AN ANALYSIS ON VARIOUS TECHNIQUES FOR DEVELOPING A DIGITAL LIBRARY ARCHITECTURE

An Analysis on Various Techniques for Developing a Digital Library Architecture

Balijepalli Bhagyalakshmi

Librarian, Apswreis, Veldurthy, Govt. of AP

Abstract – Information technology has revolutionized all walks of life and like any other profession, libraries are also at the threshold of this technology era. Libraries harness these technologies for the purpose of effective and efficient operations and services to serve their users in a better way. In this regard, Indian academic libraries are witnessing tremendous transition from manual system of access to automated system of access to information. As a result, the automation in Indian academic libraries has become the top most priority. Nevertheless, the selection of appropriate software package for effectively automating library operations and services has been a challenging task for librarians in India. Even though, a number of library software packages were developed in this direction by research and development libraries in India, however, only few of that software survived, while the rest did not last long due to their shortcomings in compatibility with the international standards.

Against this background, a number of Indian libraries switched over to foreign software packages to rise above these shortcomings. Notably, the research and development libraries and large academic libraries in India are using standard library management software to enhance their operations and services.

In this backdrop, the researcher has described the development of library automation in detail from its historic time to modern period. The landmark developments have been discussed both at Indian as well as at global level.

Initially, library automation received cold response in Indian academic libraries. However, with the falling price of hardware's, easy availability of software and ever-increasing interest among the library professionals, automation in academic libraries picked up the momentum. The present research effort is an attempt to examine the current performance of library software and automated library system of the central libraries of two institutes of national importance and one of the first ISO Certified University libraries in India.

-------------------------◆----------------------------

INTRODUCTION

Digital libraries are large, organized collections of information objects. Well-designed digital library software has the potential to enable non-specialist people to conceive, assemble, build, and disseminate new information collections. This has great social impact because it democratizes the dissemination of information. In particular, it will revolutionize the way in which education is conducted and educational materials are prepared.

The emergence of the World Wide Web is changing society's view of information by making unprecedented volumes of information freely available. Of course, it is an unreliable source of enlightenment, and undiscriminating use is dangerous – and, unfortunately, widespread. Nevertheless, the web abounds with accessible, high-quality information. Many educational establishments, international organizations, social groups, non-profit societies and charities make it their business to create sites on which they collect and organize information.

Viewed as an educational resource, however, the web exhibits serious deficiencies: uneven and erratic coverage, transience and unpredictability (will this piece of information still be there tomorrow?), and manifest dangers (will my students encounter inappropriate information?). But a far greater tragedy is that whole segments of society become disenfranchised – for while most family homes in rich countries have some degree of access to the Internet, only a tiny minority of citizens in the developing world can tap this wealth of information. Digital libraries address these problems by providing reliable sources of appropriate material. They empower educators to create collections specifically for their students, collections that mix information from different sources. They permit alternative means of distribution (e.g. CD-ROM/DVD, a very practical format in developing countries).

This course is about the use of digital libraries in education, including emerging areas of application and current and future technologies for creating and distributing digital libraries. It shows educators how to

build their own digital library collections for use in the courses they teach. It touches on large-scale national and international digital libraries for education, but is more strongly oriented towards low-budget methods of building and maintaining digital libraries by creative individuals and by self-organized communities of educators, at levels ranging from the personal to the individual institution.

In 2007-2008, for the first time in history, more than one-half of the Library's total collection budgets were spent on digital materials, a trend that is expected to continue. During the same year the Library undertook a New Directions process that culminated in several initiatives being chosen for development. One was to "Create a Digital Library Program that addresses Library roles and responsibilities for the full lifecycle requirements of digital assets managed or licensed by the Library.

The Digital Library System design allows individual organizations to include their own material in the Digital Library System or to take advantage of network based information and services offered by others. It includes data that may be internal to a given organization and that which crosses organizational boundaries. This document presents a plan to develop such a system on an experimental basis with the cooperation of the research community. Finally, it addresses the application of a Digital Library System to meet a wide variety of user needs.

The productivity gains from having access to a Digital Library System are easily as large as those derived from internal combustion engines and electric motors in the early part of this century. Just as a car on an interstate highway is vastly more effective than one on a rutted dirt road, computer-based information "vehicles" can be made dramatically more effective given the proper operating environment. Computer and communications technology has made it possible for old fashioned, slow retrieval methods to be replaced by virtually instantaneous electronic retrieval. Each user of this technology can anticipate enormous potential benefit, but we lack the natural infrastructure to support this capability on a widespread basis today. This absence of infrastructure represents both a barrier and an opportunity of dramatic proportions.

