Open Archives Initiatives Protocols for Metadata Harvesting & Z39.50 in Library Software: A Model

R K Joteen Singh Th. Madhuri Devi

Abstract

The objective of library networking is to promote resource sharing and free flow of information. Rapid technological changes have brought a drastic change in the resource sharing activities, specially Open Archives Initiatives (OAI) model enables the dissemination of research output at international, national and regional levels thus removing the restrictions placed by the traditional scientific model. Z39.50 protocol has provided a common platform which help to develop a Union catalogue. These new developments have made resource sharing more practical. In such a scenario, the best approach is to focus on how software will make it possible for the library to utilize networked resources for the maximum benefit of their users. This paper presents a design of library software, which is interoperability in nature and support networking and exploitation of web resources apart from automating library operations.

Keywords: Open Archives Initiatives, Z39.50, Interoperability, XML, Metadata harvesting

1. Introduction

Information centres are now well recognized as important social institutions for dissemination of knowledge and information. Knowledge is playing a predominant role for propelling growth and advancement in almost every sector of a nation. According to Malhan[1] in the last two centuries industrial production have determined the economic advancement of a nation and in the present generation, management and use of knowledge resources is becoming crucial in creation of wealth. Today most of the countries have well recognized that information is a resource, which is an essential input to the effective pursuit of national policies on economic, scientific, technological and social development. It is also felt that, information infrastructure and modernization of the library system are essential to a well informative and organized country. Kaul[2] discloses that, poor countries differ from rich ones not only because they have less capital but also because they have less knowledge. Knowledge based societies have mainly been created by the information technology revolution. Because of globalization, growing competition and speedy access to vast global information resources, one can witness a spurt in knowledge activities and an enormously accelerating speed in work and action. In the knowledge based societies slow and steady losses the race. It is because of speed of information accessibility that information users now prefer to go for “Search Engines” instead of browsing through the boo ks and other printed materials.

In this changing scenario, effective management of knowledge is now playing a key role in wealth creation and world’s most strong economies are no more given emphasis on industrial production

but rather on becoming powerhouses of knowledge. Satpathi[3] has identified several organizations all over world at the national and international levels engaged in creating new knowledge through Research & Development activities particularly in the field of science and technology, which has resulted in "information explosion".

The exponential growth of literature and ever increasing cost of published materials have put great financial constraint, that has resulted in shrinking collections on the library and information centres. On one side user demands are increasing and librarians are at their wits end to satisfy the needs of their users. This problem is more pronounced in developing and underdeveloped countries. According to Salgar & Murthy[4] the only viable solution to meet users demand is to make optimum use of available literature. This is being done through pooling and sharing of resources. Sharing of resources is also not an easy task however, the changes in technology have opened up new opportunities for sharing data, and knowledge. During the last couple of decades, the communications world has witnessed several new developments, two of which are the Open Access (OA) and Open Archives Initiatives (OAI) whose main objectives are to improve transfer and exchange of research publications. Open Access is described as a free availability of information on the Internet cloud, permitting users to read, download, copy, distribute, print full texts of scholarly and scientific articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without any financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. The Open Access model allows for the widest dissemination of research output, and maximum visibility, while, at the same time, reducing barriers common to traditional diffusion methods of scientific literature. It provides the means for researchers to avail themselves of full text content, using two paths: open access publishing journals, which make articles openly accessible immediately upon publications, and Open Access archiving which allows authors to deposit a digital document in a publicly accessible web site, preferably an OAI – complaint open archive. The Open Archives Initiatives(OAI) develops and promotes interoperability solutions that aim to facilitate the efficient dissemination of content. The initiative is built on the principle that interoperability will federate the distributed open archives, thus encouraging the development of value-added services such as portals, subject gateways, and specialized search engines, with the overall benefit of increasing the visibility of the Open Archives[5]. The ultimate objective is to overcome existing barriers to interoperability by using open archives for all digital materials.

