"on-line information retrieval systems are characterized as time-shared computerized communications networks. . ."

Information Storage and Retrieval Systems

by

Loida P. Faustino

Introduction

Information is a great national and international resource. Like all resources, however, it has to be processed and distributed before it can be put to productive use. Scientific and technological knowledge has been rapidly expanding and the production of this knowledge is very costly. It, therefore, becomes a must that this knowledge not be wasted. Further, knowledge already gained from previous research forms the basis for all future research efforts. The effectiveness of future work in universities, governmental and industrial laboratories depends on the availability of past knowledge, which in turn depends on the existing channels of information transfer.

Information science is an interdisciplinary science that investigates the properties and behavior of information, the forces governing the flow of information, and the means of processing it for optimum accessibility and usability.

Information system is the combination of human and computer-based capital resources which results in the collection, storage, retrieval, communication and use of information.

Management information systems (MIS) were developed to monitor everyday operations of organizations for the purpose of efficient management (planning, decision-making, reporting and control). MIS process mostly clerical data and other inputs from standard processes in an organization's work stream. Statistical reports may be generated automatically or on command by standardized report generators. Such systems usually require large scale computers to be used in manipulating large data bases. With MIS, if the complexity of processing is not too great, i.e., only a limited number of tightly-defined queries is allowed, many users may be serviced.

Information storage and retrieval systems (IS & R), such as those used in libraries and scientific documentation centers, require large data bases with greater flexibility than is usually allowed in MIS. Some of the data in IS & R

systems may be stored in more than one file in the system to facilitate searching and ensure adequately fast response to on-line queries. By providing immediate or almost immediate response to queries, users may redefine requirements depending on data already received.

Wherein data in MIS is usually endogenous (generated within the organization), the bulk of data in IS&R systems is exogenous. It is obvious, therefore, that even the largest IS&R systems cannot hold on-line all information that a user may wish to search. This is where complex selection and purging techniques are employed to ensure that only the most frequently used materials are stored in the rapid access files.

Some Characteristics of On-line Retrieval Systems

An on-line information storage and retrieval system is one in which a user can, through the computer, directly interrogate a machine-readable database of documents or document representations, e.g., title or abstract. In an on-line system, there is two-way communication between the computer and the user by way of input-output devices such as a teletypewriter or a CRT display. Although an on-line system may be operated in a dedicated mode, it is more often implemented in a time-shared environment. An on-line time-shared system operates through a number of independent concurrently usable terminals, giving each terminal user processing time when he needs it and creating the illusion (at least most of the time) that he is the sole user of the computer.

Another expression associated with on-line retrieval systems is real time. Applied to information retrieval, real time implies that the computer responds quickly enough to interact with a user's heuristic search processes. It is contrasted with delayed time, which is a characteristic of batch processing systems.

The first mechanized information retrieval systems were designed for an off-line batch processing mode of operation. The major disadvantages of this mode of operation are:

1. There is very little possibility for browsing.

2. A search strategy cannot be developed heuristically. The searcher has essentially one chance to conduct a successful search and must therefore think in advance of all likely approaches to retrieval.

3. The search must be delegated to an information specialist. The user of the information service cannot conduct his own searches. This usually causes problems because users often have difficulty describing what they want or search analysts may misinterpret a user's requirements.

4. There is a time delay.

Figure 1 illustrates the steps occuring in the information retrieval process from the time the user first approaches the system tc the time the system delivers some response. For each step in the cycle, the most important factors affecting the success or failure of the search is indicated.

The on-line search system has none of the disadvantages enumerated above. In an on-line system where the user himself is interrogating the data-base directly, problems of misinterpretation and miscommunication are avoided. Even for delayed searches, the on-line mode has the advantages of rapid response and the capability for interaction, browsing, heuristic searches. However, other problems are likely to come up in this situation. For example, the user who is not an information specialist is less familiar with system vocabulary and with indexing policies and procedures.

