40 Years of Database Distribution and Use: An Overview and Observation

1999 Miles Conrad Memorial Lecture
NFAIS Annual Conference
February 23, 1999
Philadelphia, PA

Charles P. Bourne

ABSTRACT
A review is made of some major events in the history of the development of the capability for computer searching of bibliographic and full-text databases, particularly with regard to NFAIS member organizations. Comments are also made regarding some continuing issues.

INTRODUCTION

Thank you for the invitation to share my thoughts with you on the occasion of your annual meeting. I'd like to use this occasion to review some of the major events in the history of the development of computer searching of bibliographic and full-text databases, and to make some comments about some continuing issues.

I've been a participant and observer of the information industry for over 40 years, and that provides a modest qualification for me to make some observations on activity during that time period. I've also recently completed the research for a manuscript on the history of the pre-1977 online industry and technology, and that effort served as a useful refresher. However, I retired from the industry in 1992, and have not been an active participant since then, consequently I am less qualified to speak from direct experience about the current scene.

HISTORICAL PERSPECTIVE

Some Milestones for Computer Searching

Apparently, the first serious consideration of the use of a general-purpose computer for literature searching was initiated by James Perry in 1951, as an MIT faculty member. Perry is known to many of us from his later work at Western Reserve University on the WRU Searching Selector. Phil Bagley, an MIT graduate student after hearing a presentation about the operation of Vannevar Bush's Rapid Selector at MIT, approached Perry with the idea of doing some improvements to the Rapid Selector as his Master's degree thesis topic. Perry, as a person interested in chemical documentation, and as a leader in the ACS Chemical Literature Division, suggested instead that Bagley explore the use of the Whirlwind computer for bibliographic searching. Perry knew that the world's largest and most powerful computer at that time (the IBM Whirlwind) was available on the MIT campus for their use, and that Bagley was experienced in programming that machine.

Bagley took the suggestion, did that work, wrote his very thoughtful and analytic thesis on that topic in 1951, but did not implement any such system.' There was no follow-up work at MIT on that topic, and Bagley's work was largely ignored by others, even though Perry was so impressed with Bagley's effort that he personally paid for the printing of another 200 copies of Bagley's thesis for further distribution. Bagley's conclusion in 1951 was that computer searching of large bibliographic files was infeasible at that time because of the current technical limitations. A subsequent analysis by UNIVAC researchers in 1953 similarly reported that it was essentially economically infeasible to use the UNIVAC system at that time for bibliographic searching of large files.

The first demonstration, and the first implementation of an operational computer-based literature search system, was made in California in 1954. Library staff members at the Naval Ordnance Test Station (NOTS) had been working with Mortimer Taube of Documentation, Inc. to install Taube's Uniterm indexing system for their document collection, and Taube encouraged Robert Bracken and Harley Tillit at NOTS to implement that system with the IBM 701 computer at NOTS. They agreed, and the NOTS library started to provide a regular computer-based search service to its in-house constituency at that time. Searching started with a file of about 15,000 bibliographic records, indexed only by the Uniterms, and search output was limited to report accession numbers. The task was made

even more difficult by the fact that the IBM 701, a scientific calculator, did not have any built-in character representation.

The technology for computer searching improved dramatically in 1957, courtesy of IBM. This was because of IBM's introduction of direct access storage (the IBM 305 RAMAC system with a disk drive storage unit consisting of a spindle of 50 fixed disks that provided a total capacity of 5 million alphanumeric characters), along with its first experiments to demonstrate computer searching of full text, its distribution of search and SDI software, and its extensive publicity of this activity. From 1958 onward, there were several reports of operational batch-mode computer search systems searching bibliographic records for technical report collections, primarily for in-house constituencies.

