Yu, S. C. (2006). Study on digital archives standard for library automation system. In C. Khoo, D. Singh & A.S. Chaudhry (Eds.), *Proceedings of the Asia-Pacific Conference on Library & Information Education & Practice 2006 (A-LIEP 2006), Singapore, 3-6 April 2006* (pp. 288-293). Singapore: School of Communication & Information, Nanyang Technological University.

STUDY ON DIGITAL ARCHIVES STANDARD FOR LIBRARY AUTOMATION SYSTEM

SHIEN-CHIANG YU

Associate Professor and Director of Library,
Dept. of Information & Communications Shih-Hsin University
No.1, Lane 17,Sec.1, Mu-Cha Rd, Taipei, Taiwan, 11604, R.O.C
E-mail: ysc@cc.shu.edu.tw

Abstract. With the development of internet and the trend of information system technology, the object of digital library research has extended to the application of digital archives. Basically, digital libraries storage huge amounts of data, including text, image, map audio, video and illustration via electronic formats. Further more, digital libraries could be conveniently accessed through the Internet. As the research intention of network information systems, the critical technology in digital library research could be how to let users effectively harvest correct information from the digital library. Digital library users could discover, present, and organize knowledge among these data of digital libraries. The traditional library automation system, related applying technologies and protocols, such as MARC, Z39.50 and ISO 2709, could not totally match the requirement of digital archives. The purpose of this research is to find out how to effectively manage and apply the related technology of digital archive to handle the existing operation processes in library and the management requirement of digital archives. This paper discusses an evolution model of the related technology of library automation systems.

Introduction

The development of digital archives raises the application of technology of digitization and the value add of content. It relates to the application of digital libraries as they cover the domain of digital archives. Regardless of whether focusing on digital libraries or digital archives, the research emphasizes the electronic mode to store and manage massive materials (including metadata, image, map, sound, image, illustration and other digital objects), and operating through networks in an effective way. Therefore, the research on the digital archives can be equated to the research on network information systems. The crucial technical agenda, including manufacturing and storing of digital objects, and spreading rich and correct information content effectively, requires systems that allow the user to search, present and reorganize knowledge in all kinds of data through the Internet.

The mission of the library, including collection, exhibition, research, and education, is the same as reservation or archive institution. The scope of library automation system processing archives also contains book, non-book material and other various publications. Essentially, it does not differ too much from the applied purpose of the digital archives. However, the operative items of existing library automation systems work and relative application technology, protocol (for example MARC, Z39.50, ISO2709, and so on), are not appropriate for the applied requirements of digital archives in their entirety. On the other hand, the application systems of international digital archives projects are also not appropriate for the library automation system. They result in issues of traditional library services integrating with digital archives.

This study analyzes library automation and digital archives systems and the possibility of their integration; to effectively manage and employ appropriate technology, perform existing library tasks, and satisfy the information management demand of digital library/ digital archives systems.

Motive

Due to quick development of digital archives, sunrise electronic publication and popularization of Internet, collection policies of institutions obtaining digital information are changing. Therefore, consideration of the system development mainly emphasizes the presentation and management of digital objects

Users have different levels of information demands, such as retrieval, management, preservation demand, and so on. The different historical space-time of information has different special characteristics of documents, such as descriptive method, terminology, type and so on. It has different document structures, utilization habits, media characteristics, and so on. Therefore, digital archives must employ various metadata and digital object techniques to represent the diversity of documents.

Digital archives even slice into reservation level (detailed description for metadata, highest resolution for digital object), electronic commerce (key description for metadata, well resolution for digital object) level and public information level (brief description for metadata, low resolution for digital object).

At the beginning of digital library development, systems tended to go their own ways because of the need of a particular community. Intended to be quick solutions to urgent community needs, variation of the retrieval interface, system structure, and the management policy all emerged (Suleman and Fox, 2001). In order to vertically integrate or horizontally link related documents and content with digital archives, it is necessary to provide the function of union cataloging to interflow the archives.

To sum up above mentioned, there are three differences from library automation to digital archives systems:

1. The document must simultaneously contain both of the description data (metadata) and the digital object (multimedia).
2. The demand of data structure for digital archives.
3. The mechanism of interoperability among systems or content.

