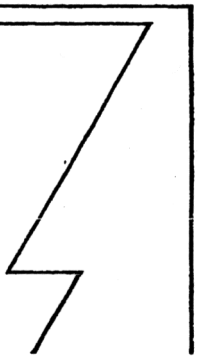
2

The Database Industry



The early history of information retrieval by computer has never been well documented. In fact, it is not altogether clear which system can legitimately be regarded as the first computer-based system for information retrieval. Among the earliest true computer-based systems were those established at the Naval Ordnance Laboratory in Silver Spring, Maryland in 1959 and the system put into operation by Western Reserve University for the American Society for Metals, circa 1960.

It is probably safe to say, however, that the first major information retrieval systems in the United States emerged in the Federal Government in the early 1960s. Perhaps the most important were those initiated by the Armed Services Technical Information Agency (later the Defense Documentation Center and now the Defense Technical Information Center); the National Aeronautics and Space Administration during 1962; and the National Library of Medicine, whose MEDLARS service was launched in 1963. These agencies must be regarded as the pioneers of large-scale bibliographic processing by computer, although many other organizations have followed in their footsteps.

In the 1960s, it was difficult to justify economically the dedication of a computerized system solely for retrospective searching of bibliographic records. Most of the very large bibliographic systems were justified by their virtuosity. They tended to be—and still are in many cases—multipurpose, generating a range of products or services from a single input operation. Many of these systems were developed as an outgrowth of the automated publishing process and the need to manipulate citations in machine-readable form for error checking, sorting, formatting, and computer typesetting. The machine-readable tapes produced from this activity could then be used to generate additional publications and to offer further services. The major service that was made possible by the machine-readable database was a retrospective search service

on demand (demand search service), although the database was also used in current awareness activities (selective dissemination of information [SDI]).

OFFLINE SYSTEMS

The characteristics of the operational computer-based retrieval systems of the 1960s were very similar. They were offline batch-processing systems that used magnetic tape as the storage medium and that, by and large, were searched serially. Search strategies were matched sequentially against document representations, and a printed bibliography was produced. Retrospective searches of the entire database were intended to be performed once, and the result consisted of all documents in the system that matched the search request. For SDI, stored search profiles were periodically processed against recent updates of the database, and results were mailed to subscribers.

Computer retrieval systems of the 1960s offered many advantages over their predecessors, including the following:

1. Through batch processing, many searches could be conducted at the same time.
2. Many access points to a document could be provided very economically.
3. Complex searches involving large numbers of terms in complex relationships could be handled.
4. Output could be generated in the form of a bibliography, and a high-quality publication could be produced by interfacing the retrieval system with a photocomposition device. Output could also be made directly to microfilm (computer output microfilm [COM]).
5. Management data on how and how much the system was used could be collected, on a regular basis and as a by-product of normal system operations.
6. Many outputs and services could be produced from a single input—a general printed index, specialized indexes, retrospective searches, and SDI searches.
7. The database, once captured in machine-readable form, could be duplicated simply and cheaply; it was easily shipped and thus could be used to provide information services by different information centers.

Despite their many advantages, the offline batch-processing systems also had disadvantages. They were essentially "one-chance" searching systems in which the searcher had to think of all the possible search approaches in advance and construct a search strategy that, when matched with the database, was likely

to retrieve all the relevant literature. In other words, they were noninteractive and nonheuristic, and they did not provide any real browsing capability.

Another major disadvantage of the offline systems was that the search results were substantially delayed—it was not possible to get an immediate response. At best, it took hours; at worst, as in the case of searches processed by a large national information center, it might take several days or weeks.

The search in an offline system was generally "delegated"; that is, the individual who needed the information had to delegate the responsibility for preparing the search strategy to an information specialist, with no opportunity to conduct his or her own search. Nondelegated searching is not invariably better that delegated searching, but the process of delegation is tricky. It is obvious that a search will produce very poor results if, in the process of delegation, the requester is unable to explain clearly what he or she is seeking or if the information specialist misinterprets the real needs of the user.