The fundamental technological issues pertaining to digital libraries are the acquisition of content1, the search of a large collection of data, and the representation of the returned results. To support efficient search, a digital library architecture must face the following challenges: data storage, content categorization and indexing, and result delivery in digital format. To be practically useful, a digital library must provide the user with a simple yet powerful interface, present the results of a search in a meaningful format, and allow the user to reformulate and iteratively refine queries.

Fundamental work in the general area of digital libraries was conducted under the sponsorship of the Digital Libraries Initiative (DLI), and numerous projects and initiatives have since contributed to advancements of the field. Today, digital libraries organize heterogeneous and diverse media types, ranging from video, to photographic images, textual documents, audio, medical data, geographically referenced data, and satellite images.

Given their ability to effectively manage this diversity of data types, we believe that digital libraries are ideally suited to manage scientific datasets. Satellite images of the earth, astronomical and planetary images, medical data, and geological measurements used in the petroleum industry, for example, all share the following characteristics: the volume of acquired data is extremely large, the density of information is high, the analysis (and retrieval) is usually performed using low-level descriptors that can be captured by a computer (rather than high-level semantics that are very hard to characterize algorithmically), the acquisition process is expensive, the analysis (by a human expert) is complex, and the data itself is valuable.

Consider, for instance, satellite images of the earth surface, which are being produced at an exponentially growing rate. The acquisition of this type of data is expensive, since it involves launching an earth-observing satellite into orbit, deploying a number of ground stations, and warehousing the tens of gigabytes produced daily by the orbiting instruments. The large datasets produced are difficult to manage: each individual image is too large to be easily transmitted over the existing Internet infrastructure, and finding images for particular purposes out of many thousands of candidates can be daunting.

Yet, environmental protection agencies, developers, real estate agents, the Forest service, farmers, paper companies, and educational institutions, just to mention a few, can significantly benefit from using remotely sensed data. What is needed are mechanisms allowing the user to effectively identify, among the terabytes of available data, the "interesting" portions of the few images relevant to the task at hand. Thus, digital libraries are the natural candidate tool to manage scientific and technical data.

BACKGROUND

The term library is understood to have been reaped from the Latin word 'liber' connoting book(s) placed in an organized manner for the purpose of use. The term traces its lineage back to preservation of written records to preserve human communication and is considered to have begun in the historic age. The library in historical terms is assumed to be collection of graphic materials arranged for easy use, taken care of by an individual(s), familiar with that arrangement and available for use by at least some

persons. However, with the trends of time libraries turned to an organized collection of general information

and literature for the purpose of their utilization. Libraries were assumed to be responsible for harboring the functions of collection, organization and circulation of their meager sources and services, but the phenomenal progress in technology and exponential growth in information made it obligatory for librarians to adopt new techniques and technology in order to cope up with the emerging circumstances.

Like any other field, libraries were also at the threshold of information communication technology. The developments of which made the role of otherwise unused libraries, multi focal and multipurpose. Since centuries, Libraries have been most vibrant agents in transforming knowledge usually organized in print form; however, the application of machines to process and transfer this knowledge in a better way represents modernization of libraries.

Modernization, nonetheless, with itself brings the new challenges that creates issues and concerns and lay added stress on any questionable area of prevailing structures. Although, there is high degree of cooperation between the prevailing and current technology primarily dominated with the computers, but the later one have an advantage over the old as it is equipped with qualities that better accommodate current circumstances.

The introduction of computers in various sections of libraries such as acquisition; cataloguing; circulation and periodicals offered libraries a new look. The primary interest of automating such operations does not designate elimination of man in favor of machines, but the choice of both in the tasks best suited to each. Computer technology offered hardware and software and conceptual foundations to build automated library system, where optimum use of man and machines is exercised to enhance the library operations and services (Heiliger & Henderson, 1971). Thus technology, hardware and software only provide a sketch that frames automation, but does not imply elimination of manual effort in libraries for disseminating knowledge.

Automating rather, is primarily aimed at effective and efficient utilization of resources in minimum duration with least cost and effort, yielding maximum output and benefits both for library and their patrons. Since, technology has been producing wonders and will perhaps continue to renew and redefine every discipline and activity. Similarly, the unending development of computer technology brought library world to a new horizon where user friendly and economically flexible software packages are readily available; thereby, making librarians judicious to compare the strengths and shortfalls of different packages and choose the package for installation that is most appropriate to suit the requirement of both librarians and their end users.