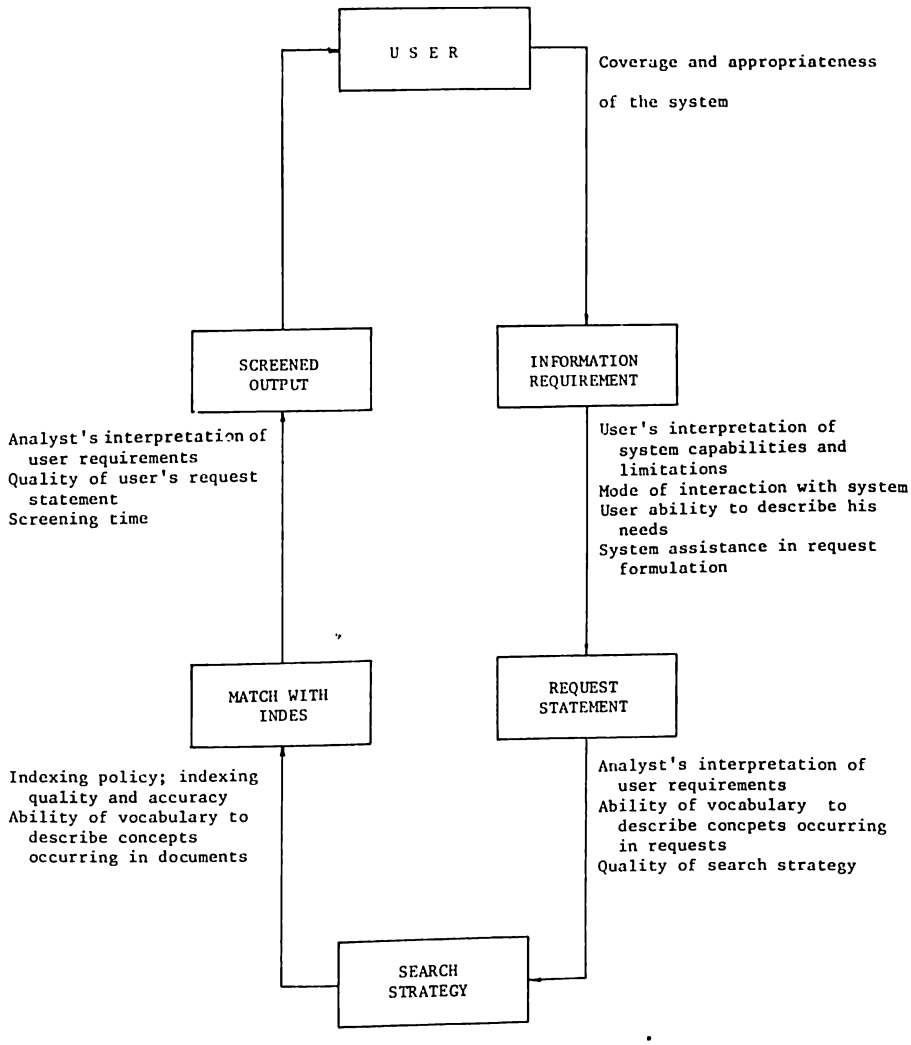

Figure 1 Steps occurring in delegated information retrieval

Equipment for on-line retrieval

In general, on-line information retrieval systems are characterized as time-shared computerized communications networks. These networks are designed to meet the information retrieval requirements of a variety of users who require access to a central database from remote locations. Furthermore, response to the user is required within a few seconds or a minute or two at most, approximating real-time systems.

A typical example of an on-line remote access system is the reservation system used by most airlines. Although these systems are not information retrieval systems, they are very similar from a systems standpoint. Both on-line information retrieval and airline reservation systems provide on-line access to a central database for a number of remote users, both operate in time-shared mode and both provide response in real-time. The major difference between the two lies in the nature of the database and the type of retrievals performed.

Figure 2 is a representation of the basic configuration of a computer communications network required to support on-line information retrieval systems.

Central Processing Facility (CPF).

The CPF is the heart of the system. It supports and controls the operations of the various users. It consists of three major parts, namely:

1. The computer or central processing unit (CPU) performs all the operations required to add information to the database, maintains the information in its files, and performs the operations needed to retrieve information from the database upon receipt of the proper instructions from a local or remote installation.

2. The peripheral equipment of the CPF consists of all auxiliary equipment needed to support the CPU. These include equipment such as high-speed printers, on- and off-line storage devices (e.g., disks, drums), etc.

3. The communications equipment serves to link the CPF with the rest of the network.

Remote Processing Facility (RPF).

