A Comparative Report on an On-Line Retrieval Service Employing Two Distinct Software Systems†

DONALD J. HUMMEL

11140 Powder Horn Drive, Potomac, Maryland 20854

A subjective evaluation comparing the National Aeronautics and Space Administration RECON and National Library of Medicine ELHILL 2 on-line information retrieval systems is presented as they relate to the delivery of a major service dealing with a collection of toxicological bibliographic references (TOXLINE). Comparisons of file structure, response time, command language, system environment, and other relevant features are made based upon actual experience with both systems.

INTRODUCTION

The National Library of Medicine (NLM) provides a variety of on-line information retrieval services dealing primarily with biomedical information.[1] Currently, the two major offerings are MEDLINE (MEDLARS On-Line) and TOXLINE (Toxicology Information On-Line).[2] Although both bibliographic information data bases are accessed by the same software system, the nature of each file is appreciably different. The MEDLINE citations are indexed using a controlled vocabulary (Medical Subject Headings or MeSH) whereas TOXLINE is basically oriented toward the free text of abstracts associated with the citations. MEDLINE is consistent in style and format following rigid standards established by the NLM. TOXLINE, being a compilation of references and abstracts from multiple secondary sources, contains no standardization for the file as a whole.

NLM on-line experimentation began in late 1967 using the ORBIT programs of Systems Development Corporation (SDC). The MEDLINE prototype (AIM-TWX) became operational in June of 1970. Whole text search experimentation on a limited collection of toxicological material was begun in mid-1970 using the Mead Data Central system since it was specifically designed for whole text search and retrieval. The success of this latter experimentation led to a competitive procurement for a whole text information retrieval system in early 1971. Mead Corporation, Informatics, Inc., and Battelle Memorial Institute responded to the request for proposal, but surprisingly both SDC and Lockheed submitted a no-bid. Evaluation of the proposals led to a contract being let to Informatics, Inc., in late March of 1971. The system proposed by Informatics was the National Aeronautics and Space Administration (NASA) version of the STIMS/RECON package with systems enhancements to facilitate free text search. During the succeeding year, RECON was modified and the TOXLINE films were built. The TOXLINE service was publicly announced and demonstrated in April of 1972, and service instituted in October of that same year. A policy decision by the NLM in December of 1973 led to the transfer of TOXLINE to the NLM system. Service was offered from NLM on April 1, 1974 and discontinued through Informatics on April 25, 1974. The NLM information retrieval system consists of a modification of the SDC ORBIT programs and is identified internally as ELHILL. All comparisons drawn in the remainder of this presentation are based upon a relatively brief period of one and one-half months of operation at NLM and a substantially longer period of RECON operation.

† 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.

TOXLINE FILE CONTENT

TOXLINE, at the time of this evaluation, consisted of six discrete data files which constituted subfiles of the master data base. The subfiles included the *Chemical-Biological Activities* material from Chemical Abstracts Service, *Toxicity Bibliography* from the NLM, *Abstracts on Health Effects of Environmental Pollutants* from BioSciences Information Service, *International Pharmaceutical Abstracts* from the American Society of Hospital Pharmacists, *Health Aspects of Pesticides Abstract Bulletin* from the Environmental Protection Agency, and a special collection of materials gathered by Dr. W. J. Hayes, Jr. All files were arranged so that they were searched simultaneously in response to a single query. The TOXLINE data base was comprised of approximately 300,000 bibliographic citations, 75% of which contained full text abstracts and 20% with MeSH indexing. The average record length, exclusive of field and record identification internal to the retrieval system, exceeded 800 characters per record. Because of storage space limitations, the ELHILL file was constructed without the abstracts available for the material prior to 1970 and without author addresses, thereby reducing the linear or header file by over 85 million characters or approximately one-third less than the RECON version. Additionally, five fields were concatenated into two fields, further reducing the need for internal overhead characters. The major difference between the files existed in the inverted file structure, the number of searchable fields, and the term inversion procedures.