ONLINE SYSTEMS

The batch-processing systems of the 1960s were followed by the online interactive retrieval systems of the 1970s and beyond. These were made possible by advances in hardware, software, and telecommunications.

In online systems, data are stored on magnetic disk. Generally the system consists of both a linear file (containing each full record in the system) and one or more inverted indexes (often called index files) created from the linear file. Exhibit 6 shows a simplified example of a few records from a linear file with accompanying inverted index entries. Each element in the inverted index consists of a value or element from a database record in the linear file (for example, an author name or a keyword) along with a unique key element (usually an accession number) that is used to retrieve the records in which that value can be found. The linear and inverted files are stored on magnetic disks, where information can be accessed randomly—hence, the ability to perform interactive searches in real time.

Innovations in telecommunications also had a major impact on the online industry. Even though online systems were available in the late 1960s and early 1970s, their use was not widespread, particularly in libraries, because of the necessary expense of accessing them over long-distance telephone lines. Packet-switching networks such as TELENET and TYMNET lowered communications costs substantially, since the network was activated for a particular user only when a message was ready to be sent, rather than maintaining a permanent connection during the entire course of a communication. It was after the introduction of these networks that online retrieval found a widespread market.

LINEAR FILE

Chan, Lois Mai
Pollard, Richard C.
Thesauri used in online databases
Greenwood Press:us
1988
United States
Language: English
Subject heading: Thesauri/Bibliography
Subject heading: Information systems/Directory
BLIB88009087
Monograph

Instructional materials used in teaching cataloging and classification
Chan, Lois Mai
Cataloging & Classification Quarterly 7:131-44 Summ '87
Language: English
Subject heading: Cataloging/Teaching
Subject heading: Surveys/Library science literature
Subject heading: Textbooks
BLIB87009368
Article

INVERTED INDEXES

Author

Chan, Lois Mai	88009087
	87009368
Pollard, Richard C.	88009087

Subject

Bibliography	88009087
Cataloging/Teaching	87009368
Directory	88009087
Information systems	88009087
Information systems/Directory	88009087
Library science literature	87009368
Surveys	87009368
Surveys/Library science literature	87009368
Teaching	88009087
Textbooks	87009368
Thesauri	88009087
Thesauri/Bibliography	88009087

EXHIBIT 6 Sample inverted and linear files. SOURCE: *Library Literature* database. Reprinted with permission of The H. W. Wilson Company.

Online retrieval systems have all the advantages that apply to batch-processing systems but avoid all the major disadvantages. They are heuristic and interactive, permit browsing, provide rapid response, and can be used in a nondelegated search mode.

Virtually all of the early online retrieval systems operated as depicted in Exhibit 5. Primary documents were acquired by an organization (a database producer) where document representations were generated, consisting of an appropriate citation and often index terms and an abstract. These representations were entered into machine-readable form and stored on magnetic tape. The information on tape was processed in various ways to produce printed indexes and sometimes also was maintained by the producer to do offline batch processing of search requests. The information on magnetic tape was also loaded onto magnetic disk, and the representations were processed to create the inverted indexes that were necessary for online, interactive searching. This was done either by leasing the tape to another organization (the database vendor) or locally by the database producer. In either case, the organization that processed the database for online searching also had to provide appropriate interactive search software.

The systems that provide bibliographic data for library catalogs underwent a similar evolution; however, rather than having producers that lease databases to vendors, member libraries input records to a centralized database that is owned and operated by an organization known as a bibliographic utility. Such an organization—for example, the Online Computer Library Center (OCLC)—then provides the appropriate services and technical support for libraries to receive, from the centralized database, copies of records that correspond to items in their own collections. These records, in either paper or machine-readable form, can then be added to an existing catalog.