The RPF includes all portions of the system that are not part of the CPF. Basically, an RPF must support on-line data entry and reception, and must permit on-line interaction with the CPF. These functions may be performed using a single piece of equipment such as a CRT terminal.

DOMESTIC: *A Minicomputer-Based Information Storage and Retrieval System*

DOMESTIC (Development of Minicomputers in an Environment of Scientific and Technical Information Centers) is a turnkey system for handling the various aspects of information storage and retrieval. The system is the result of a joint German-Israeli research and development project seeking to demonstrate the feasibility of using minicomputers for all aspects of library and information work. It provides facilities for the creation and updating of data bases, formulation of data base searches and output printing. Besides handling information retrieval from textual and other data bases, DOMESTIC is also capable of meeting the acquisitions, cataloguing, circulation and statistical needs of an information center.

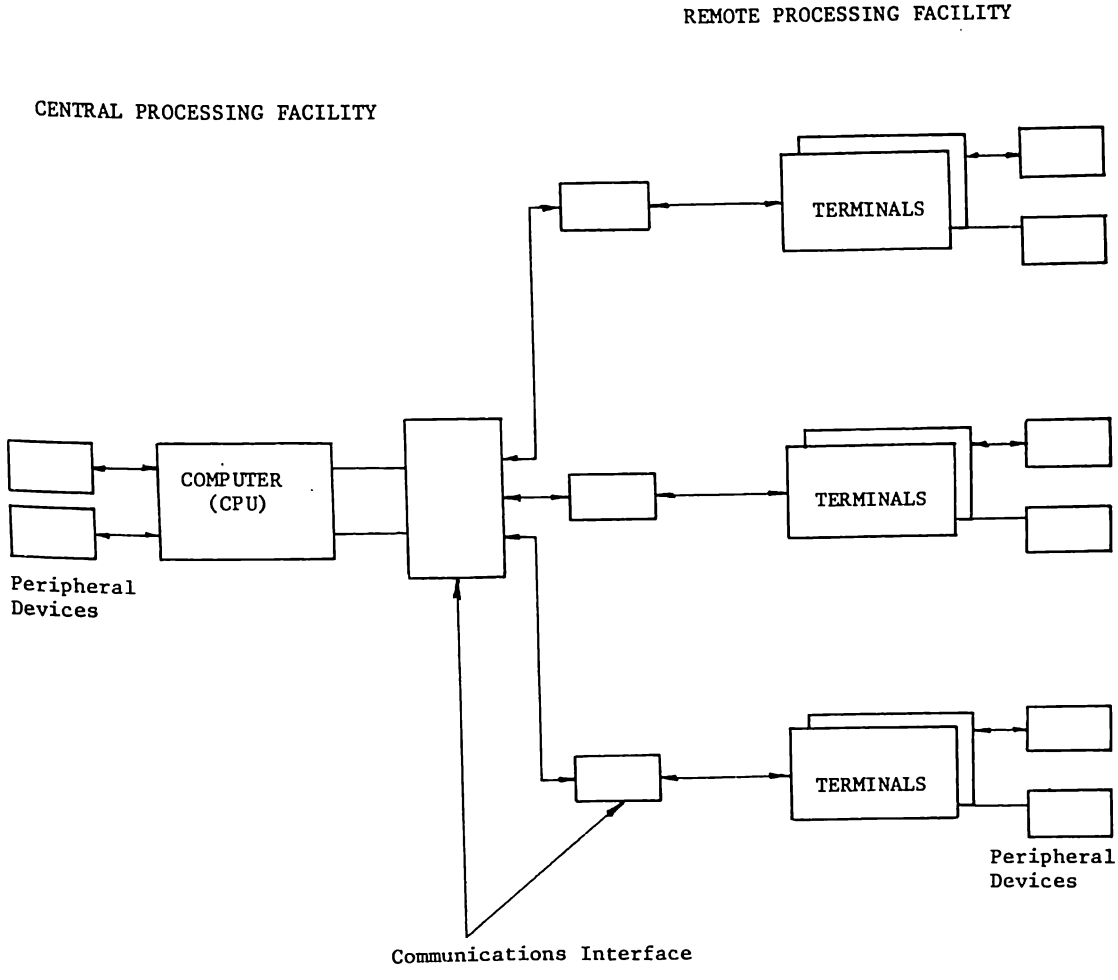

Figure 2: Basic Configuration of a Computer Communications Network.