INVERTED FILE STRUCTURE AND SEARCH FIELDS

The ELHILL inverted file structure is very simple and straightforward. All search elements from all fields are contained in one alphabetically arranged, heterogenous listing in a paired file structure consisting of an index file and a postings (linear file record pointers) file. The index file contains the term in a 39-character field, a count of the number of postings and either the posting when there is only one occurrence of the term or an address of the postings list for multiple occurrences of the term. All postings are four-character records. For textual terms one and only one posting is carried in the inverted file for multiple occurrences of a term in a single record. The six classes of search elements or fields available under ELHILL are: authors, language, journal coden, year of primary publication, text terms from both the title and abstract fields, and Chemical Abstracts Service (CAS) Registry numbers.

The RECON inverted files were hierarchical in structure consisting of a masterfile index, a cylinder index, and a track index with a further subdivision of the actual postings by subfile and field where appropriate. The masterfile index was divided into sections for identifying each unique

linear or inverted file, with a sufficient number of pointers to identify each of the cylinders of storage required for the file. Each masterfile pointer consisted of a 36-character key and an address field of the device and cylinder containing the next level track index. The maximum size of a term in the actual inverted file was a maximum of 32 characters and the minimum was two characters. The RECON system contained twelve unique inverted files for TOXLINE including secondary source identification, authors, primary journal title, journal coden, page number, year of primary publication, number of references, language, document type, MeSH classification number, CAS Registry number, and text terms. The text term inverted file was further subdivided to identify either the title or abstract field as the source of the term. Each posting was a four-character field except for text term postings which were eight characters in length.

The ELHILL structure permits a very rapid response to a query requiring only a postings or record count using the imbedded count in the index file. A query of this nature does not move any postings to the user work area but simply extracts the count from the index file. RECON, in response to a similar inquiry, not only must go to the postings file to count the entries but then moves all postings to the users work area, translating eight-character text postings to twelve-character postings during the move. RECON requires much more time to respond and uses up appreciably more of the central processing unit (cpu) resource.

TERM INVERSION PROCEDURES

All six searchable elements in ELHILL and eleven of twelve in RECON were inverted in a similar manner, each yielding a four-character posting associated with the inverted term. A major area of departure between the two systems existed in the handling of text terms which consisted of all terms in both the title and abstract fields. For ELHILL, terms from the title and abstract fields are pooled, and only one occurrence of a term in a record is carried in the inverted files. The associated posting identifies only the source record and the fact that the term came from the text. In RECON every occurrence of a term in a record was carried in the inverted file. Additionally, the posting contained information regarding the field (title or abstract), the sentence within the field, the word position within the sentence, and the physical location of the first character of the word within an internal boundary defined by the system. All of this intelligence was compressed into four characters and combined with the standard four-character record identifier yielding an eight-character posting for permanent storage. When executing a term retrieval the posting was expanded to twelve characters to take advantage of half-word operations in the IBM systems.

The facilities gained by the expanded posting in RECON allowed word proximity searching, term highlighting in the output mode, and discrete selection of terms from either the title or abstract portion of the records. Word proximity is sometimes critical in a free text system, and the ability to specify a distance relationship often precludes false drops and thereby improves precision.[3] Unfortunately, a heavy price was paid in processing textual postings owing to the increased length (double in permanent storage and triple in temporary working storage) of the postings processed plus the additional requirement to translate the eight-character storage postings to a twelve-character working posting. The ELHILL approach, while requiring very little processing time, fails to provide for either word proximity relationships and highlighting of terms. Additionally the heterogeneous admixture of all searchable items into a single inverted file listing in ELHILL created data retrieval problems not experienced in RECON.

