Automated Literature Alerting System

The Automated Literature Alerting System (ALAS) was developed at Eli Lilly and Company and implemented in January 1967 as a successor to the *Chemical Titles* tape program. The magnetic tapes received weekly from the Institute for Scientific Information contain title information from about 1800 journals which produce an average of 5600 article titles per week. However, during processing, about 600 nonpertinent journals are deleted.

The program began with an initial group of 58 *CT* subscribers, and new users are continually being added. Each user has been receiving from 0 to 400 citation cards per week, depending upon his profile of key words. The median is about 70 to 80.

The system consists of eleven programs for profile and key word generation and maintenance, indexing, searching, and printing, and uses an average of 3.2 hours of machine time a week on an IBM S/360 F30. With 82 users currently in the system, this amounts to about 2.3 minutes of computer time per user per week. Searching is based on a key word system which allows the use of words, authors, prefixes, and suffixes, and employs AND, OR, and NOT logic.

PAM BARNEY BRANNON, DONALD F. BURNHAM RICHARD M. JAMES, † and LEE ANN BERTRAM ‡

Eli Lilly and Company
Indianapolis, Indiana

• Introduction

The selective dissemination of information (SDI) system at Eli Lilly and Company began in the summer of 1962 with the processing of *Chemical Titles* KWIC (key word in context) tapes. This service notified users of the current publication of literature items in their particular fields of interest (*1*). In 1966, we decided to expand the coverage of our SDI service by using the source data tapes from the Institute for Scientific Information (ISI). The Automated Literature Alerting System (ALAS) was developed at Eli Lilly and Company and implemented in January 1967 as a successor to the *Chemical Titles* tape program.

The Institute for Scientific Information publishes *Current Contents, Index Chemicus,* and *Science Citation Index*. The magnetic tapes which we receive contain data used by ISI in generating the Source Index for the *Science Citation Index* (*2*) and for word questions in the *Automatic Subject Citation Alert* (ASCA) (*3*). This data is mailed weekly on magnetic tapes provided by Eli Lilly and Company. The tapes contain titles, authors, journal names, and other bibliographic information from about 1800 journals.

† Scientific Systems Programming, Eli Lilly and Company.
‡ Scientific Library, Eli Lilly and Company.

The citations covered include articles, abstracts, reports, errata, discussions, editorials, items about individuals, technical papers, letters, proceedings from meetings, technical notes, and reviews or bibliographies. There is very little editing of key words, Greek letters are spelled out, apostrophes are deleted from all words, and hyphens may or may not be missing. Previously, foreign titles were not translated; however, beginning in January 1968, all foreign titles are being translated into English by ISI.

The following comparison of the *CT* and ISI tapes shows the difference in data coverage.

	CT	ISI
Source of information	Titles and Authors	Titles and Authors
Number of journals covered	650	1800
Frequency	Biweekly	Weekly
Coverage	Chemistry	Science and Technology

The source data tapes are processed on an IBM System/360 Model F30 with 65,535 bytes of core storage, four tape drives, two disk drives, and one 1403 printer with standard and decimal instruction sets. The system was written and debugged under the IBM Disk Operating System (16K DOS) in Assembler Language.

During processing the system deletes about 600 jour-

nals not pertinent to our interests, leaving about 1200 journals. During 1967, we received a weekly average of 5600 source citations from ISI. Of these, a weekly average of 3300 citations was used for current searching. Each user has been receiving an average of 71 titles a week, depending upon his user profile. (The individual average ranges from 2 to 400 among users.) Machine time averaged 3.2 hours per week, including print time. With 82 users in the system, this amounted to 2.3 minutes per user per week.

• User Participation

The program began with a group of 58 *CT* subscribers, and new users are continually being added. There are from 750 to 775 scientists at Lilly with bachelor degrees and higher who are eligible for participation in the program. Currently there are no business-oriented users.

A senior librarian helps each user define a key word profile which attempts to anticipate pertinent article titles. His profile or request may be modified by the senior librarian upon his request. There is no limit to the number of profiles per user. The system currently has 320 separate profiles for its 82 users.

