foundation for the basic over-all system for production, storage, and access to information.

2. On-line capability, that is, to tie input devices directly into the computer, including the study of various modes of entering information into the system, whether they be manual keyboarding or optical scanning.

3. And finally, *man-machine capability*, whereby *all* information input to the system will become a part of the information system and can be revised, updated, or rearranged at any time during the computer handling cycle.

What has been described here is development actually being designed or tested by those concerned with the handling of primary information. These steps are illustrative of the type of detail which must be worked out and tested before implementation into the over-all system. It is much easier to write about than to do. The traditions associated with the journal system are firmly set, and it will take strong evidence of the value of any innovations before we can win approval for their testing, let alone adoption.

ACKNOWLEDGMENT

Much of the ACS work on computer-based systems is supported by the National Science Foundation.

LITERATURE CITED

(1) Kessler, M. M., "Some Very General Design Considerations," in "TIP System Report," Appendix H, Massachusetts Institute of Technology, Cambridge, Mass., Oct. 1967.

(2) Cairns, Robert W., "ACS Responsibilities for Communication," *Chem. Eng. News*, pp. 48–52 (Nov. 11, 1968).

(3) National Academy of Sciences–National Academy of Engineering, "Scientific and Technical Communication. A Pressing National Problem and Recommendations for its Solution." A report by the Committee on Scientific and Technical Communication (SATCOM), Washington, D. C., 1969.

(4) Brown, W. S., J. R. Pierce, and J. F. Traub, "The Future of Scientific Journals," *Science* **158**, 1153–1159 (1967).

(5) Swanson, Don R., "Scientific Journals and Information Services of the Future," *Am. Psychologist* **21**, 1005-1010 (1966).

(6) Kuney, J. H., B. G. Lazorchak, and S. W. Walcavich, "Computer-Aided Typesetting for the Journal of Chemical Documentation," J. CHEM. DOC. **6**, 1-2 (1966).

(7) *Ibid.*, "Computerized Typesetting of Complex Scientific Material," Proceedings, Fall Joint Computer Conference, Vol. **29**, Sparten Books, Washington, D. C., 1966.

(8) *CIS Newsletter*, Composition Information Services, Los Angeles, Calif., Oct. 15, 1968.

Subject Indexing as a By-Product of Electronic Composition*

E. R. LANNON
Consumer Protection and Environmental Health Service, Public Health Service, U. S.
Department of Health, Education, and Welfare, Washington, D. C. 20204

Received June 17, 1969

Electronic composition in the federal government has progressed rapidly since it was instituted in October 1967. During the first year of operation, it saved $900,000 in composition, printing, and binding costs, and reduced the bulk by 900 tons. Using electronic composition allows production of an index with reduced keystroking, a decreased time cycle of publication, a more meaningful index, and an established data base for additional processing. An algorithmic approach to indexing is now being developed.

Within the federal government, we talk of electronic composition as the output of composing systems using electronic or electro-mechanical components that produce typographic quality composition at speeds greater than 600 characters per second for eight-point characters. By "typographic quality" we refer to a resolution of not less than 800 lines per inch and proportionally spaced characters. We established the speed factor of 600 characters per second by determining the theoretic composing speed we could obtain from a state-of-the-art computer line printer through to a film output suitable for preparation of a printing plate. I am not aware of any operational line printer system that can achieve a theoretic speed of 600 characters per second through to film. However, it was possible to produce a camera system that would process hard copy output from a computer line printer to film at such speed. In actual operation, a line printer operating at 1100 lines per minute has an effective character rate in the range of 200 characters per second through to film. The drop-off is a function of the normal photographic process.

Indeed, it was this fact of life that was instrumental in bringing into being the first high-speed composing system as a part of the MEDLARS system of the National Library of Medicine in 1964.

Now, are we talking about the world we live in, or that better world computer salesmen are prone to say lies in the immediate future?

