Dear Mr. Harrod,

I should be interested to know how other indexers cope with the method of indexing by paragraphs instead of by page numbers. I have been asked to do this and am complying with the request but find several difficulties.

The text in question is a correspondence course divided into a number of units each of which comprises, for instance:

*Unit 1*

- Introduction  pp. 4-12
- paragraphs I.1— 10.1 pp. 13-50
- Appendix I  p. 51
- Appendix II  p. 52-3
- Appendix III  p. 53
- Appendix IV  p. 54
- Appendix V  p. 56

Glossary and References not to be indexed.

As can be seen, there is no way of indexing the Introduction or the Appendices in both of which occur subjects which are indexed in the body of the book.

As there are sixteen Units, some of which are bound separately and some together, it is essential to have the Unit number clearly distinguished—this results in such entries as:

- Access 1 4.13; 10 2.6; 11 3.18; 12 App. B
- and survey skills 3 4.2, 5.6; 10 2.16

(Unit numbers in bold)

(This is copied from another index sent me as an example). In this book there are five appendices occupying pp. 59-83. Appendix B (above) takes four pages, and the only mention of access I could find in the text was in a subheading *Media Access Skills* under heading *Planning Skills*. Also in this index there is a long introductory note and I imagine there are quite a number of students (and others) who hardly glance at introductory notes. Also in the paragraphed sections there are some paragraphs which extend over four or more pages.

I should be interested in the comments of indexers who use this method.

Yours sincerely,

V. C. Findlay.

[Any letters addressed to Mrs. Findlay, c/o the Hon. Editor, will be forwarded.]

---

**Indexing the Great Soviet Encyclopedia**

**Emil Pocock**

Creating an annual cumulative index programme for the Great Soviet encyclopedia in translation presented several problems not usually encountered in indexing a multivolume encyclopedia. First, the GSE is being published serially; when work on the index was begun, only seven volumes had been published in the original Russian and only one in English. Second, the decision to issue annual interim cumulative indexes to follow publication of every group of five volumes, in addition to the final index, posed special problems of organization and cumulation. Time was also an important factor, since the interim indexes have to appear as soon as possible after the GSE volumes they cover. Finally, facility had to be made for editing and revision of early index volumes, for quick and foolproof sorting and cumulation of index entries, for cross-reference control, and for speedy and accurate composition. This article provides an overview of how such problems were solved, using both conventional techniques and computer assistance.

In the fall of 1973, I was hired as chief indexer for Macmillan's translation of the 30-volume *Great Soviet Encyclopedia*, the in-progress third edition of the Soviet Union's major national reference work. The page proofs of the English Volume 1 were all that was available to me, some 3,000 articles and 700,000 words. The project when completed in 1979 would be massive in both size and scope, 21 million words covering all knowledge from the Soviet point of view.

The requirement of indexing one volume at a time without knowledge of the full scope of the work at first appeared only a minor annoyance, then a handicap, and finally a major consideration in tackling the whole indexing project. The organization patterns of the Russian original, although they could be basically formulated within Volume 1, would, of course, emerge more fully as the translations proceeded; in addition, there was the certainty that the Soviet editors, like encyclopedia editors in the Western world, would refine and change their concepts as their monumental work went on.
A further problem was caused by the necessity of translating and publishing volume-for-volume right behind the Soviet publication. We could not wait until the entire Russian-language edition was completed and then realphabetize in English, because then, by the time we published, some of the early material would be nearly two decades old. Since we had to publish serially while translating from another alphabet, each of our volumes could potentially run in English from A to Z. Although access to the text was available through a complete list of articles printed in the front of each volume, it was important that a comprehensive index be prepared to re-create the cohesiveness and rational organization of the Russian original. This was doubly important because we could not follow the Russian system of internal cross-references, which depended upon a completely alphabetical encyclopedia.

It was decided by the editorial staff, in consultation with the advisory board and librarians, to create not only an index but a formidable index service—that is, a series of cumulative indexes, to be published after every five volumes of the text, with each index to include all entries to date. Thus, there would be five interim cumulative indexes, and a final, sixth, index covering 30 volumes, about 1,000 pages long and containing as many as half a million entries.

The only feasible way to carry out a plan of this scope would be to make use of some sort of computerized indexing system. To construct such a system I had to make the best use of my experience with computer indexing and try to utilize all the new technology possible to solve purely editorial problems.

I had recently completed work on the 750,000-entry New York Times Book Review Index (1896-1970) and had spent several additional months indexing on the New York Times Information Bank. What I had learned most—in addition to the basic construction and manipulation of computerized indexes—was the importance of flexibility and openness in any system. The system must not lock an index into organizational patterns and methods that cannot be easily changed at a later time. True, at a very early stage in the life of the index, we had to prepare its overall structure, but we knew that we would learn more about the organization of the encyclopedia and want to make changes even in the basic form of the index. Terms might be further defined and clarified. New articles with further insight into a subject might change our concept of a group of entries. There might even be surprises through clusters of unexpected articles or an unusual interpretation of some subject. Therefore we had to plan for a computer whose technical structure could be modified to meet evolving editorial needs.