Profiles consist of key words. These may be authors, words, prefixes, or suffixes; however, truncation of both ends of a term is not allowed. A prefix such as *nucleo-* retrieves *nucleoprotein, nucleotide, nucleoside, nucleonics, nucleoplasm,* and *nucleon.* The suffix term *-sterol* retrieves titles containing the words *cholesterol, ergosterol, sitosterol,* etc. Other useful suffixes are *-RNA, -nucleotide, -virus,* and *-viral.* One profile may contain from 1 to 1,000 terms; however, most have fewer than 300 terms.

In this system, logic operators connect the terms in a profile. Two types of logic—inclusion and exclusion—are applied to individual terms and to groups or sets of terms. Each set contains terms which in turn are linked through logic operators. Nesting of sets is not allowed, i.e., a set may not appear within a set. Sets are enclosed within parentheses in the examples. The logic operators used by the system are: *AND,* a must inclusion; *OR,* an inclusion which is not mandatory; and *NOT,* a must exclusion. The logic for a term immediately precedes that term. The logic for a set precedes the open parenthesis.

The following example illustrates the use of logic operators:

AND (*AND* retrieval) *AND* (*OR* computer *OR* machine) *AND* (*OR* literature *OR* information).

This request will retrieve all titles containing the word *retrieval,* either the word *computer* or the word *machine,* or both; and either the word *literature* or the word *information* or both. Because of the *AND* operators between the sets, if any one set is not satisfied the request will be aborted.

Care must be taken in building a profile, as the sequence of terms and sets influences the final result or hit list. At times it is necessary to make two requests instead of one to achieve the desired result.

In building a profile, it is important to remember that the program processes requests from left to right. In the above example, if no article titles contained the word *retrieval,* processing of that request would terminate before getting to the second set. The *AND* before the (*AND* retrieval) indicates that this set must be present in order to satisfy the request. The order of sets and terms is especially important when using *NOT* logic, as demonstrated in the following two examples:

OR (*AND* hormon- *NOT* purification) *OR* (*AND* steroid- *AND* structure) *NOT* (*OR* adren- *OR* cortic-)

This request will produce bibliographies with titles containing the prefix *hormon-* but not *purification,* nor *adren-,* nor *cortic-;* or titles containing the words *structure* and *steroid-* but not *adren-* nor *cortic-*. *NOTE:* The word *purification* may appear with the latter titles, since it is excluded only from titles with *hormon-*. It is possible to get a title with the words *hormone* and *purification* in the following situation. A title includes the words *hormone, steroidal, purification,* and *structure.* Processing the request from left to right, *hormone* would cause a hit, *purification* would throw it out, but *steroidal* and *structure* would pick it up again. To eliminate all titles with *purification,* the following format could be used:

OR (*AND* hormon-) *OR* (*AND* steroid- *AND* structure) *NOT* (*OR* adren- *OR* purification *OR* cortic-)

Journal titles, as well as terms, may be excluded; i.e., a user may want titles containing the prefix *electro-,* but may wish to exclude articles that appear in *Electronics* or *IEEE Comput.* A maximum of 40 journals may be excluded per request.

The problem of foreign titles was previously taken care of partially by the logic of the system. The following profile is one example

AND (*OR* liver- *OR* leber- *OR* fois- *OR* fegat- *OR* higad-)

This request contains foreign equivalents as well as the English term.

When the same request can be formatted in different ways, it is more efficient to use fewer brackets. For example, *AND* (*AND* ribonucleic) *AND* (*AND* deoxyribonucleic) would be better as *AND* (*AND* ribonucleic *AND* deoxyribonucleic). By using the right combination of terms and logic operators, a fairly sophisticated profile can be devised to define most problems and eliminate unwanted hits.

• File Generation and Maintenance

After the requests are designed and punched into cards, the individual user's profiles are stored on magnetic tape. This master profile tape is generated or updated, as needed, with data from punched cards. The cards con-

tain an A or a D to signify either an addition to or deletion from the file. Depending upon the A or D indicators, a record (key word) is either merged into the file or eliminated from it. Following the creation or update of the master profile tale, all, none, or selected profiles may be listed.

At the same time, the system generates a unique key word tape by collecting the individual key words appearing in all the profiles, sorting them, and removing all duplicates. This unique key word index consists of the profile key word with a one-letter code defining the word type; e.g., W=word, P=prefix, S=suffix, A=author, and J=journal (for exclusion). It is used weekly to select title terms for the source data search file.