The first phase, DOMESTIC I, was programmed in 1978-79. It includes all information retrieval modules and the batch creation and on-line corrections of all data bases. The second phase, DOMESTIC II, started in mid-1979, developed on-line data entry and thesaurus maintenance functions, as well as programs for administration, accounting and library automation and a print generator.

Hardware and Software Requirements

A dedicated minicomputer serves all DOMESTIC functions. The system is written in FORTRAN and, at the National Center of Scientific and Technological Information (COSTI) in Tel-Aviv, the system functions on a PDP11/70 minicomputer. With this facility, four interactive terminals can function simultaneously and if memory size is increased, additional terminals may be added to the system. At the KTS Information Systems GmbH in Munich, DOMESTIC is installed on a Phillips 857 computer.

Hardware. The basic configuration needed for the DOMESTIC system includes CPU, at least 64K memory for application programs, a direct-access medium for file storage (capacity depends on the size of the databases), printer communication terminals and, if data exchange is required, a magnetic tape drive.

Operating System. The operating system must be able to support a FORTRAN compiler, a technique for overlaying programs, intertask communications and terminal communications. Its file management should support indexed sequential access method (ISAM), or at least disk-handling facilities, a record length of more than 1000 bytes, and internal record access.

Database Management System (DBMS). In order to efficiently handle files of textual data that are independent of hardware and operating system constraints, a special DBMS was developed for **DOMESTIC**. It incorporates the following features:

1. Regardless of differences in logical content, the physical structure of all files on the disk is the same and they are all accessed by the same set of interfaces.
2. Variable-length records under index sequential organization.
3. Direct access to any data item in the files.
4. Processing files sequentially forward and backward.
5. Separation between the DBMS and the file contents.
6. Avoidance of file handling by application programs.
7. Minimizing file accesses during file manipulation.

The DBMS is written in assembler language so that installation of the

system in another computer would require rewriting the DBMS in the new host computer's assembler language.

Command Language

It was foreseen that not all **DOMESTIC** users would be experienced with computers. The system was, therefore, built to be as user-friendly as possible. This was achieved by developing **DIALOG**, a simple and flexible command language. **DIALOG** commands cover all necessary functions for database searching, displaying search results, file maintenance and output printing. With it, a beginner can already get results using the basic commands while an experienced user can use the parameter options for more sophisticated operations.

Other factors taken into consideration in the design of **DIALOG** include security of database documents and user directories, independence of the command language from the database structure, fast response time and generation of error messages.

Information Retrieval in DOMESTIC

The user begins a DOMESTIC session by viewing a list of databases established in the system and selecting one in which he wishes to work. He then examines the database keyterm list which gives all the words, phrases and codes that can be used to search that database. The user now builds search statements from the keyterms appropriate to the inquiry being done.

The user can then store this query program to be run at intervals to retrieve recent additions to the database. It is also possible to revise the query by adding, changing or deleting search items or narrowing a search to specified fields. Figure 3 depicts the usual sequence of commands in a DOMESTIC search session.

Entering the System. The user begins an interactive session by typing a message such as "**RUN DOMESTIC.**" The user then enters his password and after it is recognized by the system, a list of the available databases is displayed on the screen. Each line of this display contains a number identifying one of the databases followed by the database's abbreviated name.

The next step is to choose a database using the command **BASE**. A database may be denoted by its abbreviated name or by the number it carries in the list on the screen. **BASE** is also used to request a screen display of available databases. The command may be issued at any time in a session, and need not be followed by a change in databases.

Choosing the Search Terms. The expression "search term" describes any meaningful term (a single letter or number, a word or a phrase) which the user assumes to be present in the database and would like to use for retrieval. In bibliographic databases, a search term may be a word in a title or abstract, an author's name or a journal's name. The inclusion of a search term in a search

