Chemical Information—Where Is the Computer Taking Us? 1974 Symposium of the Chemists' Club Library. Introductory Remarks

BEN H. WEIL

Analytical and Information Division, Exxon Research and Engineering Company, Linden, New Jersey 07036

Received June 18, 1974

Computers have been used in chemical-information work for two decades. Their use has sometimes been accompanied by grandiose planning, utopian claims, and the rise and fall of systems and dynastic groups, but in the past decade they have found many sound applications in chemical libraries, publishing, and information alerting and retrieval. Some of these developments have been described in earlier Chemists' Club Library Symposia, but it never before appeared appropriate or necessary to devote a whole Symposium to interpreting their significance.

In the past year or so, however, it has become apparent that computers have reached levels of capacity, versatility, and low costs that have permitted many breakthroughs in information processing, storage, and retrieval, and that we are almost certainly on the threshold of even more significant developments. The papers which follow, from the 1974 Symposium of The Chemists' Club Library, were therefore planned to permit an expert review of where the computer is taking us in at least some of the major areas of importance to chemical information—conversion of primary information, alerting, information retrieval, and clarifying the role of the librarian.

The Use of the Computer in Converting Primary Information[†]

RITA G. LERNER

American Institute of Physics, New York, New York 10017

Received June 18, 1974

Computerized typesetting, in whole or in part, of primary journals is proving increasingly attractive when it also yields a computer-manipulatable product from which both printed indexes and computer-searchable alerting and retrieval tools can be obtained. The modern system of the American Institute of Physics, integrating primary and secondary services, is described to illustrate this point. Similar programs are under way at the American Chemical Society and elsewhere.

BACKGROUND

The American Institute of Physics is a not-for-profit federation of scientific societies in the fields of physics and astronomy; the Institute acts for its members in areas which promote the advancement and diffusion of the knowledge of physics. Its activities include publishing, educational and manpower studies, public relations, placement services, and a center for the history and philosophy of physics.

The principal activity of the AIP is its publishing program. It is the world's largest publisher of primary journals in a single discipline. The total number of pages published in primary journals and conference proceedings in 1973 by AIP and its member societies amounted to 110,000 pages. Table I shows the number of pages in the primary publication program of several of the major professional societies and commercial publishers in 1973.

The size of the publication program, which covers about 90% of the American physics literature and 25% of the world's physics literature, has placed the American Institute of Physics in a unique position to develop those services for individual physicists which have become increasingly necessary as the physics literature has expanded. A decision was therefore made to undertake the development of a system in which a single keyboarding effort could be made to serve the needs of both primary and secondary information services. With the encouragement and support of the National Science Foundation, such a system has been designed and implemented.[1-3] Since 1972, the "heads" of journal articles have been computer photocomposed for most of the AIP and society journals (the "head" consists of the title, author, author's location, and abstract for each article). Starting in January 1974, AIP has been computer photocomposing the complete texts of two journals. In addition, the following secondary services are either produced from the master tape or are related to the computerized services.

SPIN—a monthly magnetic tape service containing records for about 30,000 articles a year; the record includes abstracts from about 70 major physics journals.[4]

Current Physics Titles—a set of current awareness journals,

[†] Presented in the Chemist's Club Library Seminar, New York, N. Y., April 5, 1974.

issued monthly, in subsets for solid state, nuclei and particles, and atoms and waves. CPT is computer photocomposed from the monthly SPIN tape. The records on the SPIN tape are sorted and then passed through a program called WEED, which automatically consolidates headings in subject areas with a low density of articles to produce the tape from which CPT is composed. Abstracts are not included, but each entry contains the title, author and location, bibliographic information, free language key phrases, and Current Physics Microform reel and frame numbers. Within each journal, the items are arranged by category from the Physics and Astronomy Classification Scheme.[5]

Current Physics Advance Abstracts—a set of monthly abstract journals, issued in the same three subsets as Current Physics Titles. It contains articles accepted for publication in AIP journals, and appears two to four months ahead of publication of the primary journals. Authors index and type their own abstracts within a prescribed format. CPAA includes all AIP journals and many of the European Physical Society Journals.