The first computer-based search service to a constituency broader than in-house use was provided by the American Society for Metals (ASM). In 1960, with internal support and support from the National Science Foundation (NSF), the Society provided a computer-based search service to its entire membership, following several years of experiments with the Western Reserve University Searching Selector. ASM was also the first organization to provide a computer-based bibliographic search service on a fee basis. James Perry of WRU and Majorie Hyslop of ASM are two of the persons most often identified with that service.

The first demonstration of an online bibliographic and full-text search system, and remote online searching, was a system developed and demonstrated with Air Force and ARPA support at Stanford Research Institute (SRI) in 1963. This was done as an adjunct project to the Augmented Human Intellect Program that was then underway at SRI under the direction of Doug Engelbart. As Project Leader for that online effort, and working with a very talented programmer, Len Chaitin, I tried to incorporate and make use of everything that we had learned from the almost 20 operational batch search systems that had been demonstrated by that time. We felt that this online approach was a natural and logical extension of what had already been done with the batch search systems. Our system made use of programs and a database of full-text and bibliographic records that were prepared at SRI, and then installed and searched on the Q-32 computer at SDC-Santa Monica from a small computer workstation (CDC-160A) 400 miles away at SRI.2

Considering the tremendous growth and impact of online searching throughout the world today, one would naturally ask why there wasn't more follow-up and attention to this pioneering SRI development. The reasons are twofold: (1) As a nonprofit research institute, the projects went where the money was, and we could not find funding agencies that were interested in directly supporting a continuation of this effort; and (2) there was SRI and sponsor support for many more interesting and challenging projects associated with Cold War efforts at that time that claimed a higher priority for SRI and staff attention. This would include the information technology for applied projects such as military command and control systems, and intelligence data handling systems. You have to realize that this work was going on at a time period when SRI staff members and their families had space assigned to them in the basement of SRI's main building for the pre-positioned storage of food, diapers, clothing, and other essentials, for use in the event of their need of a safe assembly and living area upon notice of a possible or actual Soviet nuclear attack on the San Francisco Bay Area.

This early SRI work was followed in May 1964 by a demonstration of online bibliographic searching on the NSF-supported MIT-TIP system by Mike Kessler, Bill Mathews, and Evan Ivie. This was done with a database of very brief citations and associated cited references from the articles of a small number of major physics journals. This initial demonstration may also have been the first demonstration of multiple terminals simultaneously searching a bibliographic database.

In 1964, we saw the first computer-based search service that was operated on a regular basis by a commercial organization as a commercial enterprise. That was the Law Research Service, provided by Law Research Services, Inc. in New York with a database of over one million abstracts of New York case law. Elias Hoppenfeld is the person most responsible for the development of that service. By virtue of subsequent activities in 1966, the Law Research Service may also have been the first online search service provided on a regular basis by a commercial organization.

Another early commercial entry was Information Interscience Inc. (III), an organization started around 1967 to provide a computer-based sci-tech service to industry; with Art Elias as one of the founders. III was the first commercial licensee of the INSPEC, COMPENDEX, and Excerpta Medica databases.

The computer-based search service that has been operating for the longest time is the DATRIX service of University Mcrofilm, Inc. (UMI), operating since 1967.

System Development Corporation (SDC) gave its first demonstrations of online bibliographic searching in late 1966, using its COLEX (CIRC OnLine EXperiment) system at Wright-Patterson Air Force Base in Dayton, working with a database of over 50,000 citations to Communist-bloc technical literature of possible interest to the Air Force.

Bunker-Ramo developed and demonstrated its NASA/RECON system in 1966-67, and with the NASA database, was the first online system to provide direct online access to a file of more than 200,000 bibliographic records. Lockheed subsequently developed and demonstrated its version of a single-terminal online RECON system with that NASA database, and in 1968 received NASA support for the development of the full NASA/RECON system.

The first instance of an international online search was a demonstration with the SDC-BOLD system in June 1967, searching the Q-32 computer in Santa Monica from a terminal in Rome, Italy. The next international demonstration was with the MIT-TIP system in July 1967. ESRO/RECON, in 1969, was the first organization to provide an online search service from a computer in Europe.

