Reprinted with permission from *IEEE Transactions on Professional Communication* PC-20(2):95-9, September 1977.

Also see Introduction on p.286, Essays, Vol:3

ISI® Data-Base-Produced Information Services

E. GARFIELD, M. KOENIG, AND T. DIRENZO

Abstract—**The Institute for Scientific Information® (ISI®) is a multinational corporation that provides a wide variety of information services to scientists and librarians throughout the world. Included are the *Science Citation Index*®, *Current Contents*®, and others which depend on sophisticated computer processing for timely production. This paper describes how certain information elements are extracted from each journal article and processed through the ISI system. Examples are given of how recent computer technology has been applied to keep ISI services cost-effective as well as to improve their quality.**

THE Institute for Scientific Information® (ISI®) is a multinational corporation that provides a wide variety of information services to scientists and librarians throughout the world. From a functional point of view, the services for the journal literature of science and technology can be classified into the following five major groups:

- Current awareness
- Selective dissemination of information (SDI)
- Retrospective search
- Document fulfillment
- Acquisitions planning.

For broad current awareness, the *Current Contents*® (*CC*®)

This paper was presented by Michael Koenig at the Third IEEE Conference on Scientific Journals, May 2-4, 1977, Reston, VA.

Eugene Garfield and Thomas DiRenzo are with the Institute for Scientific Information, 325 Chestnut St., Philadelphia, PA 19106, (215) 923-3300. Dr. Garfield is President of ISI and Mr. DiRenzo is Vice-President, Communications. Michael Koenig is with the Metrek Division of Mitre Corp., 1820 Dolley Madison Blvd., McLean, VA 22102, (703) 827-6554.

services provide weekly contents-page coverage of six different disciplinary areas: life sciences; agriculture, biology, and environmental sciences; social and behavioral sciences; engineering, technology, and applied sciences; physical and chemical sciences; and clinical practice [1].

ASCA® and *ASCATOPICS*® are the Institute's SDI services. Both provide subscribers with weekly computer reports listing recent articles relevant to their specific interests. The difference between the two services is that *ASCA* uses custom, one-of-a-kind profiles to represent subscribers' interests while *ASCATOPICS* uses standard profiles [2].

For literature searches that must cover a number of years (retrospective search), the major tools ISI offers are the *Science Citation Index*® (*SCI*®) for the natural and physical sciences and *Social Sciences Citation Index*™ (*SSCI*™) for the social and behavioral sciences [3]. These large indexes are supplemented by the smaller, more specialized *Index to Scientific Reviews*™ (*ISR*™) which provides coverage limited to review articles.

To help subscribers obtain hard copies of the articles they need when they cannot get them through traditional channels, a tearsheet service, *Original Article Tearsheet Service–OATS*®, supplies items published in journals covered by ISI. This operation casts the company in the role of being a librarian's library and one of the major sources of journal material in the United States. Complementing this basic document fulfillment service is an annual directory, *Who is Publishing in Science*, which provides the names and addresses of scientists who have published during the past calendar year. Many researchers and librarians, particularly those in developing countries, have used this directory to facilitate reprint requests and other correspondence.

In the acquisitions planning area, the *Journal Citation Reports*® (*JCR*™) gives librarians and others concerned with managing journal collections a source of objective data con-

cerning the utility of specific journals. Decisions related to which journal subscriptions should be added or deleted or how far back certain journals should be kept can now be based, at least in part, on the information provided in *JCR* [4].

Production Flow

While ISI's use of citation indexing eliminates the expensive intellectual effort associated with traditional subject-term indexing, producing a data base that grows by nearly a million items a year is a massive materials handling and information processing job. It would not be possible to diagram here all the steps involved, but Fig. 1 provides a good representation of what is required to build the ISI data base and extract various services.

The job begins with the receipt and logging in of the individual issues of journals that are covered. Immediately after the journal issues are entered into the system, their tables of contents are removed and either used as is or recomposed for use in *Current Contents*. The data for the various *CC* indexes are supplied later.

Next, the journals are sent to the editing department. There, every item must be examined to determine whether it should be covered, and everything other than minor news notices and advertisements must be marked in some way to simplify entering the information into the computer and standardizing what goes in. This process involves coding each item to show what it is; identifying the first and last pages and the references of each item; noting journals whose reference formats differ from article to article; coding titles that must be translated into English; and editing titles, authors' names, organization names and addresses, and some references.

