Pure & Appl. Chem., Vol. 49, pp. 1833 - 1844. Pergamon Press, 1977. Printed in Great Britain.

BATCH COMPUTER SYSTEMS FOR RETRIEVING CHEMICAL INFORMATION FROM TEXT FILES

Margaret K. Park

Office of Computing Activities, The University of Georgia, Athens, Georgia 30602, USA

Abstract - The major steps involved in conducting a batch-oriented computer-based search of bibliographic data bases are: formulation of a statement of the information need, selection of pertinent data bases, preparation of search profiles, conduct of the search, analysis of the initial search results, and revision (if necessary) of the search profiles for subsequent searches. Important aspects of the profile preparation step include identification of the major ideas or concepts; the establishment of alternative search strategies which represent the co-occurrence of the concepts as they may appear in titles, abstracts, or indexing of published documents; expansion of the concepts using the indexing or natural language terminology as it appears in the data base; construction of the appropriate logic statement; and, optionally, assignment of weights to sequence the bibliography into a logical order or to extend the logic capabilities of the retrieval system. The types of services available from batch-oriented retrieval systems include current awareness searches (SDI), retrospective searches, macroprofiles, and computer-readable subfiles.

Batch-oriented computer systems for searching text files in chemistry, as well as other disciplines, have been in use now for over a decade. A great deal of expertise has developed in the preparation and refinement of search profiles, in the exploitation of the indexing vocabulary of the data bases, in the design of sophisticated profile aids and retrieval techniques, and in the education and training of both users and information specialists. Many papers have been published over the years by information specialists at the various European and American centers which describe the procedures used, the aids developed, and the search techniques devised. These centers have also published search manuals, or profiling guides, in which they describe the steps involved in preparing an effective profile, the characteristics of the data bases against which the searches are to be made, and the special features of individual computer-based systems which are available for optimizing retrieval results.

The purpose of this paper is to place in perspective the principal characteristics of batch-oriented text retrieval, as seen from the scientific user's point of view. The emphasis is placed on the functions that have to be performed, and why, with comparison and contrast to the more familiar manual reference searching. Little or no attention has been given to the intricate details of computer program manipulation or to nuances of individual search strategies and retrieval techniques. Information specialists and computer scientists interested in these detailed aspects of batch retrieval systems can find relevant reports in the literature and in manuals prepared by information dissemination centers. Although much of the discussion applies equally well to on-line (interactive) retrieval systems and to numerical data or chemical structural data bases, the focus has been restricted to batch-oriented bibliographic retrieval systems.

COMPUTER-BASED VS MANUAL RETRIEVAL

Even though the electronic computer holds much less mystique for researchers and academicians in the physical sciences than for their colleagues in the so-called "soft sciences," there is still a reluctance on the part of many to confront the unknown computer in the context of computer-based retrieval, particularly if it is not a familiar tool in other aspects of their work. This hesitancy often surfaces in comments like: "I wasn't sure what the computer would need"; "If I had known it was this simple, I'd have used it before"; or "It's a lot like library searching, isn't it?". A search of the literature, regardless of whether it is to be done using printed reference works, computer-based retrieval systems, or a combination of both, does require basically the same operations and considerations. The differences between manual searching and batch-oriented computer-based retrieval are more often related to when and how the functions are performed, and to what extent, rather than whether or not they occur.

Every literature search evolves out of some need for information to answer a question; and the formulation of the appropriate question, or request statement, is essential to both manual and computer-based retrieval. The main difference between the two approaches to retrieval lies in the formality and exhaustiveness with which the initial question is phrased. In manual search of printed reference tools such as the *Chemical Abstracts* indexes, the researcher can explore alternative index entries under various subject headings by turning a few pages and can scan several columns of the index until the headings used to describe the subject of interest are discovered. The question may also be revised several times based on the types of entries found during the scan. Computer-based systems, particularly batch-oriented systems, do not readily accommodate these trial explorations in the data base itself. Hours, days, or even weeks may intervene between computer searches of the individual issues or volumes, thus spreading the search out over a long period of time. The efficient and effective use of batch-oriented computer-based retrieval places a premium on careful and thorough formulation of the information need before beginning the search--what information specialists call "negotiating the question." Thus, formulating the question or information need occurs in both types of retrieval; but in manual searching it is often integrated into the search process itself, while in computer-based retrieval it needs to occur prior to initiation of the actual search against the data base.

Another area in which there are both similarities and differences between manual and computer-based retrieval relates to the language, or indexing terminology, used to identify and retrieve relevant references. In manual searching, the most important idea or concept is subjectively selected as a starting point, and the appropriate page is located in the alphabetically arranged index or card catalog. If the term or name initially used for the concept happens to correspond to the terminology used in the index or catalog, the exploratory search can begin. If not, other alternatives may be investigated, including transfer to other places in the index or catalog based on cross-references, scope notes, use of synonymous or closely related words for the concept, or the selection of other ideas or concepts in the original question. Most traditional printed reference works are organized to facilitate this progression from one alternative to another, until a successful approach is encountered. At this point the actual index entries, abstracts, or citations can be examined. With batch computer searching, trying each of these approaches in turn on successive searches of the data base requires far too much elapsed time (and expense). Consequently, several approaches are normally used in constructing the initial profile. Authority lists, thesauri, word guides, and similar search tools must be used to determine the preferred terms or names by which concepts and chemical substances have been identified. All of the concepts occurring in the question are used as possible search points, tied together with logic statements which control the way they are to co-occur in pertinent documents. Frequently, the initial profile will also reflect the relative importance of the various concepts in the question by including several strategies, or through use of numeric values assigned to the search terms on the basis of their relative importance. As was the case in defining the search question, batch-oriented computer-based retrieval involves the same basic practices and procedures as does manual retrieval, but places a premium on using all of them in constructing the search profile before beginning the search.