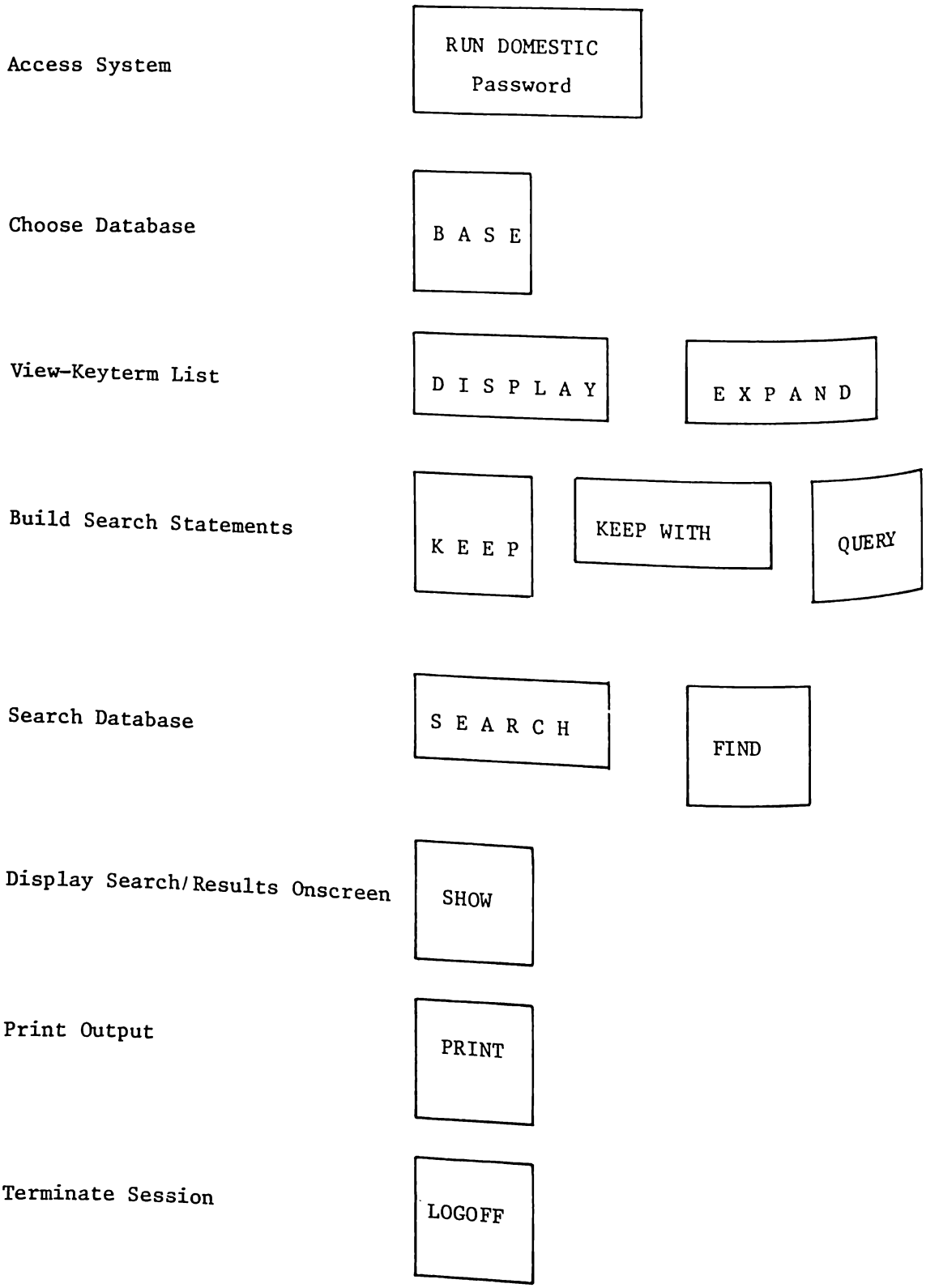

Figure 3A: Domestic I Typical Search Session

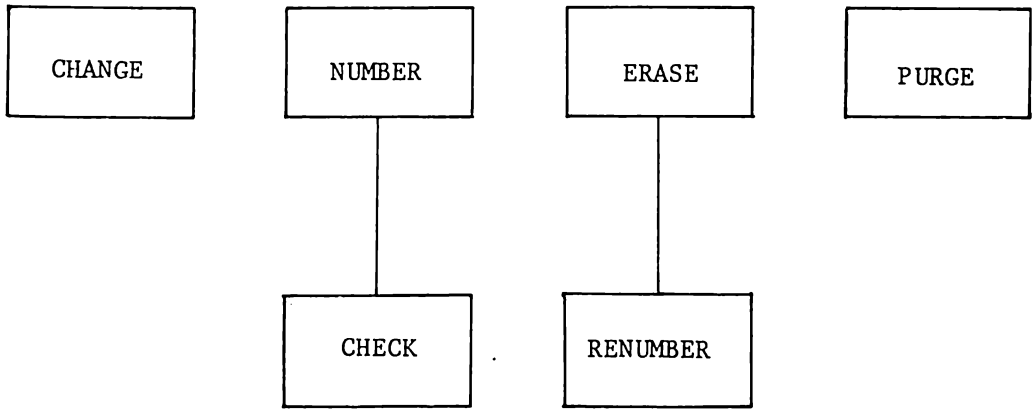

Figure 3B: Revising Search Queries

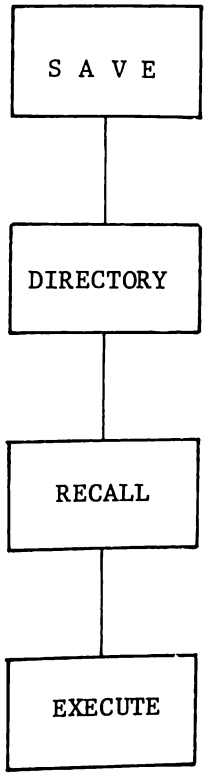

Figure 3C: Saving and Rerunning a Query

query causes the system to seek throughout the database to identify, and if appropriate, to retrieve any document which contains the search item.

Keyterm, on the other hand, refers to those words, phrases, or codes which are found in the searchable fields of a particular database, i.e., those fields defined as searchable at the time of database creation or updatings. Each database has its own keyterm list. The user specifies which part he wishes to use by typing the command **DISPLAY** followed by the selected search item. In response, a serially-numbered list of words and phrases from the keyterm list is displayed on the screen, headed by the input string. Each term is followed by a number which gives the number of documents in the database containing that term in the searchable fields.

The command **EXPAND** accesses the background information list which contains data about the hierarchichal status of a search term in the thesaurus of the database, i.e., what classification codes include this search term or are related to it.

The selection of search terms and their combination with operators is done by the commands **KEEP** and **KEEP WITH**, which produce the search statements.

A search query consists of one or more search statements. Through the command **QUERY**, the query or any subset of its statements can be displayed at any state of the session. The command **SEARCH** displays the posting (the number of documents retrieved by a search statement) for all statements of a query. The command **FIND**, by combining the commands **KEEP** and **SEARCH** for a given string, provides a shortcut in designing and executing a query.

Operators. In building a search query, the search terms are combined by using operators, special symbols defining how the keyterms and search statements are to operate on the database. DOMESTIC includes logical, word-proximity and category limitation operators.