Titles must be marked and edited to show where they begin and end, eliminate unnecessary words, add pertinent footnote annotations, and standardize punctuation, numerical expressions, and proper names. Scientific notation must be edited

to meet rules of standardization and computer processing requirements.

Authors' names and addresses must be underlined, and each name must be coded to distinguish between primary and secondary authors. Authors' names must be standardized too; this includes non-English names for which the rules of standardization are quite involved. The organizational names in authors' addresses also must be standardized [5].

References interspersed throughout the text or split between the text and footnotes, and footnotes that contain multiple references, are the toughest part of the editing job. Most often found in social science journals, these types of references require extensive editing notation to identify, integrate, and complete, and may require the help of a professional translator if they involve non-English citations.

Editing time per journal varies from half an hour to three days. Journals dealing with the social sciences are generally the most time consuming because their bibliographic standards tend to be archaic and their references are frequently complex, often citing exotic types of non-journal material, such as rare documents, legislation, and laws. It is not unusual to find references scattered throughout the text of a social science journal, which means the editor must scan the entire article. Footnotes containing multiple references are common, and the format of references, regardless of where they are found, is eclectic enough to make reformating the rule rather than the exception. The impact of these problems on productivity is great enough to justify a continuing and sizeable journal-liaison effort aimed at educating editors about the reader and economic advantages to be gained from adopting simpler, more standardized format rules [6].

The next production step is putting the edited material into the computer. This job is done by 60 data-entry operators working two shifts, five days a week, on keyboard display terminals on-line (connected directly) to a central magnetic disk

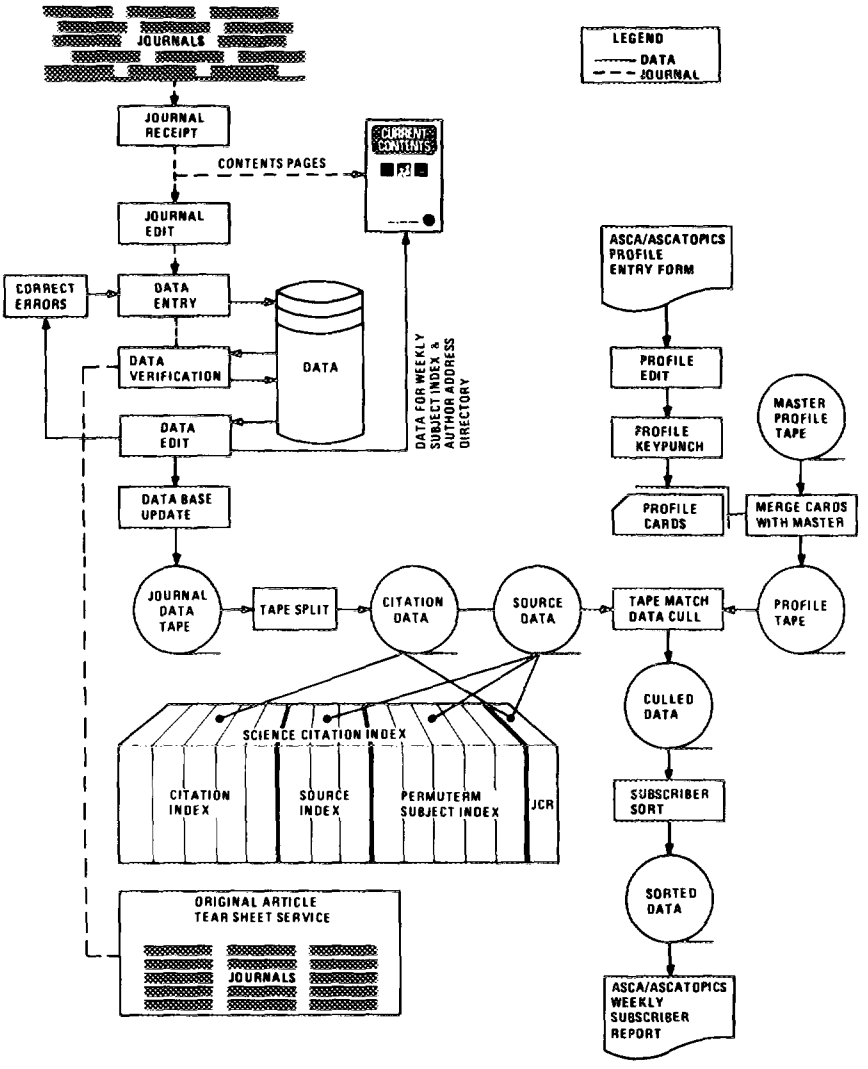

Fig. 1. Schematic of ISI® journal data flow.

memory. The journals move through this process in batches for which recording formats have been specified, and to which job control numbers have been assigned. As part of the job control procedure, individual journals are logged into the system, at the time they are assigned to a batch, by name, volume,

issue, month, year, accession date, and number, and with a status-and-date statement. After that, the status-and-date statement in the system log is updated every time the journal moves from one operation to another.

