pliers are inadequate. The industry has concentrated on teaching and learning the mechanics of information retrieval rather than the fundamentals of applying information to customers' problems.

Much of the confusion about the value of on-line searching has been created by database suppliers and on-line vendors. Lack of cooperation between vendors and suppliers works to the detriment of customers. The on-line information industry needs to adopt a strong marketing orientation that focuses on the needs of customers rather than the needs of suppliers and vendors. Until we, as an industry, focus on the needs of customers, we will not be able to give meaningful answers to questions about the cost effectiveness of on-line searching.

## REFERENCES AND NOTES

(1) The Committee on Corporation Associates of the American Chemical Society undertook a study of the cost effectiveness of information systems in the late 1960s prior to the advent of on-line services and developed a semiquantitative measure of the cost effectiveness of the systems in use at that time. See "Cost-Effectiveness of Information Systems". A Report on the Subcommittee on Economics of Chemical Information of the Committee on Corporation Associates American Chemical Society; American Chemical Society: Washington, DC, 1969.

(2) Haynes, R. M.; Erickson, T. "Added Value as a Function of Purchases of Information Services". *Inf. Soc.* **1982**, *1* (4), 307–338.

# Cost Effectiveness of On-Line Searching of Chemical Information: An Industrial Viewpoint[†]

ROBERT E. BUNTROCK

Amoco Research Center, Standard Oil Co. (Indiana), Naperville, Illinois 60566

Received October 24, 1983

Cost comparisons and the cost effectiveness of on-line searching of information are reviewed. Topics discussed include on-line vs. manual searching, charge-out of search costs, efficacy of on-line searching, on-line vs. batch computer searching, vendor system comparisons, networking, searcher productivity, telecommunications, role of the intermediary, search transmission rates and on-line charges, editing of recorded searches, and increasing cost of on-line searching of chemical information.

In the 10 years that on-line searching has been more or less readily available, there have been several publications and presentations on the cost effectiveness and cost comparisons of various aspects of on-line searching. Two papers by Almond and Nelson[1,2] are key papers for evaluation of cost effectiveness of chemical on-line searching. These papers describe formulas for calculating and evaluating on-line search performance. The formulas are for on-line usage in general, but the specific examples cover chemical databases and patent searching in particular.

Logically, the first papers to appear on the cost effectiveness of on-line searching dealt with comparisons between on-line and manual (or printed index) searching. Most authors reached the conclusion that on-line was more cost effective,[3] and most of the subsequent discussion revolved around what costs were to be included.[4]

Of course, some of the costs involved in searching of any kind vary greatly depending on place and person, so potential on-line searchers had to apply and adjust the conclusions of others to their own professional and budgetary environment. The hardest decisions to make were for those information professionals attempting to justify on-line searching in an environment where no searching service was previously provided because it was too expensive, not staffed for, or both. Those institutions that already provided searching services and charged their customers for the service probably had an easy time justifying on-line searching. Qualitative comparisons could be made by experienced searchers in the course of providing the service, and with a charge-out system already in place, the service could be paid for by the customer.

That was the situation at the Amoco Research Center. As soon as subscription fees were dropped by the on-line vendors in July 1973, a contract was signed, and the Information Services Division at Amoco has been using on-line ever since.

Although charging for on-line searching services can inhibit to some extent experimentation and learning experiences by the searcher, those organizations that charge the customer for all searching costs probably began searching on-line earlier and were better able to justify on-line searching services to their management. By all searching costs, I mean out-of-pocket costs (results of charges made from outside the immediate organization including computer service charges, both corporate and from outside the corporation) plus charges for the searcher's time (made at the current corporate rate). For those organizations that only charge for out-of-pocket costs, on-line searching may also be harder to justify because staff is probably budgeted on overhead while nonpersonnel costs may be budget-line items and more vulnerable to budget restrictions.