Operating from the beginning with the Machine Readable Cataloging (MARC) standard for storing and manipulating cataloging records (Exhibit 7), the bibliographic utilities first were used exclusively in batch mode for the production of catalog cards for member libraries. Later, in the 1970s, the databases thus compiled were also offered online so that libraries could search and modify records interactively.

Although the technological evolution of bibliographic utilities and database producers and vendors was quite similar, there are also some substantial differences between them. Bibliographic utilities exist because all member libraries that produce and enter records into the centralized database use the same standards—the second edition of the *Anglo American Cataloging Rules*, or *AACR2*—for deciding the content of document representations and the MARC standard for structuring, maintaining, and manipulating these records in machine-readable form. With few exceptions, online catalogs in libraries accept and work with MARC records. Most of the database industry, on the other hand,

```
▶NO HOLDINGS IN OCL - FOR HOLDINGS ENTER dh DEPRESS DISPLAY RECD SEND
  OCLC: 3349989   Rec stat: n Entrd: 771108        Used: 790312 ¶
▶Type: a Bib lvl:  m   Govt pub:   Lang: eng  Source:   Illus: a
  Repr:  Enc lvl:   Conf pub: 0 Ctry:   nyu Dat tp: s  M/F/B: 10
  Indx: 1 Mod rec:  Festschr: 0 Cont: b
  Desc: i Int lvl:      Dates: 1977.      ¶
  ▶ 1  010      77-77941 ¶
  ▶ 2  040      DLC ‡c DLC ¶
  ▶ 3  020      0525171940  : ‡c S17.95 ¶
  ▶ 4  050 0    GN31.2 ‡b .L43 1977 ¶
  ▶ 5  082      573.2 ¶
  ▶ 6  090       ‡b ¶
  ▶ 7  049      OCLC ¶
  ▶ 8  100 10   Leakey, Richard E. ¶
  ▶ 9  245 10   Origins :  ‡b what new discoveries reveal about the
emergence of our species and its possible future / ‡c Richard E. Leakey
and Roger Lewin. ¶
  ▶10  260 0    New York :  ‡b Dutton, ‡c c1977. ¶
  ▶11  300      264 p. :  ‡b ill. (some col.) ; ‡c 25 cm. ¶
  ▶12  504      Bibliography:  p. 257. ¶
  ▶13  500      Includes index. ¶
  ▶14  650  0   Anthropology. ¶
  ▶15  650  0   Human evolution. ¶
  ▶16  700 10   Lewin, Roger, ‡e joint author. ¶
```

EXHIBIT 7 Sample MARC record. SOURCE: Adapted from OCLC Online Union Catalog. Reprinted with permission of the OCLC Online Computer Library Center, Incorporated.

has never been standardized, and the content and structure of records varies considerably among databases. Furthermore, although the types of documents on all these systems are becoming increasingly diverse, the entries contained in bibliographic utilities and online catalogs typically tend to be bibliographic citations relating to items in the collections of libraries. The database industry, however, provides not only bibliographic data but also other types of records, including numeric data and even the full text of documents.

The scenario drawn in this section is greatly simplified, and it is, in fact, evolving in several key ways. Boundaries are no longer clear regarding the roles various organizations can play and the kind of information they provide access to, and there are many other kinds of databases and organizations involved in this arena. Much of the remainder of this chapter provides a sense of the richness and diversity of this continually changing environment.

TRENDS IN THE INDUSTRY

By every indicator, the online industry is growing continuously. A good idea of its current size and diversity can be gained by browsing through a current directory of databases, including *Computer-Readable Databases* (Marcaccio, Adams, and Williams, 1990) and the *Directory of Online Databases* (Cuadra Associates, 1992). Although this is a dynamic market—databases and vendors sometimes do not survive for long—the overall trend has been toward a net increase.