Once a journal has been assigned and logged, it goes to a data-entry operator, who verifies and updates the log to let the system know what journal is being worked on and where it is. The operator then works through the journal article by article, keying the pertinent information from each into the system in a three-part sequence.

First comes the basic information that identifies the article: its type, title, page numbers, and primary author. The middle part of the data-entry sequence involves additional author information: the address of the primary author and the names and addresses of any secondary authors. The last part of the sequence deals with the references made in the article.

When all the information about all the articles in the journal has been entered into the system, the operator lets the system know the journal is finished by updating its log. Another operator then goes through the entire data-entry sequence again, character by character, to verify the work of the first operator. After data verification, the journals are filed by their ISI accession numbers and can now be used to provide tearsheets of individual articles.

Periodically, the verified records created for batches of journals are automatically transferred, under the control of someone working at a supervisory terminal, from the magnetic disk to a magnetic tape. As the records are transferred, they also are reformatted for computer processing.

The data-entry workload can be defined by a variety of numbers, all of them large. Approximately 2,000 source articles, involving some 22,000 references are processed each day. With the record length per source article averaging 1100 characters, the total number of characters entered each day is approximately one million. (If the verification operation is

included, the total number of keystrokes per day is of the order of two million.)

At this point in the production cycle, the computer takes over. Despite the pains taken in the editing operation to identify, clarify, and standardize everything, and the character-by-character verification performed in data entry to assure keying accuracy, the first thing the computer must do to the tapes is edit them to make sure that all the records are complete and properly formatted. Some one percent are not and must be recycled through editing and data entry.

Besides checking the content and format of the individual records, which are organized by journal, the computer also checks the journals against a year-to-date file of the journal issues that have already been processed. Duplicates are recycled back through the journal-control people to work out the problems. Those that are not duplicates are copied onto the year-to-date file.

The tapes from data entry are edited this way on a daily basis and accumulated into a weekly data base, which is edited again to verify the daily checks of content and format. The edited weekly data base is then coded to show what journal records go into what service.

The records on the coded weekly data base are then sorted into two major data categories. The first is source data that include bibliographic descriptions (including titles) of the newly published source articles and names and addresses of the organizations with which the authors are affiliated. The second category is citation data, which are the brief bibliographic descriptions of the items referenced (including patents) in the source articles. These sorted records, except for the patent data, correspond to major sections in *SCI* and *SSCI*; the patent data are included only in *SCI*.

At this point, the data which make up the *Author Index and Address Directory* and *Weekly Subject Index* sections of the *Current Contents* publications are extracted onto tapes. This

is done by a routine that organizes the material into pages and specifies formats and type fonts. The tapes are used to drive an automatic photocomposition machine, which turns out reproduction-quality page proofs from which offset negatives and plates can be made for printing. These index and directory pages will join, in the appropriate *CC* edition, their counterpart table-of-contents pages which were processed earlier in the week.

Another edit of the source data file is then done to further assure the accuracy of the title information. This is done by checking every key word in every title on the file against a key-word dictionary. Words not found in the dictionary are passed on to editors who determine whether they are misspellings of valid words or are words not yet included in the dictionary. Misspelled words are corrected; new ones that are judged to be important are validated and added to the dictionary.

After the weekly title edit, *ASCA* and *ASCATOPICS* subscriber profiles, which are stored on tape, are matched against the source and citation data elements. Whenever a match occurs, the full bibliographic description of the journal item which contained the data element is recorded on another tape. After all the profiles have passed against the source and citation data, the culled items are re-sorted by subscriber number and printed as the weekly reports which are quickly mailed to users.

The rest of the computer processing is done on quarterly, trimonthly, semiannual, and annual cycles. The quarterly cycle is concerned with preparing a three-month cumulation of items covered by the *Science Citation Index*. Here again, a computer routine reformats the material and produces tapes for photocomposition. In the last quarter of the year, the material for that quarter is consolidated with what had been published in the preceding three quarters to produce a cumulative annual index.

