New Developments in Biological Abstracting And Indexing

LOUISE SCHULTZ

Current issues facing abstracting and indexing services serving scientists in biology, medicine, and agriculture bring into sharp focus the pressures which result from accelerating needs and from the impact of contemporary technology on document preparation and control. Nonetheless, the patterns of publication of scientific literature which we may describe today have been shaped during the past two hundred years. The eighteenth century saw the introduction of primary journals and the nineteenth century the introduction of secondary announcement media. In data graphed by Conrad,¹ the ratio between primary journals and abstracting services has decreased steadily during the twentieth century. The demand appears to be rising for correlations and selected announcements of portions of a given body of primary sources, and this demand can be met only by an increasing variety of abstracting and indexing services. One of the factors behind the creation and growth of such services is the interdisciplinary nature of scientific research and technological development. There is a growing interdependence between those doing research in the life sciences and those working in the physical sciences and mathematics.

A Decade of Change. We are nearing the end of a decade in which the rates of increase in the literature, in the population of scientists and technicians, and in the academic population in general have forced not only those who must cope with the literature but also those who are responsible for the investment in research itself, to recognize the role played by the literature and its handlers in the scientific advancement of the nation and the improvement of man's environment.

Accurate counts are not available of the numbers of primary journal

articles and, therefore, of the unique potential sources for the creation of abstracts or indexed representations for subsequent secondary announcement. As an indication of the changing picture, in 1957 something over 50,000 articles were represented in Biological Abstracts, whereas in 1967 something over 200,000 items will be represented there. Accumulatively, from an annual rate of 45,000 items, the ten-year period has brought under bibliographic control more than 1,000,000 documents concerning the life sciences.

During this period the attention of the executive and legislative branches of the Federal government has focused on problems of the exchange of scientific information. This attention has resulted in several comprehensive studies that have provided (1) an awareness of the weaknesses of available services, and (2) recommendations for corrections that do not ignore the delicate but pragmatic balances between information-seeking and use-behavior and cost effectiveness of investments in elaborate systems. (A comprehensive survey of the entire literature is provided in National Document-Handling Systems for Science and Technology.) Stemming from the recommendations of the Weinberg report that the scientist's role in the distribution of scientific literature must be expanded, the past two or three years have seen the creation of several non-governmental organizations whose goals include strengthening the interface between scientists and the appropriate elements of the Federal government.

In the biological sciences, the organization of the Council of Biology Editors represents an effort by the private sector to coordinate certain common aspects of the primary journals. Together with the medium for exchanging information about the problems and techniques of primary journal publication, the Council of Biology Editors has also created the opportunity for closer liaison between primary and secondary services dealing in the biological literature. Journals whose editors belong to the CBE show a discernible trend toward increased publication of abstracts in conjunction with full articles, for the use of abstracting services.

In a sense paralleling the activities of the Committee on Scientific and Technical Information (COSATI) of the Federal Council for Science and Technology, a committee on scientific and technical communication (SATCOM) has been established within the framework of the National Academy of Sciences-National Academy of Engineering. The committee is chartered to compare the activities of the private sector with those of the Federal government in the distribution of
New Developments in Biological Abstracting and Indexing

scientific information. Methods for promoting more effective relationships between information systems are to be studied. Although their charter is broadly applicable to all science, one of SATCOM's initial efforts was directed toward the major services in biology: the National Library of Medicine's MEDLARS system and the techniques and products of BioSciences Information Service (BIOSIS).*

Under the sponsorship of the Office of Science Information Service of the National Science Foundation, BIOSIS organized a conference held in late 1965 to "arrive at a clearer definition of the total information requirement of biologists; and . . . to stimulate a more enlightened and continuing interest by participants in both the problems and the potential of a unified communications system." Approximately a year later, the Council of Biological Sciences Information (COBSI) was organized under the aegis of the National Academy of Sciences; it comprises representatives of the American Association for the Advancement of Science, the American Institute of Biological Sciences, the Council of Biology Editors, and the Federation of American Societies for Experimental Biology. Other organizations active in the dissemination of biological information are being considered as possible additions to the Council.

