Papers

# LOCKHEED EXPERIENCE IN PROCESSING LARGE DATA BASES FOR ITS COMMERCIAL INFORMATION RETRIEVAL SERVICE*

R. K. SUMMIT
Lockheed Palo Alto Research Laboratory
Palo Alto, California 94304

## ABSTRACT

The last few years have seen the emergence of the information retrieval services industry. This industry consists of three tiers: data base producers, retrieval service vendors, and information centers who service the end user. The experience of Lockheed Information Systems as a retrieval service vendor is recounted as it pertains to its system and to its relationship with the data base suppliers. Government competition is seen as a potential threat to the industry.

What is the information retrieval services industry? As shown in Figure 1, the industry consists of three different segments: data base suppliers, information retrieval service companies, and information centers. Data base suppliers obtain source document information from scientists, engineers, and professionals, and produce tapes or data bases covering a variety of disciplines. Retrieval service vendors receive data bases on tapes that contain serial files of records, each of which includes a citation, abstract, and postings. These tapes are presented to create standardized formats which are then loaded onto random access storage devices and serve as the basis for information retrieval services which are offered over various data communications networks out to information centers who, in turn, service the information end user.

*Presented in the "Conference on Large Data Bases," sponsored by the NAS/NRC Committee on Chemical Information, National Academy of Sciences, May 22 – 23, 1974.