We are talking about a world that became real in October, 1967, with the installation of the electronic composing system of the Government Printing Office. This system has as its base the Linotron 1010. In terms of hardware and software, I believe this system is the most

*Presented before the Division of Chemical Literature, 157th Meeting, ACS, Minneapolis, Minn., April 14, 1969.

advanced in the world and, perhaps more significant, is economically, as well as technically, practicable. In its first year of operation, the system has produced the following savings:

1.	Reduced composition costs	$200,000
2.	Reduced printing and binding costs	$700,000
3.	Bulk shipping reduction	900 tons

All told, the electronic composing system has amortized approximately 50% of the government's capital investment of $2.2 million in the first year of operation. This was done with a utilization rate on the first machine of less than 50% of a single shift. As the first machine increases its utilization rate and the second machine, now completing its operational tests, is put into service, the amortization rate will increase.

Obviously then, from our point of view, large-scale electronic composing systems have an economic role to play in publication systems.

What perhaps is not so obvious are the possibilities that arise by having text in machine language when the costs of putting it into that language have been more than made up by savings in the actual printing process.

The master typography system developed for the Government Printing Office by the Mergenthaler Linotype Co. automates the composition of the entire page, including the generation of the page number. Accordingly, if the data to be indexed is flagged on input or flagged by stored algorithms, it is quite feasible to produce an index contemporaneously with page proof. There are many advantages to this approach. First, the keystroking required to compose the index is eliminated. Sample analysis indicates that for every four keystrokes used to flag an index term, an average of 30 keystrokes can be eliminated. In a typical textbook with 20 pages of index, we found that 7200 keystrokes for flags would produce 52,080 characters of data. Since in a typical data processing operation we would key-verify the text, we could postulate that 7200 keystrokes for index flags would eliminate a maximum of 104,160 keystrokes. On a cost basis, using typical government costs, the advantage for the text analyzed would be in the range of $84.00.

The second factor is perhaps more important, since it deals with decreasing the time cycle of publication. In today's publishing methods, the preparation of a subject index awaits the delivery of page proof to the author/editor, since only then does the author/editor know the precise page number of an index term. This time may vary considerably from 10 to 90 days. An average delay factor is about 45 days. When you consider that a textbook, particularly one in the sciences, may have a useful life of not more than two years because of obsolescence of data, any decrease in the publication cycle becomes of significance with respect to the national investment for replacement of textbooks and the currency of educational materials. The American public's investment in textbooks and related works susceptible to my thesis, according to the 1968 edition of the "Statistical Abstract of the U. S.," was $835,780,000 in 1963. Expenditure in this area increased from 1954 to 1963 by $602,041,000. It is quite safe to postulate, therefore, that expenditure in 1969 is in excess of $1,000,000,000 for publications requiring a subject index.

A third factor, which deserves much more careful analysis, deals with the usefulness of a conventional subject index. I have found from my own teaching experience that a typical undergraduate student looks in the text for the specific phrasing of the index term. The sophistication of the author can not be assumed to exist on the part of the student. The student may not know that the jargon "pump priming" may be more elegantly indexed as "deficit financing." Thus, it may well be that indexing to the specific wording of the text produces an index much more meaningful to the reader.

A fourth factor, which I must confess led me into this project in the first place, involves government publications not now indexed because of the pressure to print the subject matter without delay. Delays of up to 90 days, as can occur in the publication of textbooks, could not be countenanced. Let's take an example: the massive Congressional Record is printed overnight by the Government Printing Office. Since all federal agencies and much of the nation's industry have an acute interest in the proceedings of the Congress, the Record is required reading. To conserve the time of the readers, it is standard practice to compile a reference index of matters of specific interest to a department prior to distributing the Record within that department. According to a study undertaken by the Joint Committee on Printing of the U. S. Congress, at least $200,000 in indexers' time was being spent within the federal government for this one purpose, and the cost cited is probably underestimated by 25 to 50%.

Given the composition speed of the electronic composing system, it becomes practicable to consider converting the Congressional Record from monotype composition to electronic composition purely to produce a daily index, even if printing, binding, and keystroking costs remain constant.