The five logical operators **(OR, AND, NOT, BUTN, and XOR)** determine the Boolean relations between search terms or between statements. The word-proximity operator /ADJ/ may be used in various ways to define the adjacency, separation and order in which search terms operate. Category limitation operators apply a search statement to specific fields of the document. The search may thus be limited to titles only, or to authors, or to abstracts, etc., or any combination of categories.

Displaying and Printing the Search Results. Retrieved documents from the most recently formed search statement are displayed on the terminal by the command SHOW and printed by the command PRINT. It is possible to select an individual document, a range of documents or the whole retrieved set for display or printing. The documents are output in decreasing chronological order, with the most recently input document first and the oldest document last.

Printing or screen display of unwanted documents can be prevented by the DELETE command. To make a printed record of a single screen display or part of it, or to receive a printed list of all the statements in a search query, the command LIST can be used at any stage in the session. The command PROTOCOL causes a record to be made of every screen display during the session, and a copy to be printed out off-line after logoff, until stopped by PROTOCOL OFF.

Reformulation of Search Queries. After seeing samples of the documents retrieved by a search query, the user may wish to reformulate the query by adding new search terms, by erasing search terms or statements, or by correcting search terms or operators.

The CHANGE command is used to rebuild a statement, replacing operators or search terms by new ones. New search terms can be selected from displayed text. First, the command NUMBER indexes the lines. The desired line can then be broken up into separate words using the command CHECK. This command displays each word from a given line number or each search term of a statement, on a separate numbered line.