The computer program solved rather routinely what at first might seem to be the most difficult problems: cumulation of entries from year to year in proper alphabetical order, putting volume and page numbers (citations) in proper sequence, and positioning cross-references. A separate computer program was written to produce final pages through video composition; the assigning of folios, running heads, and continuation lines and the elimination of widows* were all handled automatically. This program also took care of the routine and predictable changes of type face. For instance, although the main entry was always in boldface, this did not have to be indicated on the card which the indexer submitted; the computer knew that boldface would continue until either a citation or a left parenthesis was reached. It was also programmed to "know" that the volume number would be boldface, the page number lightface. The only type change that the indexer had to indicate was italics, for such entries as book titles. The edited cards therefore looked like this:

| Bakunin, M. A. (Russ. rev.), *2-560b |
| Bakunin, M. A. (Russ. rev.) Russian anarchism, 1-389b |
| Bakuninism (anarchist mov.), *2-561d |
| Bakuninism (anarchist mov.), 1-388d |

What emerged from the computer is this:

| Bakunin, M. A. (Russ. rev.), 2-560b |
| Bakuninism (anarchist mov.), 2-561d; 1-388d |

*An incomplete line of type at the top of a column or page.
We decided on the following technical structure for the index: (1) entries would have a maximum of three levels; (2) there would be see and see also cross-references; and (3) citations would include volume, page, and quadrant. A citation to the major article on a subject would come first in a string of citations, out of normal sequence if need be. All of this presented no difficulty to the programmer.

However, once the basics were established, we had to take into account exceptions to rules, contradictory instructions, and eliminated options. For example, there immediately had to be a modification in the rule that entries would be boldface until a citation or a left parenthesis was reached. Some entries were exceptions since they contained parenthetical characters that were part of the full term—for example, Russian Communist Party (Bolshevik), the full name of the party that preceded the Communist Party of the Soviet Union. We solved this problem by indicating these bold parentheses on the index card as square brackets; the computer would "understand" that the square brackets meant boldface parentheses and continue in boldface until it reached its next signal to change typeface. This, of course, meant that we could not use brackets as brackets anywhere in the index, but we were able to accept such a restriction.

Another modification involved a series of checks the programmer had instituted on our index card data. It had been stipulated that no entry would be accepted by the computer unless it had a valid citation or was a see or see also cross-reference. As a check this made a lot of sense, but it did prohibit the input of notes or informative lines, such as chronological by period. Therefore we modified the program so that the coded word "LINEX" was input in place of the citation, telling the computer to accept the entry even though it had no citation. The coded word would not print in the index.

With the programmer and indexers working together, our system finally evolved as follows: (1) Indexing is done in the traditional manner on index cards. The only codes marked on the cards by the indexers were an asterisk to indicate the main citation for an entry which must be placed first even if out of sequence and a special format necessary to sort exceptional terms (explained below). (2) Each card contains only one citation or one cross-reference and exists as a discrete index unit. (3) Typists input into the computer with a minimum of coding. (4) Each index card put into the computer is also stored as an individual record, identified by a unique record number. This number remains with the entry through all cumulations, facilitating later editing and changes in individual entries without affecting any other part of the index.

The computer sorts and compiles all of the entries according to headings and various levels of subheads, sequences, and citations. It also automatically positions cross-references in their proper places.

Alphabetization is done by the computer according to specific rules which are programmed in. This index is alphabetized by first word rather than letter by letter; the computer "understands" this concept. For each index level, the computer creates in its memory a parallel sorting key that consists of a modification of the original index term based on the alphabetization rules and specific instructions for handling diacriticial marks, punctuation, and other special characters. This sort key determines the position of the index term to which it is related. An index term as it prints and its parallel sort key would look like this:

Bogdanov-Bel'skii, N. P. (Russ. painter)
BOGDANOVBELSKII, N P+RUSS PAINTER

However, a sticky problem immediately arose. Following normal index procedure we had decided to sequence identical entries in the order of people, places, and things, and in the case of identical names in descending order according to rank and title. The computer cannot, of course, differentiate between people, places, and things, and being truly democratic pays no heed to rank or title. Our sorting problem here was especially complicated because the index entries would be input over a long period of time, interim cumulations would be published at intervals, and we had no way of knowing what names might be added at later dates.

We therefore adopted the following procedure. We created an "exception sort key" which would override the computer's strict rules by manual instruction. The best way to do this was at the original input stage, which would avoid expensive corrections and additions. Therefore we had to develop a systematic sorting scheme that could be applied to every single entry that might cause trouble and would work no matter how many similar terms were eventually included and would not cause trouble if no similar term ever arose.
The special sort scheme we invented was based on numbers assigned to people according to rank and title, and to places and things. In other words, we presorted names that were likely to cause problems. The ones we missed were caught during later editing cycles, but we didn't miss too many.

