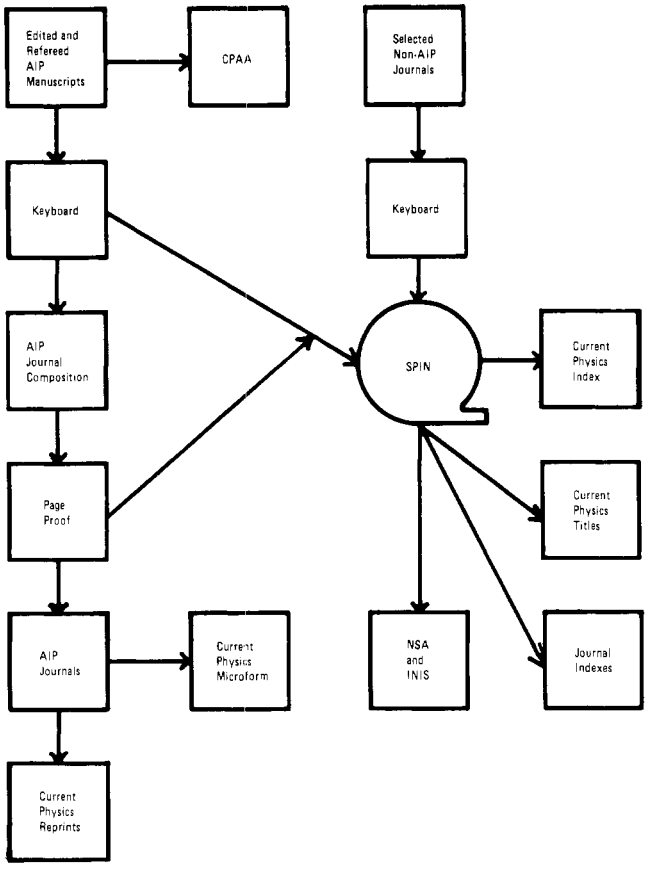


**Figure 1.** AIP Physics Information System.

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.

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

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

LEADERMART from the commercial market. Currently we have no access to LEADERMART.

The on-line systems that we are using now are those available from SDC, Lockheed, and the Chemical Data Center. We are interested primarily in scientific data bases, but we are using several business-type data bases. Our primary file is, of course, *Chemical Abstracts Condensates*, but a fair amount of use is made of MEDLINE, CAIN, and NTIS. In addition, we have used INSPEC and COMPENDEX occasionally. The two business files that we are using actively are INFORM from Abstracted Business Information, Inc., and the *Chemical Market Abstracts* from Predicasts, Inc.

Now, I would like to make some comparisons between searching on-line and searching in either a batch or a manual mode. On-line data bases are normally very current, but are limited in the size of the back files. For example, the Chemcon file from SDC often has data on-line before our library receives the printed copy of *Chemical Abstracts*. Most of the systems that we are using have about two to three years of back files on-line. The major exception is the NTIS file which covers the literature from 1964 to date.

Printed annual or cumulative indexes are usually slow to issue, and it is a cumbersome process to search through weekly or monthly index issues manually. Batch searches can be used to perform searches very similar in results to the on-line system, but they normally cover a larger back file of material.

When searching on-line systems, it is very easy to look under many terms and to intersect a variety of concepts. This facilitates complex searches. It is difficult to intersect manually a variety of ideas which require searching on multiple terms and a large expenditure of time. Multiple lookups and complex searches can be performed in batch mode also.

On-line systems are much faster than manual searches if the question is complex. Manual methods of information retrieval may be faster, if the question is well defined and requires only a few references or looking in only one place in an index. The batch method is fast, but one must wait for the run to be carried out, and this often involves a wait of anywhere from one day to one week or longer.

One of the major advantages of on-line data bases is that iterative searching is possible. This means that the search strategy can be changed as the search progresses. If very few or no answers are obtained, it is possible to broaden or expand the question and retrieve additional references. If you retrieve several hundred references and do not want that many, it is possible then to narrow your search further and restrict the number of hits received. Iterative searching is not possible in the batch mode. In batch searching, normally a question is submitted, the strategy worked out, and the question run against the file. The results are evaluated and, if they are incorrect or off the subject, the question must be reformulated and submitted for the next batch run. Thus, the ability to change search strategy while the search is being performed is a very important aspect of on-line searching.