In addition to the index, which would pay for the costs of electronic composition, a data base in machine language would be available to the Library of Congress and others interested in chronological and other summaries derived from the Record. The Canadian equivalent to our Congressional Record, known as Hansard, is already composed by computer by the Canadian Government Printing Bureau. Albeit there are significant differences between Hansard and the U. S. Congressional Record, it appears practicable to compose the U. S. Congressional Record by computer.

The following factors are our goals:

1. reducing the keystroking required to produce an index;
2. decreasing the time cycle of publication;
3. producing, perhaps, a more meaningful index;
4. establishing a data base in machine language for additional processing.

We don't have to go into the specifics of the master typography system for the purposes of this paper. It is enough to say that it is a complex set of routines using up to 140K bytes of core storage with a throughput of up to 2KC on a 360-50. I am convinced that the key to economic exploitation of high-speed composition systems lies with the computer, rather than the output device. The special purpose format processors used to drive the U. S. Air Force's Linotron 1010 have a throughput speed of up to 10KC.

The modifications introduced to the master typography system for indexing purposes are essentially a series of optional sub routines.

FEATURES OF THE MODIFICATIONS

Job Card. The first modification is the preparation of a job card, wherein the specific format to which the index is to be typeset is specified.

Function Code-Assignment Card. Within this card, the format designer designates the function code(s) to be used for the purpose of indexing. In our case, we are using two codes—i.e., the symbol for pound "#" and the lower case letter "j." Any unique character or combination of characters convenient to the format designer may, however, be used for the purpose of identifying an index term.

Detection of the Index Flag. When an index function code is detected by the precedence search routine of the master typography system, it will branch to the stored indexing routine. If the function code precedes a word group, the routine sets the index flag switch, stores the input address, and then returns to the main typography program, which proceeds to process the input data in normal fashion until the closing index function code is detected. At this point, a branch to the indexing routine again occurs. Since the index flag switch was set on the basis of the detection of the initial index function code, the following operations are performed by the indexing routine:

1. The index flag switch is reset.
2. The data appearing between the index function codes is stripped of any other function codes and the characters making up the data are translated into Extended Binary Coded Decimal Interchange Code (EBCDIC).
3. The EBCDIC characters are moved to an assigned field in a 2311 Disk Drive.
4. The data bracketed by the index function codes are moved to a second field in the assigned disk.
5. The page number generated by the master typography system that applies to the term is moved to a third field in the assigned disk.

The disk sequential feature of the disk operating system of the IBM 360 is being employed. This makes possible an alphabetic sequencing of the index terms as the job is being processed. Identical terms will be collated and the page numbers for each reference to the term will be written against the initial entry.

At this point, the master typography system resumes normal control and proceeds until another index function code is detected.

TEST OF PROGRAM

To test the hypothesis that production of a subject index as a by-product of the composition process is practical, we put it to the acid test.

A congressional hearing previously printed by the Government Printing Office was chosen for the trial publication. I personally read through the printed text and underlined the terms I believed should be in the index.

The text was then typed on an IBM MTST Model IV, using the coding conventions established for the Linotron 1010 and the index function codes previously discussed. The MTST cartridge was converted to 200 c.p.i.—seven-track tape on a Digi-Data Coupler. In essence, we then had text ready for input against the master typography system, with the index routines added without intermediate computer use to get to magnetic tape.

The tape was then run on the GPO's 360-50 system and tapes produced for input to the Linotron 1010.

The text tape was run against the master typography system three times, with the basic difference being the designated point sizes of the type. The purpose of doing so was to illustrate that the page reference numbers were automatically generated by the program.

A sample of the hard copy output of the MTST is shown in Figure 1. The figures coded for indexing are underlined. Figure 2 shows a sample page produced by Linotron 1010 in an 8 on 10 setting. Figure 3 shows the index produced by an 8 on 10 setting.