Parameters of Abstracting and Indexing. In the generalized system structure comprising input, processes, and outputs, abstracting and indexing may be considered to constitute (1) coverage and acquisitioning, (2) editing and indexing, and (3) publishing and distributing.

In discussing new developments in the abstracting and indexing services serving biologists, we recognize that current technology is most effective in support of the mechanical aspects of production. Thus, in terms of selecting items to be included in the service (coverage) and in negotiating for the acquisition of such items, contemporary data processing technology is, at present, only really useful for areas such as record keeping.

For those systems in which the full text of abstracts is entered into a data processing system, certain of the mechanical aspects of editing can be assisted by automatic data processing; techniques exist providing for automatic imposition of certain mechanical editing rules. The large services in the biological sciences have not yet widely implemented these techniques; however, the American Psychological Association is introducing them into the production of its abstracting journal.¹¹
Data processing support in publishing and distributing is also exemplified by the system being developed at the American Psychological Association, in which a photocomposition system is driven by a magnetic tape record prepared from perforated tape records of abstracts and citations. The use of computer-produced records to drive both linecasting and photocomposition equipment in the preparation of scientific literature was discussed comprehensively in the symposium on electronic composition in printing conducted by the Center for Computer Sciences and Technology of the National Bureau of Standards on June 15 and 16, 1967 at Gaithersburg, Maryland. While the capabilities of the equipment and its particular applications are not relevant here, it is worth noting that the conference was attended by more than five hundred people involved in some way with the preparation of scientific literature who recognize the potential and the need for application for such advanced techniques.

**Impact of Technology.** The demands of a data processing system for explicit, rigidly formatted information are the core of the increased interest in standards for the description of documents, machine records of which are to be prepared for some application of data processing support in the preparation or distribution of scientific literature. Whether the product is available publicly or whether it is prepared for use entirely within an organization such as a large pharmaceutical company makes no difference; the systems of the various abstracting and indexing organizations have been developed independently and without uniformity. As a result, the diverse practices and policies of the organization which does the data processing are one source of variation in the descriptions of documents being “reannounced.” Although “equivalent” or only trivially different to the human user, these efforts do not facilitate interchangeability of machine records.

The committee of the U.S.A. Standards Institute (USASI) on library work, documentation, and related publishing practices has a subcommittee whose aim is to develop a matrix “to express the bibliographic data, elements, media, and use observed ... in mechanized library systems, new codes for descriptive cataloging, bibliographic style manuals, abstracting and indexing services, [and] filing rules.” In addition to a common basis for minimum criteria in describing a document, the USASI is also interested in providing standards for abbreviation of those elements of the document description that are widely used, e.g., periodical title abbreviations.
New Developments in Biological Abstracting and Indexing

Through the medium of the National Federation of Science Abstracting and Indexing Services, relations between the principal non-profit secondary announcers of scientific literature have become closer. However, despite participation in forums concerning standardization of document descriptions, and creation of exchange agreements providing for use by one of the texts of abstracts published by another, the differences in indexing techniques and vocabularies remain and are the reason for the relatively insignificant actual exchange that takes place.

Where one or more of these services is supported by data processing systems, any opportunity for exchange of the machine record depends critically upon the joint design of the standards for the machine record. One of the few examples of the exchange of a machine-readable record has been that of the Fisheries Branch of the Food and Agriculture Organization in Rome, Italy, and the Aquatic Sciences Information Retrieval Center at the University of Rhode Island. Manuscript for the Current Bibliography for Aquatic Sciences and Fisheries was prepared on tape-perforating equipment in a format facilitating data processing manipulation of the entries.14