OPERATING ENVIRONMENT

The operating environment of the two systems was also quite different. The RECON software was resident in equipment provided by a commercial timesharing company acting as a subcontractor to Informatics, Inc. The ELHILL software is operational in the NLM computer center. Both systems were accessible by direct telephone dial-up and through the Tymshare, Inc., network (Tymnet). The communications link for RECON access was *via* a minicomputer combination of COM-10 to COM 40 to the main computer, while at NLM an IBM 3705 communications controller is in use. RECON ran in an IBM 360/65 while ELHILL is operational in an IBM 370/158. The message handling was under the control of a proprietary system (ALPHA) for RECON while ELHILL employs a standard IBM TSO (time-sharing option) package. Although not a direct function of the respective systems, RECON was capable of handling data transmission up to 120 characters per second while ELHILL is limited to a 10 to 30 cps rate. Data storage at the timesharing company was double density Ampex drives and disks, and 2330's at NLM. Internal system priority was variable at the subcontractor facility, but nominally was at the same level as the highest level timesharing job. At NLM the ELHILL programs run at the level of a systems task.

RECON as structured for the delivery of TOXLINE services was limited to ten simultaneous users. With a user community of 80 organizations, this limitation did not pose a major problem during this time. ELHILL is capable of serving 50 simultaneous users and supports both MEDLINE and TOXLINE. Fortunately, the MEDLINE service has a second service center in New York, and therefore saturation of resources at NLM seldom occurred. Informatics offered 14 hours of service per day with a half-day on Saturday, while NLM offers only 8 hours of access time per day and no Saturday service. The reduced hours of service apparently affected the West Coast users primarily, although several other users indicated that they preferred to do their searching during late, nonpeak hours.

COMMAND LANGUAGE

Two major differences between the systems from the standpoint of the user are the command language syntax for each, and the implied search capability and data manipulation afforded by the available commands. From a cosmetic standpoint, the difference between entering commands is that ELHILL requires all commands to be identified by a descriptive term and enclosed in quotation marks except for a search command while RECON requires a preceding mnemonic of a minimum of two characters for every command. Numerous other requisites involving the use of special characters and data entry conventions exist for both systems at the individual command level. The use of such characters and conventions is not necessarily consistent from command type to command type in either system which leads to the conclusion that the human engineering factor for both systems leaves a great deal to be desired.

Both systems have a rather comparable armamentarium of commands to accomplish different classes of objectives. Housekeeping commands for starting and ending a session or releasing prior search strategy to recover work space are very similar in format and function. User support commands which provide on-line tutorial and explanatory materials exist in both systems but are more extensive in ELHILL. Data output capabilities have many similarities such as three standard formats, limited optional formatting, and defaults to the number of documents printed if not otherwise specified. Generally the on-line and off-line output methodology used by RECON is superior to ELHILL in that there is greater flexibility in formating rec-

ords, no limitation of the maximum number of records requested, a method to halt output in progress, a feature permitting cathode ray tube (CRT) terminal paging, and an ability to produce continuous hard-copy output at a terminal without interruption. RECON also offers an on-line sorting option not available in ELHILL which is extremely useful in working with the TOXLINE files. A feature of mutual availability permits the association of a thesaurus to the data base; however, this is of no importance to TOXLINE.

There is a great deal of difference between the structure of the search commands and the capabilities offered by these commands. Three functions are normally required in searching free text to obtain reliable results, namely, the identification of the terms available, a selection of these terms and the associated records (or surrogate postings), and combination of the terms or records to yield a unique collection of records meeting the objectives of the search. RECON employs three commands (EXpand, SELECT, and COMBINE) to accomplish the objective while ELHILL requires only two commands (NEIGHBOR and the undefined command which will select and combine in one step).