The weekly ISI Source Index Tape is also converted to a key word out of context (KWOC) index; however, only the citations we "buy" are indexed. The 11-character journal code of each reference is checked against a core resident journal delete list. If a match is found, processing continues with the next reference. If a match is not found, the reference is "kept" and written on a new magnetic tape for further editing.

The authors and title terms of each reference are scanned for key words. Each title term is compared against a core resident key word stop list which contains articles, prepositions, and other words occurring too frequently for indexing. If a match is not found, the reference term and the article number are used to build the KWOC index. After an alphabetic sorting and editing procedure, the KWOC index records consist of the key word, its length code,[1] the word type (A or W), a string of associated source article numbers (not to exceed 20), and a sequence indicator. The latter identification flags the duplicate records which are needed to accommodate title key words appearing in more than 20 articles. The sequence indicator of the first record is always zero. Any subsequent records containing the same key word are numbered serially, beginning with 1. A single term may have a maximum of 100 records (2000 article numbers); however, there are rarely more than 10 records per key word.

The unique key word list generated from the user's profiles selects terms from the title KWOC index to form the final KWOC search index. The two alphabetic files are compared, and matches are stored in a subindex on the disk. In addition, a table of suffixes generated from the unique key word list is checked with the end of each term from the title KWOC list. Any title term containing one of the desired suffixes is listed, as the suffix, in a suffix index. After comparison of the two files, the suffix index is sorted and merged with the subindex to form the search index. This resides on a disk in the indexed sequential mode with the key word, length code, word type, and sequence indicator forming the record "key." This "key" is used in the search program for retrieval of the key word and its article numbers.

[1] Length code: The word length minus 1. This value is used by some program instructions during processing.

• Search and Print

Up to now, we have been discussing the preparation of the data base and user profiles for processing. The next step in the system is the actual searching for the articles of interest to each user and printing these article titles on "hit" cards. This is done by serially processing each profile on the master profile tape. The individual key words in the profiles become the "keys" for searching the index. When a key word is found in the search index, the string of article numbers associated with that key word becomes a list for the profile. After the second key word list is found, the logic for the second key word determines the method for combining the lists. An *AND* logic operator indicates that an article number must appear in both lists in order to satisfy the request. An *OR* operator merges the two lists, while a *NOT* operator removes any articles that appear in the second list from the first list. The program aborts any set or request at the earliest possible time. This occurs when an *AND* (indicating a must inclusion) term or set cannot be satisfied, or when a *NOT* term is found. Processing then continues with the next profile.

The lists for each user are condensed so that no article is printed more than once for any one user. The user request number which causes the hit is carried with each article number. Since all redundant article numbers are eliminated from the list, the first request which contains the duplicated article number is the only request that keeps it; e.g., if the article number 21650 occurs in request lists 02, 05, and 07 at 02-21650, 05-21650, and 07-21650, only 02-21650 is kept.

After the user's hits are condensed, the resultant list of article numbers passes to the print section of the program. This section needs two files for printing the hits. The bibliography file contains the title information received from ISI. ISI uses an 11-character code for the source journals. This code is expanded in the journal expansion file to the American Standards Association abbreviation. The file also contains an indication of whether or not Lilly subscribes to the journal. Both files are on disk in the indexed sequential mode. The article numbers from the users' lists form the key for searching the bibliography file for the proper title information. The reference is read into core, and the journal code forms the key for searching the journal expansion file. If no expanded name is found, the code is used on the printout and is typed on the console so that the journal code can be added to either the deletion or the expansion file. The bibliographic information is then listed on preprinted cards. (See Fig. 1.)

In addition to printing the hits, the program constructs a communications tape. This consists of the user number (employee number), his name and mailing address (department number), the number of hits he receives, and, in some cases, a special message. This information is printed on header cards. (See Fig. 2.)

Only the user's first 400 hits are printed. Any articles in excess of this are listed, along with the corresponding