This same basic procedure is used to prepare the *Social Sciences Citation Index* on a four-month cycle. The *Index to Scientific Reviews* is produced on a semiannual cycle. On an annual basis, other computer routines create the tapes which produce *Who is Publishing in Science* and *Journal Citation Reports.*

In the preparation of the printed indexes there are a number of additional checks for accuracy conducted in the final stages of production. The first is a detailed check of the first statistically significant batch of pages produced by the photocomposer. The accuracy of the weekly data bases and the effectiveness of the computer routine that merges them are checked by matching a random sample of articles that should be covered in the initial pages against the page proofs. The effectiveness of other key computer programs is also checked in this initial sample by looking for discrepancies and omissions in names, cross references, and formats. If everything is correct, the rest of the pages are produced. These too are checked, but for such things as print quality, the number of columns per page, and the sequence of columns—all things that can go wrong in the photocomposition stage of production. Only then is the job released to the printer. The printer's work, too, is spot-checked for all the things that can go wrong in the printing process.

The Promise of Technology

Exploiting the potential of computer technology to achieve improved cost effectiveness in creating the ISI data base is a matter of continually searching for production efficiencies among the technological advances. Some of the efficiencies are built into the lower cost per unit of processing offered by succeeding generations of equipment and can be realized merely by upgrading the equipment periodically. Other, more significant efficiencies call for the ability to innovate from the improved functional base provided by the new equipment. The

impact that key-to-disk data entry equipment has had on production is a case in point.

For a data entry operation as big as the one involved in the ISI data base, key-to-disk systems are more efficient than the older keypunch. Job control procedures are easier to implement; keying is done at electronic, rather than mechanical, speeds; and a lot of punched card handling is eliminated. In addition, each terminal operator has access to a central, disk memory, around which ISI has built a production innovation that increases efficiency far beyond the level made possible by the superior speed of key-to-disk systems.

The innovation consists of using the shared-disk memory to store an historical file of reference citations. The increase in efficiency comes from reducing the amount of keying necessary to enter and verify reference citations. Instead of keying the full citation, the operator keys in a 14-character code abstracted from the full citation. Each of the citations on the historical file has attached to it the same sort of coded identifier. If the code the operator enters matches one in the file, the full citation is brought up in the terminal display, where it is verified visually and entered on the disk with a single keystroke.

Every time an operator matches a reference citation against one in the historical file, the number of keystrokes required to enter the citation is reduced from an average of 70 to 14. And the keystrokes normally required for verification are eliminated completely.

The match-rate achieved depends on the number of citations in the historical file, which is limited by the size of the central disk memory available with the system. Initially, the file contained enough citations to produce a match-rate of 75 percent on the references that cited journal material. Some changes in the design of the file increased the utilization efficiency of the central memory enough to push the match-rate to 85 percent. Whether this rate can be raised still higher is uncertain, depend-

ing upon available memory capacity and how efficiently it is used.

The role of technology in the production of a data base goes beyond cost cutting into quality improvement. In some cases, the two can be combined. More often than not, however, quality improvements don't go hand in hand with cost reductions; they must be important enough to justify an increase in production cost. A computer-based system that automatically monitors the arrival of journals and tracks them through processing according to a planned schedule, for example, is more expensive than doing the same thing manually. But it does a better job, which produces the important qualitative benefit of increasing the timeliness and comprehensiveness of the data base.

Looking ahead, ISI plans to improve and expand its information services in the sciences and social sciences. Additionally, it has announced plans to offer a line of information services for the literature of the arts and humanities. To a large extent, the degree to which these plans are fulfilled depends on ISI's success in applying modern technology to building and exploiting its data base.

---

[1] E. Garfield, "Primary Journals, *Current Contents* and the Modern System of Scientific Communication," *Interdisciplinary Science Reviews* (in press).

[2] E. Garfield, "The Role of Man and Machine in an International Selective Dissemination of Information System," presented at the 35th Congress on Documentation of the International Federation for Documentation (FID), September 14-24, 1970, Buenos Aires, Argentina.

[3] M. Weinstock, "Citation Indexes," in *Encyclopedia of Library and Information Science*. New York: Marcel Dekker, 1971, vol. 5, pp. 16-40.

[4] E. Garfield, "Significant Journals of Science," *Nature*, vol. 264, pp. 609-15, 1976.

[5] E. Garfield, "An Address on Addresses," *Current Contents*, no. 28, p. 5, July 14, 1975.

[6] E. Garfield, "How Services from the Institute for Scientific Information Aid Journal Editors and Publishers," presented at the First International Conference of Scientific Editors, April 24-29, 1977, Jerusalem, Israel.