In searching on-line the output can be sorted to suit your needs. The references can be sorted in date order, alphabetically by author, or possibly by items which have the most hits. This sorting capability can be very useful when preparing a bibliography or literature search. On-line searching usually does not require a double lookup as is the case in most manual systems. After the number of postings are given for the index terms, the complete bibliographic data can be printed out on-line. Often the abstract is available on-line also. When doing a search manually with a printed index, first the index is checked and the abstract reference number obtained. Then the reference is looked up in the printed section of the abstracting and indexing tool.

On-line searches can be run also by the scientist, and this obviates the need for an intermediary. This can be very important because the scientist is normally a specialist in the area of interest and has a knowledge of the words that best define the question. Batch searches can only be handled by an intermediary and, of course, manual searches can be performed in either a delegated or nondelegated mode.

On-line searches can usually be carried out the same day that they are requested, depending on whether the files are up all day or only in certain windows. For instance, at Lockheed most files are available all day, five days a week, while SDC data bases are up in specified three-hour windows that begin at eight o'clock in the morning. Batch searching, of course, must be performed when the system is scheduled, and then there exists the possibility of long delays. Manual searching can be done immediately if the printed indexes are available on site.

The cost for on-line searching depends on the amount of connect time which is used. Connect time is calculated from the time the search service is dialed until the telephone connection is broken. The connect time is a function of the complexity of the question, the number of other users who are simultaneously accessing the system, and the number of citations printed on-line. The complexity of the question affects the connect time since more terms must be input and this takes more typing time. Actually, the time that one spends inputting the question is probably greater than the time that one spends getting answers back from the computer, because it is much slower when you are typing than when you are receiving. We have noticed this particularly on tape recordings we have made of searches. While typing, one thinks he is working rapidly while in reality the typing is very slow compared to the output from the computer. We normally conduct all our searches at 30 cps because it is faster and more economical than at 10 cps.

For the on-line operation at Rohm and Haas, two people were trained to operate the terminal and currently a third person is being trained. Our terminals are a portable typewriter-type at our Spring House location and a cathode ray tube at our Bristol location. Charges are billed back for our out-of-pocket computer expenses, and the charges for the information scientist are billed back also if one hour or more is spent.

Searches are available to all company personnel. Requests are received by phone or mail. After the request is received and analyzed, the search is discussed with the requestor. If possible, we try to have the requestor present during the on-line searching. In this way, the requestor can understand our mode of operation better and can decide also on the relevancy of the search results and suggest possible modifications.

One way we have devised to make it possible for people from other sites to be present during an on-line search, without physically driving to Spring House, is to set up a conference call. By setting up such a telephone call, my terminal, the user's terminal at a remote location, and the central computer can be connected together. In this way the items that are being sent to the computer and the information the computer sends back to my terminal are printed out also at the remote terminal. To facilitate this, full duplex mode is used. At the same time, I am connected by another telephone line to the person who is observing the search remotely and we can discuss the search as it proceeds and decide whether the results are relevant. We have found this technique to be very useful and to save the time that would be required to travel to Spring House when the subject is difficult to define.

When a request is received, it is discussed with the requestor. It is ascertained which data bases he would like to have searched. The key terms are discussed with the requestor and it is ascertained which terms he feels would be the best ones to use in the search. The requestor is asked

the time span to be covered. After determining these points, a search strategy is formulated.

When working with a free text type of system, such as *Chemical Abstracts,* the following information must be considered in formulating the search strategy. Chemical compounds must be searched for by all their synonymous chemical names, as well as any trade names. Word endings such as s, ing, ed, etc., must be considered, too, since the computer only searches for the character string exactly as it is entered. In addition one must enter the abbreviations that are used by Chemical Abstracts Service and also the spelled-out word, because CAS is not consistent in their use of abbreviations in the keyword index.

To arrive at these abbreviations, synonyms, trade names, etc., a number of tools are used. First, the *Index Guide* which is a portion of *Chemical Abstracts'* volume index is consulted. This is very useful for finding trade names and other synonymous chemical names for the compound being sought. Another item used is the microfiche put out by System Development Corporation, which lists all of the keywords found in about three volumes of *Chemical Abstracts Condensates.* In addition to the alphabetical listing of keywords, there is also a frequency listing that indicates how many times each keyword has appeared in these volumes. A third tool that we use, but that is no longer in print, is the *Search Guide,* from Chemical Abstracts Service. The *Search Guide* is structured like a thesaurus and is very helpful in coming up with additional synonyms or related terms. In addition, other tools, such as the *Merck Index, Roget's Thesaurus,* Frear's *Pesticide Index,* etc., are consulted.