Current Physics Microform—a monthly set of microfilms on reels or cartridges, containing the full text of articles published the preceding month in AIP journals. Reel and frame numbers are included on SPIN and in the Current Physics Titles journals.

Current Physics Reprints—reprints of full text of single articles from AIP journals. Same-day air mail service is supplied for articles already in print; articles announced in Current Physics Advance Abstracts are sent as soon as they are available in print.

Primary Journal Indexes—subject and author indexes for primary journals are generated from SPIN by a program which selects the subject and author fields and then produces sort keys. For the author indexes, the sort keys are the authors' names, arranged with surname first. For the subject indexes, the headings used are either from the Physics and Astronomy Classification Scheme or are chosen by the journal editor; the sort key is a set of numbers associated with the terms. Every change of sort key produces a display heading for the index.

Current Physics Index—a new computer photocomposed publication which will appear quarterly starting in 1974, containing abstracts of approximately 4,200 articles per quarter published in AIP journals, arranged by index terms. Each issue will have an author index. The annual cumulative index will omit abstracts, but will include complete titles and authors with the bibliographic citation in both the subject and author indexes.

Services to other publishers—AIP is providing computer photocomposed abstracts from AIP journals to Nuclear Science Abstracts for republication in NSA; it is also providing monthly tapes of the same material which is merged with other records to provide the United States input to the International Nuclear Information System (INIS). These articles are indexed and formatted by AIP to meet Nuclear Science Abstracts and INIS requirements.

Except for SPIN and CPM, all of these services are priced for individual members.

INPUT METHODS

Both AIP and non-AIP journals are keyboarded, with over 50% of the material coming from journals published by AIP and its member societies. The keyboarding is accomplished in three streams: one for full text of journal articles, one for all other AIP journals, and one for non-AIP journals.

The full text of two AIP-published journals, the *Physics of Fluids* and *Reviews of Modern Physics*, is completely computer photocomposed; this effort started in January 1974. The keyboarding and photocomposition are done by a contractor, who types the full text of the manuscripts, converts the text to magnetic tape *via* OCR, and processes the tape to produce camera-ready copy using Harris Intertype equipment.[6] AIP subsequently strips off this tape those data elements which are needed for its secondary services.

Table I. 1973 Page Counts[a]

Pergamon Press, Inc.	175,000
American Institute of Physics	110,000
Academic Press, Inc.	105,000
American Chemical Society	45,000
John Wiley & Sons, Inc.	28,400
Institute of Electrical & Electronic Engineers	27,000
McGraw-Hill, Inc.	24,000
American Mathematical Society	16,200
American Society of Civil Engineers	15,000

[a] AIP, ACS, and IEEE have $8\frac{1}{2}$ in. × 11 in. page size; other publishers use varying page sizes.

All of the other AIP-published journals are handled inhouse, with the keyboarded data for each record consisting of

Bibliographic citation
Title
Author
Author's location
Abstract
Indexing information
Microfilm reel and frame number for AIP journals

Keyboarding is done from the author's manuscript, in upper and lower case, using Datapoint 2200 processors. These processors have 8K memories, a standard typewriter keyboard, and a screen capable of displaying 12 lines of 80 characters each. The data are collected on cassettes capable of storing 1,000 lines of data or 10 to 20 "heads." Characters which do not appear on standard typewriter keyboards or line printers are handled by means of an escape key and codes indicating the name of the character; for example, the Greek letter α is keyboarded as Ga and appears on the SPIN tape as Greek-alpha.