The ERASE command erases search statements or entire search queries. Gaps left in the numerical sequence by ERASE can be closed by using the RENUMBER command. The command CORRECT takes care of errors detected in the retrieved documents. As in the CHANGE command for revising search statements, the user types the faulty string and the string which is to replace it. An additional command REDISPLAY makes it possible to see the original form of the present screen. The command CANCEL restructures a search strategy by deleting all unexecuted commands.

Winding-Up a Session. When the user has built a search query, it can be preserved for later reference through the command SAVE, which also gives the query a symbolic name. A saved query is entered in the user's library. The list of all the stored queries is called up by the command DIRECTORY. The command RECALL displays all the statements of any stored query.

Stored queries can be reactivated from time to time by the command EXECUTE which runs the query or queries again to provide only those new documents which may have been added to the database since the last search. The printing of these documents is carried out automatically in batch, after the session is terminated. The command PURGE deletes a query from the library.

The user can request information regarding the amount of time spent on the current session, the costs, etc., by using the command STATUS. At the end of the session the command LOGOFF provides a summary of the accounting information and terminates connection with DOMESTIC.

The Database Package Stairs

STAIRS/VS (Storage and Information Retrieval System/Virtual Storage)

is a database package written and developed by IBM. It offers a program package to enable the user to search documents in an interactive dialogue with the system, to display or print documents on various external devices, to create databases automatically and to extend and maintain databases. STAIRS/VS provides a range of search capabilities, as well as flexible and varied output options for retrieved documents. Privacy is provided at many levels to ensure the integrity of both the system and the data to which it has access.

STAIRS/VS consists of two groups of programs, namely:

1. Batch Database Creation and Maintenance Utilities.

2. On-line Retrieval System AQUARIUS (A Query and Retrieval Interactive Utility System).

System Environment

To execute the STAIRS/VS on-line modules, a supervisory control program is required which may be any one of the following:

For STAIRS/VS Release 1 –
IBM Customer Information Control System (CICS/OS)
IBM Information Management System (IMS/360) with Data Communication

For STAIRS/VS Release 2 –
CICS/OS/VS
IMS/VS with Data Communication

STAIRS/VS On-line Modules Under CICS. CICS is an interactive terminal-based system with the capability of limited message switching and administrative message handling. It serves as the interface between the operating system and STAIRS/VS on-line modules. STAIRS/VS makes use of the following CICS services:

1. **Terminal Management.** This involves the automatic polling of terminals and, upon request, writing on and reading from the terminal.

2. **File Management.** The reading and writing of direct access data sets.

3. **Transient Data Management.** The reading and writing of data and messages that are to be queued for later processing.

4. **Task Management.** The multiprogramming control for transaction processing.

5. **Program Management.** Inter-program communication and control.

6. **Storage Management.** The completion of main storage requests or the queuing of requests until storage is available.

CICS supports the IBM 3270 Information Display System and the IBM 2740 and 2741 communication terminals for which device-dependent message formatting is provided by STAIRS/VS.

STAIRS/VS On-line Modules Under IMS. IMS is designed primarily for inquiry and update of databases by transaction processing programs with limited interaction requirements. The STAIRS/VS on-line modules run in a batch message processing region. They use the IMS Data Communication Services (Terminal Management) for communication with the terminals. Device-dependent message formatting is provided by STAIRS/VS for the IBM 3270 Information Display System and for the IBM 2740 Communication Terminal with the station control feature.

The other functions used to control multi-user processing (Task Management, Program Management, Storage Management) and to handle transient data (Transient Data Management) are provided by an interface, an Extended Macro Services module. This module processes all requests from the methods from the STAIRS/VS on-line modules. For Data Management, the access method BSAM, BDAM and ISAM are used.

The On-line Retrieval System Aquarius

AQUARIUS provides the information retrieval facilities for the use of communication terminals.

Privacy. Great care is taken to avoid unauthorized access to databases in the AQUARIUS system. A master terminal function monitors the terminal activity when AQUARIUS is operating. The master terminal operator must sign on before any user can sign on. The master terminal operator is the only operator authorized to correct user registration notes when a privacy violation has occurred.

The AQUARIUS user is allowed two attempts to sign on correctly. If he fails twice, a privacy violation notice is written into his user registration record and he cannot sign on until the master terminal operator has corrected this record.

The databases are privacy-protected at the database, document, paragraph and formatted field levels. The checking performed depends on the function requested. For example, a user cannot have certain paragraphs of a document printed if he has no privacy clearance for the paragraphs requested.