Several factors have contributed to this growth. Williams (1988, 1992) states that the number of databases offered in online and batch modes doubled from 300 to 600 between 1975 and 1981 and quintupled from 600 to 3,000 between 1981 and 1985. One million searches were performed in 1975 and 15 million in 1985. And the number of database records increased from 52 million in 1975 to 1.68 billion in 1985 (Williams, 1980a, 1988). Williams (1992) also states that by 1991, 34.5 million searches were performed in 7,637 databases containing 4 billion records.

The July 1992 *Directory of Online Databases* (Cuadra Associates, 1992), which includes only online sources, lists more than 5,300 online databases developed by 2,158 producers and made searchable by 731 vendors. These databases and online services are offered all over the world. In fact, the internationalization of the online industry has been a growing trend that is likely to continue (Landau, 1988), with companies developing overseas and foreign investments being made in many areas, including the United States.

Bibliographic utilities and online catalogs show a similar evolution. In 1976, there were only 2 million cataloging records in the machine-readable database maintained by OCLC, the largest bibliographic utility. In 1991, there were more than 23 million records (OCLC, Inc., 1991). In the 1970s, there were only a handful of academic institutions with an economic and technical infrastructure rich enough to develop and support online catalogs. Today, technological developments, particularly the introduction of affordable online catalog software and hardware, have made it possible for all sizes and types of libraries to create and maintain an online catalog. There are undoubtedly thousands now in existence.

The online industry is not only large, it is also very complex. Compare Exhibit 5 in Chapter 1, which shows database producers and vendors and their roles, with Exhibit 8, which considers these organizations along with others involved in the database-use chain (Williams, 1986). The roles of the producer (to create databases) and vendor (to process these databases and make them searchable online) have already been mentioned. Gateway services came into being more recently (around the mid-1980s) and were introduced specifically because of the proliferation of databases and online services. They allow users

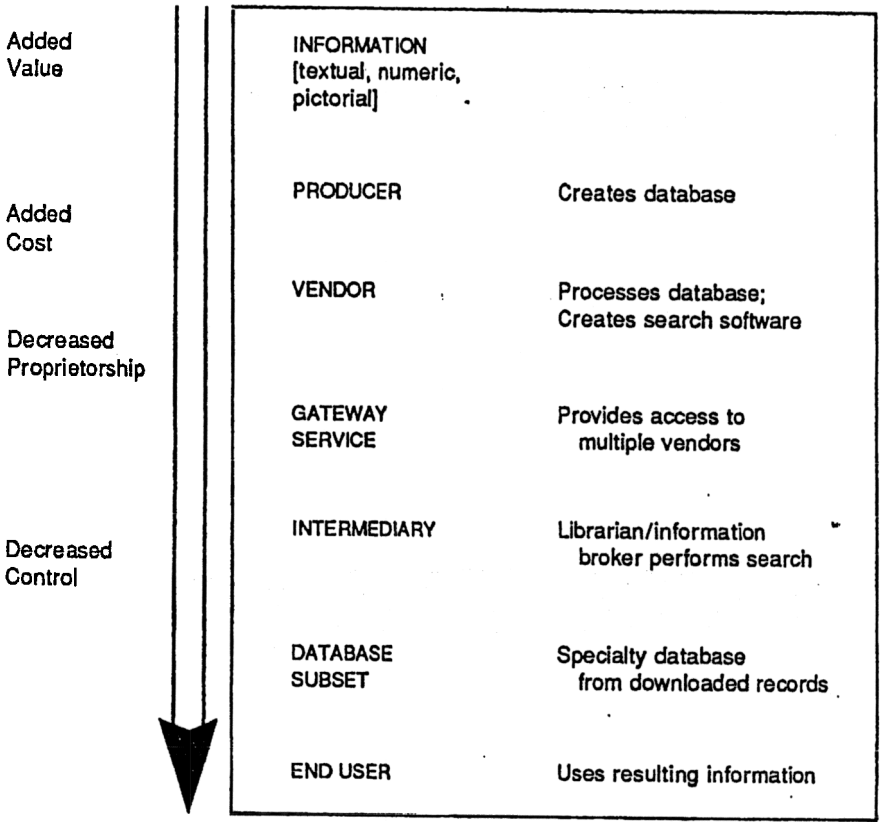


EXHIBIT 8 Database-use chain. SOURCE: **Adapted from Williams (1986) by permission of John Wiley & Sons, Inc.**