I believe we have demonstrated what we set out to prove, namely,

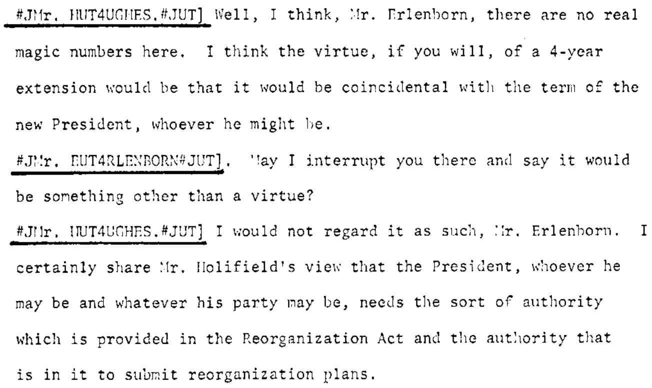
1. a subject index can be produced by computer as a by-product of input;
2. the time cycle of producing a subject index can be reduced; and
3. previously unindexed publications, such as the congressional hearing we used, can be indexed with no basic impediment to the timeliness of publication.

I should like now to discuss in very general terms, since I do not profess to be anything but an interested layman in this area, the algorithmic approach currently under development.

The work to be discussed will be done by James L. Dolby of San Jose State University and Howard Reznikoff of Rice University. Drs. Dolby and Reznikoff are quite experienced in the application of the digital computer to the composition process. The first real algorithm for the hyphenation of words was their product.

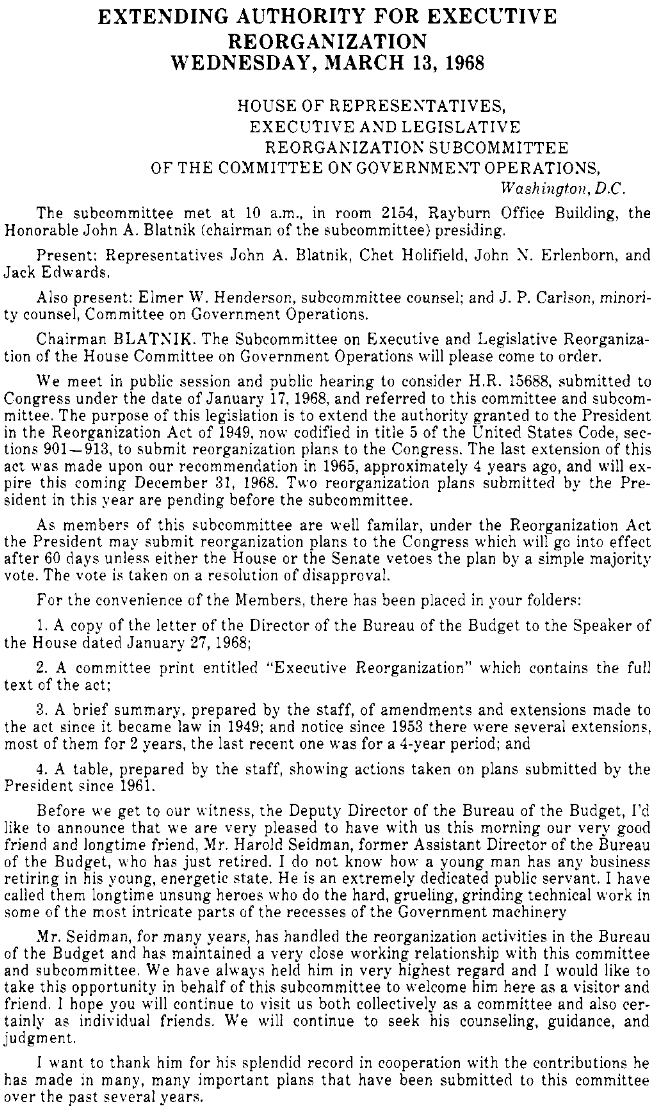
I asked Dr. Dolby, "Don't we have a new dimension in applying information storage and retrieval techniques to text when the text has been prepared for computer typesetting?"