A third area in which manual and computer-based retrieval can be compared and contrasted is the analysis of identified references or citations for their relevance to the question. When using printed indexes and abstract journals, it is quite common to scan the entries as part of the process of locating the appropriate index terms, then to narrow the focus to include the secondary concepts of the question. In manual searching there is a tendency to think of retrieved references as those references important enough to be copied off to cards or notepaper for further investigation. Thus, the results of the search seem to be at or near the one hundred percent relevance level because of the intellectual judgment applied in selection of relevant references. All the index entries or citations scanned and dismissed as irrelevant are seldom considered false retrievals since they were never copied to the notepaper (i.e., "selected" into the retrieved set). These expectations of near perfect retrieval have been carried over into the computer-based retrieval environment, often resulting in disappointment when the computer system retrieves references which are obviously irrelevant to the question. It would be far more realistic to view the computer-based retrieval system's task as selecting a reasonable subset of the whole corpus for subsequent human evaluation. The intellectual task of scanning possible entries or citations for those which are directly relevant is essential to both manual and computer-based searches. The difference is that the intellectual judgment occurs *before* reference retrieval in manual searching and *after* reference retrieval in computer-based searching. The key to effective computer-based retrieval comes in obtaining the optimal subset to be scanned to achieve a workable bibliography, a goal which usually requires analysis of the initial search results so that the profile can be revised, if necessary.

With this brief comparison of manual and computer-based retrieval as background, attention will now be directed to the major steps involved in constructing and revising search profiles for batch-oriented retrieval systems. Emphasis has been given to the purpose of these steps, rather than specific details associated with individual search systems or subtle differences among data bases. Researchers who wish to prepare their own profiles will want to obtain the manuals and documents from the appropriate information dissemination centers. However, many

centers have information specialists, profile analysts, or reference librarians on their staff to handle all of the intricate details of profile preparation. In such cases, a general familiarity with batch computer-based retrieval at the level described in this paper is all that is required to work effectively with the centers' staffs.

In the discussion which follows, a hypothetical research study has been used to illustrate the steps involved in preparing profiles and in obtaining search results. The example was constructed specifically for this purpose, and any similarity with actual searches performed by any information dissemination center is strictly coincidental. The research study and associated information query have also been formulated to illustrate retrieval problems which can subsequently be accommodated through profile revision. Experienced information specialists would be aware of such situations at the outset and take them into consideration in preparing the initial profile. However, for purposes of the illustration they are shown as part of the revision process.

MECHANICS OF COMPUTER-BASED RETRIEVAL

The mechanics of computer-based retrieval can be described as a series of six major steps or functions, beginning with the identification of the information needed from the literature. Subsequent steps include the selection of the data bases which are likely to include the relevant information, the preparation of one or more profiles to be used to search these data bases, and the conduct of the computer search (usually by an information center). The bibliography obtained from the search may be immediately useful for the research project, or revision may be needed for improved relevance of the retrieved references on subsequent search. This revision is particularly necessary when large numbers of irrelevant references are being retrieved which could be eliminated by judicious modifications to the profile, or when no relevant references are retrieved. Most centers operating batch-oriented retrieval systems expect to make at least two or three revisions to a profile before it is "fine-tuned" for continued processing, for current awareness, or for extensive retrospective searches.

Identify the information need

Identification of the information needed is, perhaps, the most important step for effective retrieval. It is also one of the most difficult steps, as reference librarians have long been aware. The difficulties seem to arise in differentiating between the research study per se and the specific problems for which published reports are desired. The research study and the information needs are obviously closely related, but they are seldom identical. A description of the research study might be stated as "I am working on a study of ...", while the statement of an information need would begin "I need to know ..." or "What is already known about ...." For example, researchers investigating the toxic effects of mercury and lead in bodies of water might have any number of information needs related to this study, such as: potential sources of mercury and lead, techniques for determining mercury and lead in aqueous solutions or in animal tissue, toxicity data for these elements in various species of animals and plants, the progression of lead and mercury through the deposition and food chains, and many other possibilities. The information needs arising out of a research study may also vary, depending on the background of the researchers as well as the phase of the study in which they are involved. At the beginning of the study, the research team may want general or review articles which provide a global view of the state-of-the-art, while at later stages in the work they may be more concerned with techniques, procedures, or specific data.

In identifying the information needs before beginning the search, it is usually advisable to prepare a written statement describing the search interests. This can be done by completing the introductory phrase, "I am interested in articles (books, patents, etc.) about ...." This initial statement is often fairly broad in scope and can be conveniently circumscribed by additional sentences beginning something like "Specific topics of interest are ...." The illustrative request statement which will be used as the example throughout the remainder of this paper is "I am interested in articles related to the toxicity of mercury and lead and their compounds as pollutants in natural bodies of water. Specifically, I am interested in reports on the toxicity of these chemicals to shellfish."