to access a variety of remote online services through a single "gateway" computer and often provide a common interface to these services as well, a very useful feature that eliminates the need to learn the different commands needed to access each system. Databases can be searched by intermediaries (information professionals in libraries or information brokerage firms). Results are sometimes used to create a new database subset, often referred to as an "in-house" database, which can then be searched locally in the library or information center. Finally, the end user interprets and uses the resulting information.

Although all these players currently exist in today's information industry, they are not always operating every time a search is performed. Furthermore, the same organization or individual can occupy more than one role in this

chain. The National Library of Medicine (NLM), for example, both produces and sells access to its MEDLINE database.

As one moves down the database-use chain, value is added by the processing or screening of data. This often results in an added cost, which is usually passed on to the user. Each step along the chain increases the potential for access to the stored information, but this is achieved at the expense of less control. Moreover, adding value increases the danger of distorting the original message—something important might be left out, changed, or added.

Furthermore, as the distance between the published information and the customer increases, and more and more organizations provide access to this information, it becomes increasingly difficult to decide not only who should assess the user's needs and meet them, but which user's needs are more usefully and appropriately met by a given information product or service—those of the intermediary user or the end user. It also complicates the choice for the user who is assessing options for gaining access to information sources.

The database-use chain also can describe the roles played by bibliographic utilities, libraries, and online catalogs. The utilities can be considered both the producers and vendors of cataloging records in the MARC format. Intermediaries in this case are librarians who search, modify, and request copies of MARC records from the database. The online catalog then effectively becomes a database subset, processed, indexed, and made searchable in ways deemed appropriate for that library's end users.

INFORMATION PRODUCTS AND SERVICES

The most important information product is the machine-readable database, which can be described by its scope or subject matter, the physical form of its contents, or its various uses.

Because no database contains all the world's information, all databases are somehow restricted in their scope. One very common restriction is subject area. The first databases were restricted to the scientific and technical information required by their sponsoring governmental agencies. When the commercial sector began to develop databases in the 1970s, abstracting and indexing services extended the offerings to other domains: first to engineering and the applied sciences, then to the social sciences (including business and economics), and finally to the arts and humanities. Virtually all these databases contain information that is primarily of interest to the scholar, businessman, or engineer. Later, however, databases were developed that focus on material of interest to the ordinary citizen—from consumer information for a wide variety of products to practical information on child care, drugs, and similar everday concerns (Williams, 1985).

Databases can also be grouped by the form of the data they contain and the uses to which they can be put. A number of dichotomies are presented in the literature:

1. *Word-oriented versus number- or picture-oriented.* Databases are viewed in terms of whether the data they contain are primarily textual or in some other format, such as numeric or graphic (Williams, 1985).

2. *Bibliographic versus nonbibliographic.* Databases containing citations to the literature are placed in one group and all other databases are placed in another group that includes directory, full-text, and numeric databases (Borgman, Moghdam, and Corbett, 1984).

3. *Reference versus source.* Reference databases are bibliographic as well as databases containing entries from sources such as directories of establishments, individuals, and software. Source databases are all other databases containing primary data, such as numeric information and the full text of documents (Harter, 1986).

Databases in machine-readable form can thus be placed in one of five different categories reflecting their content, purpose, and scope:

1. *Bibliographic/reference/word-oriented databases* contain citations to the primary literature. They are used to perform retrospective and SDI searches, usually in support of research and scholarly activities. They are sometimes restricted by the subject or the form of documents they provide access to (for example, technical reports, monographs, periodical articles). Exhibit 9A shows a sample record from *America: History and Life*, which contains citations dating from 1964 to the full range of U.S. and Canadian history, area studies, and current affairs literature.