Regular online service to a nationwide network of users on a regular, non-experimental basis started in 1968 with the SDC-COLEX and SUNY-BCN services. This activity, 31 years ago, was really the start of the age of online services.

In 1967, with Air Force support, Data Corporation in Dayton, Ohio, demonstrated its online bibliographic and full-text search system, which continued to develop as the Data Central system and then the Mead Data Central system. The first large-scale use of online full-text searching began in 1969 with Data Central's OBAR (Ohio Bar Automated Research) system. The subsequent acquisition of Data Corporation and the Data Central search system by Mead Corporation eventually led to the development of the Mead LEXIS system. Few people realize that Mead, as a paper company with a strong interest in printing technology, bought Data Corporation for its advanced printing and image processing technology, and didn't even know that it had a powerful computer search system.

Online commercial service to any interested organization (not just to a contracted community such as the online services to the ERIC or NASA communities) began with Battelle's BASIS and Lehigh's LEADERMART services, both in 1971. They both began public service before the SDC and DIALOG public search services started, but BASIS soon stopped offering service because of some issues with its operating charter as a nonprofit organization, and LEADERMART stopped for other reasons.

The first search system to provide immediate (i.e., without time windows) online search access to over 25,000 records was MIT-TIP in 1964; Bunker-Ramo was the first to reach the 200,000 record level (in 1966); SUNY-BCN was the first to reach the 500,000 record level (in 1968); and BASIS was the first online search service to provide direct and immediate access at any time to an online collection of more than one million records (in 1973).

In 1974, SDC Search Service became the first online search service to provide a publicly available online document ordering service. And in 1975, Dow Jones News/Retrieval may have been the first online bibliographic or full-text retrieval system to provide file updates on a real-time basis.

Online searching was first incorporated into the regular library school curriculum in 1975. Making use of student laboratory workbooks, and one laboratory facility with several online terminal workstations, an Online Laboratory course was established and offered as a regular feature at Berkeley at the University of California School of Librarianship. Separate lab workbooks with a series of previously tested student exercises were prepared for both the Lockheed DIALOG and SDC ORBIT search

services, and made use of the ERIC database. I was greatly assisted in this effort to develop the lab workbooks and to run the lab by my research and laboratory assistants, Barbara Anderson and Jo Maxon-Dadd. They both subsequently went to work at DIALOG, and are well known to many of you here today. The three of us also participated in the development of DLALOG's first Online Training And Practice (ONTAP) files. This was the beginning of DIALOG's Classroom Instruction Program.

Some Milestones for Database Development

A discussion of computer searching can't be done without also discussing the databases that went with that searching. Before 1960, anybody considering the possibility of computer-based searching was usually thinking in terms of an in-house or custom-made database and customized software. With the exception of the cooperative ASM/WRU efforts, bibliographic databases did not become available for second-party machine processing until 1960, when Majorie Hyslop of ASM offered to make copies of ASM's retrospective backfile tapes of 12,000 records from its Review of Metal Literature available to others. However, nobody took ASM up on that offer at that time. NERAC, with Dan Wilde in 1968, was the first search service to use the ASM tapes.

Some technical advances came in 1962 to improve the utility of the database records, namely the introduction of the 120-character IBM print chain, and a mechanism to provide the corresponding machine-language character representation. Up until this time, the databases had been limited to upper-case alphanumeric and a few special characters. This new freedom of expression was first made widely visible in 1962 in the new American Chemical Society (ACS) index publications Chemical Titles (CT) and Chemical-Biological Activities (CBAC). Surprisingly, even into the 1980s, some database suppliers were still keying their records into upper-case-only representations.