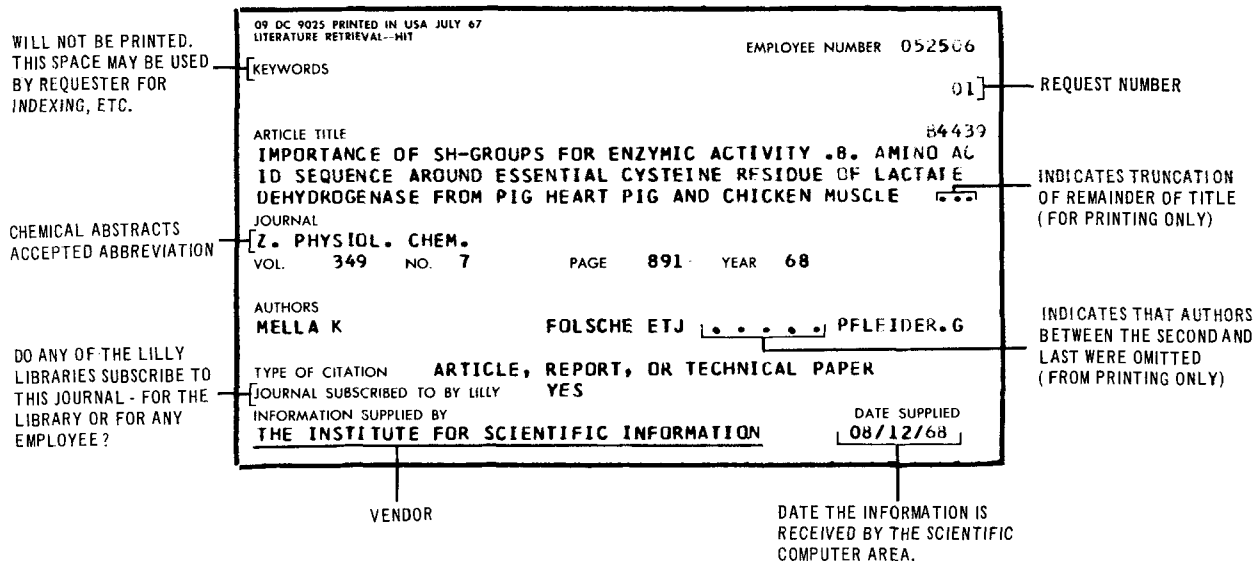

FIG. 1. ALAS bibliographic card

user and request numbers, on a temporary tape. A message indicating this accompanies the user's hit statistics on his header card. The temporary tape is kept for one week. Any user who desires a listing of the rest of his hits must so indicate by returning his header card to the computer area before the next week's run.

• Conclusion

ALAS is strictly a current altering system. We have no programs to retrieve data from previous tapes, and we do not have any definite plans to develop this area. However, we are keeping the bibliographic data for possible future use in a retrospective search system.

Although the system functions in a reasonable amount of time, several improvements can be made. These would reduce run time and possibly increase the facilities of the system.

At this stage, index sequential files are used to access the bibliography files. Provided enough direct access file space is available, a direct access method could be used to reduce the access time from two seeks, three reads to one seek, one read per record. The bibliography file could be organized sequentially by source article number and the address computed from the source article number for retrieval.

Currently, the printer uses most of the computer run time. There is essentially no overlap. By sharing a file between two core partitions under IBM/360 DOS, printing could be done in one partition and searching in another. Suitable communication techniques would have to be worked out between the partitions. It appears that searching is fast enough to queue answers and stay ahead of printing.

The method of formatting requests could be altered to facilitate user creation and change without the intervention of the senior librarian. The easiest method of accom-