After deciding on the terms that should be included in the search strategy, the Boolean Logic statement is prepared. Also utilized, if required, is the full text searching capability which can search for words or character strings in titles, abstracts, or other portions of the record. With the requestor present, either in person or *via* conference call, the search is run according to the predetermined plan. If the results appear to be relevant, we continue. Otherwise, the search strategy is modified to give the desired results. After obtaining the number of postings, several of the entries are printed on-line with their associated keyword index phrases. By examining the keyword phrases, it is possible to determine if there are other meaningful words which have not been used in the search strategy and which might be useful. The keyword entries are useful also in forming a "mini-abstract" that is of additional help to the requestor in deciding if a reference is pertinent or not. Since *Chemical Abstracts Condensates* does not contain abstracts, it is necessary to consult either the printed abstract in *Chemical Abstracts* or the original material if pertinency cannot be determined from the title and the keyword phrases.

If new terms are found that appear to be useful in the search, these terms are built into a revised search strategy. This cycle may be repeated more than once. The output is normally printed on-line if there are less than twenty hits. If more than twenty hits are obtained, they are usually printed off-line. Off-line prints are usually received by mail in two to three days from California and are more economical than on-line printing unless the search results are needed immediately. We have found that with *Chemical Abstracts Condensates* two to three references can be printed on-line per minute depending on the number of fields printed. Off-line references are printed at a cost of $0.08 each.

When searching in a system with a structured or hierarchical vocabulary, the search strategy is slightly different. We have acquired, in hard copy, a number of thesauri, such as MESH, the Predicast Hierarchical Dictionary, and SHE. The question, as posed by the requestor, is translated into the appropriate subject headings which are found in the thesaurus. After determining these headings, the search strategy is prepared, and it is decided if it appears worthwhile to STRINGSEARCH or full text search on titles and/or abstracts. When searching the index terms, the first several hits are printed on-line to determine if there are other equally useful headings that may have been missed in formulating the search. Otherwise our approach is very similar to the one just described with *Chemical Abstracts Condensates.*

After establishing on-line searching capabilities at Rohm and Haas, a variety of techniques were utilized to introduce it to our personnel. Seminars were given at our metropolitan Philadelphia locations and written notices announcing the service were distributed. Also notices are placed in our monthly library acquisitions bulletin describing new data bases as they become available. In addition we rely heavily on word of mouth advertising from satisfied users.

I am now spending about one third of my time performing on-line searches and attendant tasks. This includes operating the terminal, formulating strategies, keeping up with new data bases, reading manuals, and the like.

Average search times in *Chemical Abstracts Condensates* run about fifteen minutes and cost about a dollar a minute plus $0.08 for each reference printed off-line. Searches for compound types may possibly run slightly longer and may be more complex, since one must consider the variety of ways a chemical may be named.

Searches in the NTIS file tend to run slightly longer than those in *Chemical Abstracts Condensates.* This is due to the necessity of displaying the alphabetical listing of the vocabulary on-line to determine which indexing phrases have been used and are relevant to the search. NTIS searches probably average close to twenty minutes in length, but they cost about the same as *Chemical Abstracts Condensates* searches since the charges are less for this file. The average NTIS search cost runs about $15 and $0.10 is charged for each citation printed off-line, if the abstract is included.

Normally searches are run *via* WATS line if it is free. Tymshare is used when the WATS line is unavailable. WATS line usually is available to us only early in the morning or at the lunch hour. The use of Tymshare's communication network adds an additional $10 per hour to the search costs.

Our on-line searching capabilities have been used to develop SDI profiles. The strategy is worked up on-line and the profile run against the last update for the data base. This is repeated with several updates until the strategy appears to have been shaken down and to give good results. The output from each update run is given to the requestor for checking and suggestions as to possible changes. The search results are examined also by the information scientist. When the profile seems to be satisfactory, it is submitted to a center, such as the University of Pittsburgh or The Institute of Paper Chemistry, for routine running on an SDI basis. The first results received from the outside center are normally checked against the results that we receive for that update on our terminal. In checking, only the abstract numbers are printed at our terminal and these are checked against the center's results to see that they are present.

We have found that we have obtained very good service from all of the contractors that we are using for on-line searching. Prices seem to be comparable and there are only minor differences in the searching capabilities. One of the major differences is that Lockheed has most of its files up all day, while SDC has windows during which certain data bases are available only. Lockheed has full text searching capability and SDC has what is called STRINGSEARCH and SENSEARCH capabilities. These options are not exactly comparable but, in some instances, may be used to achieve the same end result.