Select the data bases

The second step is to decide which data bases include within their coverage articles likely to be pertinent to the information need. If only key papers are required, this selection may be simply a matter of selecting the single most relevant data base for the search. However, when an exhaustive search is required, it is usually necessary to search several data bases for questions which relate to two or more discipline areas. Data bases may also differ in their coverage of different forms of published literature. Some cover patents, while others do not. Some include technical reports, while others either do not include technical reports or include only certain types of reports. Many of the available data bases, of which there are several hundred now, correspond to printed reference works which are already familiar to users of the traditional library resources. But, there are also data

bases for which there are no corresponding printed abstracting or indexing services, as well as data bases which include more references in the machine-readable form than in the printed form. Most of the large abstracting and indexing services provide descriptions of their selection policies, which are very helpful in making judgments as to the literature coverage. These should be consulted whenever there are questions about the coverage of a data base related to a particular search question.

The search request on the toxicity of mercury and lead as pollutants affecting shellfish obviously overlaps the fields of biology and chemistry. Thus, two data bases very likely to be relevant to this search are CA-CONDENSATES, corresponding to *Chemical Abstracts*; and BIOSIS PREVIEWS, corresponding to the printed publications *Biological Abstracts* and *BioResearch Index*. However, there are also other data bases which may be of interest. COMPENDEX and INSPEC both cover the engineering literature and could be expected to have references to papers dealing with pollution, especially in the applied environmental applications areas. Also, the NTIS Bibliographic Data File--corresponding to the publication *Government Report Announcements* which announces U. S. government reports--could be expected to include pertinent technical reports. There are also other smaller and more specialized data bases, such as POLLUTION ABSTRACTS and OCEANIC ABSTRACTS, which could be considered for exhaustive coverage. For purposes of illustration, four data bases have been used: CA-CONDENSATES, BIOSIS PREVIEWS, COMPENDEX, and NTIS.

Prepare the profile(s)

Once the information need has been identified and the pertinent data bases selected, attention can be turned to preparing the search profiles which will be used to retrieve references via the computer-based retrieval system. A search profile, in this context, is a list of the index terms or classification codes for the search question which are to be matched against the corresponding index terms or codes for the documents on the data base. Logic operators, such as "and" and "or," are also used by most computer-based retrieval systems to control the manner in which the terms or codes co-occur in the documents.

The preparation of the search profiles can be viewed as consisting of five principal steps. First, the statement of the information need is analyzed to identify the major concepts, or ideas, which are to be used as retrieval keys. Secondly, one or more search strategy statements are constructed, which represent the types of document references to be retrieved. In the third step, the concepts occurring in these strategies are converted into the nomenclature or codes of the various data bases--what has been called "indexing the question." The fourth step is to construct a logic statement, which describes the manner in which the indexed concepts are to occur for document references to be retrieved. And, finally, most retrieval systems have some type of weighting scheme which can be used to control the sequence of the bibliography, to refine the logic capabilities, or both.

Identification of the major concepts is primarily a matter of identifying the important nouns and noun phrases in the statement of the information need--another reason for reducing the information request to writing. For the example, the major concepts are toxicity, mercury and lead as elements and in chemical substances, pollution, and shellfish (Fig. 1). There

```
1  TOXICITY & (HG or PB) & SHELLFISH
2  (HG or PB) & POLLUTANTS & SHELLFISH
3  (TOXICITY or POLLUTION) & AQUATIC ENVIRONMENT
4  (HG or PB) & SHELLFISH
5  (TOXICITY or POLLUTION) & SHELLFISH
6  WATER POLLUTION & (HG or PB)
```

Fig. 1. Possible search strategies for the illustrative search question.

are quite a number of possible search strategies which can be constructed from these concepts. Certainly, papers reporting the toxicity of mercury or lead (and their compounds) to shellfish would be of interest, as would papers which report the pollution effects of mercury or lead on shellfish. Generalizing the retrieval to consider mercury or lead toxicity, or pollution in the aquatic environment would undoubtedly retrieve papers which reported data on other organisms, such as fish or waterfowl, as well as shellfish. Similarly, there may be articles which report the uptake of mercury or lead by shellfish without specifically mentioning toxicity or pollution. There may also be general papers dealing with toxicity or pollution effects in shellfish caused by a wide range of chemicals (e.g., pesticides), without specifically naming mercury or lead in the title or abstract. These chemicals may also be referred to as members of the class of heavy metals. The last search strategy shown is indeed a general one and one which would retrieve extremely large numbers of references--water pollution due to mercury or lead. This strategy is included primarily to illustrate the outer bounds of generalization from the specific statement of the information need. Unless exhaustive recall were essential, it would not ordinarily be advisable to generalize the retrieval request to this extent. It would retrieve large numbers of references of very

marginal interest, at best, while probably picking up very few references which would not have otherwise been retrieved by some of the more specific strategies.

When concepts have been identified and the search strategies determined, the next step is to expand each of these concepts into the terminology used in the data base to represent the same idea. This is the point at which differences between data bases become apparent, since the abstracting and indexing services use different terminology in their data bases, just as they do in the corresponding printed indexes. To illustrate the expansion of the concept into the appropriate terminology for each of the four data bases, consider the toxicology concept.