The EXPAND and NEIGHBOR commands both allow the user to examine the index entries to the data base, but there the similarity ends. In RECON the user specifies the desired index (text, author, journal coden, etc.) and the term desired. The system responds with a ten-item list of terms alphabetically adjacent to the defined term indicating the number of associated postings. Each term is identified by a sequence number. The user may page down the list up to 99 terms. Any or all terms in this list may then be selected by a mixture of sequence number ranges and unique numbers. The ELHILL NEIGHBOR command requires only the definition of the term and responds with a list of five terms alphabetically adjacent to the specified term as they appear in the consolidated index with an identification of the source field (author, text, journal coden, etc.) and the number of associated postings. The user may move up and down the list but cannot select terms directly from the list.

RECON, by employing distinct SELECT and COMBINE commands, permits extensive, multilevel nesting of term combinations which have been previously collected into sets by use of the SELECT command. ELHILL permits only one level of nesting in its most rudimentary form in a single command. Additionally, the RECON SELECT command can be used to select multiple terms to be combined by a logical Boolean "OR" operator to produce a single set or document collection. The SELECT command is also the vehicle for proximity search using the Boolean "AND" operator. Proximity of the two words on either side of the "AND" operator may be defined as up to eight words on either side or in the same sentence. Multiple proximity relationships may be stated in the same SELECT command separated by the "OR" operator. The proximity search feature which is a most effective capability to have in reducing false drops in a free text data base is not directly available in ELHILL. However, a special STRINGSEARCH command does exist which can accomplish a limited amount of the facility offered by RECON.

Both systems support a stem or root word retrieval capability where the root is the first characters of the word. In RECON the index to be searched must be specified and the result is an unambiguous set containing the desired records. In ELHILL the response to a root search may encounter two difficulties; multimeaning and/or record overflow. Multimeanings basically indicate the root specified is not unique to one data field (*i.e.*, fish may be the root for both text terms and authors). Multimeanings can be avoided or handled if encountered, but for TOXLINE users they represent an undesirable hurdle. The record overflow is a problem unique to ELHILL, and, although seldom encountered, cannot be overcome without extensive effort.

The limitation upon the number of search statements in ELHILL is 16 (25 in the new version of the programs) while RECON will support 98. The fact that multiple terms may be selected and combined in one statement in ELHILL reduces the impact of this restriction. Furthermore, both systems provide mechanisms for releasing search statements to conserve both the operating area internal to the system and the number of available statements to be used.

RESPONSE TIME

All of the factors dealing with file design, inverted file structure application system efficiency, and operating environment have an appreciable effect upon system response time. These items and overall system throughput are the concerns of the data processing manager. Response time to an end user, however, is the elapsed time between sending a message and receiving a response. The user is not normally concerned with the various elements which influence the extent of elapsed time, only the actual delay itself. From an empirical sense the RECON response time to direct dial users seldom fell below eight to ten seconds on any command and most often exceeded this range by a factor of two or three. ELHILL response time normally fell in the four to seven second range and generally remained below ten seconds even with a load of over 30 simultaneous users. Both systems on occasion experienced problems which would cause the response time to become excessively slow with response times exceeding one minute. Problems creating this situation were not always directly related to either retrieval system but may have involved TSO or ALPHA, the operating system, conflicting requirements of other applications contesting for the computer resources, the communications hardware, or to a lesser extent the communications network itself.

A technical evaluation of response time was conducted using both software and hardware system measuring devices, and modeling techniques. Based upon a model created for SCERT (a proprietary simulation program of Comress, Inc.)[4] with data collected from the subcontractor's system using Dynaprobes (Comress, Inc.)[5] and file definitions compiled by analysts, the type of response time indicated above was predictable and indicated areas requiring attention to alleviate some of the response time problems in RECON. By December of 1973 RECON demonstrated a more rapid and stable response time and an improved degree of program reliability. Major problems with the total system reliability were often identified as interface problems between software external to RECON or the communications controller. Using a SCERT model of TOXLINE running under ELHILL, a normal 4.5-second optimal response time was predicted. A comparison of the modeled performances revealed what was already understood but not quantitated, namely, that RECON input/output and cpu requirements consistency exceeded those of ELHILL, the actual values varying from command class to command class. Background system operating, not considered an integral part of the modeling exercise, created response problems in the afternoon for ELHILL, slowing response to as much as 30 seconds. ELHILL system reliability was very good over the period observed; however, there were a number of hardware problems encountered. A new, more efficient and effective ELHILL program has been written and is in testing. The TOXLINE files will not be transferred to the new system until sometime in 1975.

