A Client/Server System for Interactive Access to a National Database Over the Internet

M.G. Sreekumar

Indian Institute of Management,
Kozhikode,
Regional Engineering College,
Calicut — 673 601

Abstract

This paper describes the development of a Client/Server prototype for the design, creation and interactive interfacing for a national database on the Internet World Wide Web (WWW). Interactivity on the Web is achieved through an HTML form that collects information (query criteria) from users and via a link to a Common Gateway Interface (CGI) script on a server, processes the information and returns the result. This dynamic website facilitates medical practitioners and biomedical researchers throughout the world, to build queries online and helps in getting a near comprehensive account of the topic of search. System design considerations include: systems analysis & planning, data capturing, subject indexing, selection of database management system (DBMS) and database creation; development of a user-friendly search interface; creation of the Web Site; and connecting the database with the Web Server. Users are provided with a copy of the national database's home page by furnishing the Uniform Resource Locator (URL) of the Web Site. The search interface is a dynamic link and the database is searchable online. The bottom line is that users are able to access the database for search and retrieval from any part of the world at the cost of a local telephone call.

Keywords: Bibliographic Databases, Electronic Databases,National Databases, Internet, World Wide Web, WWW, Hypertext Mark-up Language, HTML, Client server design, Tuberculosis, India

Introduction

Internet, often referred to as THE NET, is the name for a vast, worldwide system consisting of countless numbers of people, information sources and computers joined in a symbiotic confluence. It is actually a collection of tens of thousands of computer networks spanning the globe. Put simply, the Internet allows millions of people, all over the world, to communicate and to share. It is the first global forum and the first global library. Electronic mail (E-Mail), List Servers, USENET/Newsgroups, FTP, Telnet, Gopher, WWW etc. are just a few of the several services offered by the Internet. Among these, the World Wide Web or the WWW is relatively a new service and it is a graphical user interface (GUI) of the Internet. Web browsers are a relatively new class of client/server[1] applications. They are used on the front end - the user interface - to a web page. Client/server is a collection of service provided by servers. These services are provided on request to clients. The Web is a large system of servers which offers all kinds of information to anyone on the Net. To access information on the Web, a client programme called a browser is to be used. Servers are programmes that provide resources and clients are programmes that are used to access the resources. Internet has a lot to offer to the biomedical community. Internet access to biomedical databases is a reality. An example is the "Internet Grateful Med", the U.S. National Library of Medicine's MEDLINE accessible free over the Net with "http://igm.nlm.nih.gov" as its Uniform Resource Locator (URL). This paper describes the prototype developed for Internet access to India's National Database on Tuberculosis and Chest Diseases (URL-"http://tbindia.info.nih.gov"). The fascination in Internet databases is that it avoids the involvement of exorbitant telecommunication costs otherwise required in online searches using long distance telephone lines.

Background

The results of a pilot study conducted by the author at the Tuberculosis Research Centre (ICMR), Chennai, India revealed that the coverage of Indian biomedical literature in international indexing systems such as MEDLINE, EMBASE etc. is very less. This prompted to

submit a proposal to the Indian MEDLARS Centre of the National Informatics Centre (NIC), New Delhi, India towards creating a national database on tuberculosis and chest diseases comprising Indian biomedical literature, for the benefit of physicians and biomedical research community. The proposal has been approved and the database project is in progress in India. The author received six-month advanced training in the U.S. National Library of Medicine under the U.S. Fulbright visiting research program towards creating the database in the model of MEDLINE with subject headings prepared according to MeSH (Medical Subject Headings), the controlled vocabulary designed and developed by the Library, and also in the planning, development and publishing of the database on the Internet.

Rationale

Current trends in the handling of scientific information shows evidences of a major shift from the analog and physical form (voice, print & audio visual) to the electronic form. This is characterized by the fact that the moment information is transferred into electronic form, it becomes mobile and dynamic, rather than static. It can then be processed, transmitted, stored, retrieved and disseminated at enormous speed and convenience. By the advent of the modern communication technology, this information can be transmitted across space.