Information for Industry (IFI), in 1962, was the first organization to start regular distribution of bibliographic records on computer tape, with its Uniterm Index to U.S. Chemical Patents. In 1965, after several years of experimental trials, and with NSF and NIH support and much publicity, the next database subscription offer came from Chemical Abstracts Service (CAS) for its Chemical Titles (CT) and Chemical-Biological Activities (CBAC) tapes. NASA, DDC, and NLM started distributing their databases in 1963 and 1964, but only to selected organizations. Several more secondary services offered their databases after that. By the end of 1976, hundreds of databases with a total of over 30 million records were available for second-party use.

In 1963, Pete Luhn and Steve Furth of IBM were part of the Conference Committee for the 1963 annual American Documentation Institute (ADI) meeting in Chicago. In that capacity, they took an innovative approach to the production of the conference proceedings. They keyboarded the preprints for the short-paper portion of the conference, and then with computer composition techniques, they produced one of the volumes of the conference proceedings for distribution at the conference. This was a collection of 60 short papers (128 pages total), and was the first instance of a collection of full-text articles produced with computer composition techniques. ADI offered to make available the machine-readable records for that collection of full-text articles to any interested party for $100; it had no takers, but it gained the bragging rights to be able to claim to be the first organization to make available a database of full-text material for second-party use. Other full-text databases followed shortly thereafter, starting with Horty's collection of Pennsylvania state statutes.

CCM (Crowell Collier Macmillan) was the first database producer to execute a license for the online use of a database. Dick Kollin, as the developer of the PANDEX database, was the producer associated with that license. In 1971, PANDEX was the first database produced as a commercial venture that was made publicly available by an online search service.

SIGNIFICANT CHANGES OVER TIME

From a macro view, several noticeable changes can be seen in computer searching and database activity over the last 40 years, including shifts in the following:

Computer Resources, Capabilities, and Availability

Location of Search Engines

Away from in-house searching to major online search services

Then away from major online search services to in-house online services (site licenses)

Then to distributed/personalized searching on Internet and World Wide Web, and a return to in-house searching

Search Mechanism

Initially entirely Boolean searching, then adding natural language searching, including word proximity searching

Then more searching by relevance ranking

Nature and Extent of Database Content for Users

Expansion of search output from limited sets of upper-case characters on impact printers to full upper- and lower-case character sets with many special characters and fonts on several kinds of output media.

Expansion of search output from only accession numbers, to citations, to citations plus abstracts, to facsimiles of original publications

Availability of more backfiles of full-text and non-bibliographic databases

Source of Database Content

From entirely in-house content to external databases from secondary abstracting and indexing services

Then more database linkage and direct access to document delivery sources

Then from mostly secondary services' databases to more use of primary publishers' databases

Then to more use of databases prepared by individuals and non-publisher organizations, and a return to in-house databases

Extent of Indexing Quality

Initially only a search capability of a few data fields (e.g., title, subject heading), then expansion to provide indexed access to many data fields (e.g., cited references, CODEN, publication date, 10th author of a paper)

Initially little subject access to the database (e.g., title word searching only, or assigned subject headings only)

Then shift to a reliance on less controlled indexing of file content [e.g., raw source text searching in lieu of controlled indexing and authority control; less precision in index construction (e.g., generally no distinction of "Weed" as personal name vs. company name vs. object name vs. place name)]

Identity of the Searchers

Initially from proposed professional end-user searching to delegated searching by information professional

Then shift from delegated searching to end-user searching

Extent of Support to the Searchers

Initially from zero to considerable support for delegated searchers

Then shift back to minimal support to both the delegated searchers and the user searchers

Revenue Sources and Sinks

Initially from zero revenue to positive revenue for database suppliers and search services

Then shift of revenue share away from search services to the secondary service database suppliers

Then shift of revenue share away from search services and secondary service database suppliers to primary publishers

Generation of new revenue sources for Internet and other service providers (e.g., advertising revenues)

Generation of new revenues for primary publishers

CURRENT AND CONTINUING ACTIVITY

Data Communications

After 40 years of growth and experience, we've seen major changes in our approach to computer searching. Much of this change came about because of the changes in data communications technology and capability. The high data communication speeds that are currently taken as a given, didn't start to appear until the mid-1980s. ARPANet, the first packet-switched data communications network, began operation in 1969, and was the precursor to the Internet. The TYMNET data communication services started to provide public service in 1972, followed by the TELENET service in 1975, both with 30 cps service. Speeds higher than 120 cps didn't happen until after 1982.