As shown in Fig. 2, there are a number of terms which have the root word "toxic." Most retrieval systems have a convenient shorthand method for indicating word stems, illustrated

TOXIC		ILL EFFECT*
TOXICANT		BAD EFFECT*
TOXICATION		UNTOWARD EFFECT*
TOXICITY		LETHAL
TOXICOGENIC	TOXIC*	HAZARD*
TOXICOLOGICAL		
TOXICOLOGICALLY		CA Section 004
TOXICOLOGY		CA Cross Ref. 004
TOXICOSIS		
POISON		BIOSIS Cross Code 22506
POISONING	POISON*	
POISONOUS		NTIS COSATI Code 57Y

Fig. 2. Term expansion of the toxicology concept.

by the use of the asterisk following the stem (i.e., toxic*). The asterisk (or its equivalent character, depending on the search system) indicates that any character will be accepted in the designated position. Thus, the stem will match any of the terms shown to the left. Most retrieval systems also have the capability of indicating other types of stemming--usually called "truncation"--including prefix (e.g., *mercury to match "dimethylmercury"), suffix (e.g., merc* to match mercurial), and combinations (e.g., *terpen* to match sesquiterpenoid). Truncation is particularly helpful for search terms which are to be matched against the words which occur in titles or abstracts, since many different forms may have been used. However, a word of caution is advisable. Truncation can sometimes be very dangerous, causing matches against words which have the designated stem but which are not relevant to the concept. For example, the use of "sea*" in the list of terms related to the aquatic environment concept would retrieve not only articles dealing with sea, seas, seawater, and seashores, but also sealants, seams, search, seasons, and seats (among others). Word frequency lists and truncation guides, such as the Key-Letter-In-Context publications produced by UKCIS and other organizations, should be consulted by the information specialists or researchers who prepare the profiles before using truncation on short word stems.

A number of the data bases also have classification codes or controlled indexing vocabularies, which can be used as search terms in the profiles. CA-CONDENSATES, for example, includes the CAS section numbers; and these numbers can be used as search terms where they correspond to a concept in the search question. Section 4 of *Chemical Abstracts* covers Toxicology, and the code 004 can be used as a synonym for this concept when included as the section number and as a section cross reference. In the BIOSIS PREVIEWS data base, CROSS Codes are used to represent broad subject categories. CROSS Code 22500 applies to the broad field of Toxicology, and there is a more specific code (22506) which applies to environmental and industrial toxicology. The COSATI Codes included in the NTIS Bibliographic Data File are very broad subject classifications, which can also be used as search terms when appropriate. For example, Toxicology is represented by the code 57Y under the broad classification of medicine and biology (57).

Expansion of each of the other concepts in the question proceeds in the same manner, using a combination of word stems, classification codes, and controlled vocabulary. A completed term list for the sample question is shown in Fig. 3. The group numbers at the left tie together the synonymous or closely related terms and codes for each concept. These group numbers are also used in the logic statement to indicate the way in which terms from these groups should occur in the data base in order to retrieve a reference. For the example, several strategies have been combined. The logic statement indicates that at least one term from the toxicity concept or the pollution concept, and at least one term related to either mercury or lead, and at least one term related to shellfish or the aquatic environment must be assigned to a document in the data base before it will be retrieved in the bibliography. That is, relevant documents will be indexed with at least three search terms from the profile --one related to toxicity or pollution, one related to mercury or lead, and the third related to shellfish or the aquatic environment.

GROUP	TERM	TYPE	WEIGHT	TERM
G001	1	TXT		TOXIC*
G001	2	CXC		22506*
G001	3	CAS		004*
G001	4	CCR		004
G001	5	COC		57Y
G001	6	TXT		POISON*
G001	7	TXT		ILL EFFECT*
G001	8	TXT		BAD EFFECT*
G001	9	TXT		UNTOWARD EFFECT*
G001	10	TXT		LETHAL*
G001	11	TXT		HAZARD*
G001	12	EIT		INDUSTRIAL POISON*
G002	13	TXT		POLLUT*
G002	14	TXT		CONTAMINA*
G002	15	CXC		37015*
G002	16	COC		68D
G002	17	CAS		060*
G002	18	EIT		WATER POLLUTION*
G003	19	TXT	1000	MERCUR*
G003	20	TXT	1000	HG
G003	21	EIT	1000	MERCURY COMPOUNDS*
G004	22	TXT	500	LEAD
G004	23	TXT	500	PB
G004	24	TXT	500	PLUMBATE
G004	25	TXT	500	PLUMBIC
G004	26	TXT	500	PLUMBO*
G004	27	TXT	500	PLUMBYL*
G004	28	EIT	500	LEAD COMPOUNDS*
G005	29	TXT	50	SHELLFISH
G005	30	TXT	50	SHRIMP
G005	31	TXT	50	PRAWN*
G005	32	TXT	50	LOBSTER*
G005	33	TXT	50	MALACOSTRACA
G005	34	BTC	50	75112
G005	35	TXT	50	CLAM
G005	36	TXT	50	CLAMS
G005	37	TXT	50	OYSTER*
G005	38	TXT	50	PELECYPODA
G005	39	BTC	50	61500
G006	40	TXT		WATER
G006	41	TXT		WATERS
G006	42	TXT		FRESHWATER*
G006	43	TXT		SEAWATER*
G006	44	EIT		SEAWATER*
G006	45	TXT		SEA
G006	46	TXT		SEAS
G006	47	TXT		OCEAN*
G006	48	TXT		OCEANOGRAPHY*
G006	49	TXT		HARBOR*
G006	50	TXT		BAY
G006	51	TXT		BAYS
G006	52	TXT		ESTUAR*
G006	53	TXT		LAKE
G006	54	TXT		LAKES
G006	55	EIT		LAKES*
G006	56	EIT		INLAND WATERWAYS*
G006	57	TXT		RIVER
G006	58	TXT		RIVERS
G006	59	EIT		RIVERS*
G006	60	TXT		STREAM
G006	61	TXT		STREAMS
G006	62	TXT		RESERVOIR*
G006	63	EIT		RESERVOIRS*
G006	64	CAS		061*
G006	65	CCR		061
G006	66	COC		47D