2. Interoperability

The technological changes and development of different new systems creates compatibility issues in different hardware and software platforms. Interoperability is a property referring to the ability of diverse systems and organizations to work together. It is the ability to software and hardware on different machines from different vendors to share data. Interoperability includes the exchange of data, records, and messages between computer systems across different hardware, operating

systems, and networks. Interoperability is sought to be achieved by establishing standards that different vendors of software and hardware can adopt so that they can share data and information. In case of software the term interoperability is used to describe the capability of different programs to exchange data via a common set of exchange formats, to read and write the same file formats, and to use the same protocols[6]. The lack of interoperability can be a consequence of a lack of attention to standardization during the design of a program. Standards like OAI-PMH and Z39.50 are the emerging protocols that promote interoperability.

2.1 Open Archives Initiatives – Protocol for Metadata Harvesting

The Open Archives Initiatives – for Metadata Harvesting (OAI-PMH) is one of the mechanisms used to achieve the interoperability between digital repositories. It provides a system to facilitate the harvesting, sharing and discovery of distributed resources. This allows materials within repositories to be accessed by a greater number of users via external services. In addition, data harvested via OAI-PMH is now being used for a range of other repository applications such as reporting, enhanced user interfaces for direct searching of local repositories, and assisting with ingest of data into other systems. OAI was formed originally with a focus on disseminating content from research archives however with increasing number and types of repositories it is now being used for other types of digital material collections[7]. Its main founders are the Digital Library Federation, the Coalition for Network Information and the National Science Foundation. The OAI-PMH is based on the Hypertext Transport Protocol (HTTP) and Extensible Markup Language (XML) open standards.

2.1.1 HTML and XML

HTML is mainly for displaying text in a desired format and XML is designed to do for data exchange. There are certain limitations for both the cases such as sometimes, XML won't be up to a certain task, just like HTML as sometimes not up to the task of displaying certain information. But when it comes to display, HTML does a good job most of the time, and those who work with XML believe that, most of the time, XML will do a good job to communicate information[8]. The basic difference between HTML and XML is:

- HTML is designed for a specific application -to convey information to humans through a web browser.
- XML has no specific application -it is designed for whatever use that need it for.

An XML document can be created and retrieve information from the document by any XML parser. HTML and XML are so popular for information display and exchange because they are standards. That means that anyone can follow these standards, and the solutions they develop will be able to interoperate. XML is platform and language independent, which means it doesn't matter that one computer may be using, for example, Visual Basic on a Microsoft operating system, and the other is

Unix machine with Java code. When it is necessary to communicate between the different systems, XML is a potential fit for the exchange format. Apart from these there are pretty advantages for using XML standards.

2.1.2 Advantages of XML

- It supports Unicode, allowing almost any information in any written human language to be communicated.
- It can represent common computer science data structure such as records, lists and trees.
- Its self documenting format describes structure and field names as well as specific values.
- The strict syntax and parsing requirements make the necessary parsing algorithms extremely simple, efficient, and consistent.
- XML is heavily used as a format for document storage and processing, both online and offline.
- It is based on international standards.
- It can be updated incrementally.
- It allows validation using schema languages such as XSD and schematron, which makes effective unit-testing, firewalls, acceptance testing, contractual specification and software construction easier.
- The hierarchical structure is suitable for most types of documents.
- It is platform-independent, thus relatively immune to changes in technology.
- Forward and backward compatibility are relatively easy to maintain despite changes in DTD to Schema.
- Its processor, SGML, has been in use since 1986, so there is extensive experience and software available.

2.2 Z39.50

Z39.50 is the American National Standard Information Retrieval Application Service Definition and Protocol Specification for open Systems Interconnection. The National Information Standards Organization (NISO), an American National Standards Institute (ANSI) accredited standards developer that serves the library, information, and publishing communities, approved the original standard in 1988 (Version 1). NISO published a revised version of the standard in 1992 (Version 2). Z39.50 defines a standard way for two computers to communicate for the purpose of information retrieval. Z39.50 makes it easier to use large information databases by standardizing the procedures and

features for searching and retrieving information. Specifically, Z39.50 supports information retrieval in a distributed, client and server environment where a computer operating as a client submits a search request to another computer acting as an information server. Software on the server performs a search on one or more databases and creates a result set of records that meet the criteria of the search request. The server returns records from the result set to the client for processing. The power of Z39.50 is that it separates the user interface on the client side from the information servers, search engines, and databases. Z39.50 provides a consistent view of information from a wide variety of sources, and it offers client implementers the capability to integrate information from a range of databases and servers.

