# DIGITAL LIBRARIES

**M.L.I.Sc., Semester – II, Paper-II**

**Lesson Writers**


**Prof. K.C.Das**
PG Department of Library
& Information Science
Utkal University
Vani Vihar
Bhubaneswar
Odisha

**Mr. Kunwar Singh**
PG Department of Library
& Information Science
Utkal University
Vani Vihar
Bhubaneswar
Odisha


**Editor**


**Prof. K.C.Das**
PG Department of Library & Information Science
Utkal University, Vani Vihar
Bhubaneswar, Odisha



**Director**
**Dr. NAGARAJU BATTU**
*MBA., MHRM., LLM., M.Sc. (Psy).,MA (Soc)., M.Ed., M.Phil., Ph.D*
**CENTRE FOR DISTANCE EDUCATION**
**ACHARAYA NAGARJUNA UNIVERSITY**
**NAGARJUNA NAGAR – 522 510**

***Ph: 0863-2293299, 2293214, ,Cell:9848477441***
***0863-2346259 (Study Material)***
***Website: www.anucde.info***
***e-mail:anucdedirector@gmail.com***

# M.L.I.Sc.,

**First Edition: 2021**

**No. of Copies:**



**Published by:**
**Dr. NAGARAJU BATTU,**
***Director***
**Centre for Distance Education,**
**Acharya Nagarjuna University**

***Printed at:***

## *FOREWORD*

***Since its establishment in 1976, Acharya Nagarjuna University has been forging ahead in the path of progress and dynamism, offering a variety of courses and research contributions. I am extremely happy that by gaining 'A' grade from the NAAC in the year 2016, Acharya Nagarjuna University is offering educational opportunities at the UG, PG levels apart from research degrees to students from over 443 affiliated colleges spread over the two districts of Guntur and Prakasam.***

***The University has also started the Centre for Distance Education in 2003-04 with the aim of taking higher education to the door step of all the sectors of the society. The centre will be a great help to those who cannot join in colleges, those who cannot afford the exorbitant fees as regular students, and even to housewives desirous of pursuing higher studies. Acharya Nagarjuna University has started offering B.A., and B.Com courses at the Degree level and M.A., M.Com., M.Sc., M.B.A., and L.L.M., courses at the PG level from the academic year 2003-2004 onwards.***

***To facilitate easier understanding by students studying through the distance mode, these self-instruction materials have been prepared by eminent and experienced teachers. The lessons have been drafted with great care and expertise in the stipulated time by these teachers. Constructive ideas and scholarly suggestions are welcome from students and teachers involved respectively. Such ideas will be incorporated for the greater efficacy of this distance mode of education. For clarification of doubts and feedback, weekly classes and contact classes will be arranged at the UG and PG levels respectively.***

***It is my aim that students getting higher education through the Centre for Distance Education should improve their qualification, have better employment opportunities and in turn be part of country's progress. It is my fond desire that in the years to come, the Centre for Distance Education will go from strength to strength in the form of new courses and by catering to larger number of people. My congratulations to all the Directors, Academic Coordinators, Editors and Lesson- writers of the Centre who have helped in these endeavors.***

***Prof. P. Raja Sekhar***
*Vice-Chancellor (FAC)*
*Acharya Nagarjuna University*

# DIGITAL LIBRARIES
## Syllabus

**Objectives:**

1. To make the student understand the concept of digital libraries and major digital library initiatives
2. To create an awareness on management of digital resources
3. To make them familiar with digitization techniques and their application

**UNIT 1**

Concept of Digital Libraries - Transition of Libraries from Traditional to Digital – Definitions, Characteristics, Components, Theoretical Fundamentals, Merits and Demerits and Challenges

**UNIT 2**

Digital Library Management – Design and Organization of Digital Libraries – Architecture – Protocols – Metadata – Standards – SGML, Z39.50

**UNIT 3**

Digital Resources : Nature and Management – Digital Library Evaluation – Digital preservation – Digital Archiving - Need and Strategies

**UNIT 4**

Overview of Major Digital Library Initiatives – Digital Library Initiatives in India – Open Source Initiatives, Open Archive Initiative (OAI)

**UNIT 5**

Building the Digital Library – Digitization – Process and Methods – Planning for Digitization - Institutional Repositories – open source Software for digital libraries : GSDL; DSpace; EPrint— Future of Digital Libraries.

**Books for study and reference:**


1. Balakrishnan, Shyama & Paliwal, P.K. Library Digital Technology. Delhi, Anmol, 2001
2. Brogan, Martha L. A survey of Digital Library Aggregation service. Washington, Digital Library Federation, 2003
3. Brogan, Martha L. Contexts and Contributions: Building the distributed library. Washington, Digital Library Federation,2003
4. Deegan and Tanner. Digital Futures. London, L.A., 2002
5. Ganguly, R.C. Digital libraries: Challenges and prospects. Delhi, Isha books, 2007
6. Hughes, Lorna M. Digitizing Collections: strategic issues for the information manager. Newyork, Neal Schuman Pub., 2004
7. Iorna and Hughes. Digitizing Collections. London, Facet, 2004
8. Pedley, Paul. Digital Copyright. 2nd ed. London, Facet, 2009
9. Singh, Ram Shobhit. Encyclopaedia of digital libraries. 2 Vols, Vol.1&2. New Delhi, Anmol Pub, 2008
10. Chowdhury, G.G. and Foo,Schubert, Eds. Digital Libraries and Information Access: Research perspectives. Facet pub, 2012.

# DIGITAL LIBRARIES

## CONTENTS

LESSON - 1

# CONCEPT OF DIGITAL LIBRARIES:TRANSITION OF LIBRARIES FROM TRADITIONAL TO DIGITAL

## OBJECTIVE

After reading this chapter, students will be able to understand:

- Basic fundamental concepts of digital library
- Transition of Libraries from Traditional to Digital Libraries
- Benefits of digital library
- Characteristics of digital library

## STRUCTURE



## 1.1 INTRODUCTION

A digital library is an online collection of digital objects, of assuring quality, that are created or collected and managed according to internationally accepted principles for collection development and made accessible in a logical and sustainable manner, supported by services necessary to allow users to retrieve and exploit the resources. A digital library forms an integral part of the services of a library, applying new technology to provide access to digital collections. Within a digital library collections are created, managed and made accessible in such a way that they are readily and economically available for use by a defined community or set of communities. A collaborative digital library allows public and research libraries to form a network of digital informationin response to the needs of the Information Society. The systems of all partners in a collaborative digital library must be able to interpret. A digital library complements digital archives and initiativesfor the preservation of information resources.

## 1.2 CONCEPT OF DIGITAL LIBRARY

A 'digital library' may be understood in different ways, and may be named differently; Chowdury and Chowdury (1999) and Borgman (1999) draw attention to both complementarily and contradiction in various definitions. Terms such as electronic library, virtual library, hybrid library, gateway library, library of the future, and library without walls are used, sometimes synonymouslywith digital library, sometimes to denote a subset, or a superset, of it, sometimes to denote a rather different concept.

The basic concept under laying the digital library is not new. In 1945, Dr. Vannevar Bush ofthe U.S. Office of Scientific Research and Development discussed a device called a "memex". He envisioned this device being used by individuals as "a sort of mechanised private file and library". It would be able to store large amounts of books, pictures, periodicals, newspapers, correspondence, and so on, with material being indexed for easy retrieval. According to Saffady, the Bush vision is "one of the most influential and frequently cited precursors" of the modern digital library concept. He continued to note that although the digital library seems a revolutionary development, the concepts and technologies involved are more accurately described as evolutionary.

Although not a recent concept, in terms of actual development, digital libraries are still relatively new. Because of this, there is as yet no universally agreed terminology in place. In the literature, the digital library may also be called the library without walls, virtual library, electronic library, e-library, desktop library, online library, future library, library of the future, logical library, networked library, hybrid library, gateway library, extended library or information superhighway.

These many terms, digital library, virtual library, hybrid library and electronic or e-library are most common. Just as there is no universally agreed upon terminology for digital libraries, neither is there a common definition for this concept. In the 1990s, terms such as digital library, virtual and electronic library became widely used, but considerable uncertainty remains about what they actuallymean.

## 1.3 TRANSITION OF LIBRARIES FROM TRADITIONAL TO DIGITALLIBRARIES

The shift from traditional libraries to the digital is not merely a technological evolution, but requiresa change in the paradigm by which people access and interact with information.
A traditional library is characterized by the following:

- Emphasis on storage and preservation of physical items, particularly books and periodicals.
- cataloging at a high level rather than one of detail, e.g., author and subject indexes asopposed to full text
- browsing based on physical proximity of related materials, e.g., books on sociology arenear one another on the shelves
- passivity; information is physically assembled in one place; users must travel to the libraryto learn what is there and make use of it

By contrast, a digital library differs from the above in the following ways:

- ❖ emphasis on access to digitized materials wherever they may be located, with digitizationeliminating the need to own or store a physical item
- ❖ cataloging down to individual words or glyphs
- ❖ browsing based on hyperlinks, keyword, or any defined measure of relatedness; materialson the same subject do not need to be near one another in any physical sense
- ❖ broadcast technology; users need not visit a digital library except electronically; for themthe library exists at any place they can access it, e.g., home, school, office, or in a car.

### 1.3.1 Everything can be Stored

The total number of different books produced since printing began does not exceed one billion. (The number of books now published annually is less than one million.) If an average book occupies 500 pages at 2,000 characters per page, then even without compression it can be stored comfortably in one megabyte. Therefore, one billion megabytes are sufficient to store all books. This is $10^{15}$ bytes, or one peta byte. At commercial prices of $20 per gigabyte, this amount of disk storage capacity could be purchased for $20 million. So it is certainly feasible to consider storingall books digitally.

### 1.3.2 Very Large Databases

A database of a billion objects, each of which occupies one megabyte, is large but not inconceivable. Once one is comfortable with sizes of this kind, it is feasible to imagine a thousand such databases, or to envision them all as portions of the same global collection. This amount of storage is sufficient to house not only all books, but all of the following:

- ✓ photographs
- ✓ legislative material, court decisions
- ✓ museum objects
- ✓ recorded music
- ✓ theatrical performances, including opera and ballet speeches
- ✓ movies and videotape

### 1.3.3 Distributed Holdings

When information is digitized and accessible over a network, it makes little sense to speak of its “location,” although it is technically resident on at least one storage device somewhere, and that device is connected to at least one computer. If the information is available at multiple mirror sites, it is even less meaningful to speak of it being in a “place.” While traditional libraries measure their size by number of books, periodicals and other items held, the relevant statistic for a digital library is the size of the corpus its users may access. This means that digital libraries will want to expand their “holdings” by sharing digital links with other libraries. Unfortunately, there seems to bevery little sharing of this sort taking place at present.

How can we understand the unwillingness of libraries to share content? The question goes back to the old measure of the size of a traditional library¾the number of books it holds. When a library expends funds to assemble digitized works, it loses a portion of its prestige (or thinks it

does) by allowing other libraries to copy or access its data. Ultimately, however, *all* material should be accessible from *every* library.

## 1.4 DEFINITION OF DIGITAL LIBRARY

- "Digital libraries are organized collections of digital information. They combine the structuring and gathering of information, which libraries and archives have always done, with the digital representation that computers have made possible" (Michael Lesk).

- "An informal definition of a digital library is a managed collection of information, with associated services, where the information is stored in digital formats and accessible overa network. A crucial part of this definition is that the information is managed. A stream of data sent to earth from a satellite is not a library. The same data, when organized systematically, becomes a digital library collection" (William Arms).

- "A DL contains digital representations of the objects found in it - most understanding of the "DL" probably also assumes that it will be accessible via the Internet, though not necessarilyto everyone. But the idea of digitization is perhaps the only characteristic of a digital libraryon which there is universal agreement" (Harter).

- "In its most basic form a DL should encompass two functions: a) Provide digital content to virtual, geographically dispersed users, and b) Pull in digital information electronically from outside sources irrespective of location."

- Digital library is "a focused collection of digital objects, including text, video, and audio, along with methods for access and retrieval, and for selection, organization, and maintenanceof the collection." (Ian Witten and David Bainbridge).

- "Digital libraries are different [from traditional library automation] in that they are designed to support the creation, maintenance, management, access to, and preservation of digital content." (Bernie Hurley, the Director for Library Technologies at U. C. Berkeley. Quoted in Digital library technology trends. Sun Microsystems. August 2002)

- Sun Microsystems defines a digital library as *the electronic extension of functions users typically perform and the resources they access in a traditional library*. These information resources can be translated into digital form, stored in multimedia repositories, and made available through Web-based services. The emergence of the digital library mirrors the growth of e-learning (or distance learning) as the virtual alternative to traditional school attendance. As the student population increasingly turns to off-campus alternatives for lifelong learning, the library must evolve to fit this new educational paradigm or become obsoleteas students search for other ways to conveniently locate information resources anywhere, anytime.

## 1.5 REQUIREMENT FOR DIGITAL LIBRARIES

The Internet and World Wide Web provide the impetus and technological environment for the development and operation of a digital library. The Internet provides the TCP/IP and or its associated protocol for accessing the information and web provide tools and technique for publishing the information over Internet. In the digital environment it is reasonable to say that a central back up or archive should be created at the national level, which will store information output of the region

as well as information from outside the country. Some of the requirements for digital libraries are:

1. Audio visual: Color T.V., V.C.R., D.V.D., Sound box, Telephone etc.
2. Computer: Server, P.C. with multimedia, U.PS. Etc
3. Network: LAN, MAN, WAN, Internet etc.
4. Printer: Laser printer, Dot matrix, Barcode printer, Digital graphic printer etc
5. Scanner: H.P. Scan jet, flatbed, Sheet feeder, Drum scanner, Slide scanner, Microfilming scanner, Digital camera, Barcode scanner etc
6. Storage devices: Optical storage device, CD-ROM, Jukebox etc.
7. Software: Any suitable software, which is interconnected and suitable for LAN and WANconnection.

## 1.6 FACTORS OF CHANGE TO DIGITAL LIBRARIES

The limited buying power of libraries, complex nature of recent document, storage problem etc are some of the common factor which are influencing to change to digital mode, some other factors are –

1. Information explosion
2. Searching problem in traditional libraries
3. Low cost of technology: When we consider the storage capacity of digital document and its maintained then it can be easily realize that the cost of technologies is much more less than that of traditional libraries.
4. Environmental factor: the use of digital libraries is the cleanest technologies to fulfill the slogan“Burn a CD-ROM save a tree”
5. New generation needs

## 1.7 BENEFITS OF DIGITAL LIBRARIES

Digitalization can offer many advantages to libraries as well as to their users. The benefits mentionedby T.B. Rajashekar are the following:

- Digital libraries make it needless for the user to go somewhere. A user can get full informationat home or at work, whenever there is a PC and a network collection.
- Information can be updated continuously much more easily. It easier to keep the informationcurrent.
- An important benefit offered by digital libraries is searching and browsing in material. One can optimize searching and simultaneously search the Internet, commercial databases, and library collections. Then one can save search results and conduct additional processing to narrow or qualify results, or click through search results to access the digitized contentor locate additional items of interest.
- Information can be shared with others more easily. By placing digital information on a serverconnected to the World Wide Web makes it available to everyone.
- Duplicating of information is easy and cheap, whereas duplication of paper material wouldbe very expensive.

- Digital libraries compared to conventional libraries allow collaboration and exchange of ideas.
- Arising new forms of information: information in digital form can support features and possibilities not given in print form.
- Digital libraries are cost-saving, since expensive building, professional staff and maintenancedemanded by conventional libraries not needed anymore.

***William Arms ([Arms02]) mentioned further benefits:***

- Information access is not limited by geography, it does not matter, where in the world is the document located. There is no need to replicate the document because of its geographic availability.
- The components of digital libraries are declining rapidly in price. Digital libraries are also expensive, but it is expected that digital libraries will become much less expensive than they are now, and much less expensive than the traditional libraries.
- Possible other ways of storing information, like database or mathematics library.
- Extended possibilities for creation of informative objects. “Even when the formats are similar, materials created explicitly for the digital world are not the same as the materials originally designed for paper or other media”.

## 1.8 CHARACTERISTICS OF DIGITAL LIBRARY

Cleveland (1998) describes some characteristics of digital libraries that have been gleaned from various discussions about digital libraries (DLs), both online and in print:

- DLs are the digital face of traditional libraries that include both digital collections and traditional, fixed media collections. So they encompass both electronic and paper materials.
- DLs will also include digital materials that exist outside the physical and administrative bounds of any one digital library
- DLs will include all the processes and services that are the backbone and nervous system of libraries. However, such traditional processes, though forming the basis digital library work, will have to be revised and enhanced to accommodate the differences between newdigital media and traditional fixed media.
- DLs ideally provide a coherent view of all of the information contained within a library, no matter its form or format
- DLs will serve particular communities or constituencies, as traditional libraries do now, though those communities may be widely dispersed throughout the network.
- DLs will require both the skills of librarians and well as those of computer scientists to be viable.

## 1.9 SELF ASSESSMENT QUESTIONS

1. What is digital Library?
2. What are the benefits of DL?
3. Explain the Characteristics of DL.

## 1.10 REFERENCES

1. http://www.ifla.org/files/assets/digital-libraries/documents/ifla-unesco-digital-libraries-manifesto.pdf
2. Jindal, S. C and Vijay Lakshmi (ed.). Digital libraries; New Delhi: Isha Books, 2004.
3. http://www.ncsi.iisc.ernet.in/raja/is214/is214-2006-01-04/topic-1.htm
4. http://www.wtec.org/loyola/digilibs/02_03.htm

5. Rajashekar, T.B.: "Digital Library and Information Services in Enterprises: Their Development and Management". Year 2002. http://144.16.72.189/is214/214-2001-2002/topic-1.htm . Requested on January 10, 2004.
6. Arms, William, Y.: "Digital libraries", 2001.

7. http://www.tlu.ee/~sirvir/Information%20and%20Knowledge%20%20Management/Integration%20of%20digital%20libraries%20in%20e-learning/characteristics_of_digital_libraries.html

LESSON-2

# MERITS, DEMERITS AND CHALLENGES

**OBJECTIVE**

After reading this chapter, students will be able to understand:

- Merits and demerits of digital library
- Challenges of digital library

**STRUCTURE**



## 2.1 INTRODUCTION

A digital library is not confined to a particular location or so called building it is virtually distributed all over the world. The user can get his/ her information on his own computer screen byusing the Internet. Actually it is a network of multimedia system, which provides fingertip access.The spoken words or the graphical display of a digital library is again having a different impact fromthe words that are printed. In the new environment owing a document will not be problem for thelibrary because the user will pay for its uses. The merits of digital libraries as a means of easily andrapidly accessing books, archives and images of various types are now widely recognized by commercial interests and public bodies alike.

Traditional libraries are limited by storage space; digital libraries have the potential to store much more information, simply because digital information requires very little physical space to contain it. As such, the cost of maintaining a digital library is much lower than that of a traditional library.

. A traditional library must spend large sums of money paying for staff, book maintenance,rent, and additional books. Digital libraries may reduce or, in some instances, do away with thesefees. Both types of library require cataloguing input to allow users to locate and retrieve material. Digital libraries may be more willing to adopt innovations in technology providing users with improvements in electronic and audio book technology as well as presenting new forms of communication such as wikis and blogs; conventional libraries may consider that providing online access to their OPAC catalogue is sufficient. An important advantage to digital conversion is increasedaccessibility to users. They also increase availability to individuals who may not be traditional patrons of a library, due to geographic location or organizational affiliation.