Logic: (G001|G002)&(G003|G004)&(G005|G006)

Fig. 3. Sample profile for the illustrative search question (using the University of Georgia profile coding conventions).

Weighted values have been assigned to several of the groups in the sample profile, which will cause the references in the bibliography to be grouped. The values assigned are arbitrary, and the specific conventions may differ with individual search systems. For example, assigning a weight of 1000 to the mercury terms and a weight of 500 to the lead terms will cause the references dealing with both mercury and lead to be listed first (a summed weight of 1500), followed by the references related only to mercury (weight of 1000) and then by the references dealing with lead (weight of 500). Since the principal interest is in articles related to shellfish, a weight of 50 has been assigned to terms in this group, thus causing articles dealing with shellfish to be listed first within each of the three broad categories related to the chemical substances.

The codes or conventions used to indicate the particular fields to be searched in the data base records will differ, depending on the computer-based retrieval system used by individual information centers. In the example, the three-letter code TXT is used to designate search terms which are to be matched against the titles, abstracts, and uncontrolled language keywords, such as those used in CA-CONDENSATES and BIOSIS PREVIEWS. The remaining acronyms correspond to specific index term fields or classification codes: BTC for the Biosystematic (taxonomic) and CXC for the CROSS codes of BIOSIS PREVIEWS; CAS and CCR for the section numbers and section cross references of CA-CONDENSATES; EIT for the Subject Headings for Engineering (SHE) terms used in COMPENDEX; and COC for the COSATI codes in the NTIS Bibliographic Data File. Specific codes for term types and conventions for recording the group and term weights should be obtained from the appropriate profile preparation manuals for profiles prepared by the researchers themselves. Otherwise, all of these details will normally be provided by the information specialists on the centers' staffs and need not be of concern, except for some general understanding of their use and meaning.

Conduct the search

Once the profile has been constructed, the next step is to actually conduct the search, comparing the terms in the profile with the terms associated with the document references on the data base. For most of the computer-based retrieval systems in use, this is a straight-forward matching of the character strings represented by the search terms with the words occurring in the titles, abstracts, keywords, or designated indexing and classification fields on the data base. It is also an exact match in most cases. Every comma, space, period (full stop), letter, etc., is significant. For text retrieval purposes, the American spelling of color and the British spelling (colour) are different. The inclusion of only one version in the profile will fail to match on the other when the data base includes both spellings in its titles or abstracts. Similarly, the word X-ray, when written with and without a hypen (e.g., XRAY and X-RAY), are different for retrieval purposes. Cases like these must be taken into account in preparing the list of search terms.

There are also other differences between individual retrieval systems to be considered when deciding which center to use for a particular search. Some centers limit the number of search terms which can be included in the profile, or limit the number of references which will be printed as answers, while others do not place such limits on their services. Some centers provide only current awareness (update) search services, while others also make available retrospective searches of the past volumes of the data bases. Suffice it to say that differences do exist among centers. Researchers availing themselves of computer-based retrieval services may want to consult several centers to determine the one most appropriate for their own needs.

Analyze the search results and revise the profile

The result of a computer-based search is a bibliography of references which satisfies the requirements of the search profile. Ideally, these references will also satisfy the initial statement of the information need. In practice, the correspondence between the information need and the bibliography is seldom perfect; and it remains for the researcher to determine whether it is sufficiently close to his needs to be useful, or if it will require some modification to improve the performance.