On-line searching should be justified on overall effectiveness as well as cost effectiveness. At the Spring 1975 ACS meeting, for example, Buntrock compared searching *Chemical Abstracts Condenstates* on-line and using the printed *Chemical Abstracts* indexes.[5a]

Those who have only searched on-line, and have never searched manually, often seem to think of on-line searching as an end rather than a means to the end. On-line searching should be considered a tool, albeit a very valuable tool, in providing searching services. I believe a typical chemist or engineer should first consult "arms-reach" references to answer everyday questions, books on their shelf, the colleague next door, etc., and then consult secondary sources, especially on-line, only when the material cannot be found in the readily available sources or if it is known that it cannot be found there.

On the other hand, some are not yet aware of all that can be found in on-line sources, including the answers to many so-called reference questions. A real-life example at Amoco was a request for the papers by Smith at Stanford on artificial intelligence. Although one approach would be to use reference works such as *American Men and Women of Science* and *Directory of Graduate Research*, one readily finds that only an on-line search of *Chemical Abstracts* works. Dennis Smith has never appeared in *American Men and Women of Science*,

[†] Presented before the Division of Chemical Information, "Symposium on Cost-Effectiveness of Online Searching of Chemical Information", 185th National Meeting of the American Chemical Society, Seattle, WA, March 21, 1983, and at the 17th Middle Atlantic Regional Meeting (MARM), White Haven, PA, April 6, 1983.

0095-2338/84/1624-0054$01.50/0

and his works appear scattered throughout the Chemistry Department, if at all, in the *Directory of Graduate Research*, because he was not on the Chemistry Department teaching staff. A manual search of *Chemical Abstracts* will not work because "corporate authorship" in the printed *Chemical Abstracts* indexes is limited to patents, a restriction that does not apply to the on-line versions.

Although there are many things that cannot be found on-line, there are even more things that can only be done on-line because of inherent or necessary limitations of printed indexes. Also, several different concepts can be coordinated in on-line searching, whereas two or three concepts at most is probably the practical human limit of manual searching.

Some other early studies involved the comparison of the cost effectiveness of on-line and batch computer searching. Buntrock (Amoco) and Mullvihill (American Petroleum Institute) compared on-line and batch searching of the American Petroleum Institute (API) files.[5b] The two methods of access were found to cost either approximately the same (API) or on-line to cost somewhat less (Amoco). An entire symposium at an American Society for Information Science meeting dealt with this topic, but the discussion was limited to the comparison of on-line- and batch-generated SDI.

High prices were the prime reason that batch retrospective computer searching was never widely patronized. I remember an informal quote, for *Chemical Abstracts Condensates*, of $100 per year per "macrosection" (biochemistry, organic chemistry, etc.)! However, Kaminecki described IITRI's operations in 1975,[6] and batch retrospective searches were being run at prices more competitive with on-line searching.

Some user organizations maintained in-house loadings of various files for batch retrospective searching, but most of these were dropped when the files became well established on-line. However, if a batch loading can be justified (including the maintenance costs), searches requiring large amounts of printing will probably be cheaper than on-line with batch files.

At Amoco, Information Services suspended running American Petroleum Institute (API) files and two portions of Derwent, because internal charges (Computer Services also charges out at 100%) were roughly comparable to on-line charges. In the Chicago area, the sole remaining Amoco loading of a commercially available search file is the Comprehensive IFI U.S. Patents file. Comparisons are now being made between the new on-line loading of the IFI Comprehensive file on DIALOG and our batch loading.

The next type of comparison, that of the cost effectiveness of loadings of the same file on two or more on-line systems, actually began very early. At the Spring 1975 ACS meeting, a symposium on "User Reactions to CAS Data and Bibliographic Services", Prewitt compared searching ORBIT and DIALOG loadings of *Chemical Abstracts Condensates*.[7] The debate has continued, both on matters specific (often *Chemical Abstracts* or other chemistry-oriented files) or general. In 1979, Pemberton wrote in an editorial in *Online* magazine[8] that he would publish only one more price comparison because, within reason, searchers tended to use the system they liked best and which gave results that their clients liked best. The editorial preceded an article by Hoover[9] described by Pemberton as his "next to last price comparison article". Interestingly, the "last" article has not yet appeared.

