypunching. This xeroxed

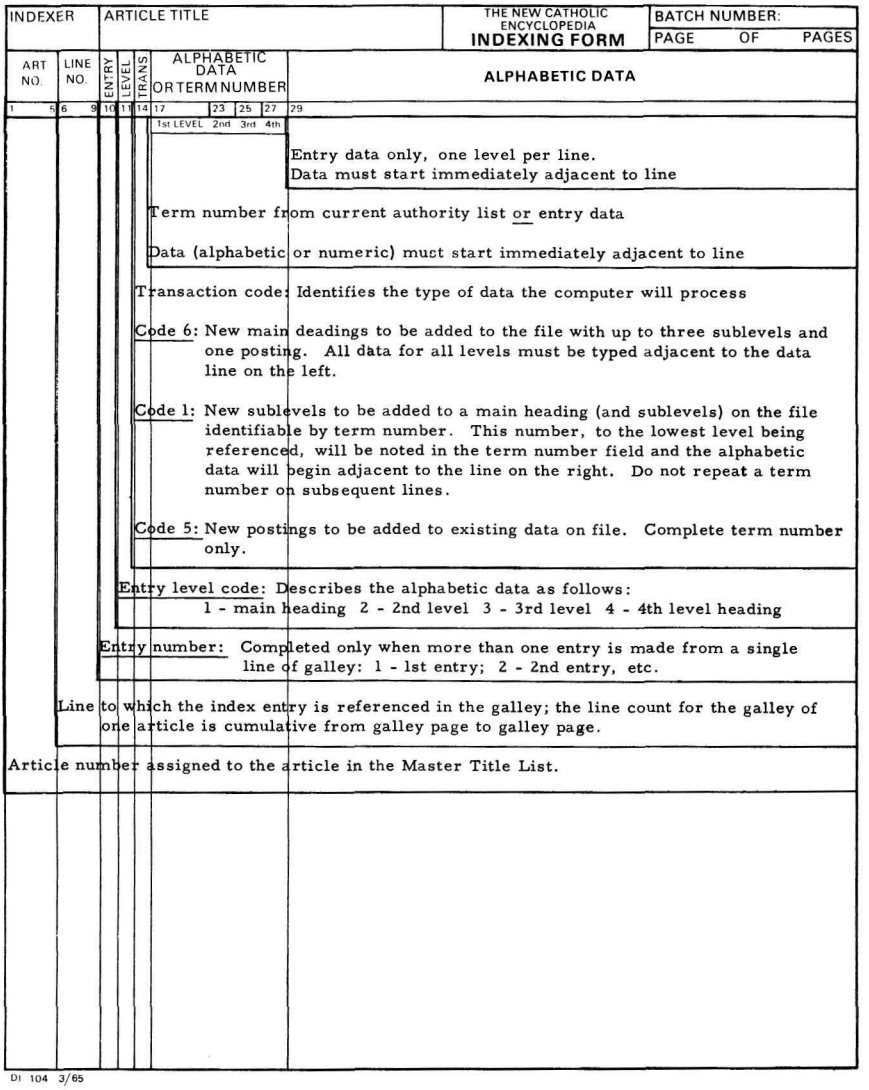
copy insured that there would be no variance in marking between the official copy and that sent to the computer firm.

The first print-out (unsorted) was returned on June 9, 1965. Once the print-out of the MTL had been corrected and approved, schedules were set to start shipping index copy. Since the MTL headings were key-punched but had not yet gone on the computer, they had no sequence numbers. A double column was therefore set up so that the indexers could post to the MTL entries but hold the key-punching until term numbers had been assigned. The forms used were specially designed for the NCE and indicated the indentions to be used. The typing looked rather complicated, but actually it was simple once the typists had mastered the general form. Typing on sheets is undoubtedly simpler and easier to handle than typing on cards. The NCE "indexing form" is shown on the facing page.

Revised schedules and indexing goals were set up along the way with every effort being made to increase speed because of the sheer quantity of material to be handled. From the very first, quantity was the one great problem. No matter what step of the project was underway, it always took just so long to accomplish the job: there were always 17,000 process slips to handle, always 17,000 galleys to process, 17,000 indexed galleys to edit, and 17,000 indexed galleys to type, proof, and prepare for the computers. Then there were the numbers. With a corps of 7 indexers, copy readers, and proofreaders increasing to a total of 24 people handling copy and using numbers to represent topics, errors in transcription were a great danger. Actually, as was later proved, the numbers were no problem, but the spelling sometimes was. All were warned to write term numbers in units of three and to read them back immediately in units of two. Error is almost impossible if this ruling is adhered to.

By original agreement, key verification of copy had been eliminated because it would have doubled the cost. In the first shipments the record was only one-half of one percent keypunch error, but as quantity increased the errors increased and it soon became necessary to ask for transaction sheets for every shipment of copy. This permitted errors to be corrected before they reached the electronic tape, but it also increased editing and proofing time. Here is an example of a transaction sheet:

659550064 1 6 Kliment SmolΔiatich, <see> Klim SmolΔiatich
087410000 1 6 KNIGHTS of [St.] GEORGE
087420000 1 6 KNIGHTS of [St.] JAMES
087430000 1 6 KNIGHTS of [St.] JOHN



NCE INDEXING FORM

(The obvious error in the spelling of "headings" in the explanation of Code 6 was not detected before this reproduction was prepared.—ED.)