Figure 4 shows some typical output from the search of the odd-numbered issues of CA-CONDENSATES for volume 84. Terms which occur in the search profile have been underscored for visibility in showing the "hit" terms. The information printed as part of the reference and the specific format used in the bibliography will vary, depending on the center which runs the search. However, most references will contain the title of the article, the names of the authors, the location at which the work was done, the name of the journal and the appropriate citation (volume, issue, date, and pagination), and the reference to the abstract in *Chemical Abstracts*. The keywords and section numbers may also be printed as part of the reference, as they are in this example. For those data bases which include abstracts in the computer-readable file, it is usually possible to obtain the abstract as part of the search results, as well as the bibliographic information.

```
PROFILE # 003020-002     ALT PRF# IUPAC          SEP. 13, 1976
ACCOUNT #          THRESHOLD WT.  00000

DATA BASE SEARCHED:  CA V 84 ODD
**************************************************************************
    003020-002              SEP. 13, 1976              WEIGHT    +1500
STRUCTURAL ALTERATIONS IN FISH EPIDERMAL MUCUS PRODUCED BY WATER-BORN
E LEAD AND MERCURY
    VARANASI U;ROBISCH PA;MALINS DC;   ENVIRON. CONSERV. DIV., NATL.
    MAR. FISH. SERV., SEATTLE, WASH.
    NATURE (LONDON)(NATUAS) 1975, 258(5534) 431-2
       CA-CONDENSATES(CHABA8) 1976, 084(17) 116489; ACS COPYRIGHT
    CAS: 004003
       KEYWDS: MUCUS TROUT LEAD MERCURY
**************************************************************************
    003020-002              SEP. 13, 1976              WEIGHT    +1050
MERCURY CONCENTRATIONS IN FISH, NORTH ATLANTIC OFFSHORE WATERS, 1971
    GREIG RA;WENZLOFF D;SHELPUK C;   MIDDLE ATL. COASTAL FISH, CENT.,
    NAT. MAR. FISH. SER., MILFORD, CONN.
    PESTIC. MONIT. J.(PEMJAA) 1975, 9(1) 15-20
       CA-CONDENSATES(CHABA8) 1976, 084(05) 026577; ACS COPYRIGHT
    CAS: 004003
       KEYWDS: MERCURY FISH LOBSTER PLANKTON SEDIMENT ECOL
**************************************************************************
    003020-002              SEP. 13, 1976              WEIGHT    +1050
ACCUMULATION, TISSUE DISTRIBUTION, AND ELIMINATION OF MERCURY-203 CHL
ORIDE AND CHLOROMETHYLMERCURY (MERCURY-203) IN THE TISSUES OF THE AME
RICAN OYSTER CRASSOSTREA VIRGINICA
    CUNNINGHAM PA;TRIPP MR;   DEP. BIOL. SCI., UNIV. DELAWARE, NEWARK
    , DEL.
    MAR. BIOL.(MBIOAJ) 1975, 31(4) 321-34
       CA-CONDENSATES(CHABA8) 1976, 084(07) 039438; ACS COPYRIGHT
    CAS: 004003
       KEYWDS: MERCURY METAB OYSTER
**************************************************************************
```

Fig. 4. Sample references from the search of the odd-numbered issues of CA-CONDENSATES Volume 84.

Table 1 summarizes the results of having searched the illustration profile against approximately 9 months collection for each of the four data bases. The relevance has been judged with respect to two criteria. The first criterion is the specific interest related to shellfish, while the second has been broadened to include all aquatic organisms. The number of references retrieved for the test period ranges from 51 for the odd-numbered issues of

Table 1. Analysis of retrieved references for sample search question [(toxicity or pollution) & (Hg or Pb) & (shellfish or aquatic environment)]

Data Base	Relevant				Irrelevant		Total
	Shellfish		Other Organisms				
BA Vol. 61 & V. 62:01-08	16	21%	20	26%	41	53%	77
BRI Vol. 76:01-09	15	22%	20	29%	34	49%	69
CA Vol. 84 (Odd)	10	20%	37	73%	4	8%	51
CA Vol. 84 (Even)	0	-	2	3%	57	97%	59
EI Vol. 76:01-08	1	2%	12	23%	39	75%	52
NTIS Vol. 76:01-19	0	-	11	10%	96	90%	107

CA-CONDENSATES to 107 for the NTIS data base. The odd-numbered issues of CA-CONDENSATES (CACon) contain considerably more relevant references than do the even-numbered issues, as would be expected, since the toxicology and pollution sections occur in the odd-numbered issues (i.e., Issues 1, 3, 5, etc.). The percentage of relevant references is also much higher in CA-CONDENSATES (odd) and BIOSIS PREVIEWS (BA and BRI) than in the COMPENDEX (EI) and NTIS data bases, as would also be expected since the chemistry and biology data bases are oriented more toward the research literature, while the engineering and technical report literature leans more toward the applied technologies.

It is seldom possible with bibliographic retrieval based on titles and abstracts, or even controlled indexing, to achieve 100% relevance in the references without risking the loss of some relevant articles. If only a few, highly relevant articles are desired, then the profile can usually be made highly specific. However, broader, more exhaustive searches will inevitably have to retrieve some irrelevant references as part of the bibliography in order to locate the more general articles of interest. For the example question, omitting the group of search terms related to water and the aquatic environment will get rid of all of the irrelevant references; but it also results in the loss of 99 of the 101 general interest citations and 3 of the 42 articles dealing specifically with shellfish (Table 2). By deleting

Table 2. Analysis of retrieved references excluding the aquatic environment concept [(toxicity or pollution) & (Hg or Pb) & shellfish]

Data Base	Relevant				Irrelevant		Total
	Shellfish		Other Organisms				
BA Vol. 61 & V. 62:01-08	16	94%	1	6%	0	-	17
BRI Vol. 76:01-09	15	94%	1	6%	0	-	16
CA Vol. 84 (Odd)	7	100%	0	-	0	-	7
CA Vol. 84 (Even)	-	-	-	-	-	-	0
EI Vol. 76:01-08	1	100%	0	-	0	-	1
NTIS Vol. 76:01-19	-	-	-	-	-	-	0

this one group, the total number of references retrieved from all data bases has been cut by a factor of 10 from 415 to 41. Whether this is a desirable or undesirable revision would depend on the importance of retrieving as many references as possible dealing with the effects of mercury and lead pollution on shellfish.