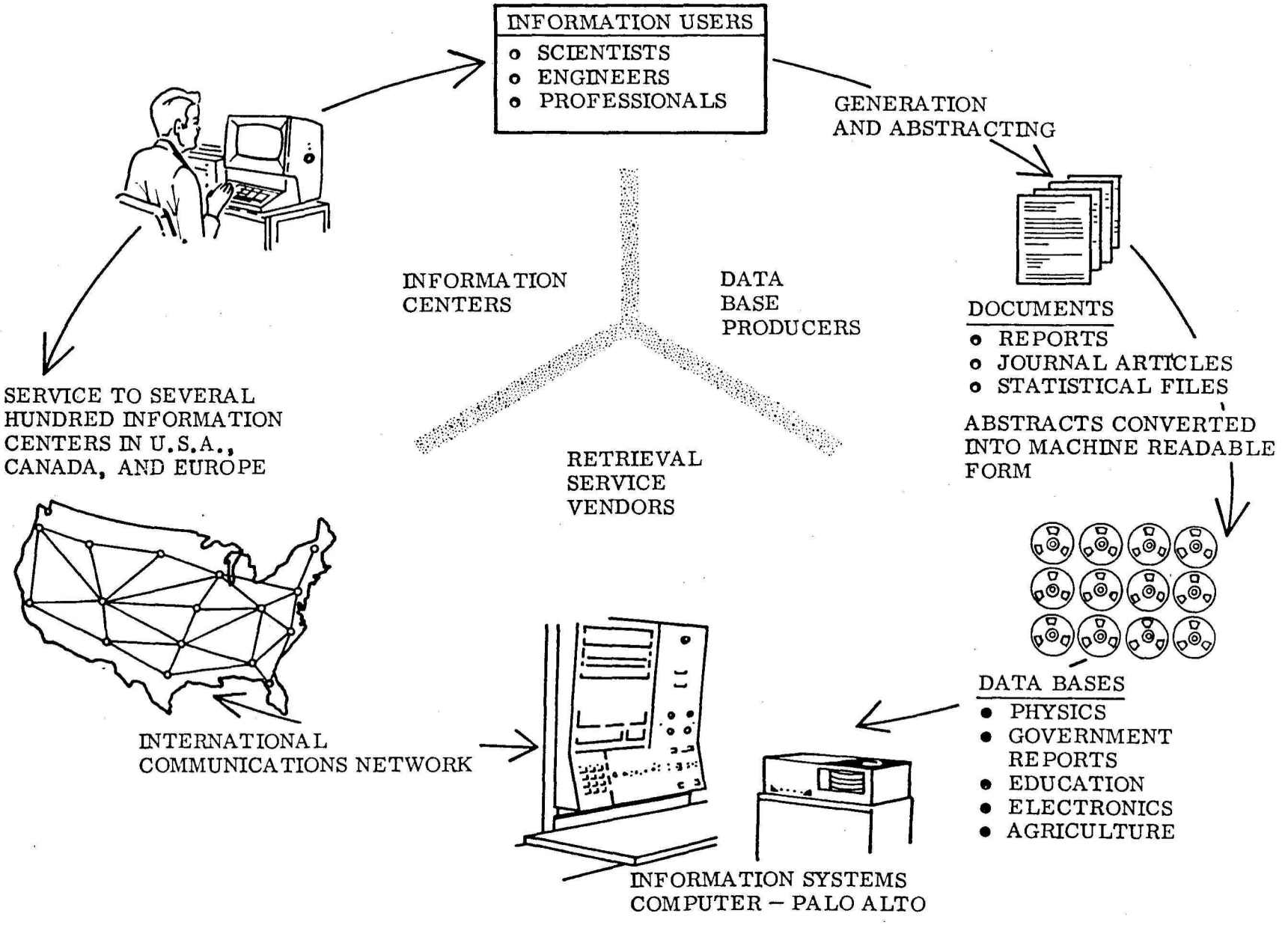


Figure 1. Information Retrieval Services Industry

Those of us who started to design these types of systems a decade or so ago designed them for end-user use. For the most part, however, our experience has been that a technical information specialist learns to operate the system and, in turn, serves the end user. Several now obvious reasons may be cited for this finding, such as intermittent use, data base unfamiliarity, and inconvenience.

Lockheed entered the online information retrieval business in 1964. Initially we started the development of DIALOG as an independent research project at Lockheed. We were awarded a development contract from NASA which resulted in the NASA/RECON system, first installed in 1969 and operating today at most NASA centers. We stayed in development and implementation work for a few years and then in 1971 decided to initiate commercial services on a limited basis. In 1972 we were awarded more service contracts, and we extended the commercial business with additional data bases. In 1973 we added a significant number of data bases that extended the discipline areas already covered to include chemistry, business, and agriculture. All of the major data bases we offer are online for the full working day; the result is some 3 billion bytes online storage and over 3 million abstracts that can be accessed. I mention this figure to give you an idea of the vast size and variety of data with which we must cope from a unified file structure.

## FILE ORGANIZATION

We have a conventional file organization similar to several which were described yesterday and today. In operation, the search process begins with a term coming into the inverted index. Out of the inverted index we get a pointer which goes into the inverted file. This file produces a string of accession numbers, any one of which can go into a linear file index for a pointer, which goes into a linear file. Actually this is not the sequence in which searching is done, but it shows the linking arrangement from one file to another. File organization of the inverted and linear index uses an indexed sequential access method (ISAM) arrangement. The inverted file itself

utilizes a basic sequential access method (BSAM) organization, as does the linear file. This file organization differs somewhat from that of NASA/RECON. Although we developed the initial access files for RECON, the software was later made to interface with the STIMS file system which was developed independently from RECON software, and is similar to what Don Hummel described yesterday.

Now let us go through a search. If we do an EXPAND command with a term, the term enters the inverted index and we come out with a display showing a listing of alphabetical near terms in the index, together with the number of postings which occur in the file for each of those terms. In some cases, if there is a thesaurus associated with the file, the display shows also the number related terms which appear in an associated thesaurus. In selecting these terms, one may simply indicate displayed reference number for terms as a list, or a range, or a list of ranges to select contiguous or noncontiguous terms from the display. To obtain related terms for a displayed term, we can EXPAND the associated reference number and produce related terms in a similar format. As a result, we can browse through the index or browse from the index to the thesaurus, and then browse through the thesaurus. An example of a simple search on the energy shortage is shown in Figure 2 (this is the search summary returned with the offline printout). This search was processed against five data bases: National Technical Information Service (NTIS), Predicasts Inc. Chemical Market Abstracts, National Agricultural Library CAIN Database, Abstracted Business Information INFORM, and Engineering Index. Figures 3 through 7 show examples of hits from each of these searches. It is interesting to note the impact of the energy crisis in the different areas represented by the data bases.