087580000 1 6 KNOW/–/NOTHINGISM
139020024 1 6 KNOW/–/NOTHINGISM
139020024 2 6 [St.] Francis College (Loretto, Pa.)
139450073 1 6 KNOW/–/NOTHINGISM
139450073 2 6 [St.] Louis University Medical School
158860162 1 6 KNOW/–/NOTHINGISM
158860162 2 6 trusteeism
165530015 1 6 KNOW/–/NOTHINGISM
010310298 1 6 Korn+08+, Alejandro (critic)
024910604 1 6 Kuan Yin
024910604 2 6 Buddhism
154460040 1 6 Kuan Yin
154460040 2 6 temple dwelling
724910604 1 6 Kuan Yin
724910604 2 6 × Kwan-yin
824910519 1 6 Kuan Yin
824910519 2 6 <see also> Avalokit¹esvara
057740044 1 6 <¹Education+13+ sentimentale, L'> (Flaubert)
654350001 1 6 Educational television (ETV), <see> Television, Educationa
654350001 1 61 l
143870040 1 6 Edward+08+, Denis
143870040 2 6 Scranton University
733140142 1 6 Edwards+09+ Bello, Joaqu¹in (auth.)
733140142 2 6 × Bello, Joaqu¹in Edwards

Galley indexing was completed by July 15, 1966, but the indexing of the illustrations and maps still remained. Since these were to be entered by volume and page, they had to be indexed from page proof. This process, therefore, was dependent on the receipt of the page proof volumes. The last volume of the encyclopedia came to the index department on September 2, 1966; the last maps arrived on September 9; all copy was due at DOC INC by September 14. The deadline was made, but DOC INC was given more material than could be keypunched to meet the original schedule of September 23. Material was run on September 29 and 30; the final unconverted print-out including all indexing data was delivered on October 5.

Throughout the last months of indexing, conversion data were also being compiled. As the completed volumes of page proof came in, they were routed to staff members for the writing up of the conversion sheets. Article numbers were listed on forms giving volume, page, and line on which the article began, and identifying any illustrations by line location and length. Any ozalid changes

were also referred to this department for comparison with galleys and necessary correction.

On October 17 all corrected copy was returned to DOC INC for incorporation with the conversion run. The conversion process was a complete success. Index entries went on the computer and were incorporated with the material already on tape, the conversion program was run, and the print-out came through with all the article and line numbers converted to volume, page, and quadrant.

This copy—some 3,000 pages of a single column print-out—was edited as far as time permitted. Corrections were again sent in, the file was updated, and on November 19 the master tape and one copy of the print-out were sent to Computer Composition, Inc. (Flourtown, Pa.), for Linofilm reproduction.

Now that the indexing has been completed and the last print-out has been run, we are in a position to look back and evaluate the program. In terms of our original objectives, the delay in deciding on and getting the contract through, as well as the time for programing, reduced the anticipated 2-year period to 14 months for indexing. Nor did automation, in our experience, require fewer people for the job as had been suggested. While it is true that fewer typists and clerks were needed, other jobs were created that rquired just as much time. The transaction sheets, for example, and the activity and error lists had to be checked by people understanding the transaction numbers as well as the copy. Photon, which had originally been considered as a possibility for the final copy of the index, was found unsatisfactory because of the problem of hyphenization and was abandoned in favor of Linofilm though both eliminate proofreading of typeset material. The computer composition firm successfully programed for Linofilm reproduction, but while we did not read proof against copy we did find it necessary to read galleys for occasional mechanical failures and for coding errors that had escaped us in the final review.

Because of changes in programing, mainly to reduce the total cost, we cannot now convert the final index file to a machine vocabulary of indexing and retrieval terms suitable for computer manipulation for information storage and retrieval. Nor was it possible for the computer to break down the final copy of the print-out into a volume-page-quadrant sequence, ideal for a final check on indexing, without reprograming at an additional cost of $10,000. A bit too high, it seemed to us, just to make sure we were right.

On the credit side, we are definitely of the opinion that we could not have completed the indexing job for the *New Catholic Encyclopedia* without the aid of the computer, mainly because it gave us

the option of indexing from galley. Secondly, in spite of undetected errors there will at least be fewer than if the work had been done manually. Thirdly, the transfer from tape to type did eliminate the line-by-line proofing that would have been necessary in a typeset job.

Among other incidental advantages was the convenience, as well as the security, of having up to 24 copies of the print-out to work with—certainly a great improvement over a single card file. Copy, too, was much easier to edit when seen in an IBM print-out rather than on individual cards.

If the index were to be done over, so as to take advantage of the present advances in the field, optical scanning, for example, should replace keypunching—the greatest weakness and the greatest expense in the total process. This would not only insure greater accuracy; it would also make possible correction of parts of an entry rather than deletion and reentry of the whole item for even a minor error. Further, the schedule should permit more time for final editing.

Not enough can be said for the people with whom we worked at Documentation Inc. From the very first we found them making every effort to understand our problems and to do all they could to solve them in the best and most economical way. I should like to pay tribute especially to the late Mortimer Taube, for his deep personal interest and for the understanding which made the contract possible; also to Don Hummel, who directed the work and always found a way out when we ran into difficulty.

Speaking as a librarian, I would say that as a professional group we have been slow to see the possibilities in automation. We need to be better informed on the basic principles involved. My recommendation to anyone starting on a similar venture would be that a pilot project be run for all processes involved. This would give better opportunity for becoming familiar with the terminology, routines, and hazards of the process before quantity and pressure increase to the degree where the objective is almost lost sight of. Automated techniques have opened a wide field and great possibilities with plenty of room for creativity and imagination for those who have the courage and patience to investigate and experiment with them.