As an example of how far we've come, in 1976, 1 made the first online search in Egypt. This is something that would currently be done without giving it a second thought. But at that time, it took almost 24 hours of repeated Telex dialing just to grab access to one of the four very busy 15 cps Telex lines out of Cairo to reach the computer in California. Similarly, the first online bibliographic search in India was not made until 1976, with a leased line to an ESA/IRS computer facility in Europe.

As we all know, the increased data transmission speeds changed the behavior of the searchers, as well as the revenue and pricing decisions of the database suppliers and search services.

Computer Equipment

The capabilities of both large-scale computers and small personal computers have grown remarkably since the 1980s, and have completely changed the game for all parties. In the 1960s, when I worked at SRI on techno-economic studies for various computer manufacturers of the market feasibility of new large computers, we had some upper-bound tests of whether the new large computers could be taken seriously for large applications. One information retrieval test was to examine how well the proposed system could handle three well-understood benchmark problems of file storage and maintenance that involved what we thought at that time were very large files: (1) a Los Angeles property title file (millions of land records and associated legal documents); (2) the California

Department of Motor Vehicles Registration Files; and (3) the Library of Congress files. Very few systems could be considered for those problems then, but systems can support those applications now.

The database and search service industry has reaped benefits from continued improvements in the capability and cost-effectiveness of new computer and telecommunications equipment. Because increased computer and storage capacity became available for the same or lower cost, the major online services were generally able to add more simultaneous use capacity, and offer increased file capacity for new files, file updates, additional index and text capacity for existing files, and backfile additions, without directly increasing the service prices. All of this helped to grow the volume of usage of the services and the databases.

The ready availability of CD-ROM publishing equipment and associated software in the late 1980s also had an impact on the database suppliers and search services, and on the organizations that were doing a lot of searching.

CLOSING COMMENTS

We've seen tremendous changes in the evolution of database searching, with increased accessibility and speed, large increases in the number of services, databases, and searchers, and much more money to be made or lost by the participants and investors. Some of this change and increased capability is really dazzling. But on some fronts, it's discouraging to see places where we've not advanced. Some things haven't changed. Early online system operators, not wanting to tie up their ports and working memory space, used to monitor terminal transactions, and issue logoff commands to searchers who went for some time (e.g., three minutes) without hitting the keyboard. Some systems still have such timed logoff protocols, but now the searcher can get special PC software to essentially tap the keys for you, to keep you from being logged off.

Primary publishers, secondary services, and online service providers are still fighting the battles of revenue distribution, as well as the pricing and information distribution practices of federal agencies. Users and their advocates are still fighting the issue of fee versus free services, and copyright versus contract use controls.

There are instances where groups have reinvented the wheel, or forgotten or didn't know that there was a wheel. Systems are faster and flashier, but the search result may not be any better than they were ten years ago.

Features such as truncation or word proximity searching that were the norm on the major online services of the '80s are now only available as "special features" on some of the Internet search services. The ability to do a thesaurus lookup for subject searching, and the ability to easily incorporate the results of such a lookup in your search formulation is a capability that is absent in many of the current search systems.