Comparisons of various on-line systems continue to fluorish, especially qualitative comparisons. Quantification is usually made more difficult because one is usually comparing "apples and oranges". Even though the major systems seem to be in a state of convergent evolution, authors still stress the differences. New systems are constantly appearing, especially in Europe, but the authors of most comparisons, as well as proponents of new systems, fail to account for the cost of learning a new system as well as maintaining proficiency on that system. In the early days of on-line, veteran searchers told new users to "learn both systems" (ORBIT and DIALOG). However, after the advent of on-line users groups, new users started asking "what does it cost to learn a new system?" Significantly, within the Chicago Online Users Group (COLUG), these new users tended to come from the financial rather than the technical community.

One way out of the dilemma of learning all of the new systems is the use of an intelligent terminal or system interface that makes all files and host systems appear the same. However, these devices (and intelligent networks for that matter) must be programmed to accommodate all of the host systems involved, so the job is never complete.

An example of such a system interface is Sci-Mate, a program for microcomputers from the Institute for Scientific Information (ISI). Sci-Mate consists of two parts: an easy-to-use on-line searching program plus a database-management program. The search aid program permits automatic dialing and logon to five search systems: DIALOG, BRS, SDC, NLM (National Library of Medicine), and ISI's own search system. Once connected, Sci-Mate permits standardized searching via a menu-based system or by direct searching of databases on these five systems (or any other accessible system) by means of the normal command language. Search results can be downloaded to the database-management portion, which can also receive and store keyboarded information.

In addition, there is promise for "smart" networks such as CSIN (Chemical Substances Information Network). CSIN facilitates on-line access to many files on several systems and enables transfer of data between files and systems. For some types of search questions, searching is almost fully automatic including autodialing, automatic file access, and entry of search strategy. Several systems allow the transfer of data between files on the same system: CAS ONLINE/the CA File, DARC QUESTEL, SDC ORBIT (with "Print Select"), and DIALOG (with "MAPRN"). However, there is great potential here for acquiring information from various sources, in ways that work best on that source, and combining and manipulating that information with material contained in other systems by using that system's own set of strengths. For example, I previously suggested to Chemical Abstracts Service (prior to the loading of the CA File) that they participate in CSIN with CAS ONLINE and the abstract text file and not reinvent the wheel by developing their own bibliographic and indexing file. The ability to use retrieval as search-entry material among several systems, as CSIN allows, is a very desirable and productive capability.

Concerning enhancement of searcher productivity, I would like to mention some items that are counterproductive. In my opinion, the golden age of searching is past, and things are going to get worse before they are going to get better. The main reason for my pessimism is the worsening state of telecommunications. Service on the major networks tends to be erratic at best and nonexistent at the worst. When either of the networks fails (or slows down so much that users leave), everyone on-line at the time tries to use the other network or tries direct-dial access. Both of these procedures degrade performance on the remaining links. At least one of the major services had extensive trouble for over a year with direct-dial access. At least some information groups are willing to pay the price premium for direct-dial access if they get increased performance, which we did in the "golden age". However, if the service is just as poor (line noise seems to be common to all modes) or if one cannot even use the service, direct dial is a poor backup. Between telecommunications degradation and degradation of vendor software response time at certain times of the day, many of us in the Midwest find it nonproductive to search in the midmorning to late morning and in the early afternoon. Colleagues have told me that marketing people from the networks do not even seem to be aware that there is a problem and that it is getting no better.