Z39.50 can be implemented on any platform. This means that it enables different computer systems with different operating systems, hardware, search engines, database management systems. A Z39.50 implementation enables one interface to access multiple systems providing end users with nearly transparent access to other systems[9]. Users access multiple systems with the familiar commands and displays of their own local systems. New commands and search techniques do not have to be learned. The results of the search are presented on the local system again, in the formats and styles users are accustomed to. One of the strengths of ANSI/NISO Z39.50 is that it is an American National Standard. NISO developed and maintains Z39.50 using consensus procedures approved by ANSI, the principal coordinator of voluntary standardization in the United States. Z39.50 is not a proprietary standard and will continue to be responsive to the needs of the implementers that use the standard and the information consumers that benefit from its implementation.

Search and Retrieve Information Through Z39.50

The basic technology of the search and retrieval of information based on Z39.50 standard is shown below:

- A query is typed into the distributed search screen (Coming from the Z39.50 server) using a web browser;
- The browser passes the query to the Z39.50 server;
- The Z39.50 server distributes the request to member library servers, with Z39.50 client installed;
- The Z39.50 clients responds with a result passed back to the initiating Z39.50 server;
- The Z39.50 server delivers pooled results to the initiating browser client.

3. Basic Design of a Library Software

Providing access to a variety of information resources residing on different computer systems with different platforms in several parts of the world to a number of users of differing natures and needs

is a major challenge for software designers. Library software, should work on different platforms, enable to harvest metadata and store locally, make it possible to search multiple disparate library catalogues and other resources in one search, and bring back one set of results. This involves a number of complex issues related to integration and seamlessness.

Fig. 1 shows the basic design of library software. As this figure shows, users of a library may have access to a range of information resources and there are different modes of getting access to them. Alternatively, users may choose one or more resources or collections and then formulate just one query, which is passed on to the various resources or collections by the software interface, results are brought back after the search is carried out. The user does not need to search the resources one by one, so this is better approach from the user's perspective because he or she formulates only one search query and sets results from all the different sources. However, technologically this approach is more challenging and a number of technical issues need to be considered in order to build this model.

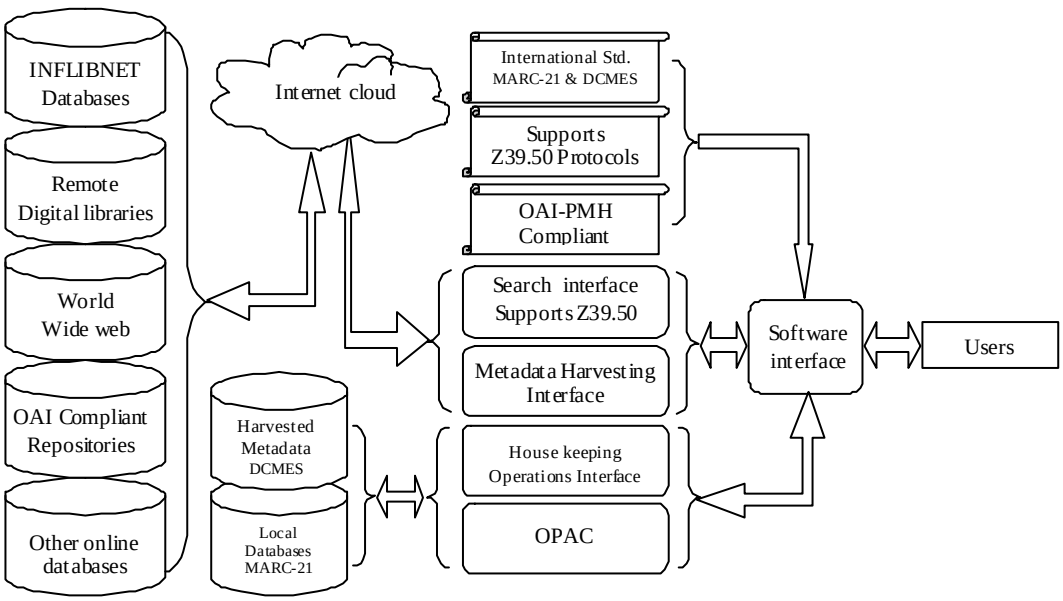

Fig. 1 Conceptual design of a library software

3.1 Suggestions

The software requirements at various levels like union catalogue compilation and database maintenance and for individual library automation at local, such as circulation and serials control operations are clearly mentioned. For application in information retrieval, federated searching and metadata harvesting the software should incorporate the following special features:

- The software package should be an integrated one, to support library automation and database construction and information retrieval.
- It should support international data standards such as MARC-21 and DCMES.
- It should work in multi-user and network environment.
- It should support copy cataloguing authoritative MARC-21 cataloguing sources via the web.
- It should provide high-level language interface to the database for the user to write any special routines to manipulate the database.
- It should facilitate federated searching to different databases using an interoperability standard such as Z39.50
- The software should allow the building of an OAI-PMH compliant institutional repository of self – archived materials.
- It should allow harvesting metadata from other OAI-PMH compliant repository.

3.2 Interoperability Scheme

One of the major problems facing to develop library software is the issue of interoperability. How to get a wide variety of computing systems to work together and/ or to talk to one another for access to and retrieval of information? Interoperability and standardization are the most important considerations for library software designers[10]. Interoperability of library software can be achieved by a number of means, such as through adopting.

- Common user interfaces
- Uniform naming and identification systems
- Standard formats for information resources
- Standard metadata formats
- Standard network protocols
- Standard information retrieval protocols
- Standard measures for authentication and security, and so on.

4. Conclusion

The technological changes and development of numerous library networks using contrasting hardware and software platforms asked library software interoperability. The technology also made wandering treasured and heterogeneous knowledge in the field of science, arts, and commerce in the internet cloud. The basic thing what we need is to use an interoperable library software which can exploit web resources and able to search multiple disparate databases in different network and platforms apart from the local repositories and databases. If we develop a software in the line of the above

designed it will promote software interoperability and compatibility. It will also have the capability to integrate all the libraries of the world to a single global library.

References

1. Malhan, IV. "Library resource sharing in a networked environment". Proceedings of Seventh National Convention for Automation of Libraries in Education and Research (CALIBER-2000), Ahmedabad, India, Feb. 2000.
2. Kaul, S. "Information Resource Sharing Models in Developing Countries: A Network Emerging from the World Bank Supported Environmental Management Capacity Building Project". INSPEL, Vol.35 No.1, 2001, p 9-26.
3. Satpathi, JN. "Resource Sharing Among Health Science Libraries in West Bengal: A Case Study". Proceedings of Forty-ninth FID Congress and Conference, New Delhi, 1998.
4. Salgar, SM. and Murthy, TAV. "Enhancing Access to Information Through Document Delivery Systems – INFLIBNET's Approch". Proceedings of Sixty-eighth IFLA Council and General Conference, Glasgow, Aug. 2002.
5. Subirats Imma, Onyancha Irene, Salokhe Gauri and Keizer Johannes. "Towards an architecture for open archive networks in Agricultural Sciences and Technology". available at: www.ftp.fao.org (accessed 14 May 2008).
6. Haravu, LJ. "Standards in Library Automation and Networking". available at: www.drtc.isibang.ac.in (accessed 4 February 2008).
7. Sahoo, BB. "Need For A National Resource Sharing Network in India: Proposed Model". available at: www.drtc.isibang.ac.in (accessed 26 February 2008).
8. Hunter David, Cagle Kurt, Gibbons, Ozu Nikola, Pinnock Jon, Spencer Paul. "Beginning XML". United Kingdom: Birmingham, Wrox Press Ltd. 2000, p18-23.
9. Moen, W. "Information Infrastructure". available at: www.cni.org (accessed 20 April 2008).
10. Chowdhury, G G. "Introduction to modern information retrieval". Great Britain: Facet Publishing, 2004. p38-40.

About Authors

Dr. R K Joteen Singh, Information Scientist in Manipur University, Imphal. **Dr. Th. Madhuri Devi,** Reader & Head of Department, Library and Information Science, Manipur University. Imphal.