A group of cards, all entries for "Charles", we would want in the following order:

Charles, St.
Charles V (Holy Roman emp.)
Charles II (Eng. king)
Charles (Aus. archduke)
Charles (river, US)
Charles (Fr. frigate)

In order to accomplish this we assigned exception sort keys to the various Charleses on the cards; these keys were based on a numerical table that correlated a unique two-digit number to each Charles according to our formula. Thus, the sort key for St. Charles is based on CHARLES 02, the digits 02 being assigned to saints, who are high on the list. The sort key for the Holy Roman Emperor Charles is based on CHARLES24, since emperors are assigned the digits 24. Kings also have high-ranking numbers, followed by archdukes, places (the Charles River), and things (the French frigate). No matter how many entries called "Charles" are later added, they will always be sorted in proper relation to one another.

A still further refinement of this procedure involves sorting according to nationality (French kings precede Hungarian kings) and number (Pope John XXII will precede Pope John XXIII).

Another efficient device we are using depends on the computer's automatic ability to expand abbreviations. The computer has been programmed so that if @ precedes a certain abbreviation on the card, like USSR, it will check the term against an abbreviation table that has been programmed in and expand it to Union of Soviet Socialist Republics. If we wish to keep the abbreviation USSR, we omit the @ sign. Exceptional sorting instructions are also part of the abbreviation, so that many frequently used subheads under historical headings, which we want in chronological rather than alphabetical order, will automatically appear in proper sequence.

After the original input is printed out by the computer, it is edited and proof-read in much the same way any proofs are handled. Changes in individual entries, even major changes affecting large groups of entries, are relatively easy to make and can be implemented without affecting the structure of the index as a whole. A single citation is moved to a new heading by merely changing the heading itself; the computer will resort and recompile the changed entry, while the old heading with other citations remains unaffected. Subheads can be added to existing entries by a single instruction; again, the computer will automatically reposition the citation and resort according to the changes made. Entries may be deleted entirely from the index file or may be retained in the file but prevented from being printed in final pages until some later time. This latter feature is controlled by a simple code that can be applied or removed at any time. Opportunities for editing occur after each input cycle, roughly one cycle per volume of GSE text, and after each cumulation of all entries.

The complete system and procedures worked according to plan, from production of cards to production of final pages. Editors familiar only with traditional typesetting did go through a period of adjustment to the new methods of compiling and printing the index, especially when it came to making changes to our compiled listing. A line was not moved with an arrow, as in linotype, but "resorted"; citations next to variant headings were not moved to the correct index term, but the heading itself was corrected.

Through use, several refinements were suggested and have been incorporated. While most are of a technical nature, one interesting addition is the computer-generated cross-reference checklist. The computer is called upon to compile a separate list of cross-references in alphabetical order according to the term referred to. Instead of the cross-reference tracers on index cards (which would be impossible in our system) or a separate cross-reference card file, the computer compiles such a list each year. Control over the references is left to the indexers, who must check the list against the index headings, although the computer could be called upon to make the comparisons and print out any invalid crosses.

The computer system has solved the major problems of producing serially published cumulative indexes and has also allowed the inclusion of many other time-saving and convenient devices. Many of the tedious components of indexing, from the production of cards to composing pages, have been taken over by the computer. This system has left in the hands of the indexers and editors only those aspects of indexing that right now can be handled only by people,
while those functions that are tedious and routine are given to the computer. While the system does not incorporate the most sophisticated hardware and techniques available, such as full on-line indexing, it is appropriately advanced for the large size and long life of this major indexing project.

**EXTRACTS FROM REVIEWS**


'It is comprehensive in its scope, is well indexed, and the bibliographies accompanying each chapter cover literature up to 1970.'


'The book is spoiled to some extent by the complete absence of an index, a list of illustrations or a bibliography.'


'The reports are supported by a name index which includes even those mentioned in the references, author index, names and addresses of all participants at the Symposium. An extremely detailed subject index is also included which one considers unusual for a text which is not really a reference work but rather a collection of papers.'


'A major benefit of the book is the concise index, which enables the reader who is unable to read all of this enormous book to pinpoint his immediate areas of concern.'


'A particularly detailed index helps the manager to use the book as a source of reference.'


'This booklet is intended as a basic introduction to a range of activities covered by personnel management. The areas considered are personnel management as a specialist function, corporate planning, organization structure, manpower planning, manpower development, remuneration and employee relations and services. There is no index to the text but the contents page gives a detailed breakdown of the subjects dealt with under each of these headings.'


'There is also a useful index to the text.'


'The format is a spiral-type notebook printed in typescript with accessibility somewhat restricted due to no index and an insufficiently concise table of contents.'


'For a case that sets out to make legal history, the absence of an index or even chapter headlines is baffling in the extreme.'

*International directory of little magazines and small presses*, 10th ed. 1974-75. Editor/Publisher, Len Fulton (Paradise, Calif.: Dustbooks, $95).

'There is undoubtedly a librarian somewhere with a need to publish or perish, or a slightly mad bibliophile, who could provide a rudimentary subject index to the ten directories (Hartman: see Sadakichi Hartmann Newsletter). The American Society of Indexers should be receptive to such a goal. The present absence of an index in the Directory means that to find a newsletter about Nin, Anais, it is necessary to read from A-U until you encounter Under the Sign of Pisces: Anais Nin and Her Circle.'

---

*The Indexer* Vol. 9 No. 4 October 1975