The question of whether or not the search requester should be present at the time of the search still seems to be asked. In a paper by Buntrock,[10] it was pointed out that searchers at the Amoco Research Center discouraged the presence of the requester in that they did not encourage it. Although several reasons were given that were thought to be valid and several other controversial topics were addressed, this one point (user present or not present) triggered most of the discussion following the paper, and several people probably left continuing to believe that there is no such thing as a good search without the user present. I feel that this opinion is probably held more in medical school libraries, in academia in general, or anywhere else where the typical search requester may not take the searcher into their confidence, views the searcher as nonprofessional, or both. There are times when the requester's presence is essential, but there are also times when the requester's presence costs the time of two professionals rather than one. There is more to on-line searching than running a terminal, and it is a pretty rare searching professional who searches more efficiently with someone observing.

One of the advantages of 1200-baud searching over 300 baud that most searchers began with is the ability to print much more retrieval on-line for a given price. This is especially valuable for those searchers that are needed urgently. Boyce and Gillen reported in 1981[11] that it is more cost effective to print on-line rather than off-line at 1200 baud. They commented that vendors would probably have to adjust their pricing if they wished to maintain the attractiveness of off-line printing. However, they confined their studies to shorter formats, which are less interesting to many, including us at Amoco. Stewart mentioned that Amoco searchers had been testing the cost effectiveness of on-line printing[12] but the results had never been published. In this work, Buntrock and Stewart timed the on-line printing of typical output with a stopwatch in several formats, especially long formats with or without abstracts. In summary, it was found that long formats from most files could be printed on-line for about the off-line cost or slightly less. If abstracts were printed in addition, the on-line print cost was somewhat higher than off-line. However, that was with 1979–1980 prices before the institution of on-line hit charges. One of the concepts developed in our study was that of the "breakeven time" or that time required to print an average record with the same cost in on-line connect time as the price of an off-line print. For CA Search, that time was about 8.5 s for either ORBIT or DIALOG long formats. Now however, the breakeven time is determined by the relative cost of the off-line print, namely, the difference between the off-line print charge and the on-line print charge. We have determined that the breakeven time is now about 4.4 and 6.4 s for CA SEARCH on ORBIT and DIALOG, respectively. This means that only shorter formats can now be printed cost effectively on-line. The change in pricing predicted by Boyce and Gillen took the form of on-line print charges and was apparently induced by the database suppliers, and not just the vendors.

For some time, searchers have noted the potential efficiency of being able to record search results in digital form and use these recorded searches to produce more usable search reports by word-processing techniques. The appropriate hardware and software have been available for a few years, and results and experiences have been described by groups from American Critical Care[13] (formerly Arnar-Stone), Exxon,[14] and Amoco.[15] Clients definitely like the new reports and bibliographies, and the cost effectiveness of the searcher–user system definitely increases. The next procedures to be explored include sorting and merging of output from more than one file. For most searching questions, several searchers have found that multiple-database searching is a necessity. However, proper treatment of the inevitable duplicates is not cost effective as yet. For example, typical petroleum company information groups need to integrate search results from Chemical Abstracts Service, API Literature and Patent files, and Derwent. Full cross-file capability on all file elements and good sort-and-merge programs are a necessity for productivity improvements.

Although costs of on-line searching were supposed to decrease with time (along with decreasing hardware costs), the opposite seems to be true lately. Several vendors have increased the royalty charges for use of the file (and thereby the on-line connect time) or instituted on-line print charges, or both. Chemical Abstracts Service is a case in point. Both forms of price increase have been instituted, and royalties continue to increase annually. The price increases have been especially sharp for the CAS Registry System based files, or the resulant "chemical dictionaries". Others have speculated that CAS was raising the prices to bring the cost of a search on outside vendor systems up to the initially rather high price for a search on CAS ONLINE. Buntrock pointed out in 1981[16] that many compound or substructure searches in the "chemical dictionaries" could be carried out for only a fraction of the cost of a search on CAS ONLINE (then, about $100). Since then, prices for some CAS ONLINE searches, for example, specific compound searches, have decreased considerably, but the fact still remains that many of these searches can still be run cost effectively on ORBIT and DIALOG.