## 2.2 NO PHYSICAL BOUNDARY:

The user of a digital library need not to go to the library physically, people from all over the worldcould gain access to the same information, as long as an Internet connection is available.

1. **Round the clock availability**: Digital libraries can be accessed at any time, 24 hours a day and 365 days of the year
2. **Multiple accesses**: The same resources can be used at the same time by a number of users.
3. **Structured approach**: Digital library provides access to much richer content in a more structured manner i.e. we can easily move from the catalog to the particular book then toa particular chapter and so on.
4. **Information retrieval**: The user is able to use any search term bellowing to the word or phrase of the entire collection. Digital library will provide very user friendly interfaces, giving click able access to its resources.
5. **Preservation and conservation**: An exact copy of the original can be made any numberof times without any degradation in quality.
6. **Space**: Whereas traditional libraries are limited by storage space, digital libraries have thepotential to store much more information, simply because digital information requires very little physical space to contain them. When the library had no space for extension digitizationis the only solution.
7. **Networking**: A particular digital library can provide the link to any other resources of otherdigital library very easily thus a seamlessly integrated resource sharing can be achieved.
8. **Cost** - The cost of maintaining a digital library is much lower than that of a traditional library. A traditional library must spend large sums of money paying for staff, book maintains, rent, and additional books. Digital libraries do away with these fees.

## 2.3 DEMERITS OF DLS

The computer viruses, lack of standardization for digitized information, quick degrading properties of digitized material, different display standard of digital product and its associated problem,health hazard nature of the radiation from monitor etc. makes digital libraries at times handicap.

1. **Copyright**: - Digitization violates the copy right law as the thought content of one author can be freely transfer by other without his acknowledgement. So One difficulty to overcome for digital libraries is the way to distribute information. How does a digital library distributeinformation at will while protecting the copyright of the author?

2. **Speed of access**: - As more and more computer are connected to the Internet its speed ofaccess reasonably decreasing. If new technology will not evolve to solve the problem thenin near future Internet will be full of error messages.

3. **Initial cost is high**: - The infrastructure cost of digital library i.e. the cost of hardware, software; leasing communication circuit is generally very high.

4. **Band width**: - Digital library will need high band for transfer of multimedia resources but the band width is decreasing day by day due to its over utilization.

5. **Efficiency**: - With the much larger volume of digital information, finding the right material for a specific task becomes increasingly difficult.

6. **Environment**: - Digital libraries cannot reproduce the environment of a traditional library.Many people also find reading printed material to be easier than reading material on a computer screen.

7. **Preservation**: - Due to technological developments, a digital library can rapidly become out-of-date and its data may become inaccessible.

## 2.4 CHALLENGES OF DLS

The Library of Congress has identified ten challenges that must be met if large and effective digital libraries are to be created during the 21st century. The challenges may be grouped under the following broad categories:

**Building the Resource**

- Develop improved technology for digitizing analog materials.
- Design search and retrieval tools that compensate for abbreviated or incomplete cataloging or descriptive information.
- Design tools that facilitate the enhancement of cataloging or descriptive information by incorporating the contributions of users.

**Interoperability**

- Establish protocols and standards to facilitate the assembly of distributed digital libraries.

**Intellectual Property**

- Address legal concerns associated with access, copying, and dissemination of physical and digital materials.

**Effective Access**

- Integrate access to both digital and physical materials.
- Develop approaches that can present heterogeneous resources in a coherent way.
- Make the National Digital Library useful to different communities of users and for different purposes.
- Provide more efficient and more flexible tools for transforming digital content to suit the needs of end-users.

**Sustaining the Resource**

- Develop economic models for the support of the National Digital Library.

## 2.5 SELF ASSESSMENT QUESTIONS

1. Write the merits and demerits of DLs.
2. Enumerate the major challenges of Digital Library?

## 2.6 REFERENCES

1. Amavizca, M., Sánchez, J. A., & Abascal, R. (1999). 3DTree: Visualization of large and complex information spaces in the Floristic Digital Library, Proceedings of Segundo Encuentro de Computación (ENC'99). Pachuca, Hidalgo, México.
2. Arms, W. (1999). Report of the NSF Science, Mathematics, Engineering, and Technology Education Library Workshop, July 21-23, 1998 (NSF 99-112): National Science Foundation,Division of Undergraduate Education. [Online at]: http://www.dlib.org/smete/public/report.html

3. Arms, W. (2000a). D-Lib Magazine. {Online at]: http://www.dlib.org
4. Arms, W. Y. (2000b). *Digital Libraries*. Cambridge, MA: MIT Press.
5. Baldonado, M., Chang, C.-C. K., Gravano, L., & Paepcke, A. (1997). The Stanford DigitalLibrary Metadata Architecture. International Journal on Digital Libraries, 1(2), 108-121.
6. Barceinas, A., Sánchez, J. A., & Schnase, J. L. (1998). MICK: A KQML inter-agent communication framework in a digital library, Memorias del Simposium Internacional de Computación (CIC'98) (pp. 66-79). Mexico City.
7. Belkin, N. J., Kantor, P., Fox, E. A., & Shaw, J. A. (1995). Combining the Evidence of MultipleQuery Representations for Information Retrieval. Information Processing and Management,31(3), 431-448.
8. Birmingham, W. (1995). University of Michigan Digital Library Project. [Homepage online at]: http://http2.sils.umich.edu/UMDL/
9. Borgman, C., & Fox, E. A. (2001). Proceedings of the First Joint Conference on Digital Libraries. June 24-28, 2001. New York: ACM Press.[Online at]: http://www.jcdl.org
10. Borgman, C. L. (1999). What are digital libraries? Competing visions. Information Processing and Management, 35(3), 227-243.
11. Bush, V. (1945). As We May Think. Atlantic Monthly,176, 101-108.
12. Can, F., Fox, E., Snavely, C., & France, R. (1995). Incremental Clustering for Very Large Document Databases: Initial MARIAN Experience. Information Systems, 84, 101-114.
13. Cassel, L., & Fox, E. A. (2000). ACM Journal of Education Resources in Computing. [Online at]: http://purl.org/net/JERIC/
14. Chen, S., & Fox, E. A. (1996). Guest Editors' Introduction to Special Issue on Digital Libraries. Journal of Visual Communication and Image Representation, 7, 1.
15. Das Neves, F. A., & Fox, E. A. (2000). A study of user behavior in an immersive Virtual Environment for digital libraries, Proceedings of the Fifth ACM Conference on Digital Libraries: DL '00, June 2-7, 2000, San Antonio, TX (pp. 103-111). New York: ACM Press.
16. Dublin Core Community. (1999). Dublin Core Metadata Initiative. [Online at]: http://purl.org/dc
17. Dumais, S. T. (1998). References to LSI Papers. [Online at]: http://superbook.bellcore.com/ ~std/lsiPapers.html
18. Fox, E. A. (1998b). Digital Library Courseware. Virginia Tech, Department of Computer Science: Blacksburg, VA. [Online at]: http://ei.cs.vt.edu/~dlib/
19. Lagoze, C., & Davis, J. R. (1995). Dienst: An Architecture for Distributed Document Libraries. Communications of the ACM, 38, 47.
20. Lagoze, C., & Payette, S. (1998). An Infrastructure for Open-Architecture Digital Libraries (TR98-1690): Cornell University, Computer Science.
21. Lesk, M. (1997). Practical Digital Libraries: Books, Bytes and Bucks. San Francisco: Morgan Kaufmann Publishers.
22. Licklider, J. C. R. (1965). Libraries of the Future. Cambridge, MA: MIT Press.
23. http://liswiki.org/wiki/Digital_library
24. http://memory.loc.gov/ammem/dli2/html/cbedl.html
25. https://repository.unm.edu/bitstream/handle/1928/1711/chapter4.html?sequence=7

## LESSON - 3

# DIGITAL LIBRARY MANAGEMENT

### OBJECTIVE

After reading this chapter, students will be able to understand:

- Fundamental concepts of digital library management
- Principles underlying the functionality of digital libraries
- Core competencies expected of digital librarians

### STRUCTURE

## 3.1 INTRODUCTION

It goes without saying that, in modern times, the development and proliferation of digital libraries is giving rise to momentous transformations in the generation, access, utilization, dissemination and also the management of information resources. The introduction of a novel technology, such as digitization of information resources, tends to warrant a number of training requirements for the earmarked staff, due to demands for effective and efficient management of the new technology. Training needs may arise either directly from the knowledge, skills and attitudes needed to operate the new technology or from the 'spin-off' effects that may have changed working practices and influenced social interactions and relationships. Thus a need to change the management style and specific roles within the organizational structure would be necessary, in order to promote the openness and speed that normally accompanies such new technologies, of which digital libraries are included.

## 3.2 DEFINITION OF DIGITAL LIBRARIES

Over a period of time, many proponents have forwarded various definitions of digital libraries and still more continue to emerge each day. According to Waters (1998) the partner institutions in the Digital Library Federation (DLF) realized in the course of developing their program that they needed a common understanding of what digital libraries are if they were to achieve the goal of effectively federating them. So they crafted the following definition, with the understanding that itmight well undergo revision as they worked together:

*"Digital libraries are organizations that provide the resources, including the specialized staff, to select, structure, offer intellectual access to, interpret, distribute, preserve the integrity of, and ensure the persistence over time of collections of digital works so that they are readily and economically available for use by a defined community or set of communities".*

Arms (2000, 2) defined the digital library as:
*"A managed collection of information with associated services, where the information is stored indigital formats and is accessible over a network".*

** Note the crucial part of the aforementioned definition, which recognizes that fact that the information is managed.

Bawden and Rowlands (1999), on the other hand, defined the digital library as:

*"...a library / information space, located in either a physical or virtual space, or a combination of both, in which a significant proportion of the resources available to users of that service exist onlyin digital form".*

However, Chowdhury and Chowdhury (2003, 8) recommended that the definition given by Gladney et al. (1994) was the most comprehensive definition of the digital library.

The definition, which came up in the course of the 1994 Institute of Electrical and Electronics Engineers Conference on Artificial Intelligence Applications (IEEE CAIA), emphasizes both the technical and service oriented aspects of digital libraries. It states as follows:

*"A digital library is an assemblage of digital computing, storage and communication machinery together with the content and software needed to reproduce, emulate and extend the services provided by conventional libraries based on paper and other material means of collecting, cataloguing, finding and disseminating information. A full service digital library must accomplish*

*all essential services of traditional libraries and also exploit the well-known advantages of digital storage,searching and communication".*

## 3.3 PRINCIPLES UNDERLYING THE FUNCTIONALITY OF DIGITAL LIBRARIES

According to Lesk (1997, 1-2), digital libraries apparently give their users powers that they did not previously possess in as far as traditional libraries are concerned. With digital libraries, it is now possible for one to search for any word or phrase and it can be accessed over the world and reproduced without error! Digital libraries address traditional problems of finding information, of delivering it to users and of preserving it for posterity. In so doing, digital information takes less space than paper-based information and thus may help libraries to reduce cost.

Deegan and Tanner (2002, 22) proposed some general principles characterizing digital libraries. They comprise the following:

i. Digital libraries are managed collections of digital objects;
ii. Digital objects are created or collected according to principles of collection development;
iii. Digital objects are made available in a cohesive manner, supported by services that are necessaryto allow users to retrieve and exploit the resources just as they would any;
iv. Digital objects are treated as long-term stable resources and appropriate processes are appliedto ensure their quality and survivability.

***So, what does it take to establish a fully functional digital library?***

### *3.3.1 Knowledgeable staff:*

Any organization that intends to establish a digital library must have a sustainable arsenal of suitably knowledgeable and skilled staff. The specific details of their knowledge and skills will be duly expounded in later section of this paper. However it suffices to say that they must be hardworking and committed individuals, loyal to their parent organization, ready and eager to continuously learn new activities pertaining to information technology with particular emphasis on digital libraries andmust have the tenacity to apply whatever they learn in their workplace. Above all, they must be compassionate and extremely sensitive to the information needs of the clientele they serve.

### *3.3.2 Provision of digital content:*

The digital library must contain information resources. It may either be new material prepared digitally from scratch (i.e. born digital), or it may be old material, converted into digital form (i.e. digitized). It may be bought, donated or converted locally from previously purchased library stock. Digital content then needs to be stored and retrieved. Information is widely found as text stored as characters and images acquired using optical scanners. These images are frequently scans of printed pages, as well as illustrations or photographs. More recently, audio and video, plus interactive material are accumulating rapidly in the digital form both newly generated and converted from older material. Copyright aspects also have to be carefully considered, at this stage, and everything has to be carried out without contravening the existing laws on fair use of information resources, in this regard.

### *3.3.3 Searching of digital information resources:*

After storing information in a digital library, mechanisms ought to be in place for one to accurately identify and locate the piece of information sought. According to Wikipedia (April 2006),most digital libraries provide a search interface, which allows information resources to be found. These resources are typically deep Web (or invisible Web) resources, since search engine crawlers cannot locate them. Some digital libraries create special pages or sitemaps to allow search enginesto find all their resources. Digital libraries frequently use a protocol developed by the Open Access Initiative namely, the Open Access Initiative Protocol for Metadata Harvesting (OAI-PMH) to exposetheir metadata to other digital libraries and search engines like Google can also use the same tofind the deep Web resources. There are two general strategies for searching a federation of digitallibraries:

(a) Distributed searching, and
(b) Searching previously harvested metadata.

Distributed searching typically involves a client sending multiple search requests in parallel to a number of servers in the federation. The results are gathered, duplicates either eliminated or clustered, and the remaining items sorted and presented back to the client. Scalability (the capability of the system to increase total throughput under increased load emanating from added resources, typically hardware) and performance issues tend to plague distributed searching for large federations of digital libraries. Protocols like Z39.50 (a client server protocol for searching and retrieving information from remote computer databases) are frequently used in distributed searching. Searching over previously harvested metadata requires the pooling of metadata collected from every digital library in the federation. This solution scales better than distributed search, but it introduces the problem of data freshness; digital libraries need to be re-harvested on a periodic basis to discover new and updated resources. OAI-PMH is frequently used by digital libraries for harvesting metadata.

### *3.3.4 Retrieval of digital information resources:*

Retrieval systems are necessary for users to obtain the information they require from the digital libraries. Though it is relatively straightforward in the case of textual information resources,it may be a subject of research for pictures sounds and videos. Whatever the case, retrieved information must be delivered to the user. A digital library must also have some form of preservation mechanism. In that way, there has to be a means of ensuring that what may be available today isstill available tomorrow.

### *3.3.5 Interoperability:*

Digital libraries are at the heart of interrelationships between several information service-related disciplines like library management, archives management, museum management, document management, knowledge management and e- commerce systems. This brings about the need for the different kind of systems to talk to each other.

Miller (2000) defined the term ‘interoperability’ as:
*“... the ongoing process of ensuring that the systems, procedures and culture of an organization are managed in such a way as to maximize opportunities for exchange and reuse of information, whether internally or externally”.*

Interoperability, in digital libraries, allows organizations and communities to retain their specialist practices, while putting high-level standards and protocols in place for sharing information. Achieving interoperability is difficult as it requires resources creators, users, funding agencies, systems and resource managers to agree on the development of standards and formats for information interchange that may not map exactly onto their established practices.

### *3.3.6 Sustainable funding:*

Digital librarians need to establish the financial wherewithal to pay for and sustain digital libraries. One can unequivocally declare that finding a way to fund digital libraries is the single most frustrating obstacle facing librarians in the present day. Digital libraries are bound to change the social system by which information is collected, transformed, managed, disseminated and preserved, both in the present day as well as in the future. Libraries will most certainly be at the helm of actualizing that phenomenon and therefore it is imperative that they strive to build sufficient capacity, in terms of financial backing. This means that digital librarians must also be reasonably adept in the dual skills of fund-raising and fund management, if at all their libraries are to survive in the medium and long term.

## 3.4 RATIONALE FOR DEVELOPING DIGITAL LIBRARIES

The growing impact and subsequent user awareness of information and communication technologies (ICTs), Web technologies and database technologies have compelled libraries to resort to digital library technology in order to render more effective information services. With the exponential growth of e-resources, it has become imperative for library and information professionalsto redefine their function of disseminating information to the respective users (Ongus and Nyamboga, 2004). The following are some of the encouraging factors that have nudged the information societytowards developing of digital libraries, as purported by Hariharan et al. (2002) and also Deegan andTanner (2002, 30-35):

- Information can be saved digitally and therefore renders immediate access to high demandand frequently used items;
- There is enhanced intellectual control along with new finding tools and support searchingcapabilities;
- Links are provided to access bibliographical tools;
- Manipulation of text and images is improved by enabling the enhancement of digital images interms of size, sharpness, colour, contrast noise reduction and so on;
- Duplication of digital resources is easy;
- The new potential of conserving fragile/precious originals while presenting surrogates in moreaccessible forms, is provided;
- The potential for integration into teaching materials by presenting the critical mass of readingmaterials is provided;
- The burden or cost of delivery is reduced;
- There is encouragement of use by providing enhanced resources in the form of widespread dissemination of unique collections;
- There is ‘virtual reunification’, allowing dispersed collections or materials that are related to oneanother to be brought together, even if they are scattered among many locations;
- More than one user can make use of use a single information resource, simultaneously;
- Provides timely access;
- Saves physical storage space;
- Capable of supporting and creating multimedia information resources thus allowing thesimultaneous integration of different media (i.e. images, graphics, sounds, videos and so on);
- No mutilation of pages due to high use or otherwise can occur;
- There is an increased use of library resources and Web-based resources through Internet or Intranet, making it easy to receive or transfer information both from as well as to any part of

theInternet, instantly;

- Supports resource sharing among libraries by providing efficient and seamless access to materials held remotely;
- There is the capability of keeping an electronic archive/history of resources previously accessed.
- There is the possibility of several libraries forming a consortium or consortia of access to bibliographic databases, abstracts, full text journals and even e-books online, by spending only a nominal amount.

As the case with everything in life, disadvantages always accompany advantages of any given issue. The following can be listed as the setbacks that are likely arise when creating digital libraries:

- The initial cost of digitization and preservation files is prohibitive. This comes about due to therelatively high cost of purchasing the required hardware and software for setting up the library;
- Special training is required and thus special skills to set up and maintain the digital library;
- The user has to accept the media, thereby making user sensitization a crucial factor to be considered. This is because the authenticity and credibility, hence acceptance of the digitizedinformation resources may have a lot to be desired;
- Bandwidth problem in accessing multimedia resources and full-text journals is a major communication barrier (particularly in a large majority of the third world countries);
- Scanning and electronically storing the original documents of the entire paper based collectionis time consuming and labour intensive;
- Intellectual Property Rights (IPR) issues maynot be clearly interpreted or correctly appliedand enforced in different parts of the world;
- Some librarians are wary of the new technology and hence may be reluctant to adopt changes.

## 3.5 NEED FOR SPECIALIZED STAFF TO MANAGE DIGITAL LIBRARIES

As previously mentioned, the most crucial component of any digital library is its staff. Although the endeavor to build a team of knowledgeable and skilled staff who are capable of managing a successful digital library may be a one-time investment, it is bound to be a time consuming project. It has become more essential than ever that librarians understand the general principles of creating and managing Web content, for instance. As digital gatekeepers, the librarians' expertise must match or even surpass those possessed by the user. We live in an age whereby information usersnot only have the knowledge but the capacity of generating information on their own, as well. Therefore there has to be a unique set of professionals who are specially trained to distinguish between information that is palatable to any given set of users, from that which is not. Librarians fit this role perfectly and being information gatekeepers and gateways, they already have the know-how of matching user needs with information resources, predominantly in traditional libraries. These are the people on whom initial attention should be concentrated, by being provided with additional tailor-made training to effectively transform them into digital librarians. Along with that, schools oflibrary and information science should be proactive enough to include the digital libraries module intheir respective curricula. This should be done with the view of churning out generations of graduates who are technologically savvy and have the capability to rise up to the occasion when called upon to do so.