The existing management, standards, and protocols in library automation systems, although this may handle the above differences, still had disparities in actual application. They were unable to totally satisfy the requirements of information management. These disparities were the major factor why library automation systems could not be implement digital archives. The demand for document presentation need simultaneously contain both of metadata and digital object, may make use of the information technology such as web page. The major difference between library automation and digital archives systems is data structure. The data structure leads into different standards of interoperability and integration. This study analyses and discusses these differences, and then it proposes an integrated framework for library automation and digital archives systems.

System Scope

The distinction between digital library and physical library is that the digital library emphasizes totally digital environment. Therefore, there are huge differences in the collection. Systems emphasize the system frameworks, digital contents and collection, metadata, interflow, standard, knowledge organization, users, and copyright protection (Shiri, 2003). Digital libraries stress digitalizing ability and service, and main environment of the system operating is the Internet. Therefore, digital libraries must closely collocate with digital objects to ensure effective use. The automation systems implemented in physical libraries are utilized to support physical materials, but they cannot substitute the physical library. Those two are different in essence.

When comparing a digital library with a digital archives institution, digital libraries emphasize the collection of fully digital information. The major objective of digital archives is to digitize cultural collections; promote the development about humanities and society, industry and economics. It not only emphasizes digital procedure and the effect of digital presentation, but also merges the characteristics of the physical library which have material collections and digital libraries which apply digital dissemination.

Table 1. Comparison of system scope

System scope	Process objective	Process key-point
Library automation system	Digital description and automatic service of entity publication (books and non-book material)	1. Provide circulation and reading of material collection 2. Provide information service of bibliographic (digital description)
Digital library system	Completely digital information and service	1. Does not have material collection 2. Only provide digital objects and metadata (digital description)
Digital archives system	Digitize the material, record the description (metadata), and present or disseminate these digital objects.	1. Provide material collection to exhibit or to circulate 2. Digitalize material collection into digital object and description 3. Provide digital objects and description for service

Based on the scope of information flow, the library automation system show as Fig. 1 includes internal processing (double arrow-line), external processing (single arrow-line), and data backup/sharing.

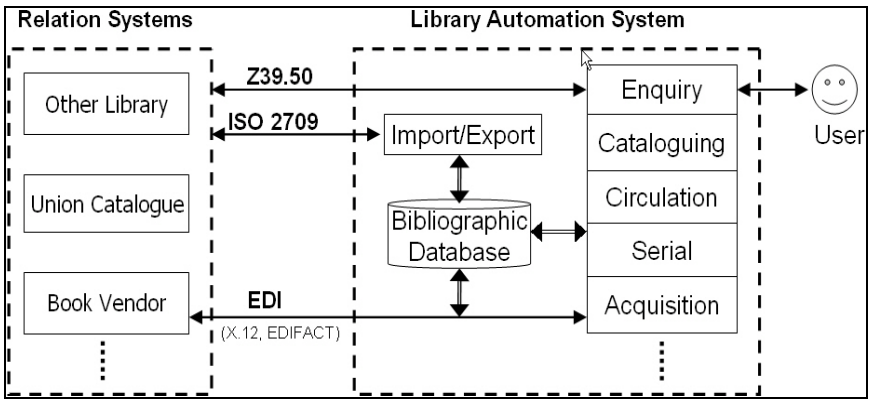

Figure 1. The structure diagram of library automation system

Internal processing

If we treat library automated information systems as individual entities, then the whole operating information flow of interior processes in the system may belong to the scope of internal processing. Data is imported by external files or are keyed in. The information processes before the system transfers the information to the internal storage form is called internal processing work. This information processing includes character-code process and data structure setup.

External processing

The system links with exterior systems belong to external processing. This work emphasizes the remote execute ability among systems. The automation environment, including the operation systems, system platform, software development kit (SDK), and data forms tends to more complex. Therefore, need to consider the method of "procedure call" (one system process another system's program or service), the transaction of cross platform (such as distributed retrieval), and exchange protocols.