Compounds retrieved from these substance searches can be coordinated with bibliographic and indexing information on ORBIT, DIALOG, DARC QUESTEL, and, now, CAS ONLINE and the CA File. This is important, because the probable majority of chemical substance searches are not only concerned with whether or not a compound exists but, in addition, if further specific information is available. CAS has released the abstracts (which are printable but not searchable) to their own system only.

Another database supplier that decided to start its own on-line system for some of its files is ISI. Additionally, Derwent started the consortium that was the first INFOLINE. Although all of these database suppliers had their reasons for being their own vendor ranging from new technology and innovative file structure to perceived lack of control with outside vendors, they must perceive the on-line business to be profitable.

It should be pointed out that lower cost alternatives to portions of the chemical dictionary files exist. The various TSCA files (Toxic Substances Control Act) are evidently considered to be as good as "in the public domain", and the previously mentioned royalty increases and on-line hit charges have not been applied to them. Also, the price for CHEMLINE on NLM, although one more than $100 per hour, has been reduced somewhat recently, probably reflecting the significant fraction of the file that is part of TSCA.

Although increased prices affect all on-line users, one can assume that they fall particularly hard on the academic sector. One academic chemical librarian said that cost recovery for on-line searching was not a problem because all of the professors had searching money built into their research grants, but such fortuitous circumstances are probably in the minority. Problems of paying the increased out-of-pocket costs for on-line searching are probably more acute in academia. If on-line searching is involved at all in the academic courses and training programs that do exist, I am also sure that paying for the capability is a big problem. In addition, although I feel that on-line searching has great pedagogical value for instruction in chemical information, both academic and nonacademic, I also feel that the training should not be just in on-line methods. A solid grounding in the use of all chemical information sources is needed, and on-line instruction should be integrated into the remainder of the course work. The ACS Division of Chemical Information has a new Education Committee, chaired by Arleen Somerville, and its mission deals with a wide range of chemical information instruction and awareness topics, both academic and nonacademic.

There are occasions when on-line services can be priced too low. Several years ago the MEDLINE/TOXLINE users community would ask NLM for various file improvements, improvements that were already featured in the ORBIT systems. Although sympathetic, the answer would often be that the on-line charge was so low the user would be able to afford the noted inefficiencies.

Probably because of a long history of cost recovery for novel services, providers of information services in the private sector have long been concerned with cost effectiveness of, and productivity in the use of, those novel services. Although not addressed previously in this paper, database quality is also of prime importance. Well-indexed and -abstracted material makes for cost-effective searching because intellectual effort spent in creating the file facilitates productive use of the file. After all, the total cost of the use of a file or service is the sum of all charges for system use plus the "people" costs associated with that use. To facilitate more productive end user or customer use of information, more work needs to be done on effective search recording, constructive sort, merge, and edit programs, and also training of end users in the use of information services.

## REFERENCES AND NOTES

(1) Almond, J. R.; Nelson, C. H. "Improvements in Cost Effectiveness in On-Line Searching. I. Predictive Model Based on Search Cost Analysis". *J. Chem. Inf. Comput. Sci.* **1978**, *18*, 13-15.

(2) Almond, J. R.; Nelson, C. H. "Improvements in Cost-Effectiveness in On-Line Searching. II. File Structure, Searchable Fields, and Software Contributions to Cost-Effectiveness in Searching Commercial Data Bases for U.S. Patents". *J. Chem. Inf. Comput. Sci.* **1979**, *19*, 222-227.

(3) Magson, M. S. "Modelling On-Line Cost-Effectiveness". *Aslib Proc.* **1980**, *32*, 35-41.

(4) Lancaster, F. W. "Some Considerations Relating to the Cost-Effectiveness of Online Services in Libraries". *Aslib Proc.* **1981**, *33*, 10-14.

(5) (a) Buntrock, R. E. "Searching *Chemical Abstracts* vs. *CA Condensates*". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 174-176. (b) Buntrock, R. E.; Mulvihill, J. G. "The American Petroleum Institute (API) Data Bases: Comparison of On-Line and Batch Searching"; ASIS Mid-Year Meeting, 3rd, Johnstown, PA, May 17, 1974.