Specifically, digital librarians are required for the purposes outlined below, among a host of emergingfunctions:

- To manage digital libraries;
- To organize the digital knowledge and information resources;
- To disseminate digital information from computer-held digital information;
- Provide digital reference services and other electronic information services;
- To provide knowledge mining from the emerging knowledge warehouses;
- To handle the tasks of mass digitization, digital storage process and digital preservation;
- To provide universal access and retrieval of digital knowledge, ultimately access to all knowledgeresources available in digital form;
- To catalogue and classify digital objects and digital knowledge.

Admittedly, the contemporary information professional has numerous challenges to contend with, in the pursuit of satisfying the ever-changing information requirements of his/her clientele. The modern emphasis is on value-for-money concepts, accuracy, and timelines in information provision, among many other issues. In another school of thought, there is a belief that the information professional who manages a digital library is, in fact a knowledge manager. In that regard, Davenport, Delong and Beers (1998) highlighted that knowledge management is carried out with a view to:

- Creating knowledge repositories;
- Improving access to knowledge;
- Creating a knowledge environment;
- Managing knowledge as an asset.

It is worth noting that in the developed world, suitably skilled professionals have attained the critical mass of expertise in constructing, stocking, managing, maintaining, evaluating and upgrading digital libraries. However the truth of the matter is that the very opposite scenario is apparent in most of the developing nations. Not much guidance, facilitation and empowerment seems to be forthcoming to the library and information service professionals in that part of the world. They are therefore rendered incapable of replicating such type of 'hi-tech' services in their own respective countries, to suit their own user populace. At best they may develop hybrid libraries, but even these are marred with problems such as underdeveloped publishing culture among the local intellectuals, lack of sufficient and sustainable funding, poor telecommunication infrastructure, unstable electrical power supply, untrained library staff, reluctant management, uninformed users as well as general lack of political goodwill, just to mention but a few. All these factors have inadvertently aggravated the broadening of the all too familiar chasm, nowadays commonly referred to as the 'digital divide'.

## 3.6 CORE COMPETENCIES EXPECTED OF DIGITAL LIBRARIANS

In the digital era, 'information objects' like books, journals, newspapers, electronic documents, images, multimedia packages, databases and so on, can be accessed in diverse ways. According to Choudhary and Chand (2002), a barrage of threats beleaguers libraries and information professionals, of which include the overwhelming fast pace of technological innovations.The threats are quite daunting and real. Subsequently, as Sreenivasulu (2000) observed, the competency of a digital librarian is represented by different sets of skills, attitudes and values thatenable him/her to work as a digital information professional or, if one may prefer, a digital knowledge worker and digital communicator. He proposed the following skills and competencies that a properly qualified

digital librarian ought to demonstrate:

• A good command of various Internet skills beyond the level of an ordinary user, including online searching and Web publishing;
• Knowledge and command of multimedia, digital technology and digital media processing techniques;

• Knowledge and skills of handling digital information systems, online and optical information involvingthe management of a CD-ROM / DVD-ROM network station and the conversion of print media intodigital media;

• Networking knowledge of using both internal and external networks, including the establishment of personal networks, intranets, external knowledge resources and extranets.

Tennant (1999) also identified several skills that are supposed to personify the digital librarian's knowledge reservoir. Some of them inevitably coincide with the ones outlined above.

### *3.6.1 Optical Character Recognition (OCR)*

Scanning a printed will capture an image but in order to make it searchable, a good knowledgeof OCR technology is required.

### *3.6.2 Imaging technologies*

Digital librarians must be aware of the various ways in which surrogates of physical items (for example, journal articles) can be captured. They must be familiar with the typical manipulationrequired to edit and save it different formats;

### *3.6.3 Markup language*

Digital librarians should have the knowledge of Hypertext Markup Language (HTML) and also a suitable combination of other Web authoring tools (such as SGML, XML, Scripting languages
e.g. JavaScript or VBScript, DreamWeaver, Macromedia Flash and so on);

### *3.6.4 Cataloguing and metadata*

Digital objects require organization and description. Digital librarians must understand the ways in which metadata can be captured. They should be familiar with standards such as Machine Readable Catalogue (MARC), Anglo American Cataloguing Rules II (AACR II), Z39.50 protocol,Dublin Core and so on;

### *3.6.5 Indexing and database technology*

Digital librarians must be familiar with a variety of tools from simple and easy indexing and searching tools to complex relational or object oriented database systems;

### *3.6.6 User interface design*

The digital librarian should be able to write the functional specifications and work with other knowledgeable professionals to achieve the desired goal of developing a user-friendly computerinterface with the library automation system, in case the library has one;

### *3.6.7 Programming*

Digital librarians need not be full-time programmers, but it would be an added advantage if they were familiar with programming languages such as C, C++ or Java. Knowledge of handling open source software such as Dspace or Greenstone Digital Library would definitely come inhandy;

### *3.6.8 Web technology*

Digital librarians must know their way around the Internet and be well versed in Web technology;

### *3.6.9 Project management*

Digital library projects need skilled management. Digital librarians should be good communicators and relate well with people both inside as well as outside the organization. Projects initiated need to be completed on time and within the stipulated budget.

## 3.7 SELF ASSESSMENT QUESTIONS

1. What is digital library management?
2. What are the core competencies of digital librarian?

## 3.8 REFERENCES

1. Arms, W. (2000). Digital libraries. Cambridge, MA: MIT Press, p.2.
2. Bawden, David and Rowlands, Ian. (1999). Understanding digital libraries: towards a conceptual framework. (British Library Research and Innovation report No. 170). London:British Library Research and Innovation Centre.
3. Choudhary, Pravin Kumar and Chand, Prakash. (2002). "Challenges for LIS professionals in the digital era". In Library and Information Networking: Papers of the National Convention on Library and Information Networking (NACLIN), held at Cochin University of Science andTechnology, Cochin (India), October 21-24, 2002. ed. by H. K. Kaul and M. D. Baby. NewDelhi: DELNET Developing Library Network, pp. 254-267.
4. Chowdhury, G. G. and Chowdhury, Sudatta. (2003). Introduction to digital libraries. London:Facet Publishing. p. 8. and pp. 285-286.
5. Davenport, T. H.; Delong, D. W. and Beers, M. C. (1998). "Successful knowledge management projects". Sloan Management Review. Vol. 39 (No. 2), pp.43-57.
6. Deegan, Marilyn and Tanner, Simon. (2002). Digital futures: strategies for the informationage. London: Library Association Publishing, p.22, pp30-35 and p.139.
7. Gladney, H. M. et al. (1994). Digital library: gross structure and requirements: reports from
a March 1994 workshop. (Accessed on 21st April, 2006) URL: http:// www.csdl.tamu.edu/ DL94/paper/fox.html
8. Hariharan, Chitra M. et al. (2002). "Developing a digital library in civil and structural engineering R&D institutions". In Library and Information Networking: Papers of the NationalConvention on Library and Information Networking (NACLIN), held at Cochin University of Science and Technology, Cochin (India), October 21-24, 2002. ed. by H. K. Kaul and M. D.Baby. New Delhi: DELNET-Developing Library Network, pp. 68-87.
9. Hastings, K and Tennant, R. (1996). "How to build a digital librarian". D-Lib Magazine. November 1996. (Accessed on 24th April 2006). URL: http://www.dlib.org/dlib/november96/ucb/11hastings.html

LESSON -4

# DESIGN AND ORGANIZATION OF DIGITAL LIBRARY- ARCHITECTURE

## OBJECTIVE

After reading this chapter, students will be able to understand:

- Basic fundamental concepts of DL architecture
- Designing principles of digital library

## STRUCTURE



## 4.1 INTRODUCTION

The concept of library is going through a revolutionary phase due to the proliferation of electronic resources. Our methods of producing, organizing and seeking information have changed

drastically with the usage of computers and databases, but our problems have not been solved. The library professional had never been exposed in the past to the changing information scenarioas they are being exposed now. Information explosion and the development of technologies and its progress are changing the previous methods of document collection, storage and dissemination. Now librarians have to face the challenge of this changing scene, otherwise they can be replaced by those, who are able to disseminate the information through CD networks, digital libraries, electronic publishing and Internet etc. So the librarians well have to fulfill this obligation as well. The level of interest regarding digital libraries has grown steadily as a greater number of institutions, including archives and Museums consider the possible implication of digital libraries while there are important unresolved digital library research and development issue, there is also a concurrentdesire to develop strategies for systematic digital library programs built upon the result of digital library Project.

## 4.2 DESIGNING PRINCIPLES OF DL

The purpose of a digital library is to provide coherent organisation and convenient access to typically large amounts of digital information. The following 10 principles are helps to design andcontinued development of any digital library system [2].

- Expect change
- Know your content
- Involve the right people
- Design usable systems
- Ensure open access
- Be aware of data rights
- Automate whenever possible
- Adopt and adhere to standards
- Ensure quality
- Be concerned about persistence

### 4.2.1 EXPECT CHANGE

It may not be apparent why the changing technology landscape is such a thorny problem for digital library projects. Consider, for example, a conversion project in which documents are converted to some digital format. If the chosen format is part of a proprietary system, viewable only through a proprietary interface, when the company that markets the interface no longer supports the system and format, the digitised documents are all but lost. Consider, too, a scenario in which a document is created in a particular word processing programme and the document is attachedto an email message sent to a notable person. Suppose the goal is to preserve all of that person'semail messages for future generations. We are all too aware of our dependence on our email technology for reading such attachments. Imagine what today's platform limitations will mean to future generations, when the content of the attachments is likely no longer accessible.

### 4.2.2 KNOW YOUR CONTENT

For users, content is the most interesting and valuable aspect of a digital library. Creators of digital libraries need to manage and make decisions about their content, including selecting the objects to be included, digitising items that exist only in analog form, possibly marking-up items using standard languages like the Standard Generalised Markup Language (SGML), and assigning metadata describing the content and other attributes of each object.

### 4.2.3 INVOLVE THE RIGHT PEOPLE

Ideally, individuals from a variety of backgrounds and offering a variety of expertise contribute to building a digital library. In practice, this may not be the case, but even when it's not, knowing that building the system requires insight from a number of fields yields a better digital library.

The two fields involved most directly are computer science and library science. Computer scientists appreciate the possibilities, as well as the limitations, of technology and are generally the ones who actually build the system. Librarians, including catalogers, indexers, and archivists, have long been the custodians of information resources, understanding not only the information needs of diverse audiences but the issues involved in preserving materials for continued access and use. Digital library research and development have meant that each group has had to understand the other groups' perspectives.

### 4.2.4 DESIGN USABLE SYSTEMS

Most digital libraries are made available over the Internet through Web technology, though, strictly speaking, this is not a necessary attribute of a digital library. However, as the advantages of the Web are so great, most library systems today are designed to be Web-accessible. The most successful Web site designs account for a number of factors, including the technical differences among computers and browsers, including speed of access, and differences among users, including Web navigation preferences. Browsers differ in the way they display information, even though they use the same basic communication protocols (such as the Hypertext Transfer Protocol, or HTTP, and File Transfer Protocol, or FTP) and standard markup languages (such as HTML and perhaps the Extensible Markup Language, or XML)[3]. Since users may change default settings, including font size and other parameters, it is always preferable to create as simple an interface as possible and avoid server-side control of the exact display of the data. Providing multiple access points not only makes a digital library more interesting, it also acknowledges the differences among its potential users.

### 4.2.5 ENSURE OPEN ACCESS

Ensuring open access is closely related to usability concerns, including access to the information in the digital library, as well as to the digital library itself. Christine Borgman defines access to information as "connectivity to a computer network and to available content, such that the technology is usable, the user has the requisite skills and knowledge, and the content itself is in a usable and useful form". Michael Lesk writes that open access to information raises a number of public policy issues, including whether or not all segments of society are given equal access to information.

### 4.2.6 BE AWARE OF DATA RIGHTS

A possible threat to open access to information arises because of intellectual property concerns. Existing intellectual property and copyright law provides economic and legal protection to publishers of physical artifacts. "Fair use" (allowing libraries to make, say, single copies of portions of books or journals) and "first-sale" rights (allowing individuals to, say, lend or resell copies of books they have purchased) have promoted greater access to physical artifacts than might be possible otherwise, but these notions are only indirectly applicable to networked information.

### 4.2.7 AUTOMATE WHENEVER POSSIBLE

Because building a digital library requires significant intellectual effort on the part of the system's creators, the more automated tools that can be built and used, the better will be the useof precious human resources. These tools need to be easy to use and incorporate real-time aids, including data validation, pull-down lists, report generation, and other time-saving devices, thereby allowing the content expert to concentrate on the intellectual tasks at hand. Content experts use the metadata entry system to add metadata to a master database, entering the information only once. Subsequently, the information is extracted and combined as needed from the master databaseto generate HTML pages, search indexes, and reports. Entering the data only once saves humantime and effort, reduces the error rate, and allows maximum flexibility. Nearly the entire Web interfaceis generated from the database, allowing regeneration whenever necessary, while adhering to the latest Web standards. The system is designed to be modular, allowing existing modules to be modified easily and new modules added for additional functionality.

### 4.2.8 ADOPT AND ADHERE TO STANDARDS

The use of standards in system building has many benefits. Applications are more readily scalable, interoperable, and portable; these characteristics are all important for the design, implementation, and maintenance of digital libraries. Using standards is especially important for the aspects of digital libraries that are most labor-intensive. Scanning, metadata entry, and document markup, all involving the evaluation and handling of individual items in a collection, are resource-intensive and best done carefully and only once. Data might still have to be migrated to other formsand formats in the future, but migration will be easier, because standards have been used consistently.

### 4.2.9 ENSURE QUALITY

Quality can be applied to all the processes and outcomes involved in creating a digital library. They are relevant to selection, metadata entry, image capture, and the overall usability of the system. Complete and correct metadata yields many benefits; incomplete or incorrect metadata affects the quality of the entire digital library. Metadata plays a vital role not only in resource discovery but in managing the collection. If, for example, subject codes are applied haphazardly or incorrectly, access could be more difficult, and attempts to generate browse hierarchies based on these codes could be foiled.

### 4.2.10 BE CONCERNED ABOUT PERSISTENCE

Digital Preservation is gaining importance today. At present there is no way to guarantee the preservation of digital information. While preservation has long been a concern of archives and libraries, it has only recently been of interest to a much larger community.

## 4.3 ARCHITECTURE OF DIGITAL LIBRARY

The architecture of the digital library as describe by Kahn and Wilensky, specifies those characteristics that apply to all types of material. To example object needs to save a name or identifier. Names are a vital building block for the digital library. Names are needed to identify digital objected, to register intellectual property in digital objects, and to record changes of ownership. They are required for citation for information retrieval and are used for links between

objects. These names must be unique. This requires a administrative system to decide who can assign them and change the objects that they identify. They must last for very long time periods, which exclude the use of an identifier tied to a specific location, such as the name of a computer. Names must persist even if the organization that named an object no longer exists when the objects is used. There need to be computer systems to resolve the name rapidly, by providing the location where an object with a given name is stored.

The corporation for National Research Initiatives has implemented a handle system which satisfies these requirements. A "handle" is a unique string used to identify digital objects. The handle is independent of the location where the digital object is stored and can remain valid over very long periods of time. A global server provides a definitive resource for legal and archival purpose, with a caching server for fast resolution. The computer system checks that new names are indeed unique, and supports standard user interfaces, such as Magic. A local handle server is being added for increased local control.

## 4.4 THE ARCHITECTURE OF DIGITAL LIBRARIES

- The architecture of a digital library is made up of four components: *user interface*, *repository*, *handle system* and *search system* (Figure 2)
- The user interface has two parts
- a standard Web browser for the interaction between the user and the library
- client services providing intermediary functions between the browser and the other partsof the library (e.g. deciding where to search, interpreting information structured as digital objects, managing relationships between digital objects and converting among the protocols used by the various parts of the system)
- The repository stores and manages digital objects and associated information
- a digital library may have different types of repositories: modern repository, legacy databases and Web servers
- The handle system provides a distributed directory service for handles of digital library resources
- input to the handle system is the handle (identifier) of a digital object of interest
- output of the handle system is the identifiers of the repositories where the digital object of interest is stored
- The search system houses various indexes and catalogs that can be searched in order to discover information before retrieving it from a repository
- searching is carried out by specially designed Web-based retrieval systems that are capable of accessing and retrieving digital objects across distributed repositories
- distributed searching involves *federating* (i.e. mapping together) similar digital objects from different sources in a way that makes them appear as one organized collection

### 4.4.1 DIGITAL LIBRARY ARCHITECTURE

In the following paragraphs an architectural approach to the digital library will be developed, which is based on taking the fundamental capabilities, introduced above, as the fundamental requirements the architecture must satisfy. I begin with the following notional architecture for digitallibraries:

- A digital library approach to information management depends fundamentally upon a distinction between data and metadata. Metadata provide external classifying and organizingrelations for data that may be unstructured, complex, or very large.

- Middleware services such as search, asset protection, and retrieval processes depend on metadata. Since metadata refers to data, which may be stored in separate hierarchical storage subsystems, integrity of reference must be maintained between metadata and data.

### 4.4.2 OPERATIONAL ARCHITECTURE

Operational architecture is an information management system represented in terms of the business processes it supports, and how information related to conduct of the business processes passes through the system's components.

The example shown in Figure 4.2 is an enterprise that conducts training by utilizing an extensive computer-based simulation system. The operational (business) processes, most obviousin the example, depend on the timely and well-organized capture of training information as it happens, and both contemporaneous and retrospective search and retrieval of information from a training event. Although the information is generated in several different enterprise domains (eight in the example), effective utilization of information often depends on cross-domain searches and retrievals. Therefore, digital library services must provide information interoperability in middleware.