2. *Bibliographic/referral/word-oriented databases* contain information about people, companies, research projects, and media such as software and audiovisual materials. They are not used to point to literature sources but to answer questions about nonprint sources; they are often used to answer ready-reference queries. Exhibit 9B shows a sample record from a typical referral database (*American Men and Women of Science*) that contains biographical data on 130,000 American and Canadian scientists in the physical and biological sciences.

3. *Nonbibliographic/source/word-oriented* databases are full-text databases that contain the texts of original documents in machine-readable form. These documents can be journal articles, newspapers, newsletters, encyclopedias, dictionaries, and other types of reference books. Full-text databases are used to answer factual questions and to retrieve citations to the literature. In the

A: America: History and Life (Bibliographic/Reference)

950008 26-8
AMERICAN HOMESTEADERS AND THE CANADIAN PRAIRIES, 1899 AND 1909.
Percy, Michael B ; Woroby, Tamara
Explorations in Economic History 1987 24(1): 77-100.
NOTE: Based on published census records and other public documents; 2 fig., 3 tables, 12 notes, ref., appendix.
DOCUMENT TYPE: ARTICLE
ABSTRACT: The out-migration of American homesteaders to the Canadian prairies is best explained by human-capital investments in wheat farming in 1899 and 1909 and by the techniques of dry farming in 1899. Canadian promotional expenditures also contributed to higher rates of out-migration. High tenancy rate and low agricultural wages in the United States were not important contributors to out-migration. (P. J. Coleman)
DESCRIPTORS: Prairie Provinces ; USA ; Homesteading and Homesteaders ; Agriculture ; 1894-1913 ; Migration
HISTORICAL PERIOD: 1890D 1900D 1910D 1800H 1900H
HISTORICAL PERIOD (Starting): 1894
HISTORICAL PERIOD (Ending): 1913

B: American Men and Women of Science (Bibliographic/Referral)

0010037
Brown, Donald D
DISCIPLINE: BIOLOGY, GENERAL (00200207)
SUBJECT SPECIALTY: DEVELOPMENTAL BIOLOGY
BORN: Cincinnati, Ohio, Dec 30, 31 MARRIED: 57 NO. OF CHILDREN: 3
EDUCATION: Univ Chicago MS & MD 56
HONORARY DEGREES: DSci Univ Chicago 76 Univ Maryland 83
PROFESSIONAL EXPERIENCE: Intern Charity Hosp New Orleans La 56-57; res assoc biochem NIMH 57-59; spec fel Pasteur Inst Paris 59-60; spec fel 60-61 MEM STAFF BIOCHEM CARNEGIE INST DEPT EMBRYOL 61 to present, DIR DEPT EMBRYOL 76 to present
CONCURRENT POSITIONS: Prof Johns Hopkins Univ 69 to present
MEMBERSHIPS: Nat Acad Sci; Am Soc Biol Chem; Soc Develop Biol; Am Acad Arts & Sci; Am Soc Cell Biol (pres 92); Am Philos Soc
HONORS AND AWARDS: US Steel Found Award Molecular Biol Nat Acad Sci 73; V D Mattia Lectr Roche Inst Molecular Biol 75; Boris Pregel Award NY Acad Sci 77; Ross G Harrison Prize Int Soc Develop Biologists 81; Feodor Lynen Medal 87
RESEARCH: Control of genes during development; isolation of genes
ADDRESS: Dept Embryol, Carnegie Inst Washington 115 W University Pkwy Baltimore , MD 21210

EXHIBIT 9 Sample bibliographic database records. SOURCES: Exhibit 9A reprinted from DIALOG File 38 by permission of ABC-CLIO; Exhibit 9B reprinted from *American Men & Women of Science 1992–1993,* 18th Edition, © 1992, by Reed Publishing (USA) Inc., p. 795, with permission of R. R. Bowker, a Reed Reference Publishing Company.