Subject authority control seems to have been an activity that has been left out of much of today's Internet searching; it's an "inconvenient" issue that tends to get in the way of the system operators. The proper treatment of this issue requires the system operators to do something more than just load another file or build a couple of lookup tables, and they seem to be reluctant to move in that area. A good review of prior and under-utilized research results associated with indexing and access to information is provided in a recent publication by Marcia Bates3, and a good review of the need for traditional content authority control in today's electronic media is provided in a recent publication by Milstead and Feldman.4

Multi-file searches, using the same search formulation, and with good handling of any resulting duplicate citations, is also an area that has not been handled well in those instances when a searcher has moved away from the single-server use of any of the major services of the 80's to the alternative of searching scores of hundreds of individual Internet search sites.

In this age of highly touted search engines, it appears that few if any of them can provide some of the capabilities that were demonstrated in the 1960's. One example would be Don Hillman's LEADER

software and the LEADERMART service in 1970, which accepted a client's request for information in a narrative text form consisting of sentences that described the client's problem, and with text parsing of the request statement and matching of the parsed database, the search output would consist of citations or textual passages from the selected documents. Such an approach allowed you to write an abstract of a hypothetical document you'd like to see, and have that used as the search input. Few of today's commercial search systems provide that kind of capability as an alternative.

Another example of overlooked early technology would be SDC's question-answering system, SYNTHEX, which in 1965, working with a full-text portion of the Golden Book Encyclopedia, successfully responded to queries of the following forms:

What do policemen do?
What is a horse?
What do horses eat?
What do baby octopuses eat?
Describe a cat.
Name four small animals with long tails.

Even after 30 years, I'm not aware of any current operational online systems that can demonstrate that kind of capability.

People thinking about designing user-friendly front ends for online search systems might benefit from a review of the 1980s work by Charlie Meadow at Drexel on the IIDA (Individualized Instruction for Data Access) instructional and diagnostic front-end software for online search systems. Current designers could also benefit from the many user interface studies done on the online public access catalogs (OPACs) and from more current work at the University of Illinois at Urbana-Champaign on interactive term suggestion and the seamless integration of thesauri and classifications into online search systems.5,6

People thinking about the coordination of multiple subject indexing systems should be aware of, and review the early and extensive studies and projects done in the 1960s and 1970s on index merging and convertibility, such as those at ASTIA, Battelle, Datatrol, and Engineering Index.

People thinking about the development of multi-system or network switching search capabilities should be aware of, and review the experiences of the previous early work such as that done by Dick Marcus and others at MIT, and David Tolliver's work at the Franklin Institute Research Laboratory with his OL' SAM system.

But all things considered, these are very interesting times. I only hope that the current participants can improve their own operations or the pace of development by paying more attention to the efforts of their predecessors.

Endnotes

1.Philip R. Bagley, "Electronic Digital Machines for High-Speed Information Searching." M.S. thesis, Cambridge, MA: Massachusetts Institute of Technology. August 1951. 133 p.

2.Charles P. Boume, "Research on Computer Augmented Information Management," Menlo Park, CA: Stanford Research Institute, November 1963. Report No. ESD-TDR-64-177. NTIS Report No. AD-432 098. 49 p.

3.Marcia J. Bates, "Indexing and Access for Digital Libraries and the Internet: Human, Database, and Domain Factors." Journal of the American Society for Information Science, 49(13): 1185-1205, November 1998.

4.Jessica Milstead, Susan Feldman; "METADATA: Cataloging by Any Other Name," Online 23(1): 24-31, January/February 1999.

5.Pauline Atherton Cochrane; Eric H. Johnson, "Visual Dewey: DDC in a Hypertextual Browser for the Library User," in Advances in Knowledge Organization, Vol. 5 (pp. 95-106). Washington, DC: International Society for Knowledge Organization. 1966. pp. 95-106.

6.Bruce R. Schatz; Eric H. Johnson; Pauline A. Cochrane; Hsinshun Chen, "Interactive Term Suggestion for Users of Digital Libraries: Using Subject Thesauri and Co-occurrence Lists for Information Retrieval." Paper presented at the Digital Libraries '96: 1st ACM International Conference on Research and Development in Digital Libraries, Bethesda, MD, March 20-23,1996.