I asked this question since most of the research done through 1967 concluded that subject indexing by computer was technically possible, but uneconomic because of the cost of converting text to machine language. I reasoned



```
#JMr. HUT4UGHES.#JUT] Well, I think, Mr. Erlenborn, there are no real
magic numbers here.  I think the virtue, if you will, of a 4-year
extension would be that it would be coincidental with the term of the
new President, whoever he might be.
#JMr. EUT4RLENBORN#JUT].  May I interrupt you there and say it would
be something other than a virtue?
#JMr. HUT4UGHES.#JUT] I would not regard it as such, Mr. Erlenborn.  I
certainly share Mr. Holifield's view that the President, whoever he
may be and whatever his party may be, needs the sort of authority
which is provided in the Reorganization Act and the authority that
is in it to submit reorganization plans.
```

Figure 1. Sample of hard copy output of MTST
Terms coded for indexing are underlined. "UT4" and "UT" are function codes for Linotron 1010.



EXTENDING AUTHORITY FOR EXECUTIVE
REORGANIZATION
WEDNESDAY, MARCH 13, 1968

HOUSE OF REPRESENTATIVES,
EXECUTIVE AND LEGISLATIVE
REORGANIZATION SUBCOMMITTEE
OF THE COMMITTEE ON GOVERNMENT OPERATIONS,
Washington, D.C.

The subcommittee met at 10 a.m., in room 2154, Rayburn Office Building, the Honorable John A. Blatnik (chairman of the subcommittee) presiding.

Present: Representatives John A. Blatnik, Chet Holifield, John N. Erlenborn, and Jack Edwards.

Also present: Elmer W. Henderson, subcommittee counsel; and J. P. Carlson, minority counsel, Committee on Government Operations.

Chairman BLATNIK. The Subcommittee on Executive and Legislative Reorganization of the House Committee on Government Operations will please come to order.

We meet in public session and public hearing to consider H.R. 15688, submitted to Congress under the date of January 17, 1968, and referred to this committee and subcommittee. The purpose of this legislation is to extend the authority granted to the President in the Reorganization Act of 1949, now codified in title 5 of the United States Code, sections 901—913, to submit reorganization plans to the Congress. The last extension of this act was made upon our recommendation in 1965, approximately 4 years ago, and will expire this coming December 31, 1968. Two reorganization plans submitted by the President in this year are pending before the subcommittee.

As members of this subcommittee are well familar, under the Reorganization Act the President may submit reorganization plans to the Congress which will go into effect after 60 days unless either the House or the Senate vetoes the plan by a simple majority vote. The vote is taken on a resolution of disapproval.

For the convenience of the Members, there has been placed in your folders:

1. A copy of the letter of the Director of the Bureau of the Budget to the Speaker of the House dated January 27, 1968;

2. A committee print entitled "Executive Reorganization" which contains the full text of the act;

3. A brief summary, prepared by the staff, of amendments and extensions made to the act since it became law in 1949; and notice since 1953 there were several extensions, most of them for 2 years, the last recent one was for a 4-year period; and

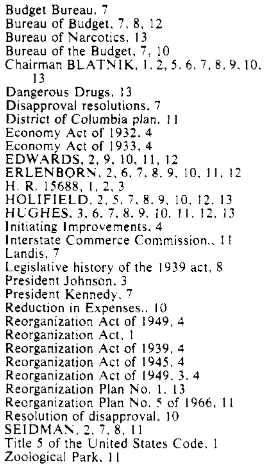
4. A table, prepared by the staff, showing actions taken on plans submitted by the President since 1961.

Before we get to our witness, the Deputy Director of the Bureau of the Budget, I'd like to announce that we are very pleased to have with us this morning our very good friend and longtime friend, Mr. Harold Seidman, former Assistant Director of the Bureau of the Budget, who has just retired. I do not know how a young man has any business retiring in his young, energetic state. He is an extremely dedicated public servant. I have called them longtime unsung heroes who do the hard, grueling, grinding technical work in some of the most intricate parts of the recesses of the Government machinery

Mr. Seidman, for many years, has handled the reorganization activities in the Bureau of the Budget and has maintained a very close working relationship with this committee and subcommittee. We have always held him in very highest regard and I would like to take this opportunity in behalf of this subcommittee to welcome him here as a visitor and friend. I hope you will continue to visit us both collectively as a committee and also certainly as individual friends. We will continue to seek his counseling, guidance, and judgment.