In dealing with several data base suppliers, we encounter several challenging problems. We occasionally have a royalty arrangement that is reasonable to a supplier

```
**********************************************************
*                                                        *
*                                                        *
*                                                        *
*          Lockheed Information Retrieval Service        *
*                                                        *
*                                                        *
*                                                        *
*                                                        *
**********************************************************

                    TITLE        ENERGY SHORTAGE
                    DATE/PILE    5-13-74/16
                    SEARCHER     R. K. SUMMIT
                    REQUESTOR    SAME
                    ADDRESS      LOCKHEED, PALO ALTO
USER    1    5-13-74

                  SEARCH HISTORY
SET   ITEMS  DESCRIPTION
  1     779  FUEL
  2     234  FUELS
  3    1355  ENERGY
  4     405  GASOLINE
  5    2228  PETROLEUM
  6    2896  OIL
  7    4997  1+2+3+4+5+6
  8     543  SHORTAGE
  9     323  SHORTAGES
 10     202  CRISIS
 11      44  SHORTFALL
 12     975  8+9+10+11
 13     468  7*12

SRCH TIME        5.37   PRINT COUNT        50 DESCS.:        10
```

Figure 2. Energy Shortage Search History

The U S Energy Crisis, The Multinational Oil Corporations and Their Relationship to U. S. Foreign Policy in the Middle East

Army War Coll Carlisle Barracks Pa (403565)

Monograph
AUTHOR: Pappageorge, John G.
C1064D1 FLD: 5C, 5D, 97G*, 56A USGRDR7314
28 Feb 73 56 p*
MONITOR: 18

ABSTRACT: America's current energy crisis consists of a growing dependence on foreign oil brought about by a continuing diminution in known domestic petroleum reserves and aggravated by a host of domestic anomolies that cry out for some sort of unified energy policy. Yet any steps taken domestically will have far reaching international effects, particularly in the Middle East. Eight giant corporations (five of them American) in effect of the consumer in the West and in Japan and, on the other hand, they are the ones who discover and pump most of the oil out of the ground in the producing countries. Hence, they have a powerful influence in the Middle East and are a contributing factor in the stability of that politically volatile part of the world. (Modified author abstract)

DESCRIPTORS: (*Petroleum industry, *Middle East), (*United States government, *Foreign policy), Industrial production, Refineries, Distribution (Economics), Natural resources, Political science, Industrial procurement

IDENTIFIERS: Energy crisis, Market analysis, A

AD-760 868 NTIS Prices: PC$3.00/MF$0.95

Figure 3. National Technical Information Service (NTIS)

*World *Crude Petroleum *Supply

The Arab oil cutbacks have injured nearly all countries. Gasoline prices soared from $1.01 to $1.49/gal in India, and to dramatize the seriousness of the shortages, Prime Minister Indira Ghandi began riding in a two-wheeled, horse-drawn gig. In the Netherlands, the Prime Minister rode a bycycle to work, and a strict ban was imposed on Sunday driving. Throughout Europe, there was a vague air of siege. Fuel prices are rising, driving restrictions have been imposed, and in Britain, ration cards have already been printed — just in case. The German Bundestag granted Chancellor Brandt's government blanket emergency powers to take whatever actions it considers necessary to hold down the use of gasoline and heating oil.

312062 Time 73/11/19 P88-95

*World *Crude Petroleum *Supply

The so-called energy crisis is not so much a crisis as a change to a very fine balance between supply of crude and requirements on a worldwide basis, according to Dr. A. W. Pearce, Chairman of Esso Petroleum, speaking in London. Three years ago the situation was one of about 10% spare capacity in crude production, in ships and refineries. Today crude production would be just about in balance with supply if everything went right and with world demand at a price fair to buyer and seller. Mr. P. Emery, Under-Secretary for Industry, said last week that the UK government is convinced that Britain is unlikely to face an energy crisis and there will be no need for emergency measures.