Several of the Datapoints are physically located within the copy editing groups, to enable the copy editors to proofread and correct on-line. Corrected cassettes are batched daily onto a standard computer tape, which is processed through a check program to locate errors such as missing fields or invalid terms. The next morning, the tape is processed through a COBOL program[7] which eliminates the field tags, converts the special characters, and inserts the PAGE-1 codes and formats. The output of the COBOL program is processed to obtain a tape for the Videocomp photocomposer; this tape is then run through the Videocomp and hard copy is returned to the copy editors within 24 hours. A total of three to five days elapses between receipt of the manuscript from the journal editor to the production of a camera-ready "head," ready to be matched up with the text of the remainder of the article. In the meantime, indexers and copy editors have been working on the manuscripts, preparing them for typewriter composition or standard typesetting methods. After the remainder of the text has been composed, the "heads" are stripped in, and galleys or page proofs are run and sent to the author for approval. When these have been returned, the article is assigned to a particular journal issue and given page numbers. The page numbers, indexing, and author's alterations are entered onto the Datapoint cassettes as corrections. Articles whose records have been completed are removed from the cassettes monthly and placed on a master tape which is used to create SPIN and Current Physics Titles.

Non-AIP published journals are keyboarded either in-house or by an outside contractor, after receipt of either page proof or journal issues from the publisher.

COMPUTER PHOTOCOMPOSITION

Except for *Current Physics Advance Abstracts, Physics of Fluids,* and *Reviews of Modern Physics,* all of the printed products discussed above are composed on a Videocomp 800. The program which produces the tape for the Videocomp is called PAGE-1, and it exists in versions for both the Spectra 70 and IBM 360 computers.

The input to PAGE-1 is prepared by a COBOL program which selects the proper data elements for each publication, and then deletes the field tags and replaces them with instructions specifying the print format for each field; parameters such as type font and print size are specified at run time. The COBOL program also translates the representation of special characters on the master tape to the representation used by the photocomposer. The output tape from the COBOL program consists of the text to be set, with control words interspersed with composition commands, preceded by definitions and values of the parameters to be used.

Using commands inserted by the COBOL program, PAGE-1 will print page numbers, running heads, center a title on a page, and indent paragraphs. PAGE-1 takes the text strings from the tape, breaks them into lines, assembles the lines into columns of predetermined length and width, and assembles the columns into pages. Under options, it can justify the right-hand side of a column and can break or hyphenate words at the end of a line. The end product, after the tape is passed through PAGE-1, is camera-ready hard copy for the printer which is identical in style and quality with that produced by conventional typesetting.

The problems presented by computer photocomposition are in the representation of special characters and mathematical constructions, the locations of certain symbols such as superscripts, subscripts, accent marks and summations, and the correction of errors.

Physics journals use approximately 1500 special characters and symbols. The COBOL program takes those characters which are not available on a standard keyboard, and, using a table look-up, composes instructions for the appropriate printed graphic symbol. Usually, the translated instructions provide the font and address of a Videocomp character. Required characters which are not available on the Videocomp must be created. For example, the symbols for direct sum ⊕ and direct product ⊗ are made by setting a plus sign or multiplication sign, backspacing by the width of the characters, and setting a circle around them. Left and right arrows (← and →) are created by setting a less than or greater than sign in type that is half the point size of the type that is being set, followed or preceded by two em dashes. The perpendicular sign is created by combining an en dash with a vertical bar. Special constructions also include the use of diacritical marks with upper and lower case characters, numerals, Greek and German letters, subscripts and superscripts. Marks may appear above, below, or through a character; examples are é, α̣, x̄̈, h̄. The COBOL program analyzes the character name and provides PAGE-1 with data on vertical displacement of the mark, horizontal spacing to center the mark over the character (which may have a different character width than the mark), and readjustment to restore the setting position, type font, and point size. Special characters which are unrecognizable, either because they are misspelled or not available, are flagged on an error listing, and a flagging character is placed in the margin of the photocomposed copy. A flag or fixed space is also inserted in the copy in place of a character which cannot be set, or when a required data element is missing. A missing character can then be composed and pasted in place on the final copy.