Another way of revising the profile is to determine how the individual search terms are performing--whether they are retrieving relevant or irrelevant citations, or some of each. Table 3 summarizes the number of references in which each profile term occurred across all data bases for certain selected search terms. Fifteen of the 66 terms occur in none of the retrieved references (e.g., ILL EFFECT, PLUMBATE, and OCEANOGRAPHY). Such terms should be carefully examined to verify that the spelling and data element designation (DEM) are correct, after which they can either be left in the profile or removed. For example, the term OCEANOGRAPHY should be deleted since it would be picked up by the truncated form of OCEAN* which also occurs in the profile.

All of the terms in group 5 dealing with shellfish seem to be performing satisfactorily. They occur only in relevant references and predominantly in those of specific interest. However, there are a number of terms which should be added to this group based on an inspection of the keywords in the retrieved references, among them MUSSEL (and MUSSELS), SHRIMPS (an alternative plural form), and MULLUSC* (also MOLLUSCA and MULLUSK*).

The group causing the largest number of irrelevant retrievals is the last one dealing with the aquatic environment, especially the search terms WATER and WATERS where the number of irrelevant references greatly outnumbers the relevant ones. Some of the references being retrieved through use of one of these terms are probably of general interest while others are probably not relevant to the information need as phrased. Examination of these and similar references suggests that the profile could be improved by deleting the terms WATER and WATERS, replacing them with more specific terms which imply water in the environment (e.g., AQUATIC, COAST, COASTAL, BEACH, BEACHES, MARINE, OFFSHORE, etc.). A few relevant

Table 3. Analysis of retrieved references for selected search terms

Search term	Relevant		Irrelevant
	Shellfish	Other Organisms	
TOXIC*	3	13	32
22506*	31	41	73
004*	9	34	3
57Y	0	2	12
POISON*	2	1	5
HAZARD*	0	0	16
POLLUT*	13	46	151
CONTAMINA*	3	7	21
37015*	28	41	71
68D	0	9	53
060*	0	0	43
WATER POLLUTION*	1	10	24
MERCUR*	37	84	139
LEAD	9	29	146
PB	0	2	18
SHELLFISH	8	1	0
SHRIMP	4	0	0
75112	14	0	0
OYSTER*	9	0	0
61500	15	1	0
WATER	8	56	205
WATERS	3	6	34
SEAWATER*	0	1	6
SEA	2	16	12
OCEAN*	0	6	8
BAY	3	3	10
ESTUAR*	4	6	10
LAKE	1	7	16
LAKES	0	6	2
RIVER	0	13	20
STREAM	0	3	15
STREAMS	0	3	7
RESERVOIR*	0	3	4

references of general interest may be missed by this refinement, but the large number of analytical and sewage treatment articles should be reduced considerably.

So far, the analysis has been directed toward terms which are predominantly advantageous or detrimental. There are also search terms which are both essential and detrimental, as is the case for the term LEAD. In the biology and chemistry data bases, the irrelevant references are due primarily to the occurrence of the metal or its compounds in a context other than natural bodies of water (e.g., lead in paint or in drinking water and its poisoning effects on humans). In the COMPENDEX and NTIS data bases, there are also a sizeable number of irrelevant retrievals which can be attributed to the use of the verb "to lead" in the abstracts (e.g., "...and may lead to errors in data interpretation."). The element symbol PB also causes irrelevant retrievals in the NTIS data base because of its occurrence in PB document numbers cited in the abstract (e.g., PB-223 693). In cases such as this, where the data bases involved are of marginal interest anyway, it would probably be satisfactory to limit the use of these search terms to the title and keywords, thus avoiding the problems with the text of the abstracts. Otherwise, such irrelevant references would have to be screened out manually since lead and its alternative forms are obviously important search terms for relevant references.

The importance of the analysis of the search results and the number of refinements needed to get the profile performing satisfactorily will depend on a number of factors, among which are the precision with which the information need was initially stated and the performance of the first version of the profile. If the profile is to be run continually against update issues of the data base as they are received on a weekly, biweekly, or monthly basis (i.e., current awareness search), then it is wise to invest some time on the

first three or four searches to analyze the references in the bibliography carefully and to make refinements to the profile. Similarly, if a lengthy retrospective search is being run via a batch-oriented retrieval system, where different volumes will be searched over a period of several days or weeks, then the results of the first two or three searches should be examined carefully and the profile modified, as necessary, to "fine-tune" the performance in subsequent volumes. A few key changes, such as those suggested for the sample question, can make the difference between a useful or an unsatisfactory computer-based search.

BATCH-ORIENTED RETRIEVAL SERVICES

There are a number of different kinds of information services available which are based on batch-oriented computer retrieval systems. Four of the most commonly used ones will be described briefly as a survey of the wider range of possibilities: current awareness (SDI) searches, retrospective searches, macroprofiles, and computer-readable subfiles.