DIGITAL LIBRARY ARCHITECTURE

Before describing specific features of the Digital Library System, it will be helpful to review some of the fundamental assumptions which strongly affect its design. Perhaps the most dominant of these assumptions are that the system is distributed, heterarchical, hierarchical, networked and strongly display-oriented. In addition, it must have an ability to interact with other autonomous Digital Library Systems that do not adhere to its internal standards and procedures.

The rationale behind the first assumption (distribution), in part, is that existing digital information sources are not physically collocated and that, as a practical matter, the Digital Library System design has to accommodate many geographically distributed components. The distributed system design does not rule out the centralization or at least concentration of resources where this meets pragmatic needs for minimizing operating costs, aggregating communications facilities, and so on. The important point is that the design forces neither centralization nor pure decentralization but accommodates both styles.

We assume that users will access the services of the Digital Library from powerful, geographically distributed and often locally networked workstations. This assumption places networking at the center of the distributed architecture. Even if all the data content of the Digital Library were centralized, its users cannot be.

Distinctions between entirely different (autonomous) library systems leads to at least one level of hierarchical structure in the architecture. Components which can interact among themselves using an internal set of conventions are distinguishable from the set of components which use an external conventions. Distributed, decentralized but hierarchically structured computer services seem to be a natural consequence of the organization of the present and foreseeable marketplace for the use of systems like the Digital Library. Computer services which cross the jurisdictional boundaries between organizations, or even between divisions or departments of one organization, require management structures for access control and accounting. Services which span multiple organizations typically exhibit two or more levels of hierarchical structure stemming from the necessity to

draw boundaries around component operating and management responsibilities.

Another rationale behind the hierarchical structure of the system is to constrain the scope of the data management problems so that system growth does not lead to exponential amounts of database updating and consistency checking activity. Similar motivations often impose structure on otherwise unstructured telecommunication networks, for example. The importance of scaling in all dimensions cannot be over-emphasized. The architecture must scale in sizes and numbers of databases, numbers of users, numbers of components, bandwidth of underlying data communication, varieties of archived content and variation in presentation media and access methods.

AN APPLICATION SITUATION

Digital libraries for scientific applications must support access to multiple data types, sophisticated mechanisms for specifying query semantics, and the ability to manage and process large amounts of information. We have addressed these issues in a recently constructed application for content-based retrieval from archives of petroleum well-bore imagery. We will describe this application and use of the retrieval tool as an introduction to the capabilities of our system.

An important component of oil exploration and oilfield management is obtaining information on the geology of the subsurface strata using data collected from already-drilled wells. Data is collected by lowering a package of instruments ("logging tools") to the bottom of the well, and slowly pulling them back to the surface. While the package is being pulled to the surface, the instruments are measuring various physical properties of the rocks surrounding the borehole, including electrical resistivity, sonic velocity, and natural and induced radioactivity.

Most measurements are "single channel"; typically every 6 inches a single measurement is made of the surrounding rock by a given instrument. Other instruments are more sophisticated. For example, the Formation Micro-scanner Imager (FMI) has four arms, each with two 'pads' which press firmly against the surrounding borehole walls. The pads on the FMI have a very high density of electrodes which can detect subtle variations in electrical resistivity. With this tool, instead of one measurement being made every 6 inches, 192 measurements around the circumference of the bore are made every 0.1 inches. The result of measuring 1000 feet of a borehole with this instrument is therefore an image 192 pixels wide, and 120 000 pixels high.
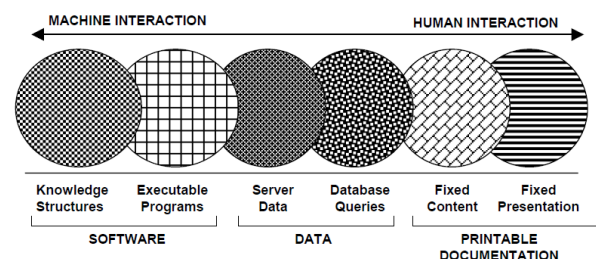
Geologists studying well data are interested in identifying strata of particular types and/or particular characteristics. For example, it might be interesting to find all the coarsely bedded sandstone, or all the finely laminated shale intervals. Bulk lithologies (i.e., rock types) can generally be distinguished using the single channel measurements. For example, sandstone is characterized by low gamma ray values, while shale is characterized by high gamma ray values. Electrical resistivity can be used to distinguish sandstone whose pores are filled with oil versus water. Once these gross lithologies have been identified, the higher resolution FMI images can be used to identify fine-scaled features within the rock. These finescale features can give the geologist clues as to the environment in which the strata were originally deposited: e.g., river, beach, desert.