A second major problem in the setting of technical text is that of line breaks. A line break is the undesirable splitting of a word between the end of one line and the beginning of the next line. PAGE-1 is designed to break text strings at a space between words or at any introduction of composition commands; RCA's more recent system, PAGE-2, will break words at forward and backspace commands, which AIP uses in creating special characters. Either system can cause unwanted line breaks to occur when there is a change in type size or font, such as a subscript, superscript, or special character; for example, a chemical formula such as CO_2 could appear with the CO at the end of one line and the subscript 2 at the beginning of the next line. In primary journal composition, we have resolved this problem by flagging each space between words to be the beginning of a trial setting; a word which overflows the line is reset on the following line. In journal indexes and current awareness journals, we must not only avoid line breaks but also avoid placing a heading at the bottom of a column and its following text at the top of another column or page. The latter requirement ties up the trial setting option; line breaks can still be prevented by having the COBOL program calculate the width of a word containing a potential line break, and then inserting commands to forward space and backspace by the width of the word. Any word which would fit on the line would not be affected by the forward and backspacing, but a forward space large enough to hit the right boundary of the column will result in the start of a new line, after justification.

Error corrections can be handled through the PAGE-1 correction facility; an alternative is to enter control cards to provide blank spaces where erroneous material would appear and then to run a special tape, carrying only corrected material, through PAGE-1. The pages of corrected data are then pasted onto the final copy.

DATA BASE DEVELOPMENT

Figure 1 shows the production system currently in use at AIP, which produces journals, microfilm, and a variety of secondary services in different formats. The development of new technologies such as OCR and computer photocomposition, and the sophisticated keyboarding and computer manipulation techniques which are now available, have been combined to make it possible to offer both primary and secondary services from a single data base. Within the next few years, the state of the art probably will have advanced to the point where we can produce high quality printing plates from microfilm, or simultaneously produce hard copy and microfilm from a single pass through a photocomposer. The systems approach to the communication of physics information has made it possible for AIP to offer secondary services at prices which individual members can afford, in contrast to the institutional price structure which many of the traditional abstracting and indexing services have been forced to adopt.

Other publishers are also investigating the systems approach. The American Chemical Society's *Inorganic Chemistry,* for example, is being typewriter-composed at Chemical Abstracts Service in preparation for computer photocomposition of the journal later in 1974 using CAS' APS4 photocomposer. *Chemical Abstracts* itself will be completely photocomposed by the end of 1975. The American Society of Civil Engineers photocomposes all of its journals, although the print tape is not reused for its secondary services; from typed pages converted by OCR to magnetic tape, the ASCE produces abstract cards for its monthly

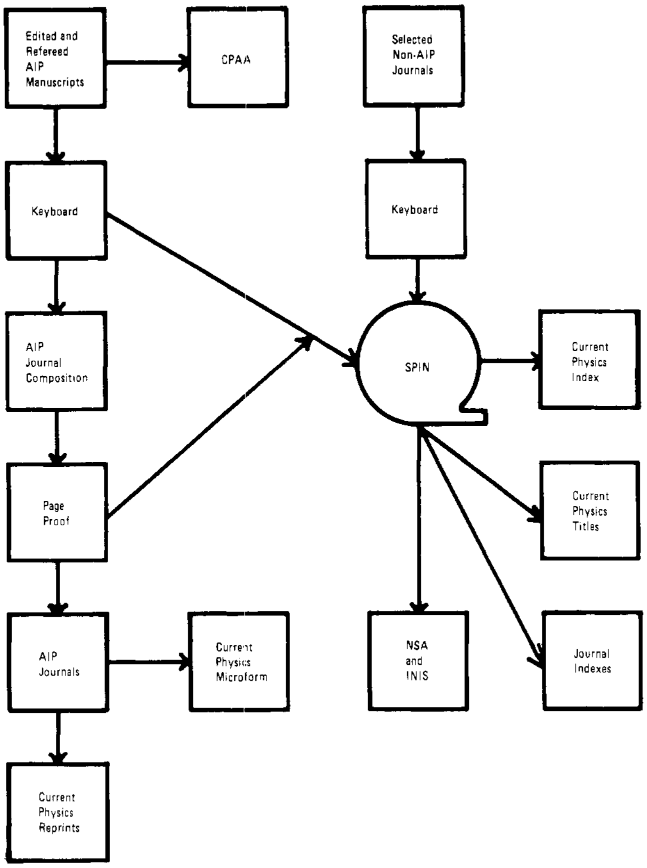

Figure 1. AIP Physics Information System.