Backup and sharing

This work includes the data format of the import/export interface. The operational item needs to consider not only the acceptable data exchange among cross-platforms, but also the open interoperability protocols. Besides, the backup procedure and the data sharing are approximately the same in information exchange processing; it is mainly to transfer the internal data format to the external data format. Data or procedure exchange must consider the information protocol, but the output of backup considers the input/output model of peripheral storage equipment.

Technique Analysis

Markup and construction

MARC as well as various metadata contain three sections; semantic, syntax and structure. Each library automation system follows semantic cataloguing rules. Individual data structure must be adopted as internal structure to store into a database. The external structure and syntax are based on ISO2709 (Cooper, 1996) - the catalog information exchange format standard (1981, 2ed.), as bibliographic cataloging standard. The problem of ISO2709 depends on control code only limit to three kinds of field, which are tag, subfield and indicator. and is unable to identify the variety of MARC (USMARC, UKMARC, Chinese MARC etc), the character code (Unicode, ASCII, etc.), and the data is unable to present directly in webpage, because of such disadvantage which is unable to be appropriate for the applicable demand in digital archives information description.

On the other hand, the XML supports language- and platform-neutral facilities and offers a unique combination of flexibility, simplicity, and readability performed manually or automatically. It uses a

reasonably concise syntax that can provide developers with an enormous amount of power. The simplicity of XML with its clear structures makes it useful for record description. XML can be messaged, manipulated, processed, fragmented, and rebuilt far more easily than MARC formats. XML enables customized markup languages to be defined with application-specific tags that represent information in such application domains as chemistry, electronics, and general business. Therefore, XML is widely used in metadata of digital archives as main syntax and format. As Fig. 2 is completely based on XML, to achieve the transparency of information process, and the merged the demand of integration of data and functions among various digital archives system

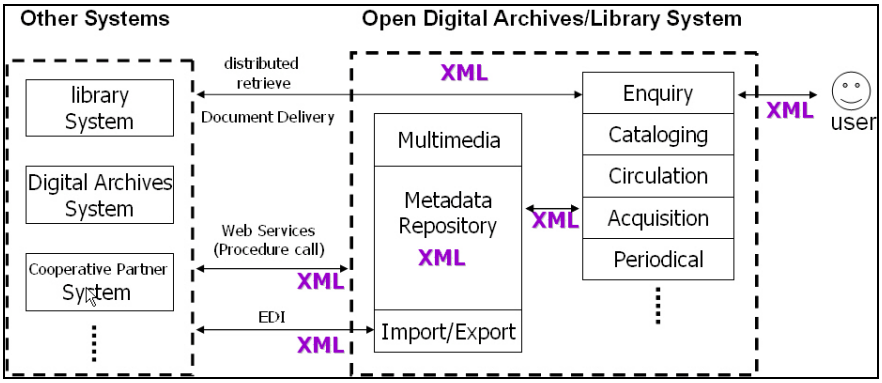

Figure 2. Integrate with functions of digital archives for open library system

For those reasons, the Lane Medical Library of U.S.A. Stanford University medical center started the Medlane Plan in 1998. The first step was planning XML MARC, and then announced the execution software in December 1999 and transferred MARC to XML document (Lane Medical Library, 2004) by comparison with each other. This plan including following goals (Tennant, 2002):

1. Provide elastic transformation function to transfer the MARC record to the XML document in accordance with the comparative table (or called mapping table) which is provided by the user.
2. Test the different format of XML document in order to decide what the more suitable structure for the library is.
3. Guarantee the transfer process could hold complete data which record by MARC, and can store it using an open software method.

Besides the syntax of data processing being based on XML, the metadata which decides the cataloging rule also has to follow the open definition standard for structure description. The structure definition of XML data recording includes XML Schema and DTD. Since 1990, the Network Development of Library of Congress (LC) and MARC Standards Office started to carry on a pilot project about MARC DTDs, to take the MARC record as a specific document type, and defined the MARC tag into element. At the beginning, the main goal was establishing a standard SGML DTD to supply the catalog data to transfer to SGML document format from MARC data structure (Keith, 2004).