Among the techniques used in indexing the biological literature, Baxendale lists prescriptive indexing in which the indexer uses “an explicit glossary of labels and maps the statements of the document onto the prescribed descriptive labels,”15 a variation of which is the “concept list”; “data type” glossaries, such as that of the Engineers Joint Council, in which types of terms are designated; and “keyword,” based on selection of words from designated portions of the document description such as the title. The application of data processing support to the announcement of biological literature has contributed to the emergence of keyword (in or out of context) indexes generated mechanically from strings of words such as titles and/or additional natural language elements.16 Baxendale sums up their development thus:

One factor leading to their [keyword indexes] development is the belated adjustment of conventional indexing and classification tools to changing patterns of need, to changing terminology, and to new technology. . . . The second factor is the relative simplicity and practicality of computer manipulation of each of these forms of indexing . . . . Although, in their present rather primitive form, these indexing methods have their inherent qualitative limitations, they are often an acceptable compromise with practical exigencies.17
Terminology control for creation of an index may be considered to be implicit in the semantics of the phrase \(^{18}\) and is demonstrated by the syntax of an entry in a conventional index and mechanically by the syntax in a title, where the index is generated from words in the title. Control is most commonly exercised in the \textit{de facto} creation of a thesaurus or glossary prescribing acceptable index tags or terms and indicating the relationships between such terms where these have been considered in the creation of the vocabulary.

Among the biological literature handling systems, the one which demonstrates the problems and techniques of creation and use of a controlled vocabulary specified by thesaurus is the MEDLARS system, in which the vocabulary is that of the \textit{Medical Subject Headings} (\textit{MeSH}) list. The \textit{MeSH} thesaurus was developed to provide for publishing \textit{Index Medicus} and is "a dynamic thesaurus which attempts to respond to new needs of usage, research, and publication. New headings are added annually; unused headings are deleted, provisional headings are maintained for use by indexers and searchers. . . . Analysts use \textit{MeSH} terms of maximum specificity. Intermediary or general terms (e.g., \textit{LIVER DISEASES} or \textit{ENZYMES}) are assigned only when there are no \textit{MeSH} headings for the specific concept discussed . . . [and] an average of eight headings are assigned to each article." \(^{19}\)

The importance of terminology control, and its connection with technology, is acknowledged in other fields. For agriculture, for example, Task Force ABLE reported: "When plans were made for issuing the preliminary edition of the subject heading list, arrangements were made for punching paper tape to be used later . . . [in the] hope that possible automation of the \textit{Bibliography of Agriculture} subject index will make available a comparable tape for use in compiling a unified thesaurus." \(^{20}\) The commitment to controlled vocabulary is also illustrated by the following:

The vocabulary method selected by the chief of the Pesticide Information Center and its staff is very important since the vocabulary structure and content are critical features of the whole . . . system. Nothing can be retrieved from either the journal or the mechanized system if it is not accurately and consistently indexed. \(^{21}\)

The entries in an index created by a machine or by machine-like selection from the string of words constituting titles or other given portions of text representing a full source document may be considered uncontrolled in terms of the system handling the entries. (In terms of the creator or user of the title, both the use of individual words and
New Developments in Biological Abstracting and Indexing

combinations of nominatives in appropriate syntactical relationships cannot be truly uncontrolled if communication is to be effected by the transfer of such strings of words.) In a very large system, the difficulties of an uncontrolled indexing vocabulary center around (1) the size of the vocabulary, (2) the inconsistency of authors in respect to the specificity or generality with which they have “described” the contents of the source document by means of its title, (3) the variations in usage of individual word tags, and (4) the dynamic aspects of synonomy in a system with a large basic vocabulary.

BioSciences Information Service compromises between these difficulties and the inflexibilities (and equally inconsistent application) of a controlled indexing vocabulary through a combination of indexing techniques. The range of “control” of the input vocabulary extends from “none” (represented by authors’ titles, in the Biological Abstracts Subjects in Context [B.A.S.I.C.], a permuted context index), through “nominal” (represented by guidelines for augmenting the contents of the latter index), to “dynamic-but-tight” (the concept classification tags represented by the names of the sections in which Biological Abstracts is organized), and finally to the essentially “prescribed” hierarchy of organisms, the names of which represent entries in the taxonomic index (Biosystematic).