journals and bimonthly abstracts journal, an annual combined subject and author index for all of its journals, and permuted subject and author indexes for conference proceedings. The IEEE photocomposes its indexes but not its journals. About 15 to 20% of the journal pages published by Pergamon Press and McGraw-Hill are photocomposed, and this percentage is expected to increase. Much of this activity is economically feasible only if multiple use is made of a single keyboarding by integrating primary and secondary services. The once clear-cut distinction between primary journal publishers and abstracting and indexing services is therefore becoming blurred, as we see secondary services composing full text for primary journals, and publishers offering abstracts on tape and in hard copy to secondary services. We are beginning to see information as a broad spectrum of services on a scale ranging from brief notations of content to full text, with the ultimate user, the individual scientist, able to select those services which best meet his individual requirements.

LITERATURE CITED

(1) Herschman, A., "Keeping Up With What's Going On in Physics," *Phys. Today,* **24** (11), 23-29 (1971).

(2) Koch, H. W., "Current Physics Information," *Science,* **174,** 918-922 (1971).

(3) Metzner, A. W. K., "Integrating Primary and Secondary Journals: A Model for the Immediate Future," *IEEE Trans. Prof. Commun.* **PC-16,** 84-91 and 175-176 (1973).

(4) Auffray, J.-P., "SPIN Technical Specifications," AIP ID72-S, American Institute of Physics, New York, N. Y., 1972.

(5) "Physics and Astronomy Classification Scheme," AIP R-261, American Institute of Physics, New York, N. Y., 1974.

(6) McQuillan, R., "The Composition Technology Book Composition System," in Proceedings of the 8th DECUS European Seminar, Strasbourg, France, Sept 1972.

(7) Alt, F. L., and Kirk, J. Y., "Computer Photocomposition of Technical Text," *Commun. ACM,* **16,** 386-391 (1973).

On-Line Searching of Computer Data Bases[†]

BARBARA G. PREWITT

Rohm and Haas Company, Research Division, Spring House, Pennsylvania 19477

Received June 18, 1974

The Research Library of Rohm and Haas Company has been searching a variety of bibliographic data bases on-line for over one year. A summary of our experiences and the merits of on-line searching is presented. A conference call technique for driving a remote slave terminal is described.

The Rohm and Haas Company Research Library has facilities located at each of the three metropolitan Philadelphia research laboratories. Also located in Philadelphia is our home office, and other facilities are found across the country and overseas. The Research Library services primarily scientific personnel of the Research Division. These people are usually chemists, but are also scientists from other disciplines such as biologists, engineers, etc.

We have had more than one year's experience in searching on-line data bases and have had essentially no formal training in any of the systems that we are currently using. All the systems that we are using now utilize Boolean Logic.

The first system we investigated was the LEADERMART system from Lehigh University. LEADERMART was of interest to us because it did not require a question to be input in a Boolean Logic form. One merely typed in the words or a sentence describing the question. The computer analyzed these words and came up with the references that provided the best match. We felt that this system was very easy for scientists to use themselves and would relieve the information chemist of the necessity of performing all searches. The system had both the *Chemical Abstracts Condensates* and *Compendex* data bases up, and both of these were of interest to our clientele. We did extensive experimentation with LEADERMART and found it to be very useful. We had just begun to teach bench chemists how to do their own on-line searches when Lehigh removed

[†] Presented in the Chemist's Club Library Seminar, New York, N. Y., April 5, 1974.