In the wake of developments and changes in technology, in February, 1998, the World Wide Web Consortium (W3C) announced the document definition standard of XML. The LC Network Development and MARC Standards Office started to research the interchange of MARC21's ISO2709 and XML, and officially announced the definition of MARC21 XML Schema in September, 2002 (McCallum Keith, 2002). It provided JAVA application program interface (API) to transfer MARC21 and XML document suitably. Taiwan also completed the XML transfer program from Chinese MARC in August, 2004 by the National Central Library that adopted XML Schema to make the definition of MARC structure.

Data exchange and distributed enquiry

The distributed retrieval function is the ability for cross systems to take remote enquiries. The core of function execution must contain the protocol of translatable information exchange, data packing, and transmission of bilateral systems. The data encapsulation stands for the message by protocol of data exchange. Digital archives data processing is based on XML, and uses DTD or XML Schema to be the

definition for description structure. Consequently, XML is a common standard in the data exchange. In addition, many institutions developed open protocols or methods for data exchange or interoperability.

OAI-PMH is an open standard that adopt XML to export and harvest metadata (this metadata refers to data that were described by definition of syntax and structure). The objective of design is to share technology for export files, provides one simple, low barrier way to establish a method for communicate, share and deliver data among systems. The major applied objective not only provides the exchange and the use of data, but also includes the following three points:

1. Simplify the content of documents to share effectively.
2. Promote the storage and use of electronic document.
3. Expand the scope of storage and use in digital data type.

This will help the service provider (client side) to harvest all related resources conveniently from the data provider (server site, which is the digital archives system). "The open standard for distributed information retrieval" of the Taiwan library information related technology standard assembles in 2001(Union Catalog of National Digital Archives Program, 2004). OAI-PMH has been listed as one of the standard suggestions of distributed retrieval standard research group.

OAI-PMH not only achieves data exchange and sharing among systems, but can also solve the disadvantages of ISO2709. MARC-specific data, which only supports three record levels (tag, indicator and subfield), unable to limit the MARC category, to recognize the character code, to present on webpage directly and so on. Therefore, systems must have the ability to apply XML and the related extended technology when using the library system integrated digital archives system.

Z39.50 is the open information retrieval protocol which has been applied to the library automation system for a long time. MARC and Dublin Core adopted its kernel for data processing. Z39.50 is a kind of peer to peer service. It can be used as a tool to build federated search systems. In such a system, a origin (or named Z39.50 Client) sends a search in parallel to a number of information targets (or named Z39.50 Server) that comprise the federation, and then gathers the results, eliminates or clusters duplicates, sorts the resulting records and presents them to the user. It has proven to be very difficult to create high-quality federated search services across large numbers of autonomous information servers through Z39.50 for following three reasons (Lynch, 2001):

1. Retrieval accuracy: different servers interpret Z39.50 queries differently, in part due to lack of specificity in the standard, leading to semantic inconsistencies as a search is processed at different servers.
2. Scaling: In the management of searches that are run at large numbers of servers; one has to worry about servers that are unavailable (and with enough servers, at least one always will be unavailable).
3. Performance: the user has to wait for a lot of record transfer and post-processing before seeing a result, making Z39.50-based federated search performance sensitive to participating server response time, result size, and network bandwidth

The OAI-PMH itself is only the protocol to export and harvest data, it does not contain any search commands. Therefore, it was not classified as a distributed retrieval protocol. When building up a platform of union category for interoperation and sharing data among systems by OAI-PMH, the system could provide the functions for distributed retrieval when collocating with indexing. Z39.50 is under reversed procedure (Z39.50 International Maintenance Agency, 2002). According to the foregoing, OAI-PMH still is hitherto the major protocol that can provide the library automation and digital archives system for data exchange and distributed retrieval among various systems, still need OAI-PMH.

Proposed Framework

To conclude we must ensure that the data processing core of library automation systems unify XML as the document indication basis of book and non-book collection descriptions. XML document however does not contain the data structure and description of present data. Based on the Metadata description standard (for example: USMARC, UKMARC, etc.), it also contains the structure definition and style-sheet language for this Metadata.

Besides changing the internal data structure into XML, existing MARC catalogue data made by ISO2709 uses the specific application, matching with MARC which is defined by XML Schema, to migrate the data from ISO2709 to the XML document format, then import into the novel framework system. The XML data exchange standard can collocate any kind of existing XML extended exchange protocol to achieve the work of data exchange or interoperation.