Current awareness searches

As implied by the name, current awareness searches are designed as specialized bibliographies of research currently being published. These searches, also known as Selective Dissemination of Information (SDI), are run against the update tapes made available by most data base producers and correspond to the weekly, biweekly, or monthly printed issues of publications such as *Chemical Abstracts*, *Biological Abstracts* and *BioResearch Index*, *Engineering Index*, and *Government Report Announcements*, to name just a few. The magnetic tape versions are usually available some time before the corresponding printed publications because of the delay involved in printing, binding, and distributing the documents. Thus, using the computer-based search services not only obviates the hours of paging through the printed issues but also obtains the selected references earlier than they would otherwise be available from the abstracting and indexing service. Costs of obtaining current awareness searches will vary according to the data base and to the center supplying the service, as will the actual services provided.

Retrospective searches

The update tapes obtained from the various data base producers can be saved and combined into data base collections, just as is done by libraries in building a collection of the printed versions. Searches offered on these collected volumes of the data base are usually marketed as retrospective search services. Such retrospective searches are available from centers who use batch-oriented retrieval systems, as well as those specializing in the on-line methods. The time required to obtain a complete retrospective search is usually much longer than with the corresponding on-line searches, but it is often possible to do much more complex searches via the batch systems. The trade-offs between batch and on-line retrospective searching would need to be explored with an experienced information specialist for any particular case. However, as a general rule, simple and relatively straightforward questions can be handled in a much more cost-effective way through on-line retrieval systems; while complex searches that require elaborate use of truncation, weighted terms as an extension of the logic, or careful delineation of search terms by data element can be more readily accommodated by most of the large batch-oriented retrieval systems.

Macroprofiles

Some information dissemination centers, particularly the United Kingdom Chemical Information Service (UKCIS), have introduced specialized services from the computer-readable data bases called macroprofiles. These profiles represent subjects or topics which are of interest to a relatively large group of users. Examples from among the 40 macroprofiles currently being published by UKCIS include *Chemical Hazards*, *Environmental Pollution*, *Fungicides*, *Prostaglandins*, *Solvent Extraction*, *Liquid Crystals*, and *Organo Fluorine Chemistry*. Considerable attention has usually been given to the profiles by information specialists to refine them until they are performing optimally, using all the sophisticated searching techniques and strategies available in the computer-based retrieval systems. The output of the search is then duplicated through photo-offset printing methods, and the bibliographies are distributed to the group of interested users at a lower cost than comparable individual searches.

Computer-readable subfiles

There is growing interest in the fourth type of service emanating from batch-oriented retrieval systems--the production of computer-readable subfiles. For this type of service, a profile is usually prepared as for the other types of services; but the search results are kept in the computer-readable form, rather than being printed on paper. These subfiles may represent company interests, for example, and may be used to build retrospective data base collections along specialized subject or topic areas. Subfiles from several data bases may be combined into a single data base to meet the needs of a mission-oriented company or government agency whose interests cut across several disciplines. Such subfiles of bibliographic data have also been used to supplement specialized data bases, such as those dealing

with chemical substances or quantitative data. Not all data base producers will permit their data bases to be used to create computer-readable subfiles, and those that do usually require a use fee and contractual arrangements governing the reuse of the information. Although there has been limited use of this type of service so far, there will undoubtedly be increasing interest in this form of output as computing facilities (especially mini-computers) become more widely available.

SUMMARY AND CONCLUDING REMARKS

This paper began with a brief comparison of manual and computer-based retrieval, then traced the major steps involved in conducting a batch-oriented computer-based search of bibliographic data bases. These steps begin with the formulation of a statement of the information need, followed by the selection of pertinent data bases and the preparation of search profiles. Important aspects of the profile preparation step include identification of the major ideas or concepts; the establishment of alternative search strategies which represent the co-occurrence of the concepts as they might appear in the titles, abstracts, or indexing of published documents; expansion of the concepts using the indexing or natural language terminology as it appears in the data base; construction of the appropriate logic statement, showing the linkages between the groups of terms; and, optionally, assignment of weights to sequence the bibliography into logical order, or to extend the logic capabilities of the retrieval system. Following the initial search, it is usually advisable to examine the references retrieved--both the relevant and the irrelevant references--to identify terms which should be added to the profile term list to make it more complete, to correct any spelling or truncation errors, or to revise the profile in more substantial ways (e.g., addition or deletion of terms or groups of terms). These revisions are particularly important when the profile will be used for continuing current awareness searches or batch-oriented retrospective searches. Finally, four of the major types of retrieval services based on batch-oriented retrieval technology were described briefly.

It has been possible to describe and to illustrate only the major features of batch-oriented computer-based retrieval. While it is quite possible to learn in a few hours how to prepare adequately performing profiles, it can take months of experience to become familiar with a large number of data bases and to become really proficient with all the sophisticated techniques available in most retrieval systems for refining search profiles. Information specialists are available on the staff of most information dissemination centers to assist researchers in preparing and refining their profiles. With over a decade of experience now available in conducting batch-oriented computer-based retrieval services, it is possible to say with considerable confidence that this method of information retrieval has come of age and should be considered one of the standard methodologies available to researchers in support of their scientific endeavors.