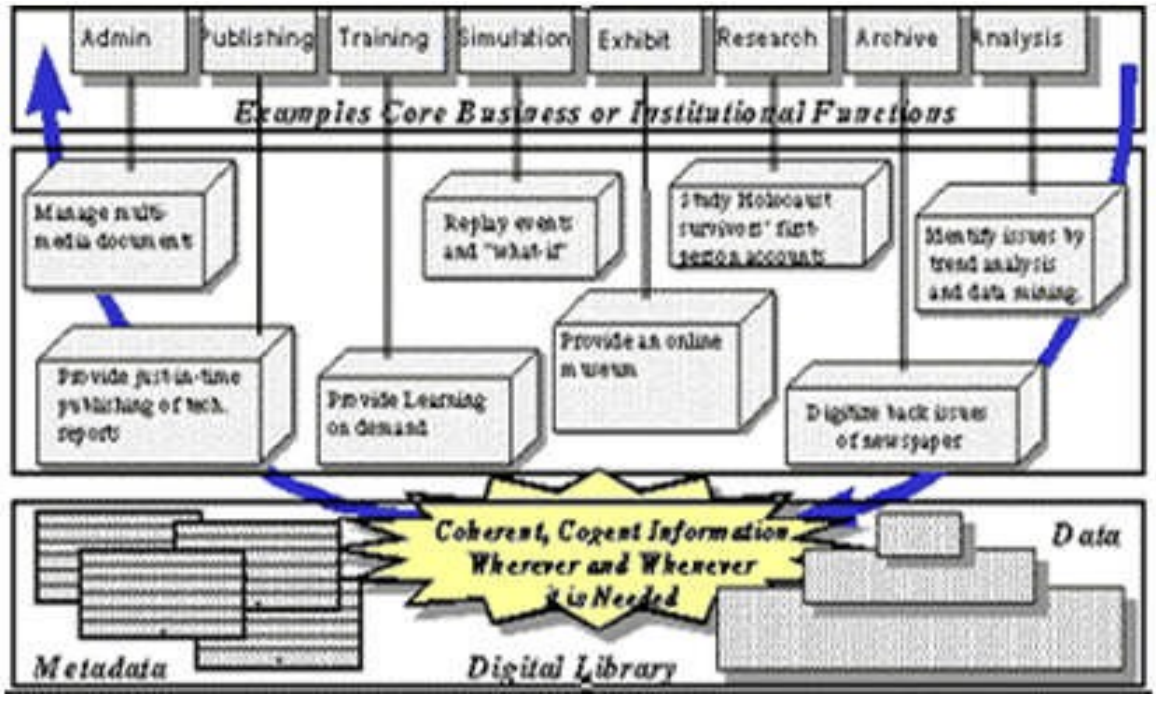


**Fig. 4.1 Notational architecture-building blocks of information to enhance existing functions and enable new operational capabilities.**

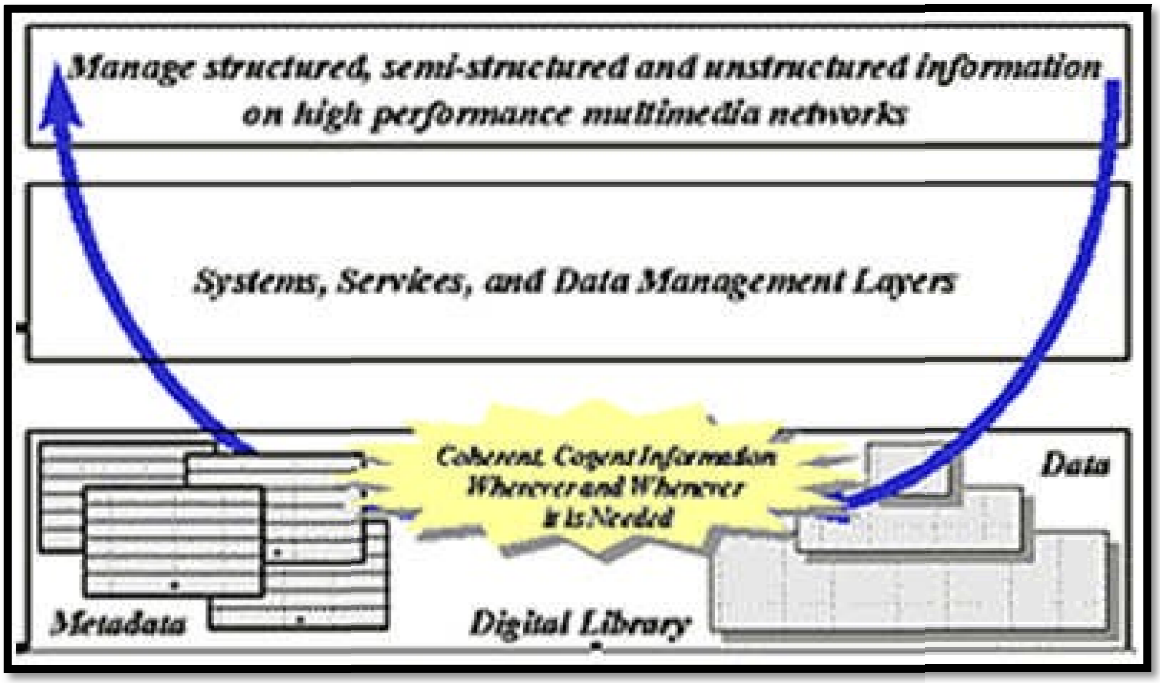


**Fig. 4.2 Operational Architecture**

### 4.4.3 TECHNICAL ARCHITECTURE

A technical architecture breaks down operational (business) processes into functional components and capabilities (Figure 4.3). Hardware and software implementations are still not resolved.

The utilization of digital library materials depends on the existence of metadata to give an efficient and accurate view of content. Metadata must be created as content is added to the digital library. Metadata and data must be bound together logically, and there must be a robust underlying technology to manage the logical connection through time, across platforms, and over geographical separations, all on a networked, distributed system.

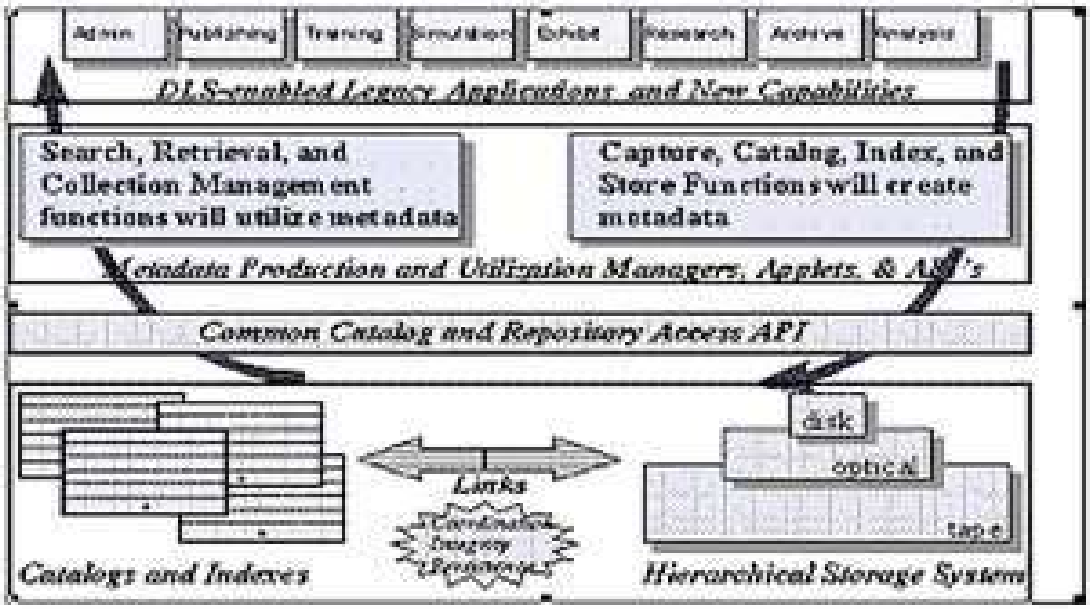


**Fig. 4.3 Technical architecture**

### 4.4.4 SYSTEMS ARCHITECTURE

Systems architecture shows the technology enablers and their inter-relationships. In Figure 4.4, the digital library is a centralized subsystem that interacts with a variety of data producers and consumers within a complex distributed system.

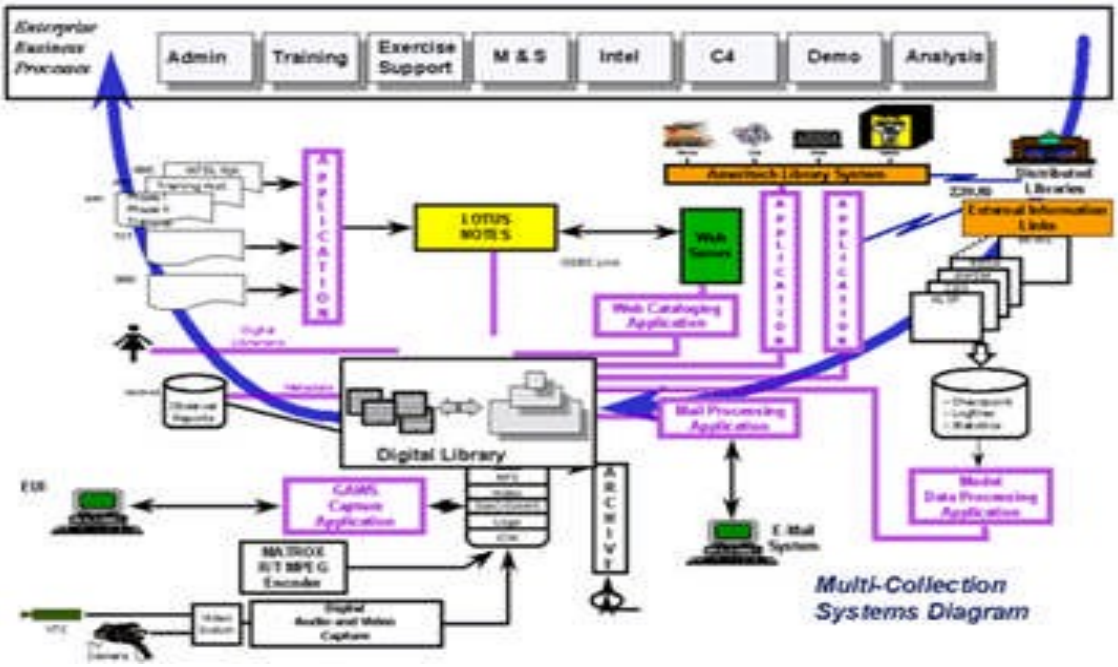


**Fig. 4.4 Systems architecture**

Fully detailed systems architecture resolves into software and hardware systems. Desirable systems properties such as scalability and extensibility can be taken into account at the systems architecture level. The systems architecture is rationalized relative to the operational and technical architectures.

## 4.5 SELF ASSESSMENT QUESTIONS

1. Write the designing principles of digital Library.
2. What is digital library architecture?

## 4.6 REFERENCES

1. Elisabetj,J.show.(2000). Building a Digital Library: Atechnology managers point of view.The Journal of Academic Librarian Ship, Vol.26 No 6, pp 394-398
2. R.K. and Vishwanthan K.R.(2001). Digital Libraries: development and principles, LibraryReview , Vol 50 No 1, pp 10-16
3. Rob kling and Margaret Elliott (1994). Digital library Design for organisational usability. SIGOISBulletin, Vol.15 No 2
4. National Information Standards Organisation. The Dublin Core Metadata Elemen Set: Draft Standard S39.85; see www.niso.org/S3985.html.Sharma
5. Lesk, M.(1997). Practical Digital Libraries, Books, Bytes, and Bucks. San Francisco: Morgan Kaufman Publishers.
6. Marilyn Deegan and Simon Tanner.(2002). Digital futures: Strategies for the informationage. London: Library Association Publishing.
7. http://www.ncgia.ucsb.edu/giscc/units/u191/u191.html

# LESSON-5
# METADATA STANDARDS

## OBJECTIVE

After reading this chapter, students will be able to understand:

- ❖ Fundamental concepts of metadata.
- ❖ Types of metadata and tools for creating metadata

## STRUCTURE



## 5.1 INTRODUCTION

Digital library technologies are by now well established and understood throughout the higher education community and the creation of digital collections, either in the form of 'born-digital' materials or the conversion of standard library materials into digital form, is now a well-established part of the activities of most higher education institutions. Making effective use of

these resourcesis dependent on the creation of good quality metadata, without which they cannot be found by users nor administered effectively by their host institutions. To move beyond the ambit of the individual repository so that digital resources can be managed usefully at a sector-wide level, allowing, for instance, collections to be searched together in a federated fashion, requires some degree of standardization of metadata. This report attempts to provide a snapshot of digital librarymetadata at a stage when such standardization has become fully practical, even if its possibilities have not yet been fully realized within the higher education sector. The bulk of the report does this by surveying a group of standards which are based on the XML (eXtensible Markup Language) markup language, and showing how they relate to each other in ways which allow them to form an integrated metadata scheme. It will be seen that their application in a scheme of this type is not entirely without its problems, most of which are caused by overlaps and redundancies between the standards, but practical solutions will be outlined for these. An assessment of future developmentsin library metadata will point to possibilities for consolidating the potential offered by these developments and indicate how they are likely to affect all stakeholders in digital libraries.

## 5.2 DIGITAL LIBRARIES IN A NETWORKED AGE

What exactly constitutes a 'digital library' has never been easy to pin down. William Saffady(1995) provided a definition in a seminal article in 1995 that remains pertinent today:-
*"...a library that maintains all, or a substantial part, of its collection in computer-processable form as an alternative, supplement or complement to the conventional printed and microfilm materialsthat currently dominate library collections." (p. 221)*

In the years since Saffady wrote these words, the Internet has become home to digital collections which are wider and more diverse than this rather narrow definition allows: most notably, institutional digital repositories have become important vehicles for the dissemination of research output, museums have increasingly mounted digital collections to complement their curatorial work, and archives have increasingly (at least as far as copyright allows) mounted the collections they hold in their custody for the wider world to access. More recently, with the advent of so-called Web 2.0 technologies (Anderson, 2007) the digital collection has become a more fluid, interactive concept, as applications such as blogs, wikis, and folksonomies become part of the information landscape. While this report concentrates specifically on the core digital library sector, its applicability in principle, at least, to these other sectors will also be considered.

The richness of the collections that have appeared on the Web from these disparate quarters has made an exceptional contribution to the higher education community, but for those hoping to use them the overall experience can be a somewhat daunting one. The overall impression can beof a messy information environment, where every collection has to be searched in different ways through a different interface, and where finding the collections themselves can be, in itself, a difficult and time-consuming process. Even the collections of a single institution created over a period of time may suffer from multiple interfaces and no facilities for cross-searching: Oxford University's digital initiatives1, for instance, present a diversity of interfaces and disparities of cataloguing standards that result in their usage being more cumbersome and time-consuming than the technologies underlying them would potentially allow.

The irony of the ever improving powers of technology failing to deliver their full potential because of an increasingly messy information environment for the user is obvious and needs to beaddressed. This is particularly so at a time when the volume of digital collections is likely to increaseexponentially now that the importance of the electronic medium for providing access to

informationand the academic record is fully recognized. To tidy up the information environment, and render itcoherent and easily approachable, requires above all an acceptance of the importance of metadata.

## 5.3 WHAT IS METADATA?

Metadata is structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage information resource. Metadata is often called data about dataor information about information. The term metadata is used differently in different communities. Some use it to refer to machine understandable information, while others use it only for records that describe electronic resources. In the library environment, metadata is commonly used for any formal scheme of resource description, applying to any type of object, digital or non-digital. Traditionallibrary cataloging is a form of metadata; MARC 21 and the rule sets used with it, such as AACR2, are metadata standards. Other metadata schemes have been developed to describe various types of textual and non-textual objects including published books, electronic documents, archival findingaids, art objects, educational and training materials, and scientific datasets.

## 5.4 WHAT DOES METADATA DO?

An important reason for creating descriptive metadata is to facilitate discovery of relevant information. In addition to resource discovery, metadata can help organize electronic resources, facilitate interoperability and legacy resource integration, provide digital identification, and supportarchiving and preservation.

### *5.4.1 Resource Discovery*

Metadata serves the same functions in resource discovery as good cataloging does by:

- Allowing resources to be found by relevant criteria;
- Identifying resources;
- Bringing similar resources together;
- Distinguishing dissimilar resources; and
- Giving location information.

### *5.4.2 Organizing Electronic Resources*

As the number of Web-based resources grows exponentially, aggregate sites or portals are increasingly useful in organizing links to resources based on audience or topic. Such lists canbe built as static WebPages, with the names and locations of the resources “hardcoded” in the HTML. However, it is more efficient and increasingly more common to build these pages dynamically from metadata stored in databases. Various software tools can be used to automatically extract and reformat the information for Web applications.

### *5.4.3 Interoperability*

Describing a resource with metadata allows it to be understood by both humans and machines in ways that promote interoperability. Interoperability is the ability of multiple systems with different hardware and software platforms, data structures, and interfaces to exchange data with minimal loss of content and functionality. Using defined metadata schemes, shared transfer protocols, and crosswalks between schemes, resources across the network can be searched more seamlessly.

### *5.4.4 Digital Identification*

Most metadata schemes include elements such as standard numbers to uniquely identify the work or object to which the metadata refers. The location of a digital object may also be given using a file name, URL (Uniform Resource Locator), or some more persistent identifier such as a PURL (Persistent URL) or DOI (Digital Object Identifier). Persistent identifiers are preferred because object locations often change, making the standard URL (and therefore the metadata record) invalid. In addition to the actual elements that point to the object, the metadata can be combined to act asa set of identifying data, differentiating one object from another for validation purposes.

### *5.4.5 Archiving and Preservation*

Most current metadata efforts center around the discovery of recently created resources. However, there is a growing concern that digital resources will not survive in usable form into the future. Digital information is fragile; it can be corrupted or altered, intentionally or unintentionally. Itmay become unusable as storage media and hardware and software technologies change. Formatmigration and perhaps emulation of current hardware and software behavior in future hardware and software platforms are strategies for overcoming these challenges. Metadata is key to ensuringthat resources will survive and continue to be accessible into the future. Archiving and preservation require special elements to track the lineage of a digital object (where it came from and how it haschanged over time), to detail its physical characteristics, and to document its behavior in order toemulate it on future technologies

## 5.5 THE NEED FOR METADATA STANDARDIZATION

Metadata is the core of any information retrieval system and so its implications for any digital library are profound: the choice of a metadata scheme underpins any such library's ability to deliver objects in a meaningful way, and greatly affects its long-term ability to maintain and preserve its digital assets. The necessity for common approaches to metadata has been acknowledged inthe library community for as long as inter-institutional co-operation has been practised. It was recognized particularly in the 1960s when the MARC (Machine Readable Cataloguing) standard and AACR (Anglo-American Cataloguing Rules) cataloguing rules were created to standardize practices into a form which would make full use of the then nascent computing technologies.

The MARC standard provided a uniform container to hold cataloguing information in a formthat would readily transfer between systems, while AACR provided consistent rules to govern whatinformation would populate the fields of the MARC record and how it would be formatted. It was asa result of the adoption of these standards that the large union catalogues and collaborative cataloguing projects that are now such a prominent part of the library world became possible. The technology of the digital library offers even greater potential for inter-institutional collaboration: notonly can multiple collections be rendered cross-searchable in the style of a union catalogue, but the objects that constitute these collections can themselves readily be integrated into inter-institutional virtual repositories. To do so effectively, however, requires standard approaches to metadata. Without these, problems rapidly arise when digital library collections reach any substantial size: intelligent cross-searching, for instance, becomes very difficult to achieve as inconsistent item descriptions rapidly render retrieval from large collections very imprecise. Without consistency of metadata practices, the often-stated ideal of a 'hybrid library' (Rusbridge, 1998), which integratestraditional and electronic resources, remains a remote possibility.

To adopt an analogy from the traditional library world, it is necessary to standardize both the containers for digital library metadata (a digital library MARC standard) and the rules for the metadata content itself (an analogue to AACR). Unfortunately, adapting these long-established standards for the digital library environment is not a feasible option per se. The metadata required for digital objects is more complicated than that required for physical library items, which is generally limited to describing the intellectual content of an item (such as its author, title or subject), and such basic administrative information (for example, shelf marks) as is required to curate it. A useful typology for digital library metadata, adopted by early key projects such as the Making of America II (MOA2)2, indicates the range of information that must be included:

- ✓ Descriptive metadata: analogous to the tradition catalogue record, this is information on the item's intellectual contents which allows it to be retrieved and its value to the user assessed
- ✓ Administrative metadata: the information necessary to curate the digital item, which includes (not exclusively):-
- ✓ technical metadata: all necessary technical information (for example, file formats) to allow the host system to store and process the item
- ✓ rights management: declarations of rights held in the item and the information necessary to restrict its delivery to those entitled to access it
- ✓ digital provenance: information on the creation and subsequent treatment of the digital item, including details of responsibilities for each event in its lifespan
- ✓ Structural metadata: information necessary to record the internal structure of an item so that it can be rendered to the user in a sensible form (for instance, a book must be deliveredin its page order).