The activities of secondary announcers and distributors of scientific literature formerly were limited to the publication of an announcement journal. Proliferation of the specialized information centers serving the needs of a small, usually well-defined, user group and capable of offering services other than conventional cumulative republication, has contributed to a re-evaluation by the large secondary services of both the potential and the problems of expanding services beyond publication. In Figure 1, the relationships among some of the principal tasks in a comprehensive, data processing-based secondary service are modelled.

Technology is modifying the techniques for creating the textual “image” to be published by offering a choice between (1) the storage of the record as character strings in machine manipulable form, and (2) storage of the image (either as copies of the printed citation and/or abstract, or else in microform). The motivation in storing the machine manipulable record may be to search it directly and in full or merely to reproduce the contents of the item selectively in preparing derivative publications, individually tailored bibliographies, or demand and current awareness retrieval services.

JANUARY, 1968 [343]
Except in the case in which the mission of the retrieval system can justify the expense of on-line cathode ray tube displays of the contents of the machine record, reproduction for human use of the contents of the announcement involves operation of a line printer, nominally at six hundred lines per minute. Where the item to be reproduced consists, on the average, of two hundred words, the cost of reproduction from an image file rather than from machine-readable files becomes more attractive in terms of cost of storage and reproduction. An image storage system offers reproduction of a page bearing some four thousand characters (including special characters, line drawings and halftone illustrations not reproducible on a computer line printer) in twenty to thirty seconds.
New Developments in Biological Abstracting and Indexing

Where the full record rather than an index or other referent is to be searched for retrieval, storage capacity and processing time constitute the principal limitations to machine-form storage of full records. That is, if the machine record is in a serial storage medium such as magnetic tape, processing time is proportionate to the size of the file. However, where the storage medium is random access or quasi-random access, processing time decreases but the cost of storage and storage maintenance rises.

Products of exploitation of the machine record fall into several categories: publications derived from the total machine record according to specifications of an identified user group; retrieval and delivery of reproductions of images of announcement records in response to individual search specifications; and analytic reports on the basis of which trends in biological research can be predicted so that literature handlers can acquire and make available the information which will be most useful to future biological research. The chief aim is to serve the individual biologist from a vastly expanding literature, at a time when needs for information are changing dynamically as new fields in the life sciences are recognized and become the focus of research. Services being offered by large secondary literature handlers in biology who are using data processing in the preparation of indexes and abstracts for publication are exemplified by the techniques for exploiting the machine record at BioSciences Information Service.

On the basis of machine selection of abstracts tagged with the taxonomic categorization of fungus and lichen, BioSciences Information Service has initiated an experimental publication called Abstracts of Mycology, being made available as a separate and produced as a derivative publication of the basic abstracting journal, Biological Abstracts.

In a separate experiment, BIOSIS is examining the possibility of a closer liaison with, in this case, the ichthyologists in assembling a rich machine index to the literature of interest to a particular group. While Biological Abstracts will continue to announce, as rapidly as material becomes available through routine channels, papers in ichthyology, the machine record created in conjunction with publishing an abstract or citation of such an article will be augmented by terms provided by the ichthyologists. The resulting enriched machine record will be available for individual search services to ichthyologists or others and will be the basis for assembling at some regular interval, perhaps yearly, a bibliography for the field to be distributed under the auspices of

JANUARY, 1968

[345]
the American Museum of Natural History in New York, the joint participant in this experiment with BIOSIS.

Assuming this latter experiment develops to the degree anticipated by BIOSIS and the AMNH, it represents a paradigm for a series of relationships between the secondary announcement service organization and individual subject specialist groups. We may consider this the contemporary expression of the commitment by the subject specialists, for their mutual benefit, to analysis of the contents of their own literature, in an environment in which control of the record of that literature has become a specialty to which is applied modern technology of publishing and machine manipulation of text.