310685 Chem Age 73/10/12 P54

Figure 4. Predicasts result

KF26.A3534 1973a ID NO.– 74-9408079 BOOK CAT– 74005092 824081
Impact of fuel shortage on agriculture; Hearings ... Ninety-third Congress, first session, June 12 and 13, 1973
U.S., Senate, Committee on Agriculture and Forestry, Subcommittee on Agricultural Research and Legislation
Washington, U.S. Govt. Print. Off. 202 p. 1973
DESCRIPTORS– Fuel
CAT CODE– 5510
SEARCH– 19730000
DOC TYPE– MONOGRAPH

1 EC7AG ID NO.– 74-9026822 621972
What can farmers do about shortages? Fuel, management.
U.S., Dept. of Agriculture, Statistical Reporting Service
Agric Situation (Washington) 58 (1): 10–11. Jan/Feb 1974
CAT CODE– 1010
SEARCH– 19740200
SOURCE– USDA DOC TYPE– ARTICLE

1 EC7AG ID NO.– 74-9026821 821971
The energy crisis and 1974 farm production
U.S., Dept. of Agriculture, Statistical Reporting Service
Agric Situation (Washington) 58 (1): 8–9. Jan/Feb 1974
DESCRIPTORS– United States
CAT CODE– 1030
SEARCH– 19740200
SOURCE– USDA DOC TYPE– ARTICLE

1 EC7AG ID NO.– 74-9025819 821969
Agriculture and the energy crisis
U.S., Dept. of Agriculture, Statistical Reporting Service
Agric Situation (Washington) 58 (1): 2–4. Jan/Feb 1974
DESCRIPTORS– United States
CAT CODE– 1030
SEARCH– 19740200
SOURCE– USDA DOC TYPE– ARTICLE

A281.9 F76FO ID NO.– 74-9025189 820433
Japanese cotton industry gears to energy shortage
Hornbeck, B M
U.S., Foreign Agricultural Service
Foreign Agric 12 (8): 5–6, 16. Feb 25, 1974
DESCRIPTORS– Japan
CAT CODE– 1930
SEARCH– 19740225
SOURCE– USDA DOC TYPE– ARTICLE

Figure 5. National Agricultural Library/Cain result

ID NO.– HOHO44017083 017086
THE APARTMENT SCENE
WELLS, H. CLARKE
HOUSE AND HOME V45 N3 MAR 1974 P80
THE ENERGY-CRISIS MEANS HIGHER OPERATING COSTS FOR OWNERS AND TENANTS, SO ECONOMY MEASURES ARE IN ORDER. APARTMENT OWNERS WHO MAKE NO ATTEMPT TO CUT DOWN ON ENERGY CONSUMPTION MAY WELL END UP WITH APARTMENTS THAT ARE PROHIBITIVELY EXPENSIVE TO LIVE IN AND/OR OPERATE. MANY MODEST ENERGY-SAVING STEPS CAN BE TAKEN. MECHANICAL TIMERS ON LIGHTS FOR RECREATION ROOMS AND OUTSIDE REC FACILITIES KEEP RESIDENTS FROM ADDING UNNECESSARILY TO THE ELECTRIC BILL. REFLECTIVE FILE ON LARGE WINDOWS HELPS MINIMIZE AIR-CONDITIONING LOADS AND IS BARELY NOTICEABLE. ADDING NEW SERVICES, SUCH AS SHUTTLE BUSES, CAN BE ENERGY-SAVING. THE GASOLINE CRISIS MAKES IT MORE FEASIBLE THAN EVER BEFORE TO INVEST IN REFURBISHING OF WELL-LOCATED OLDER COMPLEXES TO ENHANCE THEIR LIVABILITY.

ID NO.– MED044017035 017035
THE ENERGY CRUNCH – OIL COMPANIES SPEAK UP
MEDIA DECISIONS V9 N3 MAR 1974 P56–57, 94–102
CONTROVERSY CONTINUES HIGH OVER MEDIA PRACTICES IN AN ENERGY-SHORTAGE ECONOMY. GASOLINE ADVERTISERS HIT BACK WITH THEIR OWN HEADLINES IN PAID SPACE AND ON THE AIR. SPECIAL STATEMENTS OF COMPANY POLICY HAVE BEEN ADDED AND RUSHED INTO CIRCULATION. SOME LONG-TERM CORPORATE ADVERTISING PROGRAMS HAVE BEEN SWITCHED IN CONTENT BUT NOT IN MEDIA IN ORDER TO DISCUSS THE ISSUES IN THE ENERGY CRISIS. PORTIONS OF SOME REGULAR PRODUCT ADVERTISING PROGRAMS IN PRINT AND BROADCAST HAVE BEEN DIVERTED TO EXPLANATIONS OF CORPORATE POINTS OF VIEW WHERE PRODUCT SHORTAGES OR CHANGES IN SERVICE DIRECTLY AFFECT THE PUBLIC, AS WITH THE AIRLINES. THE MOST SIGNIFICANT CHANGE IN MEDIA USAGE HAS BEEN IN THE EXPANSION OF WHAT THE COMPILERS OF ADVERTISING VOLUME CALL CORPORATE ADVERTISING.