This type of metadata is necessary as an item may often be comprised of multiple (often thousands) of files - for example, the images of individual pages that make up a digitized book.

## 5.6 TYPES OF METADATA

### *5.6.1 Descriptive metadata*

A number of schemes are available for descriptive metadata, of which the two that have established themselves most securely in the digital library world are Dublin Core (DC7) and MODS (Metadata Object Description Schema8). Dublin Core is perhaps the most widely used scheme for many reasons, of which its simplicity is perhaps the primary: a set of 15 basic fields (such as creator, subject, identifier) designed for resource description and discovery in any type of media, it has formed the basis of many a digital library and underlies further important standards, such as the Open Archives Initiative's Protocol for Metadata Harvesting (OAI-PMH9). If these 15 fields provetoo broad for a given application, DC offers the option of qualifying them to allow greater precision:the creator field, for instance, may be qualified to delineate that role more specifically (such as creator. author). While this allows increased precision, it does inevitably reduce the interoperabilityof DC metadata.

MODS, covered in an earlier TSW report (Gartner, 2003), offers an alternative that effectively gets round this deficiency in DC. It offers a richer set of approximately 80 elements, which allows

a much greater degree of precision but retains interoperability by virtue of these elements being fixed and so employed without qualification. Although designed specifically for digital library objects, MODS is based on a subset of the MARC standard and therefore integrates well with metadata held in traditional library catalogues. It also incorporates the facility to extend its element set in thevery rare cases where this is needed (although this will, of course, reduce its interoperability slightly).

Both DC and MODS have committed adherents in the digital library sector, and it is difficultat present to imagine either becoming predominant. The wide constituency of DC users, and the increasing set of community-agreed extensions to it (with their concomitant support bases), both make it a safe option for the digital library designer, despite its limitations stated above. The advantages of MODS, which, despite its origin as a derivative from MARC, is not merely a bibliographic standard, merit serious consideration, particularly for metadata requirements of anycomplexity. Either, however, are perfectly sensible options and the choice of which is adopted mayoften depend on the existing expertise of implementers.

### *5.6.2 Administrative metadata*

METS divides its administrative metadata section into several smaller subsections, each of which can be populated with one or other of the schemas listed above. Specifically:

Ï% Technical metadata: the schema used here will obviously depend on the type of files included in the object; more than one type of file will require the use of more than one type of metadata schema. All necessary technical information (for example, file formats) to allow the host system to store and process the item

The most obvious choices are:-

- ✓ still images: MIX
- ✓ video: VIDEOMD (or possibly PBCore)
- ✓ audio: AUDIOMD
- ✓ text: Schema for Technical Metadata for Text or TEI Headers

Ï% rights metadata: the METS Schema for Rights Declaration for straightforward ownership declarations and access control, or XrML/ODRL for more complex rights control. The rights schema of PREMIS also fits in here. Declarations of rights held in the item and the information necessary to restrict its delivery to those entitled to access it.

Ï% source metadata: this is metadata about the original item from which the digital copy was made – it will, of course, be absent if the item is born digital. This is predominantly descriptive metadata, and so will generally use MODS or DC (although, of course, more specialized schemas may be used where required).

Ï% digital provenance metadata: information on the creation and subsequent treatment of the digital item, including details of responsibilities for each event in its lifespan. This is essentiallyan audit trail for the digital object, tracing its creation and the changes that have been made to it.The events schema of PREMIS fulfils this function and should be located here.

### *5.6.3 Structural metadata*

Information necessary to record the internal structure of an item so that it can be renderedto the user in a sensible form (for instance, a book must be delivered in its page order). This typeof metadata is necessary as an item may often be comprised of multiple (often thousands) of files - for example, the images of individual pages that make up a digitized book.

METS provides a hierarchical structural map to encode metadata on the internal structure of an item. This is simply a series of nested elements, named div (short for division) whose nesting is meant to emulate the structure of the digital object: so, for example, a digitized book would have its structure of divs arranged to match its original divisions into chapters, sections of chapters and so on. The structural map may encode either a physical or a logical structure: it may, for instance, describe the pagination structure of a volume, or the arrangement of its intellectual contents. Often these will coincide, but where this is not the case it is entirely feasible to include multiple structural maps, each describing a different type of structure, and to link them together by METS's comprehensive linking facilities.

## 5.7 XML: THE STANDARD BEHIND THE STANDARDS

All of the standards discussed here use XML as their semantic and structural underpinning. XML began life in the 1960s as SGML (Standard Generalised Markup Language), a system for tagging up electronic texts using semantically meaningful tags. In addition to marking up texts themselves, it has also come to be used increasingly as a standalone mechanism for encoding metadata for all types of objects in traditional or electronic libraries. There is, for example, a translation to XML of the MARC standard, which encodes the standard MARC fields in XML tags: for example, the title (245) field of the MARC record 245 10| a Arithmetic /|cCarl Sandburg ; illustrated as an anamorphic adventure by Ted Rand.is rendered in XML tags as follows:-

```
<datafield tag="245" ind1="1" ind2="0">
<subfield code="a">Arithmetic /</subfield>
<subfield code="c">Carl Sandburg ; illustrated as
ananamorphic adventure by Ted Rand.</subfield>
</datafield>
```

The strengths of XML as the basis of a metadata scheme have often been acknowledged, for example by UKOLN in their *Good Practice Guide for Developers of Cultural Heritage Web Services: Metadata Sharin and XML* (Johnston, 2004). It is a fully open standard registered with the ISO (International Standards Organisation), and so is independent of any given software application. It is acknowledged as the most archivally robust metadata format by, for example, the influential Commission on Preservation and Access (Coleman and Willis, 1997). It also benefits from the relative simplicity of its syntax, combined with great flexibility in the ways in which it can be used: in particular, its ability to encode hierarchical structures by the simple expedient of nesting tags allows it to represent complex relations between metadata components simply and elegantly.

XML applications are usually expressed in what are known as *schemas:* these consist essentially of definitions of the elements which make up the XML tags and of the rules which dictate how they should be used together (for instance, which should nest within which). In the

discussion which follows, the term *schema* is used with this technical meaning, in contrast to *scheme* which is used to describe a metadata system as a whole, XML or otherwise.

## 5.8 CREATING METADATA

Who creates metadata? The answer to this varies by discipline, the resource being described, the tools available, and the expected outcome, but it is almost always a cooperative effort. Much basic structural and administrative metadata is supplied by the technical staff who initially digitizes or otherwise creates the digital object, or is generated through an automated process. For descriptive metadata, it is best in some situations if the originator of the resource provides the information. This is particularly true in the documentation of scientific datasets where the originator has significant understanding of the rationale for the dataset and the uses to which it could be put, and for which there is little if any textual information from which an indexer could work.

However, many projects have found that it is more efficient to have indexers or other information professionals create the descriptive metadata, because the authors or creators of the data do not have the time or the skills. In other cases, a combination of researcher and information professional is used. The researcher may create a skeleton, completing the elements that can be supplied most readily. Then results may be supplemented or reviewed by the information specialist for consistency and compliance with the schema syntax and local guidelines.

### *5.8.1 Creation Tools*

Many metadata project initiatives have developed tools and made them available to others, sometimes for free. A growing number of commercial software tools are also becoming available. Creation tools fall into several categories:

- Templates allow a user to enter the metadata values into pre-set fields that match the element set being used. The template will then generate a formatted set of the element attributes and their corresponding values.
- Mark-up tools will structure the metadata attributes and values into the specified schema language. Most of these tools generate XML or SGML Document Type Definitions (DTD). Some templates include such a mark-up as part of their final translation of the metadata.
- Extraction tools will automatically create metadata from an analysis of the digital resource. These tools are generally limited to textual resources. The quality of the metadata extracted can vary significantly based on the tool's algorithms as well as the content and structure of the source text. These tools should be considered as an aid to creating metadata. The resulting metadata should always be manually reviewed and edited.
- Conversion tools will translate one metadata format to another. The similarity of elements in the source and target formats will affect how much additional editing and manual input of metadata may be required.

### *5.8.2 Metadata Quality Control*

The creation of metadata automatically or by information originators who are not familiar with cataloguing, indexing or vocabulary control can create quality problems. Mandatory elements may be missing or used incorrectly. Schema syntax may have errors that prevent the metadata from being processed correctly. Metadata content terminology may be inconsistent, making it difficult to locate relevant information.

The Framework of Guidance for Building Good Digital Collections, available on the NISO website, articulates six principles applying to good metadata:

- Good metadata should be appropriate to the materials in the collection, users of the collection, and intended, current and likely use of the digital object. Good metadata supports interoperability.
- Good metadata uses standard controlled vocabularies to reflect that what, where, when and who of the content.
- Good metadata includes a clear statement on the conditions and terms of use for the digital object.
- Good metadata records are objects themselves and therefore should have the qualities of achievability, persistence, unique identification, etc. good metadata should be authoritative and verifiable.
- Good metadata supports the long-term management of objects in collections.

There are a number of ongoing efforts for dealing with the metadata quality challenge:

- Metadata creation tools are being improved with such features as templates, pick lists thatlimit the selection in a particular field, and improved validation rules.
- Software interoperability programs that can automate the games that can automate the "crosswalk" between different schemas are continuously being developed and refined.
- Content originators are being formally trained in understanding metadata and controlled vocabulary concepts and in the use of metadata related software tools.
- Existing controlled vocabularies that may have initially been designed for a specific use or a narrow audience are getting broader use and awareness. For example, the Content Types and Subtypes originally defined for MIME email exchange are commonly used as thecontrolled list for the Dublin Core Format element.
- Communities of users are developing and refining audience-specific metadata schemas, application profiles, controlled vocabularies, and user guidelines. The MODS User Guidelinesare a good example of the latter.

## 5.9 METADATA STANDARDS

International standards apply to metadata. Much work is being accomplished in the national and international standards communities, especially ANSI (American National Standards Institute) and ISO (International Organization for Standardization) to reach consensus on standardizing metadata and registries.

The core standard is ISO/IEC 11179-1:2004 and subsequent standards (see ISO/IEC 11179). All yet published registrations according to this standard cover just the definition of metadata and do not serve the structuring of metadata storage or retrieval neither any administrative standardisation. It is important to note that this standard refers to metadata as the data about containers of the data and not to metadata (metacontent) as the data about the data contents. It should also be noted that this standard describes itself originally as a "data element" registry, describing disembodied data elements, and explicitly disavows the capability of containing complex structures. Thus the original term "data element" is more applicable than the later applied buzzword "metadata".

The Dublin Core metadata terms are a set of vocabulary terms which can be used to describe resources for the purposes of discovery. The original set of 15 classic metadata terms, known as the Dublin Core Metadata Element Set are endorsed in the following standards documents:

- IETF RFC 5013
- ISO Standard 15836-2009
- NISO Standard Z39.85.

Although not a standard, Micro format (also mentioned in the section metadata on the internet below) is a web-based approach to semantic markup which seeks to re-use existing HTML/XHTML tags to convey metadata. Micro format follows XHTML and HTML standards but is not a standard in itself. One advocate of micro formats, Tantek Çelik, characterized a problem with alternative approaches:

## 5.10 SELF ASSESSMENT QUESTIONS

1. Define what metadata is.
2. Write types of metadata.

## 5.11 REFERENCES


1. Anderson, P. 2007. What is Web 2.0? Ideas, technologies and implications for education.Published by JISC Technology and Standards Watch. Available online at:
2. http://www.jisc.ac.uk/media/documents/techwatch/tsw0701b.pdf [last accessed: 9th April2008]
3. Barlas, C. 2006. Digital Rights Expression Languages (DRELs). JISC Technology & Standards Watch, July 2006. Available online at:
4. http://www.jisc.ac.uk/whatwedo/services/services_techwatch/techwatch/techwatch_ic_reports2005_published.aspx [last accessed: 9th April 2008]
5. Bekaert, J., Hochstenbach, P., Van De Sompel, H. 2003. Using MPEG-21 DIDL to Represent Complex Digital Objects in the Los Alamos National Laboratory Digital Library. D-Lib Magazine,Volume 9, Number 11, November 2003. Available online at:
6. http://www.dlib.org/dlib/november03/bekaert/11bekaert.html [last accessed: 9th April 2008]
7. Candela, L., CASTELLI, D., PAGANO, P. 2007. A Reference Architecture for Digital Library Systems: Principles and Applications. Lecture Notes in Computer Science, Springer Berlin.Volume 4877/2007. DOI 10.1007/978-3-540-77088-6
8. Coleman, J., WILLIS, D. 1997. SGML as a Framework for Digital Preservation and Access.Commission on Preservation and Access. Available online .

1. Gartner, R. 2003. MODS: Metadata Object Description Schema. JISC Technology & Standards Watch. October 2003. Availlable online at: http://www.jisc.ac.uk/whatwedo/services/services_techwatch/techwatch/techwatch_report_0306.aspx [last accessed: 9th April 2008]
2. Guenther, R. 2007. Best practices for using PREMIS with METS. The Library of Congress (Draft), 9th August 2007. Available online at: http://www.loc.gov/standards/premis/best-practices-premismets-20070809.doc [last accessed: 9th April 2008]
3. http://www.niso.org/publications/press/UnderstandingMetadata.pdf
4. Johnston, P. 2004. Good Practice Guide for Developers of Cultural Heritage Web Services. UKOLN. Available online at: http://www.ukoln.ac.uk/interop-focus/gpg/Metadata/#section2 [lastaccessed: 9th April 2008]
5. Olivier, B. 2007. Having your cake and eating it: The e-Framework's Service-Oriented Approach to IT in Higher Education. Educause Review. vol. 42, no. 4 (July/August 2007), pp. 58–67. Available online at: http://connect.educause.edu/Library/EDUCAUSE+Review/HavingYourCakeandEatingIt/44596?time=1207577197 [last accessed: 9th April 2008]
6. Rusbridge, C. 1998. Towards the Hybrid Library. D-Lib Magazine. July/August 1998. Available online at: http://www.dlib.org/dlib/july98/rusbridge/07rusbridge.html [last accessed: 9th April2008]
7. Saffady, W. 1995. Digital library concepts and technologies for the management of library collections: an analysis of methods and costs. Library Technology Reports,. Vol. 31, No. 3, pp. 221-380. Available for purchase online at: http://cat.inist.fr/?aModele=afficheN&cpsidt=2485292
8. "ISO/IEC 11179-1:2004 Information technology - Metadata registries (MDR) - Part 1: Framework". Iso.org. 2009-03-18. Retrieved 2011-12-23.
9. "DCMI Specifications". Dublincore.org. 2009-12-14. Retrieved 2013-08-17.
10. "Dublin Core Metadata Element Set, Version 1.1". Dublincore.org. Retrieved 2013-08-17.
11. J. Kunze, T. Baker (2007 [last update]). "The Dublin Core Metadata Element Set". ietf.org. Retrieved 17 August 2013.
12. "ISO 15836:2009 - Information and documentation - The Dublin Core metadata element set". Iso.org. 2009-02-18. Retrieved 2013-08-17.
13. "NISO Standards - National Information Standards Organization". Niso.org. 2007-05-22.Retrieved 2013-08-17.
14. "What's the Next Big Thing on the Web? It May Be a Small, Simple Thing — Microformats". Knowledge@Wharton.Wharton School of the University of Pennsylvania. 2005-07-27.

LESSON - 6

# DIGITAL RESOURCES: NATURE AND MANAGEMENT

**OBJECTIVE**

After reading this chapter, students will be able to understand:

- What are Digital Resources and types of Digital resources
- Basic fundamental concepts of digital library protocols
- Different protocols of DL

**STRUCTURE**



## 6.1 INTRODUCTION

The global information infrastructure is transforming with the growth in full text digital resources and fast communication facilities. This is so because institutions, agencies and departments in every country are busy capturing, processing, storing and disseminating information in the digital form. One of the primary methods of digital collection building is digitization. The digitization means the conversion of any fixed or analogue media-such as books, journal articles, photos, paintings, microforms- into electronic form through scanning, sampling, or in fact even re-keying. Digital library is a library where the information is made available in electronic form and access to it is provided through computers and other media like Local area network or Internet. Digitizing information has become a boon to many including library and information professionals. It reduces storage space, increases efficiency of retrieval of information.

The digital library offers users the prospects of access to electronic resources at their convenience temporarily and spatially. Users don't have to be concern with the physical library hours of operation and users don't have to go physically to the library to access resources. Some of the Digital library initiatives undertaken so far world-over are highlighted: Medoc in Germany, Digital library projects at the National diet library in Japan, Tsinghua University Central Library in China, Biblio theca Universalis-G7 Project, Library of Congress USA, NSW Parliament News paper clippings and press releases imaging project in Australia, Internet Archives USA, California digitallibraries. In India, the major initiatives undertaken in

this direction: Million Book Universal Digital Library project, Vidyanidhi project, National Mission of Manuscripts, Traditional Knowledge DigitalLibrary (TKDL), Gyan Nidhi and Indian National Digital Library in Engineering Science and Technology (INDEST). The initiatives undertaken so far clearly indicate that there is a boom in funding for digital library projects.

## 6.2 DIGITAL RESOURCES

A digital resource is anything which is *published* in computer-readable format. The definition of *published* is not developed here in detail, but is understood to take a meaning consistent with the Library's habitual use of the term.

The term *digital* is used to distinguish between computer-readable information and analogue electronic information. In practice, the use of analogue electronic data is generally restricted to broadcasting (of television and audio) and publishing some recordings (VHS cassettes, audiocassettes, etc). Analogue electronic data is excluded from the scope of this study.

*Digital resources* therefore include publications which are produced both on paper and in computer-readable form (eg so-called "parallel" editions of journals) and those which are unavailable in any other form (eg multimedia CD-ROMs, ROM cartridges). By far the largest proportion of such resources is made up of:

- Publications available on the Internet;
- CD-ROMs

## 6.3 TYPES OF DIGITAL RESOURCES

In its simple connotation, digital resources refer to any resource, which is in digitized form.i.e. which can be read and scanned by means of electronic media. Unlike conventional form, digital resources do not require separate space in a library as these can be stored in a computer locallyor remotely. Digital collections must be selected, acquired, organized, made accessible and preserved. Digital resources include wide range of materials such as:

a) Collections in which complete contents of documents are created or converted into machinereadable form for online access.
b) Scanned images, images of photographic or printed texts, etc.
c) Scientific data sets such as protein sequences or nucleic acid sequences etc., What are digitalresources?

The definition of digital resources used here is explored in section. In summary, Internet resources, CD-ROMs and the Library's own digitised materials are considered.

## 6.4 FACTORS

A large number of effects associated with digital resources might affect reading rooms. We have to consider resources referred to from within the Library, resources referred to from outsidethe Library, free and chargeable resources, and so on. For example:

- increasing availability of digital resources on the Internet may allow readers to discover information from their own desktops without needing to visit the Library;
- availability of many CD-ROMs might increase the number of readers wishing to refer to them.

Clearly several factors need to be considered. The five factors which could have an effect are:

- Source;
- Content;
- Format;
- Usage;
- Access.

The possible natures of these factors are shown in the following table in Figure 1:

| Source | Content | Format* | Usage | Access |
|---|---|---|---|---|
| Digitised | BLMetadata | Text or data | Fact finding | From within material |
| | | | (catalogue record) | reading room |
| Free Internet Data (monograph, | | Low-resolution | Studying | Remote from resource |
| serial etc) | | image | | reading room |
| Chargeable Internet** resource | | High-resolution image | Browsing | |
| Offline resource (eg CD-ROM) | | Sound Motion | | |

*This definition of Format is intentionally not academically rigorous, Rather, it is chosen to identify the formats which are, or which could be, significantly different in terms of their effects on reading rooms.