(6) Kaminecki, R. M.; Llewellen, P. A.; Schipma, P. B. "Searching *Chemical Abstracts Condensates*, On-Line and Batch". *J. Chem. Inf. Comput. Sci.* **1976**, *16*, 125-127.

(7) Prewitt, B. G. "Searching the *Chemical Abstracts Condensates* Data Base via Two On-Line Systems". *J. Chem. Inf. Comput. Sci.* **1975**, *15*, 177-183.

(8) Pemberton, J. K. "The Inverted File". *Online* (*Weston, Conn.*) **1979**, *3*, 6-7.

(9) Hoover, R. E. "A Comparison of Three Commercial ONLINE Vendors". *Online* (*Weston, Conn.*) **1979**, *3*, 15-21.

(10) Buntrock, R. E. "The Effect of the Searching Environment on Search Performance". *Online* (*Weston, Conn.*) **1979**, *3*, 10-13.

(11) Boyce, B. R.; Gillen, E. J. "Is It More Cost-Effective to Print On- or Offline?". *Ref. Q.* **1981**, *21*, 117-120.

(12) Stewart, A. K. "The 1200 Baud Experience". *Online* (*Weston, Conn.*) **1978**, *2*, 13-18.

(13) Fortune, J.; Horwich, J.; Schwartz, R. "Use of a Word Processor Interfaced to a Mini Computer to Facilitate On-Line Searching". "Abstracts of Papers", 175th National Meeting of the American Chemical Society, Honolulu, HI, Apr 1-6, 1979; ACS/CSJ Chemical Congress; CHIF 54.

(14) Dedert, P. L. "Electronic Editing of Online Search Results—Choices and Experiences". Tri-Society Symposium (ACS-CINF, ASIS SIG-BC, SLA Chem. Div.), Columbus, OH, Oct 17, 1982; Abstr.

(15) Stewart, A. K. "Selection and Use of Equipment to Manipulate Search Output". Proceedings of the ASIS Annual Meeting, 44th, Washington, DC, Oct 25-30, 1981, pp 251-252.

(16) Buntrock, R. E. "Chemcorner". *Database* **1981**, *5*, 79-81.

# Computer Storage and Retrieval of Generic Chemical Structures in Patents. 5. Algorithmic Generation of Fragment Descriptors for Generic Structure Screening

STEPHEN M. WELFORD, MICHAEL F. LYNCH,* and JOHN M. BARNARD

Department of Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.

Received December 8, 1983

Considerations for the use of limited-environment screens for screening generic chemical structures are discussed. The general strategy and detailed procedures for the automatic generation of screens from the extended connection table representation (ECTR) of generic chemical structures are described. A bitscreen record for generic database structures and specific, generic, and substructure queries is described, and a number of screening algorithms are proposed.

## INTRODUCTION

Previous papers in this series have introduced a novel approach to the computer representation and searching of generic chemical structures in patents.[1-5] This approach employs a systematic language GENSAL[2] for the interactive input of structure information in both graphic and textual form, during which process a machine-based representation of the generic structure is automatically created.[6] This representation is called the extended connection table representation (ECTR) and has been described previously.[4]

A topological structure grammar TOPOGRAM has been proposed,[3] and its associated generative and recognitive algorithms have since been developed. The potential uses of TOPOGRAM in the context of a generic structure storage and retrieval system have been described under three areas of application. These are, first, as a means of providing for computer storage a compact description of generic nomenclatural expressions, and particularly of homologous series terms, second, as a device for generating structural fragments characteristic of these generic expressions for use in screening and subsequent structure-matching procedures, and, third, as a generalized mechanism for determining the inclusion of specific substructures within the radical classes defined by generic nomenclatural expressions.

The systematic nature of the GENSAL language enables a complete topological representation of many types of generic structural descriptions to be created automatically in the form of an ECTR. The organization of the ECTR, whose com-

0095-2338/84/1624-0057$01.50/0 © 1984 American Chemical Society