The role of bibliographic databases in front line biomedical research needs no special emphasis. A researcher first turns to the most popular database in the subject for state-of-the-art before commencing his work. Similarly a physician needs quick, accurate and articulate information about a problem/case at hand. Unfortunately in most cases celebrated international indexing systems such as MEDLINE and EMBASE or their off springs do not offer adequate support to the researchers and physicians of Asian and African world. The fact is that the coverage of biomedical journals emanating from Asia, Africa and other developing nations is lamentable. It is also not feasible for any one indexing system to cover almost all the journals in a discipline. It has therefore become inevitable for these nations to have indigenous bibliographic database/s to meet the information needs of the biomedical community. It is also an effective measure to have bibliometric control over the discipline. Moreover the Internet facilitates remote access to databases without the involvement of online telecommunication lines - with the help of a direct link to the Internet Service Provider, the user is able to access international databases on the Internet with no additional cost on telecommunication charges. This has special relevance to developing nations as they need not indefinitely wait any more for inclusion of their journals in international indexing systems; and also in recognition of the low cost, but appropriate, effective, efficient and reliable technology. Several such examples of Internet accessible databases are available such as the Internet Grateful Med (IGM), (the Internet version of MEDLINE, U.S. National Library of Medicine); the GenBank of the U.S. National Centre for Biotechnology Information-NCBI and others. Hence the lamentable situation of the developing countries could be improvised with the possible features available in the Internet environment. One such attempt is the project proposed and completed.

Objectives

Creation of India's national database on tuberculosis and chest diseases with the following targets:

(i) The database shall be in the model of MEDLINE, the world's most appreciated biomedical bibliographic database;
(ii) The subject indexing shall be done according to MeSH (Medical Subject Headings), the controlled vocabulary thesaurus;
(iii) To develop a WWW platform for remote access for search and retrieval via the Internet; and
(iv) To publish the database on the Internet.

Electronic Databases

Electronic databases (EDB) are electronic equivalents of their print counterparts. For the most part, EDBs are produced before the production of the print version. They can be categorized into two types - reference and source databases. Reference databases refer or direct a user to another source, often a document, for more details. Reference databases can be further sub-categorized into bibliographic (containing primarily citations from published information like journal articles, reports, patents, dissertations, conference proceedings and books) and directory (for e.g. listing of companies, associations or people). Source databases contain complete data or the full text of the original source information. MEDLINE, CA-SEARCH, INSPEC, COMPENDEX, MATSCI etc. are examples of EDBs.

EDBs have the unique quality of condensing the time element in the creation, storage, retrieval and dissemination process. There are two major database

services -Current Awareness Service (CAS) and Retrospective Search Service (RSS). The efficiency of such systems and services depend largely on its user friendliness (human computer interaction). This could be achieved by appropriate design of the search interface, response time, retrieval capability, display and its features, other related facilities for printing and downloading etc.

Material & Methods

Planning is essential in the design and development of a database. An in-depth systems analysis of the aims, goals, scope, coverage and the target audience of the database is a prerequisite. Selection of the most appropriate Database Management System (DBMS) and identification of excellent software and hardware systems shall determine the quality and success of the database. The systems selected should stand to the test of time and technology for several years to come and also provision should be made towards its scalability. A complete feasibility study needs to be done towards uninterrupted supply of data and a strategy oriented data capturing. The captured data requires organization and processing before it is fed to the computer. One major factor here is the "subject indexing" process which will be detailed later. Before creating the database, it is significant to determine the kinds of information that will be tracked and whether that information should be put in one or multiple databases.

A database is an organized collection of textual, numeric, date and image information pertaining to related data. It could be employee records, sales figures or a books catalogue in a library. The database under study in this project consists of information in Journal Articles, Conference Proceedings, Books, Research Reports, Government Reports etc. The database management system proposed by the sponsors of this project, viz., the Indian MEDLARS Centre of the National Informatics Centre (NIC), is "CDS/ISIS". CDS/ISIS stands for Computerized Documentation Service/Integrated Set of Information Systems. This is a non-numeric database management system produced and distributed by United Nations Educational Scientific and Cultural Organisation (UNESCO). The project is in progress in the Tuberculosis Research Centre (Indian Council of Medical Research), Chennai, India.

To create a database, it is required to define its structure, which is a list of fields (significant data elements when put together, constitutes each entity in the set considered for the database). The national tuberculosis database structure is defined as per the Common Communication Format (CCF), proposed by the UNESCO, with slight variations wherever inevitable. After the creation of the structure, the next step is to populate the database with Records. A "Record" is the basic unit of retrievable information in the database. When the database is searched, records are retrieved. Each record is composed of information in one or more of the fields defined in the structure. Most of the ground work for creation of this database were done in the Tuberculosis Research Centre, Chennai, India and a sample set of 500 records were brought to the National Library of Medicine for further studies.

Indexing or subject heading operation[2] (the process of deriving subject headings or descriptors to the document or article in hand) has become imperative, as in most cases, the title of the document/article may not explicitly bring out the thought content and also practically it is impossible to mention all these in the title. The rationale behind the process is therefore to give maximum retrieval capabilities so as to get its most use. Hence, MEDICAL SUBJECT HEADINGS (MeSH), the U.S. National Library of Medicine's controlled vocabulary thesaurus, was used in this project also.