**Strictly, commercial network services which are not presented through the Internet should be included here; the term Internet is used for simplicity.

## 6.5 FULL MODEL OF DIGITAL RESOURCE ATTRIBUTES

Each combination of the different factors' natures is different. For example,

- a digitized BL resource which is a serial in the form of text used for browsing within a reading room may have a different impact from
- a digitized BL resource which is a catalogue in the form of text used for fact finding within a reading room.

To arrive at well-reasoned quantitative estimates, each combination should be examined separately. By simple combinatorial arithmetic, there are 4 x 2 x 5 x 3 x 2 = 240 combinations. The effects of each of these 240 potential scenarios should be considered carefully and this potentially for each reading room. In each case, there would be consideration of:

- the number of readers;
- the potential population of readers;
- the levels of readers' use of materials;
- how the above may be changed by the digital format; etc. It is not appropriate to attempt an exercise on this scale in the present study, so a simpler representation is called for.

## 6.6 A MODEL OF DIGITAL RESOURCES

The model defined by the table above can be simplified by aggregating the combinations, discarding less significant aggregations until the number of scenarios is manageable. The basis for the aggregation and discarding is qualitative, applying an understanding of the likely nature of the effects, as informed by interviews conducted during the study. The final aggregation used consists of six scenarios:

- Access to the Library's catalogues (in any digital form, from anywhere);

- Use of digitized versions of the Library's collections, from within the reading rooms;
- Use of digitized versions of the Library's collections, from outside the Library (principally by means of the Internet);
- Use of Internet resources which are not digitized versions of the Library's collections, from within the reading rooms;
- Use of Internet resources which are not digitized versions of the Library's collections, from outside the Library (ie conventional network research);
- Use of CD-ROMs in reading rooms.

## 6.7 PRESERVATION OF DIGITAL RESOURCES

"Digital preservation" or "Digital archiving" means taking steps to ensure the long term access to the digital documents. Unlike the print publications, the digital preservation is more complex as one has to take care of many aspects of the documents such as content, presentation,functionality, authenticity etc. At the time of selection and acquisition only one must think of preservation of digital resources. The digital technology as well as other technologies such as Internet and Web technologies are continuously changing due to upgradations of software and hardware, proliferation of standards and protocols for file formats, network interfaces, storage media and devices etc. As a result of which there is constant danger of "techno-obsolescence". Hence preservation policy for digital resources is of prime importance and should take care of following aspects:

- Preservation of digital resources at different levels depending on its usability, functionalities.
- Continuous reviewing of the digital resources ensuring long term access to them.
- Weeding out obsolete information and invalid websites.

***There are three ways to preserve digital resources:***

### *6.7.1 TECHNOLOGY PRESERVATION*

The older technology can be preserved for viewing digital objects in their original formats but it is not feasible in long term due to cost, space and technical support requirements.
e.g. Hardware.

### *6.7.2 TECHNOLOGY EMULATION*

It refers to creating new software that copies the operations of older hardware and software thus ensuring it's originality in terms of physical presence, content and functionality. Some digital resources are highly dependent on particular hardware or software. Emulation techniques can be useful in such cases. However, emulation for preserving digital resources over the long term has not been tested.

### *6..7.3 DATA MIGRATION*

Migration covers a range of activities to periodically copy, convert or transfer digital information from a medium that is becoming obsolete or physically deteriorating to a newer one (e.g. floppy disk to CD-ROM), and/or converting from one format to another (e.g. Microsoft Word to ASCII), and/or moving documents from one platform to another (e.g. VAX to UNIX). Migration certainly preserves the physical presence and the content of the digital object. However, itmay not preserve presentation, functionality and context7

In order to avoid duplication of efforts and resources, many libraries are now working in partnerships. Examples of library consortia include the eLib (Electronic Library) project inthe U.K., the Digital Library Federation and the Research Libraries Group ARCHES (Archival Server and Test

Bed) in U.S., the NEDLIB (Networked European Depository Library) etc. These groups were founded primarily to build digital libraries, in which managing preservationwas a necessary component. The next important step in this regard is "Open Archival Information System Reference Model" (OAIS) developed by the Consultative Committee on Space Data for providing a conceptual framework and reference tool for defining a digitalarchive. It describes a specific functional model of both people and system requirements for implementing a digital archive. The NEDLIB project is implementing the OAIS model within the context of the deposit of electronic materials for archiving.

## 6.8 SELF ASSESSMENT QUESTIONS

1. What are digital resources? Explain
2. How to manage digital resources?

## 6.9 REFERENCES

1. Krishnamurthy, M: Digital Library Gateway for Library and Information science: A study. SRELS JOURNAL OF INFORMATION MANAGEMENT vol 39 (3), 2002, p. 245-254.
2. Mohan Raj Pradhan: Developing Digital Libraries: Technologies and Challenges. HERALD OF LIBRARY SCIENCE Vol 42 (2), 2004.
3. Ruth H. Miller: Electronic Resources and Academic Libraries, 1980-2000: A Historical Perspective Library Trends (Spring) 2000.

4. Dr Uma Kanjilal: Education and Training for Digital Libraries: Model for Web Enhanced
5. Continuing Education Programme, ICDL 2004, New Delhi 24-27 conference papers www.uohyd.ernet.in

**LESSON-7**

# DIGITAL LIBRARY EVALUATION

**OBJECTIVE**

After reading this chapter, students will be able to understand:

- Basic idea about evaluation of DL
- Criteria for evaluation of DL

**STRUCTURE**



## 7.1 INTRODUCTION

An evaluation is basically a judgment of worth. We evaluate a system in order to ascertain the level of its performance or its value (Chowdhury, 1999, 200). Although there have been

tremendous technological developments over the last two decades, information systems continueto be difficult to learn and use for the lay person. Special training is required and special skills needto be acquired to be able to use information systems effectively.

## 7.2 BASIC PROBLEMS IN EVALUATION

Saracevic (2000) has identified the following possible reasons for there having been comparativelylittle research on digital library evaluation:

- Digital libraries are still at the stage of evaluation; evaluation at this stage may be a bit premature and even dangerous.
- At this early stage of digital library development, informal and anecdotal ways of evaluation may be sufficient.
- Evaluation at a very basic technical level-to provide that an electronic collection is accessible-may be sufficient.
- The outcome of the performance measures of digital libraries at the early stage of development may not be very encouraging; such studies may discourage the funding bodies, which in turn slow down the pace of digital library research.

## 7.3 WHAT NEEDS TO BE EVALUATED?

A digital library may be evaluated from a number of perspectives, such as system, access and usability, user interfaces, information retrieval, content and domain, services, cost, and the overall benefits and impact. Marchionini (2000) comments that evaluating digital libraries is a bit like judging the success of a marriage where much depends on how successful the partners areas individuals as well as the emergent conditions made possible by the union.

Saracevic (2000) provides a long list of elements for the evaluation of digital libraries, and suggests that an evaluation project may select from these, clearly indicating what is included, and what is excluded, in the evaluation project. The list of elements for evaluation includes the following:

- Digital collections, resources
- Selection, gathering, holdings, media
- Distribution, collections, links
- Organization, structure, storage
- Interpretation, representation, metadata
- Management
- Preservation, persistence
- Access
- Physical networks

- Distribution
- Interfaces, interaction
- Search, retrieval
- Services
- Availability
- Range of available services, e.g. dissemination delivery
- Assistance, referral
- Use, users, communities
- Security, privacy, policies, legal aspects, licenses
- Management, operations, staff
- Cost, economics
- Integration, co-operation with other resources, libraries or services.

Saracevic (2000) proposes seven general classes or levels of evaluation, three of whichare users centred and three systems centred, while the seventh is an interface in between.

## 7.4 USER-CENTERED LEVELS OF EVALUATION

***7.4.1 Social Level-*** At this level the major objective of an evaluation might be to assess how well a digital library supports the needs, roles and practices of a society of community.

***7.4.2 Institutional Level-*** At this level, the objective might be to assess how well a digital library supports its parent organization's mission and objectives.

***7.4.3 Individual Level-*** At this level the objective of an evaluation might be to assess how well a digital library supports the information needs, tasks and activities of people as individual users or small groups.

The interface level that is in between the user-centred and system-centred levels of evaluation aims to assess how well a digital library interface provides and supports access, searching, navigation, browsing and interactions with the library.

## 7.5 SYSTEM-CENTERED EVALUATION

***7.5.1 Engineering level:*** how well the hardware, networks and related technologieswork.

***7.5.2 Proceeding level:*** how well the various procedures, techniques, algorithms, operations,etc., perform

***7.5.3 Content level:*** how well the information resources are selected, represented, organized, structured and managed.

## 7.6 EVALUATION STAGES

*Following five stages are involved in evaluation:*

1. Design of the evaluation

2. Drawing up an evaluation plan
3. Data gathering and recording
4. Data analysis ad interpretation of results
5. Presentation of findings

## 7.7 EVALUATION CRITERIA

Criteria refer to chosen standard(s) to judge things by. Criteria are then used to develop measures. (To define the differences by examples: time is a criterion, minute is a measure, and watch is a measuring instrument; relevance is a criterion, precision and recall are measures, and human relevance judgment is a measuring instrument). The importance of criteria follows from this truism: there can be no evaluation without explicitly or implicitly having some criterion or criteriafirst.

Since 1950's evaluation of IR systems uses relevance as the basic criterion for evaluation. Libraries use a variety of (more or less) standardized criteria for evaluation of components, such as a collection, or services, such as reference. Digital library efforts have not as yet developed anything similar as to evaluation criteria. There is nothing like relevance to be a basic criterion, there are no more or less standardized criteria for digital library evaluation. Several efforts that are devoted to developing digital library metrics have not produced, as yet, generalizable and accepted metrics, some of which may be used for evaluation. Thus, evaluators have chosen their own evaluation criteria as they went along. As a result, criteria for digital library evaluation fluctuate widely from effort to effort.

***A summary of most often used criteria follows:***

### 7.7.1 Usability

Usability has been used widely in digital library evaluation, but there is no uniform definition of what does it cover in digital library context. Usability is a very general criterion that covers a lot of ground and includes many specific criteria – it is a meta term. ISO defines usability "as the extentto which a product can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use." This was pretty much the umbrella underwhich usability was used in digital library evaluation.

***Here is a list of specific usability criteria applied in various studies:***

### 7.7.2 **Content** (of a portal or site)

- ✓ accessibly, availability
- ✓ clarity (as presented)
- ✓ complexity (organization, structure)
- ✓ informativeness
- ✓ transparency
- ✓ understanding, effort to understand
- ✓ adequacy

- ✓ coverage, overlap,
- ✓ quality, accuracy
- ✓ validity, reliability
- ✓ authority

**7.7.3 Process** – carrying out tasks such as search, navigate, browse, find, evaluate or obtain aresource etc.

- √ learnability to carry out
- √ effort/time to carry out
- √ convenience, ease of use
- √ lostness (confusion)
- √ support for carrying out
- √ completion (achievement of task)
- √ interpretation difficulty
- √ sureness in results
- √ error rate

### 7.7.4 Format

- √ attractiveness
- √ sustaining efforts
- √ consistency
- √ representation of labels (how well are concepts represented?)
- √ communicativeness of messages

### 7.7.5 Overall assessment

- √ satisfaction
- √ success
- √ relevance, usefulness of results
- √ impact, value
- √ quality of experience
- √ barriers, irritability
- √ preferences
- √ learning

## 7.8 SYSTEMS FEATURES

As digital libraries are systems, many traditional systems evaluation criteria were used. Some pertain to performance of technology others to performance of given processes or algorithms using technology.

### 7.8.1 Technology performance

- ✓ response time
- ✓ processing time, speed
- ✓ capacity, load

### 7.8.2 Process/algorithm performance

- ✓ relevance (of obtained results)
- ✓ clustering
- ✓ similarity
- ✓ functionality
- ✓ flexibility
- ✓ comparison with human performance
- ✓ error rate
- ✓ optimization
- ✓ logical decisions
- ✓ path length
- ✓ clickthroughs
- ✓ retrieval time

### 7.8.3 Overall system

- ✓ maintainability
- ✓ scalability
- ✓ interoperability
- ✓ sharability
- ✓ costs

So far only a few evaluation studies have been carried out and they have mostly focused on the usability aspects. Saracevic (2000) comments that the ultimate goal for a digital library evaluation is to assess how digital libraries are transforming our education, research, learning and living. In future, more digital library evaluation studies will be able to answer these questions.

## 7.9 SELF ASSESSMENT QUESTIONS

1. What is Evaluation of digital library?
2. Explain the criteria for evaluating the digital library.
3. How to evaluate the digital objects.

## 7.10 REFERENCES

1. Ghowdhury, G. G. & Chowdhury, S. Introduction to Digital Libraries. London; Facet Publishing,2003, p.267-283. (ISBN 978-85604-465-3).
2. Marchionini, G. (2000) Evaluating Digital Libraries: A Longitudinal and Multifaceted View,Library Trends, 49 (2), 304-33.
3. Saracevic, T. (2000) Digital Library Evaluation: Toward Evaluation of Concepts, LibraryTrends, 49 (2), 350-69.
4. Saracevic, T. (2004) Evaluation of Digital Libraries: An Overview

LESSON-8

# DIGITAL PRESERVATION

**OBJECTIVE**

After reading this chapter, students will be able to understand:

- Basic fundamental concepts of digital preservation
- Definition of digital preservation
- Benefits of preservation

**STRUCTURE**

## 8.1 INTRODUCTION

Information preservation is one of the most important issues in human history, culture, and economics, as well as the development of our civilization. While earliest information was recorded in carvings on stone, ceramic, bamboo, or wood, the development of civilization paved the way for new storage media and techniques for recording information, such as writing on silk or printing on paper. Eventually we were able to put photographic images on film and music on records. A revolutionary change occurred in the information storage field with the invention of electronic storage media. With the advent of high-performance computing and high-speed networks, the use of digital technologies is increasing rapidly. Digital technologies enable information to be created, manipulated, disseminated, located, and stored with increasing ease. Ensuring long-term access to the digitally stored information poses a significant challenge, and is increasingly recognized as an important part of digital data management.

Digital preservation is the active management of digital information over time to ensure its accessibility. Preservation of digital information is widely considered to require more constant and ongoing attention than preservation of other media. This constant input of effort, time, and money to handle rapid technological and organizational advance is considered the main stumbling block for preserving digital information. Indeed, while we are still able to read our written heritage from several thousand years ago, the digital information created merely a decade ago is in serious danger of being lost, creating a digital Dark Age.

## 8.2 DIGITAL PRESERVATION

The term “digital preservation” refers to both preservation of materials that are created originally in digital form and never exist in print or analog form (also called “born-digital” and “electronic records”) and the use of imaging technology to create digital surrogates of analog materials for access and preservation purposes. While this broad use of the term digital preservation can causeconfusion, data on both aspects of digital preservation were analyzed. Digital materials, regardless of whether they are created initially in digital form or converted to digital form, are threatened by technology obsolescence and physical deterioration.

Digital preservation is the set of processes and activities that ensure continued access to information and all kinds of records, scientific and cultural heritage existing in digital formats. This includes the preservation of materials resulting from digital reformatting, but particularly information that is born-digital and has no analog counterpart. In the language of digital imaging and electronic resources, preservation is no longer just the product of a program but an ongoing process. In thisregard the way digital information is stored is important in ensuring its longevity. The long-term storage of digital information is assisted by the inclusion of preservation metadata.

## 8.3 DEFINITIONS OF DIGITAL PRESERVATION

Digital preservation is defined as: long-term, error-free storage of digital information, withmeans for retrieval and interpretation, for the entire time span the information is required for. Long- term is defined as “long enough to be concerned with the impacts of changing

technologies, including support for new media and data formats, or with a changing user community. Long Termmay extend indefinitely". "Retrieval" means obtaining needed digital files from the long-term, error- free digital storage, without possibility of corrupting the continued error-free storage of the digitalfiles. "Interpretation" means that the retrieved digital files, files that, for example, are of texts, charts, images or sounds, are decoded and transformed into usable representations. This is often interpreted as "rendering", i.e. making it available for a human to access. However, in many casesit will mean able to be processed by computational means.

***There are three definitions of 'long term' digital preservation are given below:***

Digital preservation is a set of activities required to make sure digital objects can be located, rendered, used and understood in the future. This can include managing the object names and locations, updating the storage media, documenting the content and tracking hardware and softwarechanges to make sure objects can still be opened and understood.

1. "Digital preservation combines policies, strategies and actions to ensure access to reformatted and born digital content regardless of the challenges of media failure and technological change. The goal of digital preservation is the accurate rendering of authenticated content over time" (ALA 2007:2).

2. "The act of maintaining information, in a correct and Independently Understandable form,over the Long Term" (CCSDS 2002: 1-11).

3. "All activities concerning the maintenance and care for/curation of digital or electronic objects,in relation to both storage and access" (Research Councils UK 2008: 6).

## 8.4 WHAT DO WE NEED TO PRESERVE?

*Various aspects of the digital objects may be needed to be preserved.*

- The lowest level of preservation requirements includes preservation of the bit stream; this does not however ensure understandability, readability or usefulness of the digital object. The biggest risk in terms of understandability is that the meaning (and even the names) associated with values in a dataset, although known to the data producers, is not availableto the users; without this the data is essentially useless.
- Another aspect is that, even for users within the same sub-discipline, terminology drifts and meaning is lost; users in different (sub) disciplines will require even more help with the semantics of the data.
- A more complex approach may strive to preserve not only the 1s and 0s but also the meaning so that it remains readable and understandable. Such an approach requires the preservation of additional information (representation information, technical metadata etc.)
- Even more ambitious preservation approaches try to preserve understandable content in such a way that the provenance and source of the digital object also remains clear. Thusthe users can have trust that the object is authentic, accurate, complete etc.

## 8.5 WHY SHOULD WE CARE ABOUT DIGITAL PRESERVATION?

### *8.5.1 Storage media/data carrier problem*

Digital objects are much more 'fragile' than traditional analogue documents such as books or other hard copy mediums. Digital objects are fragile because they require various layersof technological mediation before they can be heard, seen or understood by people. Digital

objects are also much more venerable to physical damage. One scratch on CD-ROM containing 100 e-books can make the content inaccessible, whereas to damage 100 hard copy books by one scratching move is - fortunately - impossible. A flash memory stick can drop into glass of water or get magnetized, portable hard drive or laptop can slip from your hands and get irreparably damaged in a second.

Digital objects require pro-active intervention to remain accessible. While you can put a book on a shelf and return to it in upwards of 100 years and still open it and see the contentas it was intended by the author/publisher, the same approach of benign neglect to a digital object is almost a guarantee that it will be inaccessible in the future.

### *8.5.2 Hardware obsolescence*

Even if you returned to the digital object in five years to find the disk is in perfect condition and you have software that can open the file, but if that file is on a disc your computer doesn't have a drive for you will not be able to access it.

### *8.5.3 Software and format obsolescence problem*

Alternatively the software or file format can become obsolete for a number of reasons. For example software upgrades may not support legacy files; the format take upis low and the industry does not produce compatible software; software which supports the format may be bought by a competitor and withdrawn from the market place. Withoutthe intervention of digital preservation techniques the information contained will no longerbe accessible.