SPECTRUM OF THE DIGITAL LIBRARY SYSTEM

A large amount of information is already available in computer-based form but is not easily accessible; therefore, relatively little use is made of it. Unless one already knows how to access such information, it may not be obvious even how to get started. Exploring databases for new information is at best a highly speculative process that is often expensive and unproductive. To the providers of database services, and the suppliers of user equipment, this situation translates directly into unrealized potential. Moreover, the vast majority of information that a user may ultimately wish to retrieve surely exceeds the currently available supply by a considerable amount. Without a system for convenient and widespread access to such information by unsophisticated as well as experienced users, it may never be economical to provide it. Until it is provided, however, widespread use may be stifled. Here we see a classic chicken-egg dilemma and hence progress on both fronts moves at a glacial pace.

The spectrum of possibilities for use of a Digital Library System ranges from the tangible to the intangible, from the very specific to the vague and from the visual to the invisible. At the right-hand side of the spectrum, we denote fixed format documents intended to be read by people. These are generally assumed to be prepared for publication and have definite presentation formats. These documents are stared and retrieved in their presentation form. They are guaranteed to be reproduced as they were originally created, subject only to scale and resolution limitations of the print server.

Fixed Content, flexible format documents, shown just to the left of the fixed presentation documents in above Figure , require the user or his system to

specify how to present the information, assuming its content remains unchanged. For this class of documentation, the user might wish the text to be single spaced, double spaced, margins adjusted, page boundaries adjusted, fonts changed and so forth. This is in marked contrast with the fixed format documents, where no substantial visual changes of any kind are permitted.

In the middle of the above figure are shown database queries and data of the kind collected from sensors. The system treats sensor data along with database entries as if the were new types of objects in the library; this treatment requires understanding the semantics of objects in the library for the purpose of analysis and question/answering. When prestored answers are available without the need for searching documents, retrieval requests can be satisfied more quickly. Obviously, it will not be possible to anticipate all such questions in advance.

To the extreme left in above Figure are the two most speculative aspects of the spectrum of the Digital Library content. Although many attempts have been made to achieve reusable software, the infrastructure to reach this goal is still largely unexplored. Further, the preparation and reuse of knowledge structures in the development of intelligent systems is also virgin territory. This latter subject will be the focus of the second volume in this series. The initial version of the Digital Library System will be tailored for the domain of printable documents. However, the underlying technology will be designed to allow evolution to cover the remaining portions of the spectrum. Ultimately, we see the library system encompassing the entire range of possibilities shown.

Even with this initial restriction on content, the span of possibilities for inclusion in the library is enormous. In the implementation plan, we discuss how the library system will be developed aid how the supply of documentation can begin and expand. Most users subscribe to a given information service to retrieve highly selective pieces of information. Rarely do they learn to use the full complement of capabilities available on that or any other system. Almost all existing on-line informational services support users that connect via simple alphanumeric terminals or PCs in terminal emulation mode. Most users are able to do little more than print a received text string or view it on a screen. The power of personal computers is rarely used to exploit further processing of received on-line information. With the exception of spread sheet programs that accept certain financial data obtained electronically, and mail systems that allow for forwarding, little or no user processing of received information typically occurs.

LIBRARY AUTOMATION

Library technology like other fields, witnessed the supremacy of United States and United Kingdom. Though the lineage of term 'library automation' is traced back to 1930s, when punch cards developed by 'Hollerith' were first used in Acquisition and Circulation systems, but in actual practice computer application begun to use in libraries in 1960s. The application of computers ushered American and British libraries in the era of computerization. Against this background, different techniques were introduced during mid 20th century for streamlining circulation system, but the efforts taken during this century were by and large revolving around improving the manual system. However, some attempts were aimed to introduce mechanization to reduce the work of charging system (Martin, 1949). In addition, Texas University used photo charging system by making use of punch cards and Unit Record System of IBM which focused on machine readable data.

During the subsequent quarter century, a number of other libraries incorporated IBM's Unit Record System into their Circulation procedures. In this regard, IBM's Montclair system which resulted in most sophisticated automated Circulation system in pre computer era was an outstanding exception. But despite its proven success, the use of automated circulation, because of its non-cost effectiveness never reached great proportions in libraries. Till 1950s, the practice of automation in terms of machine readable records had not became widespread, as they were in limited use in circulation section. However, their use in Illinois Public Library made greater impact on acquisition section. The interest of computer application in libraries surfaced during 1950s and reached broader level in 1960s. Later, in this decade computer application introduced off line batch processing in libraries to enhance acquisition; cataloguing; circulation and serial control sections. By the mid-60s, more than 80 American academic libraries which shortly extended to 150 libraries, computerized their circulation system. Similarly, acquisition section was also computerized. During early 60s, several libraries like University of California and San Diego attempted to go beyond the listing of serials by implementing procedures for check-in and clearing (Salmon, 1969).