It is intended to publish the database on the Internet with a suitable World Wide Web front-end which shall facilitate search and retrieval of the database from a remote location. It is therefore essential to change the database platform to a convenient DBMS which has capability and compatibility to meet the above mentioned goal. Also the DBMS selected should be viable and feasible to the existing infra-structural facilities at the Tuberculosis Research Centre, Chennai and the National Informatics Centre, New Delhi in particular and the general support facilities in computer communication and networking prevailing in India in general. In view of this, the following were evaluated: 1. Microsoft ACCESS for Windows 95; 2. Borland Paradox; 3. Filemaker Pro; 4. INMAGIC DB/Text Webserver and 5. Cold Fusion. To this end, the above five systems were evaluated and the most appropriate system was recommended for the database. With a view to get a better understanding and general features of these products, the Web Sites of these systems were first visited and downloaded relevant information for ready reference.

Literature survey was conducted on MEDLINE, HEALTHSTAR and LIBRARY LITERATURE databases for confirmation of any such studies conducted

elsewhere on the topic. The search revealed that similar studies were conducted for evaluating the efficiency of some of the existing software packages. Interestingly, none of these studies reported of any of these packages. Search was then extended manually to the popular computer journals available in the NLM Staff Library. The 1996 issues of *"PC World"*, *"PC Magazine"*[3], *"PC Computing"*, *"Windows Magazine"* and *"Database"* were scanned for literature on the topic. The study was later focused on evaluating them individually based on their trial versions/evaluation copies. A set of parameters were identified and tested them with each of these database management systems. Towards the objectives of the database project, INMAGIC'S DB/TEXT WEBSERVER[4] meets most of the requirements of the set parameters and hence it is chosen as the DBMS for the national database on tuberculosis and chest diseases. The other major reasons for its choice include: the built-in HTML support, TCP/IP compatibility and Network file serving capability. It is robust, well supported and it has immense features in connecting the database to the HTML document. The hardware considerations and the communication requirements are the same for any normal WebSite. However, the operating system recommended by Inmagic is Windows NT Server.

In the process of building a search interface, a set of sample records were exported to a delimited ASCII (American Standard Code for Information Interchange) text file from the original CDS/ISIS database brought from India. This text file was then imported to an INMAGIC database after carefully defining the structure, field properties, indexing instructions etc. A QBE (Query By Example) screen was designed using the INMAGIC QBE Screen Designer. The fields such as Author, Title, MeSH Terms, Non-MeSH Terms, Journal Title/Conference Information and Year were added to the QBE screen. While adding each QBE box, specifications were given as to which field or combination of fields, it needs searching. Provisions were made so as to enable users perform search on free-text as well as MeSH and non-MeSH terms (a set of subject headings prepared for the purpose of this database similar to that of MeSH Structure). All the QBE boxes were also attached with Boolean Operator Buttons (AND, OR and NOT) for combination of search criteria. The Boolean "OR" shall retrieve records that meet any of the criteria, "AND" shall retrieve only those records that meet all of the criteria and "NOT" shall exclude all those referred records from the search results.

Two separate display forms, one for immediate bibliographic information and another one for full record details, were created as output forms. The QBE screen and display forms were then tested several times for any discrepancy in culling out information from the database and also in the proper display of the searched output as defined by the forms. The interface was again subjected to testing by experienced information professionals in the Library for comments and compliance.

Web Site Design

World Wide Web (WWW)[5] is the most utilized tool of the Internet. HyperText Markup Language (HTML)[6], is a mark-up, or formatting, language. Text files with HTML tags can be marked up so that they can be read over a network or locally on computer. Documents available on the Web are HTML files.

The HTML is designed as a standard way to format documents so they can be viewed on several different platforms. HTML documents are written as plain text files with codes that determine the appearance of the document when viewed through a HTML browser such as "Netscape" or the "Internet Explorer". Since HTML documents are text files, they can be created by such text editors as Windows Notepad, Simple-text for Macintosh and pico, vi or emacs for Unix. The formatting of HTML documents is accomplished through the use of tags. Tags are pieces of code surrounded by the symbols <and>. Most HTML elements come in set of two - the start-tag, and the end-tag. For example, if we use the <B></B> for bolded text, then anything placed between <B> and </B> will appear as bold text when viewed through a HTML browser. Hypertext Transfer Protocol (HTTP) is the Internet protocol use for distributing hypertext documents on the Net. HTTP is layered on top of Transfer Control Protocol / Internet Protocol (TCP/IP), which is the basic communication protocol and the foundation of the Net. All other protocols, such as FTP, Gopher etc., are also layered on top of TCP/IP.