## 8.6 BENEFITS OF DIGITAL PRESERVATION

There are some benefits of DP are given below:

1. To develop a digital preservation strategy.
2. To plan coherent digital preservation programmes.
3. To ensure and reinforce accountability.
4. To demonstrate that such funds can and will be used responsibly and consistently.
5. To ensure digital materials available for current and future use.
6. To define the significant properties that need to be preserved for particular classes resources.
7. To assist agencies in designing digitization programmes.
8. To provide a comprehensive statement on the digital preservation.
9. To provide security measures that ensures the protection of digital materials during use.

## 8.7 DIGITAL PRESERVATION STRATEGIES

In 2006, the Online Computer Library Center developed a four-point strategy for the long-term preservation of digital objects that consisted of:

- Assessing the risks for loss of content posed by technology variables such as commonlyused proprietary file formats and software applications.

- Evaluating the digital content objects to determine what type and degree of formatconversion or other preservation actions should be applied.

- Determining the appropriate metadata needed for each object type and how it is associatedwith the objects.
- Providing access to the content.

There are several additional strategies that individuals and organizations may use to actively combat the loss of digital information.

### 8.7.1 Refreshing

Refreshing is the transfer of data between two types of the same storage medium so there are no bitrot changes or alteration of data. For example, transferring census data from an old preservation CD to a new one. This strategy may need to be combined with migration when the software or hardware required to read the data is no longer available or is unable to understand the format of the data. Refreshing will likely always is necessary due to the deterioration of physical media.

### 8.7.2 Migration

Migration is the transferring of data to newer system environments (Garrett et al., 1996). This may include conversion of resources from one file format to another (e.g., conversion of Microsoft Word to PDF or Open Document) or from one operating system to another (e.g., Windows to GNU/Linux) so the resource remains fully accessible and functional. Two significant problems face migration as a plausible method of digital preservation in the long terms. Due to the fact that digital objects are subject to a state of near continuous change, migration may cause problems in relation to authenticity and migration has proven to be time-consuming and expensive for "large collections of heterogeneous objects, which would need constant monitoring and intervention.

### 8.7.3 Replication

Creating duplicate copies of data on one or more systems is called replication. Data that exists as a single copy in only one location is highly vulnerable to software or hardware failure, intentional or accidental alteration, and environmental catastrophes like fire, flooding, etc. Digital data is more likely to survive if it is replicated in several locations. Replicated data may introduce difficulties in refreshing, migration, versioning, and access control since the data is located in multiple places.

### 8.7.4 Emulation

Emulation is the replicating of functionality of an obsolete system. According to van der Hoeven, "Emulation does not focus on the digital object, but on the hard- and software environment in which the object is rendered. It aims at recreating the environment in which the digital object was originally created." Examples are having the ability to replicate or imitate another operating system. Examples include emulating an Atari 2600 on a Windows system or emulating WordPerfect 1.0 on a Macintosh. Emulators may be built for applications, operating systems, or hardware platforms. Emulation has been a popular strategy for retaining the functionality of old video game systems, such as with the MAME project. The feasibility of emulation as a catch-all solution has been debated in the academic community (Granger, 2000).

Raymond A. Lorie has suggested a Universal Virtual Computer (UVC) could be used to run

any software in the future on a yet unknown platform. The UVC strategy uses a combination of emulation and migration. The UVC strategy has not yet been widely adopted by the digital preservation community.

Jeff Rothenberg, a major proponent of Emulation for digital preservation in libraries, working in partnership with Koninklijke Bibliotheek and National Archief of the Netherlands, developed a software program called Dioscuri, a modular emulator that succeeds in running MS-DOS,WordPerfect 5.1, DOS games, and more.

Another example of emulation as a form of digital preservation can be seen in the example of Emory University and the Salman Rushdie's papers. Rushdie donated an outdated computer to the Emory University library, which was so old that the library was unable to extract papers from the harddrive. In order to procure the papers, the library emulated the old software system and was able to take the papers off his old computer.

### 8.7.5 Encapsulation

This method maintains that preserved objects should be self-describing, virtually "linking content with all of the information required for it to be deciphered and understood". The files associated with the digital object would have details of how to interpret that object by using "logical structures called "containers" or "wrappers" to provide a relationship between all information components that could be used in future development of emulators, viewers or converters through machine readable specifications. The method of encapsulation is usually applied to collections that will go unused for long periods of time·

## 8.8 PERSISTENT ARCHIVES CONCEPT

Developed by the San Diego Supercomputing Center and funded by the National Archives and Records Administration, this method requires the development of comprehensive and extensive infrastructure that enables "the preservation of the organisation of collection as well as the objects that make up that collection, maintained in a platform independent form". A persistent archive includes both the data constituting the digital object and the context that the defines the provenance, authenticity, and structure of the digital entities. This allows for the replacement of hardware or software components with minimal effect on the preservation system. This method can be based on virtual data grids and resembles OAIS Information Model (specifically the Archival InformationPackage).

## 8.9 METADATA ATTACHMENT

Metadata is data on a digital file that includes information on creation, access rights, restrictions, preservation history, and rights management. Metadata attached to digital files may be affected by file format obsolescence. ASCII is considered to be the most durable format for metadata because it is widespread, backwards compatible when used with Unicode, and utilizes human-readable characters, not numeric codes. It retains information, but not the structure information it is presented in. For higher functionality, SGML or XML should be used. Both markuplanguages are stored in ASCII format, but contain tags that denote structure and format.

A digital preservation strategy is the broad approach adopted by an organisation to ensure that the content of digital records remains in a usable form over time3. The strategy works by identifying specific triggers which will determine when digital preservation activitiesmay take place. For example:

| Function | Policy | Strategy |
|---|---|---|
| Ownership for digital preservation | “ | |
| Strategic Alignment of digital preservation | “ | |
| Coverage of digital preservation | “ | |
| Roles and Responsibilities | “ | |
| Relationship to other policies | “ | |
| Relationship to other documentation | “ | |
| Type of storage for digital records | | “ |
| Accepted / Preferred formats | | “ |
| Triggers for migration | | “ |
| Types of migration | | “ |
| Triggers for normalisation | | “ |
| Types of normalisation | | “ |

## 8.10 DIGITAL PRESERVATION REPOSITORY ASSESSMENT AND CERTIFICATION

A few of the major frameworks for digital preservation repository assessment and certification are described below. A more detailed list is maintained by the U.S. Center for Research Libraries.

## 8.11 SPECIFIC TOOLS AND METHODOLOGIES

### *8.11.1 TRAC*

In 2007, CRL/OCLC published Trustworthy Repositories Audit & Certification: Criteria & Checklist (TRAC), a document allowing digital repositories to assess their capability to reliably store, migrate, and provide access to digital content. TRAC is based upon existing standards and best practices for trustworthy digital repositories and incorporates a set of 84 audit and certification criteria arranged in three sections: Organizational Infrastructure; Digital Object Management; and Technologies, Technical Infrastructure, and Security.

TRAC “provides tools for the audit, assessment, and potential certification of digital repositories, establishes the documentation requirements required for audit, delineates a process for certification, and establishes appropriate methodologies for determining the soundness and sustainability of digital repositories”.

### *8.11.2 DRAMBORA*

Digital Repository Audit Method Based On Risk Assessment (DRAMBORA), introduced by the Digital Curation Centre (DCC) and Digital Preservation Europe (DPE) in 2007, offers a methodology and a toolkit for digital repository self-assessment.

The DRAMBORA process is arranged in six stages and concentrates on evaluation of likelihood and potential impact of risks on the repository. The auditor is required to describe and document the repository’s role, objectives, policies, activities and assets, in order to identify and assess the risks associated with these activities and assets and define appropriate measures to manage them.

### *8.11.3 European Framework for Audit and Certification of Digital Repositories*

The European Framework for Audit and Certification of Digital Repositories was defined in a memorandum of understanding signed in July 2010 between Consultative Committee for Space Data Systems (CCSDS), Data Seal of Approval (DSA) Board and German Institute for Standardization (DIN) "Trustworthy Archives – Certification" Working Group.

The framework is intended to help organizations in obtaining appropriate certification as a trusted digital repository and establishes three increasingly demanding levels of assessment:

1. Basic Certification: self-assessment using 16 criteria of the Data Seal of Approval (DSA).
2. Extended Certification: Basic Certification and additional externally reviewed self-audit againstISO 16363 or DIN 31644 requirements.
3. Formal Certification: validation of the self-certification with a third-party official audit basedon ISO 16363 or DIN 31644.

### *8.11.4 Nestor Catalogue of Criteria*

A German initiative, *nestor* (the Network of Expertise in Long-Term Storage of Digital Resources) sponsored by the German Ministry of Education and Research, developed a catalogueof criteria for trusted digital repositories in 2004. In 2008 the second version of the document was published. The catalogue, aiming primarily at German cultural heritage and higher education institutions, establishes guidelines for planning, implementing, and self-evaluation of trustworthy long-term digital repositories.

The *nestor* catalogue of criteria conforms to the OAIS reference model terminology and consists of three sections covering topics related to Organizational Framework, Object Management, and Infrastructure and Security.

### *8.11.5 PLANETS Project*

In 2002 the *Preservation and Long-term Access through Networked Services* (PLANETS) project, part of the EU Framework Programmes for Research and Technological Development 6, addressed core digital preservation challenges. The primary goal for *Planets* was to build practical services and tools to help ensure long-term access to digital cultural and scientific assets. The outputs of the project are now sustained by the follow-on organisation, the Open Planets Foundation.

### *8.11.6 PLATTER*

Planning Tool for Trusted Electronic Repositories (PLATTER) is a tool released by Digital Preservation Europe (DPE) to help digital repositories in identifying their self-defined goals and priorities in order to gain trust from the stakeholders.

PLATTER is intended to be used as a complementary tool to DRAMBORA, NESTOR, and TRAC. It is based on ten core principles for trusted repositories and defines nine Strategic Objective Plans, covering such areas as acquisition, preservation and dissemination of content, finance, staffing, succession planning, technical infrastructure, data and metadata specifications, and disaster planning. The tool enables repositories to develop and maintain documentation requiredfor an audit.

### *8.11.7 Audit and Certification of Trustworthy Digital Repositories (ISO 16363)*

Audit and Certification of Trustworthy Digital Repositories (ISO 16363:2012), developed by the Consultative Committee for Space Data Systems (CCSDS), was approved as a full internationalstandard in March 2012. Extending the OAIS Reference Model and based largely on the TRAC checklist, the standard is designed for all types of digital repositories. It provides a detailed specification of criteria against which the trustworthiness of a digital repository should be evaluated.

The CCSDS Repository Audit and Certification Working Group has also developed and submitted for approval a second standard, Requirements for Bodies Providing Audit and Certification of Candidate Trustworthy Digital Repositories (ISO 16919), that defines the external auditing process and requirements for organizations responsible for assessment and certification of digital repositories.

## 8.12 SELF ASSESSMENT QUESTIONS

1. What is digital preservation?
2. Write definition of digital preservation.
3. Explain the benefits of digital preservation.

## 8.13 REFERENCES


1. D. Woodyard, Digital Preservation: The Australian Experience, Proc. Third Conf. Digital Library: Positioning the Fountain of Knowledge, Malaysia (2000), http://www.nla.gov.au/nla/staffpaper/dw001004.html.
2. http://ip.org.au/digital-preservation/
3. D. M. Levy, Heroic Measure: Reflections on the Possibility and Purpose of DigitalPreservation, Proc. Conf. ACM Digital Libraries, Pittsburgh (1998) pp. 152-161.
4. Digital Preservation Policy Tool
5. http://www.erpanet.org/guidance/docs/ ERPANETPolicyTool.pdf
6. ALA (American Library Association) (2007). Definitions of digital preservation. Chicago: American Library Association. Available a t:http://www.ala.org/ala/mgrps/divs/alcts/resources/preserv/defdigpres0408.pdf

1. CCSDS (Consultative Committee for Space Data Systems) (2002). Reference Model for an Open Archival Information System (OAIS). Blue Book, Issue 1. Washington, DC (US): CCSDS Secretariat, January 2002. Technical report. CCSDS 650.0-B-1. Recommendation for Space Data System Standards. Available at:
http://public.ccsds.org/publications/archive/650x0b1.pdf.
2. Research Councils UK (2008). Code of Conduct and Policy on the Governance of Good Research Conduct: Integrity, Clarity, and Good Management. Public Consultation Document. July – October 2008. Available at:http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/reviews/grc/consultation.pdf
3. VERHEUL, Ingeborg (2006). Networking for Digital Preservation: Current Practice in 15 NationalLibraries. Munchen: K.G. Saur.
4. Day, Michael. "The long-term preservation of Web content". Web archiving (Berlin: Springer,2006), pp. 177-199. ISBN 3-540-23338-5.
5. Online Computer Library Center, Inc. (2006). OCLC Digital Archive Preservation Policy andSupporting Documentation, p. 5
6. http://www.ijdc.net/index.php/ijdc/article/view/50/35
7. Rothenberg, Jeff (1998). *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation*. Washington, DC, USA: Council on Library and InformationResources. ISBN 1-887334-63-7.
8. Lorie, Raymond A. (2001). "Long Term Preservation of Digital Information". *Proceedings of the 1st ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '01)*. Roanoke, Virginia, USA.pp. 346–352.
9. Hoeven, J. (2007). "Dioscuri: emulator for digital preservation". *D-Lib Magazine* **13** (11/12).doi:10.1045/november2007-inbrief.
10. http://marbl.library.emory.edu/innovations/salman-rushdie
11. Digital Preservation: Planning, Process and Approaches for Libraries Teena KapoorJaypee Institute of Information TechnologyA-10, Sector-62, Noida UP
12. SOLUTIONS WALKTHROUGH REPORT José Miguel Araújo Ferreira Department ofInformation Systems University of Minho 4800-058 Guimarães, Portugal
13. Moore, Reagan W., Andre Merzky. Persistent Archive Research Group. Dec. 25, 2003.
14. NISO Framework Advisory Group. (2007). A Framework of Guidance for Building GoodDigital Collections, 3rd edition, p. 57,
15. National Initiative for a Networked Cultural Heritage. (2002). NINCH Guide to Good Practicein the Digital Representation and Management of Cultural Heritage Materials
16. "Center for Research Libraries - Other Assessment Tools". Retrieved Sept. 6, 2012.
17. OCLC and CRL (2007). "Trustworthy Repository Audit & Certification: Criteria & Checklist".Retrieved April 16, 2012.
18. Phillips, Stephen C (2010). "Service level agreements for storage and preservation, p.13.". Retrieved May 1, 2012.

19. Ball, Alex (2010). ”Preservation and Curation in Institutional Repositories (version 1.3)”. Edinburgh, UK: Digital Curation Centre. p. 48. Retrieved June 24, 2012.
20. APARSEN Project (2012). *Report on Peer Review of Digital Repositories*. p. 10. Retrieved October 8, 2012.
21. Dobratz, Susanne; Schoger, Astrid (2007). “Trustworthy Digital Long-Term Repositories: The Nestor Approach in the Context of International Developments”. *Research and Advanced Technology for Digital Libraries*. Springer Berlin / Heidelberg. pp. 210–222. ISBN 978-3-540-74850-2.
22. Horstkemper, Gregor; Beinert, Tobias; Schrimpf, Sabine (2009). ”Assessment of Trustworthiness of Digital Archives”. *Proceedings of the Sino-German Symposium on Development of Library andInformation Services*. pp. 74–75. Retrieved October 2, 2012.
23. ”Planets project”. *website*. 2009. Retrieved 7 December 2011.
24. “The Open Planets Foundation”. *website*. 2010. Retrieved 7 December 2011.
25. CCSDS (2011). ”Audit and Certification of Trustworthy Digital Repositories, Recommended Practice”. *CCSDS 652.1-M-1. Issue 1*. Washington, DC: CCSDS, September 2011. pp. 1–1. Retrieved October 10, 2012.
26. Ruusalepp, Raivo; Lee, Christopher A.; vander Werf, Bram; Woolard, Matthew (2012). “Standards Alignment”. In McGovern, Nancy Y. *Aligning National Approaches to Digital Preservation*. Atlanta,GA: Educopia Institute. pp. 115–165 [124]. ISBN 978-0-9826653-1-2.

LESSON-9

# DIGITAL ARCHIVING: NEEDS AND STRATEGIES

**OBJECTIVE**

After reading this chapter, students will be able to understand:

- Basic fundamental concepts of digital archiving
- Common issues of digital archiving
- Principles of digital archiving

**STRUCTURE**



## 9.1 INTRODUCTION

An archive is an accumulation of historical records, or the physical place they are located. Archives contain primary source documents that have accumulated over the course of an individual

or organization's lifetime, and are kept to show the function of that person or organization. Professional archivists and historians generally understand archives to be records that have been naturally and necessarily generated as a product of regular legal, commercial, administrative or social activities. They have been metaphorically defined as "the secretions of an organism" rather than those that have been consciously written or created to communicate a particular message to posterity.

In general, archives consist of records that have been selected for permanent or long-term preservation on grounds of their enduring cultural, historical, or evidentiary value. Archival records are normally unpublished and almost always unique, unlike books or magazines for which many identical copies exist. This means that archives (the places) are quite distinct from libraries with regard to their functions and organization, although archival collections can often be found withinlibrary buildings.

## 9.2 DIGITAL ARCHIVE

Archaeology is in a special position with respect to archiving because archaeological fieldwork, which creates archaeological data, also destroys the primary in situ archaeological evidence itself. Increasingly, the digital record may be the only source of information about archaeological research materials. It is essential, therefore, that the digital records that describe archaeological resources be made accessible and that their preservation be ensured. Providing for the accessibility of archaeological data and its long-term preservation are the goals of digital archiving.

Digital archiving is different from traditional archiving. Traditional archiving practice seeksto preserve physical objects (e.g., artefacts, samples, paper, photographs and microfilm) that carry information. Digital archiving seeks to preserve the information regardless of the media on which that information is stored. Computer disks and other magnetic and optical media degrade, and the information on them is lost unless it has been moved to other media. Software and hardware change rapidly: the physical media on which digital data are stored are impermanent. Other methods are necessary to ensure wide access to and long-term preservation of digital data.

## 9.3 PRINCIPLES OF ARCHIVING DIGITAL DATA

The points below present an outline of the key issues to be considered when creatinga digital archive.

- Ensure that existing digital data are safeguarded and deposited in an appropriate digital archive.
- When creating a new digital archive, ensure that it conforms to existing standards and guidelines on how data should be structured, preserved and accessed.
- All digital archives should ideally be deposited in a digital archiving facility or collections repository where they can be properly accessed, curated, and maintained for the future.
- The key to successful digital archiving is thorough documentation of the data, how they were collected, what standards were used to describe them and how they have been managed since collection.
- If there are concerns that some data (e.g., specific site location information) needs to be kept confidential (as required by the Archaeological Resource Protection Act (ARPA) in the US), a means of easily separating these data from non-confidential data must be developed

for reports, analytical datasets, and for displaying site locations on maps. It is also essential that this process is documented and deposited as part of the archive.

- There is generally no need to preserve interim versions of final digital files. Exceptions to this include interim datasets where either data or text is subsequently discarded or decimated to final publication. These issues are discussed in the later section on Preservation Intervention Points.

- Data already held safely in paper archives do not need to be digitized, except to providea digital security copy or online access to the data. When digitizing or scanning from paper records, do not automatically discard the paper originals when complete. Offer them to relevant documentary archives.

- Although the digital, paper and archaeological resource archives may be dispersed, the integrity of the complete archive must be ensured by cross-referencing between physical collections and digital records.