I want to thank him for his splendid record in cooperation with the contributions he has made in many, many important plans that have been submitted to this committee over the past several years.

Figure 2. Sample page produced by Linotron 1010 on 8 on 10 setting



Budget Bureau, 7
Bureau of Budget, 7, 8, 12
Bureau of Narcotics, 13
Bureau of the Budget, 7, 10
Chairman BLATNIK, 1, 2, 5, 6, 7, 8, 9, 10,
13
Dangerous Drugs, 13
Disapproval resolutions, 7
District of Columbia plan, 11
Economy Act of 1932, 4
Economy Act of 1933, 4
EDWARDS, 2, 9, 10, 11, 12
ERLENBORN, 2, 6, 7, 8, 9, 10, 11, 12
H. R. 15688, 1, 2, 3
HOLIFIELD, 2, 5, 7, 8, 9, 10, 12, 13
HUGHES, 3, 6, 7, 8, 9, 10, 11, 12, 13
Initiating Improvements, 4
Interstate Commerce Commission., 11
Landis, 7
Legislative history of the 1939 act, 8
President Johnson, 3
President Kennedy, 7
Reduction in Expenses., 10
Reorganization Act of 1949, 4
Reorganization Act, 1
Reorganization Act of 1939, 4
Reorganization Act of 1945, 4
Reorganization Act of 1949, 3, 4
Reorganization Plan No. 1, 13
Reorganization Plan No. 5 of 1966, 11
Resolution of disapproval, 10
SEIDMAN, 2, 7, 8, 11
Title 5 of the United States Code, 1
Zoological Park, 11

Figure 3. Index produced by 8 on 10 setting

that having machine language paid for in the publication process eliminated the economic barrier.

Dr. Dolby agreed with my reasoning and began to develop an algorithmic approach. We put his approach into the specific context of the Congressional Record.

For a publication such as the Congressional Record, the algorithmic approach will involve three primary rules. The first rule will be involved in identifying proper names. This is not new as an idea, but as applied to initial composition, it is a new application of an old idea.

The second rule is specific to the Congressional Record, because of the formats used for headings and captions. Much meaningful information is contained in these formats, and the rule must be capable of extracting these data.

The third rule is what is truly new in application. This involves a deep index of words and terms of high syntactical value. Essentially, this approach suppresses insignificant words and leaves in core those words or word combinations which by definition are significant.

The programmed algorithms would then insert the index function codes previously mentioned, and the indexing routine described earlier would proceed to compile the index.

I should stress the point that we have not yet produced the algorithms and do not know if indeed it can be done economically. We propose to find out.

If we are successful, we will have produced the index without the keying of index function codes and without the necessity of marking terms to be indexed in the manuscript. As well, if successful, we will have produced a more complete index, since any term that satisfies the stored rules will be extracted.

By simulation of the algorithms in a text devoted to statistics, Dr. Dolby found that even the admittedly rough and preliminary algorithm extracted essentially the same terms as had the author, and that 90% of the terms so produced were generally recognized index terms in the field. Ten per cent of the terms were not generally recognized as index terms, albeit even these in large measure were meaningful. The nonmeaningful extractions represent the work yet to be done.

Under a proposed contract to be considered by the Government Printing Office, the algorithm used experimentally will be intensively analyzed to determine the extent to which more sophisticated rules can be derived to skim off the most important entries from the deep index.

ACKNOWLEDGMENT

I should like to express my thanks to the following people who made possible this paper and the programs described: John F. Haley, Joint Committee on Printing, United States Congress; John J. Boyle, U. S. Government Printing Office; Charles Sobel, Donald Rollert, and James Leithauser, Mergenthaler Linotype Co.; James L. Dolby, San Jose State University.