The OAI-PMH for harvesting and exporting metadata (such as DC or MARC) from other libraries or archives systems must coordinate with are storage in local systems and be built with those brief metadata which were harvested from the service provider (other library or archives systems). As the user retrieves the data needed, to inspect the detailed content, the system immediately harvests detailed metadata from the service provider by OAI-PMH. The efficiency of this execution is certainly not more effective than when using the Z39.50 directly. However, Z39.50 is still unable to support XML. Considering the common demands of union catalog and distributed retrieval, it is a better solution now to adopt OAI-PMH. Moreover, the OAI-PMH supports the MARC standard that had been defined by the XML standard and any self-defined metadata. Therefore, it may conveniently extend the library system from distributed retrieval function to many applications layer of the library service federals.

Conclusion

Digital libraries and digital archives systems emphasize the open information processing platform. This coincides with the basic principle of the library automation system. The tasks of libraries focus on the collection, organization, and arrangement, provided with especially outstanding discipline and service on the physical collections, and the unity of the metadata description. Since automation began, coordinating the various applications of information systems has involved the re-engineering procedure of the whole workflow and completely performs the standard operation procedure (SOP). However, these standards have been working through the change of information systems and the network environment. Partial standards do not suit the management of existing digital archives.

Up to now, the library automation system cannot directly handle data format from digital archives, or directly exchange data with other digital archives systems. The major reasons are the data structure which include metadata definition, exchange format, interoperability protocol do not support XML directly. Some standards above are under revision or have been revised to XML standard. It still is necessary to enhance the promotion and application in system service, and to raise the level of library collections and services.

References

Cooper, M. D. (1996), *Design of Library Automation Systems*, John Wiley & Sons, Inc., New York.

Keith, C (2004), "Using XSLT to manipulate MARC metadata", *Library Hi Tech*, Vol. 22 No.2, pp. 122-130.

Lynch, C. A. (2001), "Metadata Harvesting and the Open Archives Initiative", *ARL Bimonthly Report 217*, available at: www.arl.org/newsltr/217/mhp.html

Lane Medical Library (2004), "Medlane:XMLMARC", available at: laneweb.stanford.edu:2380/wiki/medlane/xmlmarc

McCallum, S. and Keith, C. (2002), "Library of Congress Announces Standard MARCXML Schema", available at: xml.coverpages.org/LOC-StandardMARCXML-ShemaAnnounce.html

Melnik, S. (1999), "Generic Interoperability Framework (GINF) Middleware", Department of Computer Science, Stanford University, available at : www-diglib.stanford.edu/diglib/ginf/WD/ginf-magic/

Nelson, M. L. et al. (1998), "NCSTRL+: Adding Multi-Discipline and Multi-Genre Support to the Dienst Protocol Using Clusters and Buckets", in Proceedings of the IEEE Forum on Research and Technology Advances in Digital Libraries (IEEE ADL '98), Santa Barbara, CA, pp. 128-136, available at : techreports.larc.nasa.gov/ltrs/PDF/1998/mtg/NASA-98-ieeedl-mln.pdf

Paepcke, A. et al. (2000), "Search Middleware and the Simple Digital Library Interoperability Protocol", *D-Lib Magazine*, Vol. 6 No.3, available at: www.dlib.org/dlib/march00/paepcke/03paepcke.html

Shiri, A. (2003), "Digital library research: current developments and trends", *Library Review*, Vol. 52 No. 5, pp. 198-202.

Suleman, H. and Fox, E. A. (2001), "A Framework for Building Open Digital Libraries", *D-Lib Magazine*, Vol. 7 No. 12 available at: www.dlib.org/dlib/december01/suleman/12suleman.html

Tennant, R. (2000), *XML in Libraries*, Neal-Schuman, New York.

Union Catalog of National Digital Archives Program (2004), "Introduction to Union Catalog of National Digital Archives Program", available at: catalog.ndap.org.tw/System/Info/Intro.doc

Z39.50 International Maintenance Agency (2002), "ZING: Z39.50-International Next Generation", The Library of Congress, available at: www.loc.gov/z3950/agency/zing/