A third area of exploitation of a machine record of biological literature is typified by individual retrieval services resulting in delivery to an individual requester of a subset of the material published, which is judged to meet his expressed needs. In early summer 1965, Bio-Sciences Information Service entered into an experimental program with the Walter Reed Army Institute of Research (WRAIR) to determine the feasibility of performing searches of the material published in Biological Abstracts, on the basis of the indexes which had been prepared in machine readable form as an adjunct of publication. These are B.A.S.I.C. (a keyword in context index), CROSS (a coordinate posted index in which the vocabulary consists of the names of the sections in which Biological Abstracts is organized), and the taxonomic categories of organisms discussed in papers announced in BA.

Progress in the experiment as it developed into an operational service has been described elsewhere. In general, the experiment with WRAIR involves Electrowriter terminals for communication of requests to BIOSIS. The Electrowriter is an electro-mechanical device that generates an electronic "signal" by the mechanical motion of a pen writing a message on the transmitting terminal. The signals are transmitted by phone line to drive a pen at the receiving terminal to replicate the message. Figure 2 illustrates the system relationships.

At BIOSIS, a biologist analyzes the message received within the framework of (1) the structure of the indexes to Biological Abstracts, Bioresearch Index, and Bioresearch Titles; (2) a reference library accessible to him; and (3) the staff of biologists at BIOSIS who represent the specialties of the life sciences. In analyzing the incoming message from WRAIR, the biologist, whom we call a strategist, determines what combination of manual and computer searches is most appropriate to identifying the subset of the entire machine-supported
file (more than ¾ million items by late 1967) that may be appropriate to the message. The available file represents items announced since September 1959. The machine file will ultimately include index records for items announced in Biological Abstracts, Abstracts of Bacteriology, and Botanical Abstracts since 1917—a total of nearly two million items.

The incoming message, referred to as the search specification, is expanded into a strategy comprising the set of computer file searches and manual searches based on the vocabulary determined by the strategist. The list of keyword access points, and the names of the files involved in performing a given search is referred to as the strategy. A listing of the strategy is made available to the requester along with delivery of a set of copies of the abstracts and/or citations identi-
LOUISE SCHULTZ

fied as responsive to his inquiry message. Corresponding to the machine file of indexes, a 16 mm. microfilm file has been prepared. Abstracts identified as appropriate are reproduced on a microfilm reader printer and packages of abstracts are post-edited by a qualified biologist, in most cases the search strategist, prior to shipment of the package to the requester.

In addition to service to three Electrowriters of the Army Medical Research and Development Command (one at the Walter Reed Army Institute of Research, one at the Letterman Hospital in San Francisco, and one at STINFO [Scientific and Technical Information] headquarters in the District of Columbia), the system is regularly servicing the Division of Biologics Standards of the National Institutes of Health. Individual searches are processed for other requesters at the prices quoted by BIOSIS after preliminary analyses of the search specifications. At BioSciences Information Service, the search service has been developing from a minimally automatic to an increasingly computer-based system. Considered essential to the system are the procedures and criteria for post-editing the results of machine or mixed manual-machine identification of abstracts.

Partly in support of the search services to a particular user group, BIOSIS has begun the machining of full abstract text (not of Biological Abstracts however) in order to study the structure and problems of automatically indexing and searching such full records. As soon as economically feasible, BIOSIS expects to create the entire contents of all its publications on devices that produce a machine-manipulable record simultaneously as a by-product of the production of the publication. With this machine record, and the experience gained before that time on the basis of the experiment with full abstract text machining, BIOSIS will implement plans for additional automatic data processing support to the generation of indexes. Further, in conjunction with the machining of full records, BIOSIS will develop a guide to the vocabulary of the biological literature, both to help the individual using the published indexes and as the basis for the machine index to the total machine file.

BIOSIS plans to study the feasibility of automatic assignments of the tags constituting the CROSS vocabulary, based on the occurrence of words in and added to the titles. The preliminary study of this problem is to be made in conjunction with a research program developed mutually by BIOSIS and an industrial chemical manufacturer to which duplicates of the tapes of the B.A.S.I.C. and CROSS indexes to BA
New Developments in Biological Abstracting and Indexing

are supplied. Already experienced in providing their chemist population with selective listings of the literature of chemistry (from Chemical Titles), this organization is interested in examining the combination indexes that include both the keywords and the concept categorization of CROSS, for services to their biologists.