One of the most important landmarks of this decade was the evolution of Library of Congress MARC format.

After witnessing great deal of progress and development that computers brought to the libraries of developed countries, a number of developing countries also envisioned the importance of computers in libraries and subsequently efforts were made for implementation of Information Technology (IT) in libraries.

In this regard, Indian Statistical Institute (ISI), Calcutta was the first to install computer system in 1955, but as far the application of computers in Indian libraries is concerned, Indian National Scientific and Documentation Center (INSDOC) was possibly the first to computerize author and subject indexes of Indian Science Abstract (ISA) in 1965. However, in 1967, INSDOC brought out roster of 'Indian Scientific and Technical Translators' with the help of computers. In 1973, INSDOC with the help of computers brought out first union catalogue by the name of "regional union catalogue of scientific serials" Bombay-Poona. In 1978, it initiated SDI service as a National Information System for Science and Technology (NISSAT) project with chemical abstracts and INSPEC databases with the use of CAN/SDI software of Indian Institute of Technology (IIT) Madras. In 1970s, many libraries ventured in preparing computerized databases. Through the initiative and financial assistance of NISSAT, many library networks like: CALIBNET; DELNET; INFLIBNET; PUNNET; NICNET; INDONET and SIRNET were notable networks that became operational.

CONCLUSION

Technology has had and will perhaps continue to have a dramatic impact on library operations and services. It is the main force for changing the core work culture of library situation. The trends in technology will certainly find their way into large academic library set up. Because, the libraries must satisfy the expectations of their end users to sustain their goal, objective and existence in present techno-oriented world.

The first effort for computerization of library work started in early 1970s in India. However, it has been seen that Indian academic libraries gained significant momentum during the last decade in automating their functions and services. Presently, the large libraries of Indian Universities and IITs are in a state of implementing integrated library systems for automating their entire operations and services. But, the library staff and user community particularly in academic universities have still limited knowledge in this regard. Therefore, the implications of studies for technological changes and improvements in academic libraries are of good value to the librarians and authorities for implementing effective and successful automated library system.

An important consideration in making a scientific digital library useful is a flexible and expressive query facility. Central to this is an appropriate paradigm for formulating queries—a paradigm that readily supports the different definitions of content that are implied by different domains, or that capture different user's intent. We have developed what we believe to be a natural approach to query formulation by combining an abstraction pyramid with an object-based model. This approach permits easy combination of a variety of specifications, including example-based and semantic label-based.

A library of definitions, both user-defined and system defined, is developed through time, with powerful facilities for combining existing object definitions in creating new object types. In addition, the system readily supports both the novice and the expert user. The former, by providing a set of pre-extracted types, the latter by providing a rich facility for composing new types of arbitrary complexity.

REFERENCES

- Arms, W.Y. (2000) Digital libraries. MIT Press, Cambridge, Massachusetts.

- Aswal, R. S. (2006). Library Automation for 21st Century. New Delhi: ESS ESS, 56-57.

- Chen, S.S. (1998) Digital libraries: The life cycle of information. BE (Better Earth) Publisher, Columbia, MO.

- Dutta, Namita (1993). The Use of PC's in Library Automation; Information Technology & Health Science Libraries. New Delhi: Medical Library Association of India, 59-66.

- Heiliger, Edward M., & Paul B. Handerson (1970). Library Information Systems: From Library Automation to Distributed Information Access Solutions. Westport: Libraries Unlimited, 333.

- Lagoze, C. and Fielding, D. (1998) "Defining collections in distributed digital libraries". D-Lib Magazine, Vol. 4, No. 11; November.

- Ranjan, Vandana (1995). Information Trends in Library for Computer Professionals. In Nair Raman R (Ed.), Academic Library Automation (p. 102). New Delhi: ESS ESS.

- Saffady, William (1989). Library Automation: An Overview. Library Trends. 37(3), 273-277.

- Smith, T.R: A digital library for geographically referenced materials. IEEE Computer Magazine 29:54–60, May 1996

- Sun Microsystems (2000) The digital library toolkit. Sun Microsystems, Palo Alto, CA.

- Voight, Melvin J. (1956).the Trend Toward Mechanization in Libraries.

- Library Trends, 195.

Balijepalli Bhagyalakshmi