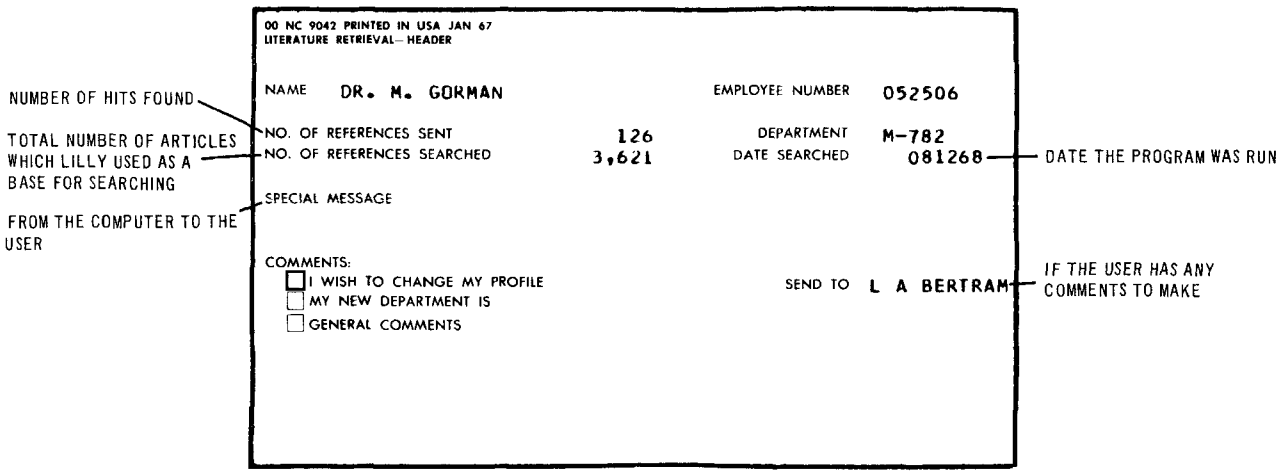

FIG. 2. ALAS header card

plishing this might be to design an easy-to-use, easily understood request format and to write a program to translate it to the present internal code. Another approach might be to reexamine the entire request processing and adopt a more efficient design.

User response to the system after the first nine months of operation was very good. Table 1 shows the results of a survey which included 57 users. Forty-one (72%) of the users scanned an average of two sources in addition to using ALAS. Of these 41 people, 21 (51%) said that they found titles in other journals which were missed by ALAS; however, the number seems to be small. (A few users reported finding one to two in a nine-month period.) Twenty (35%) of the users replied that information they received as a result of getting these citation cards altered some research plans in the following ways:

1. Presented new techniques and methods which increased speed, efficiency, and accuracy of work.
2. Gave new ideas for research projects.
3. Promoted a shift in emphasis from one aspect of a problem to another.
4. Aided in interpretation of optical rotatory dispersion (ORD) and circular dichroism (CD) data.

TABLE 1. ALAS survey

Question	Reply	Number	Percent
1. Does the ISI service significantly reduce your dependence on other means to obtain current references?	Yes	49	86
	No	8	14
2. The subject content of a paper is not always suggested by its title. Has this weakness of titles caused you to miss important references?	Yes	9	16
	No	16	28
	Not sure	31	54
	No answer	1	2
3. On the average, how many citations do you receive each week?	Range	2 to 400	
	Average	71	
Do you think this is too many?	Yes	10	17
	No	46	81
	No answer	1	2
If so, could your keyword profile be altered to reduce the number without losing pertinent titles?	Yes	7	70
	No	2	20
	No answer	1	10
4. Of all the citation cards you receive each week, approximately what percent do you look up?	Range	1% to 80%	
	Average	18%	
Of these, approximately what percent is			
a. of use in your current work?		46%	
b. of no current use but required for professional growth?		46%	
	No answer	8%	

One of the comments received most often was that using ALAS reduces significantly the time required to keep abreast of current literature. Some users find that it flags references in obscure journals which are not covered in other abstracting services.

The system works well when interests can be defined in descriptive terms; however, this is not always possible. For example, natural product names invented by authors cannot be picked up. Other language limitations have caused a user interested in image display to get hits like "Body *Image* Distortion of Obese Women," (*Psychosomatic Medicine*) and "*Image* of the African Medicine Man." When an author's name appears in a title, it is indexed as a word and not as an author. This has caused hits like "Dr. Coating's Obituary."

Most of the problems mentioned, such as "nonsense" hits and missing important hits, are inherent in automatic literature retrieval systems, especially when the system is limited by its dependence upon the author's ability to create a descriptive title. However, we feel that by spending a little time and effort in constructing profiles, one can minimize these problems and obtain a reasonably comprehensive bibliography.

• Acknowledgments

We would like to thank all those who participated in the design and development of the literature alerting system, particularly the following: Literature Task Group, Eli Lilly and Company; Dr. E. P. King, Eli Lilly and Company; Dr. Robert Korfhage, Purdue University, Lafayette, Indiana; Mr. William Muirhead, Eli Lilly and Company; and Dr. C. N. Rice, formerly of Eli Lilly and Company, presently with National Library of Medicine, Bethesda, Maryland.

References

1. RICE, C. N., A Computer-Based Alerting System for *Chemical Titles, Journal of Chemical Documentation,* 5:163–165 (1965).
2. GARFIELD, E., *Science Citation Index*—A New Dimension in Indexing, *Science,* 144:649–654 (1964).
3. GARFIELD, E., and I. H. SHER, ISI's Experience with ASCA—A Selective Dissemination System, *Journal of Chemical Documentation,* 7:147–153 (1967).