Among the technological developments expected to have an impact upon the handlers of scientific literature are the opportunities for long distance communication, particularly between data processing systems. On the assumption that such communication is technologically and economically feasible within the foreseeable future, the Interuniversity Communications Council has established the effort known as EDUCOM. Successful realization of a network between universities, for the exchange of both administrative and scientific information in digital form, suggests a potentially wider outlet for the products of a scientific literature handling organization, as well as additional direct contacts with potential users of that literature.

The establishment of regional search centers for the use of the MEDLARS computer based data file can be considered a network. Although communication from the individual requesting information is limited to his access to a regional search center, the fact that each of the search centers duplicates the information created at a central point (the National Library of Medicine) may be considered to constitute a network-like connection between the centers. Decentralization of the search service results in development of individual search programs at each individual center, to exploit the data processing equipment available there. Regional MEDLARS centers have been established at the University of California Medical Center, Los Angeles, the University of Colorado Medical Center, Denver, the University of Alabama Medical Center, Birmingham, and international center in Brussels. Copies of the machine record are also available in France and in Stockholm for local processing by the national information services of those countries.25

With the generation of a multitude of computer-processed records of the literature of biology, concern is rising about the techniques that may be required to enable users fully to exploit such records locally. The Institute for Library Research (ILR) at the University of California has faced the problem of the role of the university library system in coping with new media being produced by secondary handlers of literature. As part of the program to clarify both the policies of the university system and the research necessary, ILR conducted a symposium in May, 1967, to consider the following questions:

JANUARY, 1968
1. Does the individual university researcher need information services from the newer media such as data processing system records?

2. Does the individual campus of a large university system need to set up local service on such media, or should the campus participate in intrauniversity or interuniversity networks, or should such services be acquired from their sources in more conventional forms?

3. Is the university library an appropriate center for the management of such services?

4. If the library is an appropriate center, does it have the capability for managing such services, or what additional requirements are demanded to achieve such a capability?

5. How does the university library, as the local manager of such services, handle the problems of incompatibility between media from a variety of sources?

While it is not appropriate for a provider of these media to speculate on the conclusions to be drawn by the ILR, the library community should be aware that these questions are being examined. A report summarizing the comments of a group of users and creators of biological information and of biological information services is scheduled to be prepared as a result of that meeting.

Conclusions. A number of factors are influencing biological abstracting and indexing activities, and also the plans for new services and products. Among them are:

1. the increase in world population and, therefore, in the population of the world's biologists;
2. the attendant increase in the world's biological literature;
3. the increased competition among nations for improvement of human environment, together with exploration of new environments;
4. the focus of the Federal government of the United States on the importance of scientific literature for future advances, and resultant Federal policies;
5. increasing awareness among users that they should take the responsibility for adding their own contributions to a system, in forms suitable for transmission to their contemporary and future peers;
New Developments in Biological Abstracting and Indexing

6. the realization that information, as much as "hardware," is a support tool to be planned for, paid for, and exploited; and
7. the introduction of new equipment and techniques for handling literature, and the investment by both the Federal and private sectors in the introduction of such techniques and equipment into the processing of scientific literature.

The long-range plans implicit in the above factors represent, in themselves, a departure from the former assumption of responsibility only for compiling and publishing information about the scientific literature. With unprecedented enthusiasm and imagination, abstracting and indexing services have seized on the capabilities of automatic data processing and pressed them into the service of speedily making known work in the life sciences. Relieved of mechanical bottlenecks to document management, they are improving the ease and consistency with which the individual can select from the most comprehensive biological data base those items of particular interest in his professional work.

References

12. Ibid., p. 1041.
17. Baxendale, op. cit., p. 84.
18. Ibid., p. 88.