File Organization. Each STAIRS/VS database consists of the following data sets:

1. **Text Data Set** – A BDAM data set that contains the text documents in nearly the same form in which they are printed or displayed.
2. **Text Index Data Set** – A BDAM data set that records of which point to the documents in the Text Data Set. The Text Index contains formatted fields for each document. The formatted fields are used for the SELECT function and for sorting result strings.
3. **Dictionary Data Set** – A BDAM data set that contains every different word in the database. The word entry includes a pointer to the string of all occurrences of that word in the database. The dictionary also contains synonym pointers.
4. **Inverted Data Set** – A BDAM data set that contains strings of all occurrences of words in the database. Each word occurrence is described by its location within the document.
5. **Data Base Control Block (DBCB)** – For each database there is a DBCB record in an ISAM data set. This DBCB is used to locate the data sets of the database. The DBCB also contains the password of the database. The paragraph names and their privacy levels and the key of the Formatted Field Definition Record.
6. **Formatted Field Definition Record** – This describes the formatted fields, if any, established in the Text Index Data Set. This record contains information on the relative position of an individual field, the type of the field (numeric or alphanumeric), and the name of the formatted field. All Formatted Field Definition records are contained in an ISAM data set.

An individual paragraph code (PC) must be assigned to each paragraph. Single paragraphs or classes of paragraphs can be defined in the DBCB by assigning a paragraph name to this part of the document.

When a document is displayed during a terminal session, the user can refer to paragraphs by their names or by paragraph numbers. When the SELECT function is applied, the formatted fields used for index searching are defined by their formatted field names.

The Multi-Database Concepts of Aquarius

The number of databases which can be on-line for information retrieval with AQUARIUS is limited only by the capacity of the operating system installed, i.e., by the number of data sets that can be used for one task. Up to

sixteen databases can be concatenated by the user as if there were only one database.

An initial or master database, once created for a specific set of documents, can be extended later by adding databases without having to recreate the original database(s). Nevertheless, any of the concatenated or subordinate databases can be used individually for document retrieval.

The STAIRS/VS multi-database concept enables the user to extend databases as follows:

Documents can be added to a database by means of database creation on-line or, for mass data, in batch mode. This database can then be concatenated to the master database.

A Database Merge Program allows the merging of up to four databases into one, at the same time deleting documents which have been flagged as obsolete. Thus, a set of concatenated databases can be reorganized periodically.

Basic Specifications

The first three functions provided by AQUARIUS are retrieval functions that use the following methods:

1. Boolean logic search methods (SEARCH)
2. Index search methods (SELECT)
3. Weighted search methods (RANK)

The remaining functions – BROWSE, EXEC, HELP, SAVE AND SORT provide additional services to the user.

These functions constitute the main capabilities for the dialogue between the terminal user and the system. During a terminal session, the user controls and performs the dialogue by prompting the functions selectively according to the progress of document retrieval.

A user can invoke the functions by entering the appropriate command at the beginning and throughout the terminal session. The user selects the mode of retrieval to be used by entering one of the following commands:

. .BROWSE
. .RANK
. .SEARCH
. .SELECT
. .EXEC

By searching the whole database using one of the various search methods (SEARCH, SELECT or RANK), the system produces a listing of document numbers or document list (DL) which can be subsequently extended, reduced, resequenced or scanned by the AQUARIUS functions SEARCH, SELECT, SORT, RANK and BROWSE.

The BROWSE function does not produce a document list but causes the documents of a document list to be displayed at the terminal. The EXEC function allows reexecution of a previously saved set of SEARCH, RANK and/or SELECT queries. Saved rank are reexecuted as SEARCH queries.

The dialogue consists of the iterative use of the main functions. Each query may be reformulated, improved, broadened or narrowed during a terminal session, depending on the response of the system. Below is a summary of the different AQUARIUS functions.

FUNCTION	REFERENCE	RESULT
BROWSE	Database/DL	–
RANK	Database	DL in ranked order
SEARCH	Database/DL	DL/reduced DL
SELECT	Database/DL	DL/reduced DL
SORT	DL	Resequenced DL
HELP	Message file	Help messages
EXEC	Stored query name	Reinitiated SEARCH

and/or SELECT queries