A Web Site on the Internet is identified by its Uniform (or Universal) Resource Locator (URL) address. It is a naming, or addressing, convention used to locate a site on the World Wide Web. Example : http://www.nlm.nih.gov (URL of the National Library's home page). Among the panoply of scientific communication tools, Web publishing is the latest. Web publishing provides unique access to scholarly information because of its serendipitous nature, immediacy, its hypertext structure and its universal

appeal and accessibility. A typical Web page will feature attractive art, dynamic links to other Sites, scrolling text, tables, forms, simple animation, and possibly a few Common Gateway Interface (CGI), the specification for how an HTTP browser should communicate with server gateway programs and Java scripts.

In the Web environment, the "queries" process is online and it is a two-way process between the HTML form and back-end Web Server (Internet Server). It requires use of Gateway Scripts. Interactivity on the Web is still mostly based on forms that collect information from users and then via a link to a CGI script on a server, process the information and then return the result.

Results

A home page for the national database's Web Site has been created using HTML codes. The Web page is provided with a form containing boxes for Author, Subject, Source and Year of publication of the source document. It is the same interface developed earlier for the database (INMAGIC DBMS). With the help of CGI scripts, the query criteria fed to the boxes perform searches on the database and bring results to the Web front-end. This form acts as the search interface for the end-user. The interface is made user-friendly with the help of directions and online help facility. A post box is also attached for receiving comments and suggestions from users for further refinement and modifications.

The home page is also provided with two hot links, the general as well as the technical aspects of Tuberculosis and another, a documentation on the project. Connecting links have been provided wherever essential.

The national database Web Site has been transferred to the Internet server of the Division of Computer Research and Technology (DCRT) of the National Institutes of Health (NIH), Maryland, USA. The URL of the site is "http://tbindia.info.nih.gov".

Discussion

In summary, the web site facilitates online access to the national database over the Internet. As already discussed, the latest Internet technology and its Web hypertext protocol (the Hypertext Transfer Protocol - HTTP) is a low cost and appropriate technology most suitable for developing nations for online search of bibliographic databases. The protocol is very versatile and dependable. The client can be any popular web browser such as Mosaic, Netscape or the Internet Explorer. Users are provided with a copy of the National Database's home page by furnishing the URL in the browser (client). The search interface is a dynamic link and they can interact with the database online. Various approaches using different search elements and Boolean operators could be made towards the sharpening of the search. The online help provided beneath the search interface answers most of the real time doubts/online queries. For instance, a search against "EPIDEMIOLOGY" "AND" "TUBERCULOSIS" returns citations of all such articles on the subject published from India. The physician or the researcher is thus able to get a near comprehensive account of the topic under search. This goes a long way in helping the user community with the quick provision of accurate and articulate information for case diagnosis and decision making or for the swift furtherance of their research. The title in the result screen is again a hot link which, upon clicking, brings the abstract of the article. This is purposely done to save the time of the user, as they need not browse through the entire abstracts unnecessarily. As a by-product, the home page could be used for publicising messages, creating institution(s)'s own Web pages etc. For example, the national database home page has two hot links 1. "About Tuberculosis" - an educative-cum-publicity material and 2. "About the Project" - a project documentation. Double clicking on them will take the user to the respective page and give details about tuberculosis as well as the institution, the Tuberculosis Research Centre, Chennai and the Indian Council of Medical Research (ICMR) which is the parent organisation.

Acknowledgment

The author thanks the United States Educational Foundation in India and the U.S. Fulbright Programme for the Fulbright Fellowship awarded to him. The encouragement and support rendered by the Director, Tuberculosis Research Centre, Chennai, the Director General, Indian Council of Medical Research, the National Informatics Centre and the U.S. National Library of Medicine is acknowledged. Sincere thanks to Dr. A. Amudhavalli, Reader, Department of Library and Information Science, University of Madras for her valuable suggestions throughout. The author expresses his gratitude to Dr. Vinyshil Gautam, Director, IIM Kozhikode for kindly permitting him to present this paper.

References

1. Sid Wise. *Client/Server Performance Tuning* Designing for speed. USA: McGraw-Hill, 1997.
2. The U.S. National Library of Medicine. *Medical Subject Headings (MeSH)* 1996.
3. Rand R. Readers rate:report of the fifth annual support and satisfaction survey of PC softwares. *PC Magazine* 1996;Jul:227-45.
4. Lance N. Ulanoff, Tin Albano, Yuonne Koulouthros. From start to finish:building a web site. *PC Magazine* 1996;Sep.10:102-256.
5. Dean Scharf. *HTML visual quick reference*. 2nd ed. USA:Que,1996.
6. Ellingen D. Inmagic DB/Textworks. *Database*. 1996;Apr./May :46-51.