ID NO.– CSA044017032 017032
STORE DESIGN WOES ARE ACCENTUATED BY ENERGY-CRISIS
CHAIN STORE AGE V50 N4 APR 1974 P26–27
THE ENERGY-CRISIS, COMING ON TOP OF ESCALATING INFLATION, IS CAUSING SEVERE PROBLEMS FOR STORE ARCHITECTS AND DESIGNERS. THE VERY BASIS OF STORE PLANNING DEPENDS ON SITE SELECTION. IF THERE ARE CHANGES IN FUEL AVAILABILITY FOR SITES AS WELL AS CUSTOMER DRIVING HABITS, WHOLE SITE CONCEPTS WILL HAVE TO BE RETHOUGHT. THERE MAY BE AN INCREASE IN CENTRAL CITY ACTIVITY. THE ENERGY-CRISIS WILL ACCELERATE COMPRESSION OF STORE SIZE. AN OPEN STORE PLAN CAN CONSUME 25% LESS ENERGY THAN A SHOP OR PARTITIONED FLOOR. COSTS AND SHORTAGES OF MATERIALS HAVE CAUSED A SLOWDOWN IN CONSTRUCTION, A SLOWDOWN PREVIOUSLY UNDER WAY DUE TO THE HIGH COST OF MONEY. DELIVERY DATES HAVE BECOME UNCERTAIN. HIGH WOOD COSTS BEFORE THE ENERGY-CRISIS CAUSED A MOVEMENT IN PLASTIC, BUT THAT'S NOW A PROBLEM SINCE PLASTICS ARE PETROLEUM-BASED.

Figure 6. ABI/INFORM result

ID NO.– E1731203739 641580
IMPROVED OIL RECOVERY COULD HELP EASE ENERGY SHORTAGE.
Geffen, Ted M.
Amoco Prod Co, Tulsa, Okla
DESCRIPTORS– *OIL WELL PRODUCTION
IDENTIFIERS– TERTIARY OIL PRODUCTION METHODS
CARD ALERT–511
CODEN– WOOIAS SOURCE– World Oil v 177 n Oct 1973 p 84–88
Economic incentives plus advancement up the $left double quote$ learning curve $right double quote$ will enable the industry to produce more $left double quote$ unrecoverable oil $right double quote$. The $left double quote$ Big Four $right double quote$ Tertiary recovery methods, i.e., hydrocarbon miscible, CO/ /2 miscible, water miscible, and thermal–their advantages and limitations–are outlined.

ID NO.–E1731103598 637234
USING THE OCEAN IN MEETING THE ENERGY SHORTAGE.
Klima, Otto
GE
DESCRIPTORS–(*NATURAL GAS, *Transportation), (TANKERS, Construction and Outfitting)
CARD ALERT–512, 522, 671, 673
SOURCE– Space Congr., 9th Proceedings, Cocoa Beach, Fla. Apr 19–21 1972 p 21–24
A program is discussed that has great relevance to both the oceanic and aerospace community. It has promise for profound influence on employment, the economy, resource utilization and environmental quality. This program involves the construction and deployment of a fleet of liquefied natural gas carriers to satisfy the country's ever-increasing energy needs.

ID NO.– E1731002677 632109
IN SITU GASIFICATION OF COAL: SOLVING THE ENERGY CRISIS.
Hucka, V.; Das, B.
Lavel Univ, Que
DESCRIPTORS– (*GAS MANUFACTURE, *Underground)
IDENTIFIERS– COAL GASIFICATION
CARD ALERT– 503, 522
CODEN– MIENAB SOURCE– Mining Eng (N Y) v 25 n 8 Aug 1973 p 49–50
In-situ gasification of low-grade coals, such as lignites, could be a partial solution to the energy-ecology crisis as it can increase the exploitable reserves of coal and the coal itself never reaches the surface. Two methods of in-situ gasification $EM DASH$ borehole gasification and the underground roadway method $EM DASH$ are discussed. 5 refs.