DATA BASE MANAGEMENT

Systems differences from the standpoint of the data base

manager begin following conversion of the secondary sources data to a standard input format. For RECON this meant a standard STIMS input record which was then processed by the system to generate the file accessible by the RECON software. The update facilities of STIMS were available for file maintenance of existing records as well as adding new records. The generation of records for ELHILL was an indirect affair circumventing much of the file creation and maintenance procedures employed for MEDLARS records and therefore restricting much of the file maintenance activity. Additionally, physical record size limitations under ELHILL could represent a serious difficulty if the concatenation of abstracts for duplicate citations were designated as a method of reducing overall file size. Record handling in the STIMS/RECON does not have this limitation. Finally, the RECON inverted files are structured in such a manner that they can be acted upon by the STIMS maintenance programs. This facility permitted direct changes to be applied to the index files without incurring a requirement to totally regenerate the indexes as is required in the MEDLARS/ELHILL system following changes to the linear file.

COST

Of all the changes introduced by the switch from RECON to ELHILL, the most universally accepted was the reduction in charges levied for TOXLINE services. The $350.00 initiation fee and $100.00 per trainee charge were eliminated entirely. The on-line cost was reduced from $45.00 to $15.00 per connect hour, and off-line print was reduced from approximately $0.25 to $0.10 per page. The real reduction in on-line charge was greater than the actual dollar amounts shown because of the faster response time realized by using the ELHILL system.

What facilitated the dramatic reduction in cost was not a change in policy but the economic advantage of incorporating TOXLINE into an existing Government service, thereby eliminating the contractor and subcontractor costs of providing the service. The NLM has always assumed full financial responsibility for the creation, maintenance, and storage of the TOXLINE file, development and maintenance of the retrieval package, and general administrative costs to make the system available under the contract with Informatics, Inc. The user had paid only for that incremental cost incurred in the actual delivery of required services, namely, computer resources required, data transmission costs, mailing and handling costs, file-use fees, training at the contractor's facility, billing costs, and a portion of the personnel directly related to customer support. With the elimination of the commercial rates for these services and the assumption of customer support services by the Government at no cost to the user, the rates for service underwent the dramatic reduction.

CONCLUSION

My subjective evaluation of the two systems is based upon a long-time exposure to TOXLINE under RECON and only limited experience with ELHILL. Response time and cpu utilization requirements of ELHILL are superior to those of RECON. The ELHILL language is a bit more simple than that of RECON, but the data retrieval capability and ancillary features of STIMS/RECON are superior in almost all respects except for response time and cpu utilization. Assuming an equal operating environment (*i.e.*, contractor or Government installation) and therefore relatively comparable costs for delivery of service, I would be inclined to select RECON as the TOXLINE delivery vehicle based upon the performance of both systems at this time (May 1974).

LITERATURE CITED

(1) McCarn, D. B., and Leiter, J., "On-line Services in Medicine and Beyond," *Science*, **181**, 318–324 (1973).
(2) Kissman, H. M., and Hummel, D. J., "TOXICON-An On-line Toxicology Information Service," *Chem Tech*, **2**, 727–730 (1973).
(3) Vasta, B. M., "Proximity Searching as an Information Retrieval Tool," *J. Chem. Inf. Comput. Sci.*, in press; presented at the 166th National Meeting of the American Chemical Society, Chicago, Ill., Aug 1973.
(4) Datapro Research Corp., "SCERT," Datapro 70, December 1973.
(5) Weinstein, M., "Monitors: A Great Untapped Resource," *Computerworld*, April 3, 1974.