Figure 7. Engineering Index (COMPENDEX) result

but that may be inconsistent with the way that we have designed to charge customers for retrieval service. We account and bill on an elapsed time basis, whereas the data base supplier may wish to be paid royalties on a record access basis. As a result, we must modify the accounting system to count records displayed or typed. Another proposed royalty basis may be simply a straight subscription fee to the data base, independent of amount of use. Still another basis can be developed on a per access basis; i.e., every time a customer enters the data base irrespective of how long he stays or how many items he prints out, they are charged. On a per day access basis, the customer can come in multiple times during the day if he chooses, but he is assessed a daily access charge the first time he enters. There can also be a monthly access fee which is the same as the daily except one can come in many times in one month for the same fee. Another royalty basis can be a simple flat monthly fee. Finally, we can have any combination of these. It becomes quite a delicate negotiating problem to decide because there is some justification for each of these alternatives in terms of equitable charges between the small users and large users, cost recovery, and so on.

With regard to tape processing, we have had minimal problems processing tapes from any of the public organizations and most of the private organizations. This has not always been true. In fact, one programmer developed a whole system of folklore based on air shipment of tapes. He was concerned that upper atmosphere exposure had the effect of partially degaussing magnetic tape. Thus, whenever he encountered a tape with many errors, he concluded that it must have been shipped air parcel post, and he was usually right. Now one might question the effect of airport x-ray machines. Recently we have had some trouble reading tapes that are generated on a non-IBM computer where the skew is slightly different from our IBM-compatible drive. We

found that there is a test tape which can be obtained from IBM that the supplier can run on his machine to see whether he is producing tapes within tolerance. After providing the particular supplier with this tape, we have had much better luck with the tapes we receive.

Let us now turn from the physical processing problems to the logical processing problems of data bases. I believe that the larger the data base and the longer the period it has been accumulating, the greater the likelihood that there have been some changes in documentation or some changes in the format in the data base. This is particularly likely if the data base is normally used for SDI or publication work. Because of the emphasis on the current tape issue rather than a retrospective collection of the tapes, changes in format have relatively small impact. These small changes, if missed in the documentation, can have great impact on a file loaded for retrospective searching. A recent experience bears this out. The data file contained over 800,000 records. After examining massive documentation, we loaded a test file to see whether we had properly understood the format. We then loaded the entire file. Very shortly after bringing the file on line we discovered that in many of the earlier tapes special coding had been added to the descriptors for publication purposes. We went back to documentation and, sure enough, there was a footnote on one of the pages indicating this change. We had no choice but to reload the entire file.

A second problem we have had is with irregular delivery tapes. People using our service get used to the update cycle, and frequently have standing requests to update.

Finally, logical errors occur in creating the tapes. In one file we had a month or two of missing items, and people who were very familiar with the corresponding printed publication. It turned out that these items had been deleted from tape after

publication but the fact had not been documented or discovered in the file. In another instance we had multiple records for the same accession numbers. In such a case, you get absolutely false drops; that is to say, you search on "computers" and you retrieve an abstract on "animal husbandry." This retrieval results because only one of the two similarly numbered abstracts can be stored.

## PROBLEMS AND QUESTIONS

Data base suppliers are quite reasonably concerned with the impact that online retrieval service will have on their printed products. This is particularly complicated in the case of printed products carrying advertising because although we can pay royalties to the data base supplier, we have not yet learned how to communicate advertising to terminal customers. A second problem is the timely and economic delivery of source documents. We have reasonably automated the retrieval and access process, but for the most part document delivery services are much the same as they were 20 years ago. We have recently made arrangements with some of our data base suppliers to allow direct ordering via the search terminal. This summer we shall also originate direct ordering of original article tear sheet (OATS®) service from the Institute for Scientific Information.

A final problem for the Information Retrieval Service Industry is government competition. Government accounting methods which expense fixed assets at the time of purchase encourage the provision of computer based retrieval services from within the government. We have recently seen cases in which commercially offered data bases have been transferred to a government computer, and are now being offered to the public at large on a subsidized fee basis. Extension of this practice of competition with the private sector could have a severe impact on commercial information retrieval services.