In accordance with the principles presented above, digital archives should at least provide an index to archaeological sites, finds and paper archives and at best provide access to digital recordsof data, material, documentation, interpretation and analyses. It is recommended that the collectionor creation of digital datasets be planned at the outset of a project and incorporated into project scopes of work and specifications. It is recognized that funding agencies must acknowledge such requirements if widespread implementation is ever to be achieved.

## 9.4 ISSUES IN COMMON ACROSS THE DOMAIN

The Archives Domain has identified the following issues that remain in need of attention. They are issues common to the whole cultural community.

### *9.4.1 Issue 1 – Loss due to IT obsolescence*

Though a major aspiration is that important digital collections last over time, digital objectsand particularly digital records can easily be 'lost' and a 'digital black hole' covering several decades can easily eventuate. A key challenge of the digital age is to find ways of providing continuing access to digital objects that depend on outdated technology. Digital objects/records can be lost due to various types of IT obsolescence.

Physical storage media for digital records have been found to be unstable, fragile, prone to corruption and relatively short-lived. Thus most media requires frequent 'refreshing', ie copying onto new media. Also the standard media of the 1990s, the floppy disk, is mostly being superseded by CDs and DVDs, which are also on the way to becoming obsolete. Currently, whatever physical medium is selected to store digital objects, a migration schedule is necessary to ensure that the records on that medium are not lost.

Hardware needed to store and access digital records also becomes obsolete over time as newer, faster and larger capacity models come onto the market. Most new hardware has a usable life of only three to five years. Software applications needed to access the digital objects are also quickly becoming obsolete. Software includes both the applications needed to make and read the record, and the operating system on which that software is run. Software upgrades and new releases are as frequent as the hardware and storage medium releases and pose an equal challenge.

Along with the need to convert, refresh or migrate digital objects, there is a need to keep audit trails of any transformations, to track the workflow and to identify any changes to the integrityor authenticity of the record. Without such audit checks, data could be lost.

Presently, no archive can cope with the variety of storage formats that may come into its custody. Hence a variety of solutions have been trialled and put in place by the Domain. The Domain acknowledges that for safe long-term preservation we cannot rely on present storage media options. Other management strategies need to come into play.

### *9.4.2 Issue 2 – Loss due to lack of capture and poor practice by creators*

Loss can also occur as a result of poor practice at an organisational level. There is a perpetual risk that important digital objects or records are not captured – that they are left to sit on someone's hard drive, CD, DVD, or other storage media and not captured into a larger organisational system. Even when digital objects are captured, they may not be captured into sustainable and appropriate management systems. The Archives Domain has promoted the 'capture at creation' and 'good recordkeeping message' to alleviate this concern, and has published guidelines to help implement best practice.

The Domain remains concerned that poor digital recordkeeping may result in loss of records and possible loss of record authenticity, hence resulting in the loss of the evidentiary value of a digital collection. Furthermore, unless digital records are captured and managed in appropriate and adequate management systems, digital archiving will become unnecessarily difficult. The guidelines the Domain has produced regarding the capture and management of digital records are applicable to digital objects generally and so can be used, with slight modifications, to suit clients in other cultural domains.

### *9.4.3 Issue 3 – Digital authenticity and integrity*

It is a common requirement that digital objects retain their authenticity and integrity over time, for both historical as well as for accountability and evidentiary reasons. This is often backed up by legislative requirements, particularly so in the Archives Domain.

To ensure integrity, and to prove the authenticity of digital objects within a collection overtime, the essential characteristics of that object or record need to be captured and preserved. In this way, the context in which the digital object was created and used is captured along with the content.

Digital collections also need to be kept safe from unauthorised access, alteration or deletion. Data security, data integrity and audit requirements are all required components of maintaining accountability and integrity.

### *9.4.4 Issue 4 – Volume of digital collections*

The Archives Domain anticipates that it will be dealing with, potentially, pet bytes of digital data in the near future (a pet byte is two orders of magnitude larger than a gigabyte). The cultural community as a whole will similarly be dealing with pet bytes of digital data in the near future andso the issue of volume is a common one.

*In particular, there will be a need to:*

• sort the chaff from the wheat, i.e. to use evaluation and appraisal tools to determine what's worth keeping

• ensure that objects can be quickly retrieved from within large and distributed digital collections, ie to employ a metadata management framework.

The Archives Domain is used to dealing with large volumes of physical material and has strategies and robust systems in place that allow it to cope with such volumes (i.e. its appraisal methodologies, metadata schemas, and intellectual control mechanisms). In 2005, the ten major archival institutions collectively held almost 550 kilometres of records in their custody, equivalent to more than 68 million items (see CAARA statistics and methodology for determining item numbersat www.caara.org.au).

The Archives Domain is now transferring that expertise to managing large volumes of digital objects. For example, the National Archives of Australia now has more than 500,000 items – close to 15 million images – available online. Other jurisdictions are building up similar volumes of digitalcollections.

The expertise archivists have developed in dealing with large volumes of physical records stands them in good stead for dealing with the volume of digital records expected over the next few years.

### *9.4.5 Issue 5 – Evaluation/appraisal methodologies*

Not all digital objects need to be kept for long periods of time, let alone permanently. The ability to guide decisions on what to keep and what not to keep, through the appraisal process, has been a core function of the Archives Domain and it has methodologies and guidelines in place on how to embark on an appraisal process.

***Appraisal***

The process of evaluating business activities to determine which records need to be captured and how long the records need to be kept to meet business needs, the requirements of organisational accountability, and community expectations.

To cope with large volumes of digital objects, excellent metadata frameworks and management systems need to be in place, to enable the retrieval of digital records/objects from amongst large digital collections within acceptable time frames. Institutions and archival institutions in particular, need to have intellectual and physical systems designed, and in place, well before they can effectively take into custody large-scale digital object/record transfers. It is better for organisations to be prepared in advance, rather than try to re-engineer themselves having already created or assumed custody of large digital collections.

### *9.4.6 Issue 6 – A digital archives framework*

There is more to ensuring the longevity of a digital collection than just preserving the digital objects within it. Ensuring the long-term viability of digital collections is what the Archives Domainmeans by the term ‘digital archiving’.

***Digital archiving***

Digital archiving covers the identification, appraisal, description and tagging, storage, preservation, management and retrieval of digital records, including all of the policies, guidelines and systems associated with those processes, so that the logical and physical integrity of the records is securely maintained over time.

Digital archiving covers the spectrum of laws, policies, procedures and methodologies required to address the 'whole of life' issues of a digital object or record. Digital archiving subsumes within it the critical function of digital preservation. Though this approach to digital archiving is particularly pertinent to the Archives Domain, as their digital collections are primarily made up of 'born digital' records, the digital archiving approach is one that has valuable application across many domains and the tools and skills developed can be shared with, and used by, other domainsas appropriate.

### *9.4.7 Issue 7 – Intellectual control and description frameworks*

Intellectual control and description schemas and frameworks include appropriate metadata models and standards that ensure:

- Resource discovery
- Intellectual control and description
- Administrative control
- Preservation control.

A number of such metadata schemas are already in use across the cultural domain. Othersare Domain-specific. As we operate more collaboratively in an online environment, and make our collections available via the internet through distributed databases, the issue of interoperability comes to the fore – we need eo either match or map metadata to enable metadata harvesting and federated searching.

### *9.4.8 Issue 8 – Sustainable preservation solutions*

From the Archives Domain's point of view, digital preservation relies on good digital recordkeeping. Good recordkeeping and archival systems provide access to complete, reliable and authentic records into the future. In other words, the Domain takes a 'whole of life' approach to digital preservation. This emphasis may be less relevant where an organisation is creating a collection via a digitisation program, but the concept of 'capture at creation' would be a worthwhileemphasis of any preservation solution, whether the records are born digital or digital surrogates.Of particular concern to the Archives Domain is the ability of small to medium-sized archival programs to implement sustainable digital preservation solutions. At the moment digital preservationis seen as a prohibitively expensive activity and as such the exclusive preserve of the large government archives. There is an urgent need to develop simple and scalable digital preservation tools and strategies that can be deployed in all archival settings. The innovations developed in thelarger institutions need to be repackaged in 'lite' form with supporting training programs so that they can spread across the entire Domain.

### *9.4.9 Issue 9 – Access and delivery*

Australia's cultural institutions are responsible not only for accumulating and maintaining their individually unique collections, but also for increasing public access to these collections. A critical issue, therefore, is that appropriate accessibility infrastructure is in place. Present and future generations need:

- The ability to readily access and interpret digital collections over time
- User-friendly online interfaces to digital collections
- Access management and delivery systems to support the quick and seamless retrieval and delivery of digital objects

• Controlled access mechanisms, for privacy, security or copyright reasons.

### *9.4.10 Issue 10 – Visibility of digital collections*

At present, cultural digital collections are generally made visible on individual organisation'swebsites, and either accessed there directly, or through cultural gateways or portals. Visibility of collections and their objects, via internet search providers, is dependant on the index ranking givento an organisation's website by that provider.

A companion issue is the need to increase the public's awareness of the many cultural Domain digital collections available to them. An ongoing proactive marketing strategy is required toincrease the public's awareness of the range and utility of digital collections.

### *9.4.11 Issue 11 – Sustainable management systems*

Intellectual and physical management systems that are employed to store, manage, retrieve and deliver digital objects should, ideally, be based on open standards to ensure sustainability of the systems over time. Open standards exist for format types, for operating systems, disk drives and so on. If proprietary systems are used, digital objects could be lost or rendered uninterruptable over time. The Archives Domain is advocating that digital archiving solutions be based on open standards such as the Open Archival Information System (OAIS) Reference Model ('Blue Book' digital preservation framework – ISO 14721: 2003).

### *9.4.12 Issue 12 – Shared and trusted digital repository networks*

The Australian Partnership for Sustainable Repositories (APSR) projects (http:// www.apsr.edu.au/) are developing sustainable repository frameworks, interoperability capabilities, metadata harvesting mechanisms, standards development, and so on. Another group focusing on the requirements of trusted digital repositories is MAGDIR – Working Group on Management of Australian Government Digital Information Resources, which includes in its agenda:

- Repository functions and processes
- The organisational structure and governance requirements needed to support sustainable trusted digital repositories
- Technology requirements and sustainable system infrastructures and strategies
- Usability of information by designated target communities.

The Archives Domain supports the development of networks of shared and trusted digital repositories to provide 'single point' access to a range of digital collections hosted by cultural and other organisations in Australia.

There are, however, many different approaches that research institutions and cultural organisations can take. The Archives Domain focuses on a digital archive, as opposed to a trusted repository. Central to a digital archive is a robust and sustainable preservation functionality, which assures the longevity and integrity of the record.

### *9.4.13 Issue 13 – Skilled staff*

Digital archiving is a relatively new 'archival function' and for that reason there are few people at present who have the particular combination of skills required for the job (ie a combination of traditional archival and recordkeeping skills, as well as acute awareness of the role technology

plays in digital object management, ICT experience, and skills in system stewardship). There is an urgent need for training in all areas associated with the creation, capture, management, preservation and accessibility to digital collections. Having inadequate numbers of trained and skilled staff can seriously hinder the development of the digital archiving agendas.

There is also a need to embed into any skills development and training programs the conceptual difference between digital archiving and digital preservation so as to avert any misconception that 'digital preservation' is the total solution. Training programs need to explain the need for, and role of, 'digital archiving'. People need to understand that digital preservation is a vital subcomponent of digital archiving and that preservation on its own – without the policies and systems of digital archiving will not deliver a long-term solution.

### *9.4.14 Issue 14 – Adequate resourcing*

Resources are required for research and development into the following aspects of digital archiving:

- Framework design
- Implementation
- Application of descriptive and intellectual control metadata to large volumes of digital objects
- Digital preservation
- Appropriate access and delivery mechanisms.

Additional costs are associated with the initial development and implementation of digital archiving solutions, but possibly also with their sustainable maintenance. The extent of resourcingneeded for this may be Domain specific and will depend on the:

- Quantity of digital objects/records that require preservation
- Range and complexity of formats and content
- Extent of control needed over the digital objects
- Standard of access that is required
- Degree of standardization among the objects in a collection.

Government and other funding bodies have not, as yet, understood the consequences of not adequately funding the research and development needed to achieve sustainable digital archiving solutions. There is obviously a great need to make funding or sponsoring organisations aware of the range of issues involved in digital archiving and of the critical nature of having effective solutions in place. Though a number of substantial, one-off allocations have been made for digitisation programs, most organisations need to fund their digital records research and development, and prototyping of digital archiving solutions, from present internal funds. This is not sustainable, giventhe time critical nature of the problem.

### *9.4.15 Issue 15 – Copyright, intellectual property and fair use*

The Copyright Amendment (Digital Agenda) Act 2000 has updated copyright law for the digital environment. The introduction of this law has led to a heightened awareness of the possibility of copyright infringement as well as an awareness of the administrative cost associated with negotiating and securing copyright licences for digitised and born digital objects. The aim of the legislation is to ensure that cultural institutions: can access, and promote access to, copyright material in the online environment on reasonable terms, including having regard to the benefits of public access to the material and the provision of adequate remuneration to creators and investors. Copyright Amendment (Digital Agenda) Act 2000 (Cth), s. 3

An issue is still the extent to which copyright reforms are applied 'in situ' and the extent to which they promote or impede cultural institutions' ability to use digital technologies to

achieve their public interest missions. Further changes to copyright law are on the way. A draft exposure Bill on reforms to federal copyright law is presently being prepared by the Attorney-General's Department, with the aim of making the copyright law fairer for consumers.

Another issue facing the cultural community is licensing arrangements for copyright and the increasing demand for commercial content development.

### *9.4.16 Issue 16 – Organisational re-engineering*

The traditions of stewardship and best practice that cultural organisations have used in a print or analogue-based environment are not fully adequate in a digital environment. Many processes, workflows and management approaches have had to be re-examined and often re-engineered. There is a need to establish new best practices as well as new organisational structures that best meet the requirements of digital archiving and digital preservation.

### *9.4.17 Issue 17 – Digitisation standards*

Cultural organisations have and are continuing to invest considerable effort into digitisation programs and many are using or have set quality standards to ensure consistency of approach, the ability to exchange, deliver products and to ensure at the start of a digitisation project that thebest strategies are adopted to ensure longterm viability of image files. Digitisation guidelines prepared thus far often cover the technical, metadata and administrative components associated with digitisation.

## 9.5 SELF ASSESSMENT QUESTIONS

1. What is digital Archive?
2. Enumerate the issues involved in digital archiving
3. Explain the principles of digital archiving

## 9.6 REFERENCES

1. Glossary of Library and Internet Terms. University of South Dakota Library. Archivedfrom the original on 2009-03-10. Retrieved 30 April 2007.

2. Galbraith, V.H. (1948). Studies in the Public Records. London. p. 3.
3. A Glossary of Archival and Records Terminology. Society of American Archivists.Retrieved 7 December 2012.
4. 'What is Digital Archiving?' Edited by Kieron Niven with contributions by Mason Scott Thompson Archaeology Data Service / Digital Antiquity (2011) Guides to Good Practice

LESSON-10

# OVERVIEW OF MAJOR DIGITAL LIBRARY INITIATIVES

**OBJECTIVE**

After reading this chapter, students will be able to understand:

- Basic fundamental concepts of digital library initiatives
- Importance of digital library initiatives

**STRUCTURE**



## 10.1 INTRODUCTION

Libraries have always strived to collect process and disseminate information. But information today exists in many forms than just as printed matter and this has lead to the evolution of Digital Libraries. The concept behind Digital Libraries (DL) has its roots in what is the philosophy of libraries of disseminating 'knowledge for all'. Digital libraries break the barrier of physical boundaries and strive to give access to information across varied domains and communities. Though the term 'Digital Library' came into being with the popularity of the Web in the early 1990s, the inception ofthese can be traced back to the pre 1990s in projects dealing with automated storage and retrieval of information, library networks and online resource sharing efforts. The major fillip to precipitation of the DL idea is attributed to the overwhelming effect of the WWW in the functioning of every field.The presence of the Net as GUI and ease of use and implementation of linking resources through hyperlinks and making them available through the Hyper Text Transfer Protocol (HTTP), all addeda tremendous impact to everyday use of the Web technology to common activities of most professions. One of the profound areas where the Web had immediate and visible effects is in theonline information services area. There had been models and thought along the lines of networkedinformation for a while before the Net became the *'Web'* but it rather accelerated beyond anyone'simagination and culminated in today's digital libraries.

## 10.2 WHAT IS THE DIGITAL LIBRARY INITIATIVE?

The DLI is a five-year plan to move aggressively, but intelligently, towards the creation of a new library system. That new library is characterized most specifically by its ability to use technology to enhance information services for students and faculty, to support new instructional methodologies,and to improve access to all forms of information. While the new library will have both print and digital materials, the DLI focuses on:

- Developing a user-centered information technology infrastructure across all university libraries
- Designing more effective services capitalizing on technology
- Acquiring, organizing, and disseminating high quality digital content
- Creating new multimedia content with other faculty colleagues
- Experimenting with digital preservation
- Continuously assessing and evaluating the impact of information technology and making appropriate changes in the provision of services and information resources

## 10.3 DIGITAL LIBRARY INITIATIVE

The formal projects of digital library under the Digital Library Initiative (DLI) Phase I (http:// www.dli2.nsf.gov/dlione/), started in 1994 as a joint initiative of the National Science Foundation (NSF), Department of Defense Advanced Research Projects Agency (DARPA), and the National Aeronautics and Space Administration (NASA), in 1994. Phase I involved a total funding of US $ 24 million for a period of four years from 1994 to 1998 to six universities. The intent in the first phase was to concentrate on the investigation and development of underlying technologies for digital libraries. The Initiative targeted research on information storage, searching and access. The goalsfor the phase were set as developing the technologies related to:

- Capturing, categorizing and organizing information
- Searching, browsing, filtering, summarizing and visualization
- Networking protocols and standards

## 10.4 THE MAJOR SIX PROJECTS UNDERTAKEN IN THE DLI (1994- 1998) ARE THE FOLLOWING:

- University of California at Berkeley- Environmental Planning Library and Geographic Information Systems
- University of California at Santa Barbara-Alexandria Digital Library project; Spatiallyreferenced Map Information
- Carnegie Mellon University Project- Informedia Digital Video Library
- University of Illinois at Urbana-Champaign-Federating Repositories of Scientific Literature
- University of Michigan Digital Library Project (UMDL): Intelligent Agents for InformationLocation
- Stanford University-Interoperation Mechanisms Among Heterogeneous Services

Second phase aims at intensive study of the architecture and usability issues of digital libraries including the vigorous research on:

a) Human-centered DL architecture
b) Content and Collections-based DL architecture and
c) Systems-centered DL architecture

## 10.5 TESTBEDS ARE DEVELOPED AT FOLLOWING UNIVERSITIES:

- University of Arizona (High-Performance Digital Library Classification Systems:From Information Retrieval to Knowledge Management)
- University of California Berkeley (Re-inventing Scholarly Information Dissemination andUse)
- University of California Santa Barbara (Alexandria Digital Earth Prototype)
- Carnegie Mellon University (Million books project)
- Columbia University (A Patient Care Digital Library: Personalized Search and Summarization over Multimedia Information )
- Harvard University (Operational Social Science Digital Data Library)
- University of South Carolina (A Software and Data Library for Experiments, Simulations,and Archiving)
- Stanford University (Stanford Digital Library Technologies Project)
- Tufts University (The Perseus Digital Library Project)
- Open Archive Initiative (Digital Library Federation, the Coalition for Networked Information,and from National Science Foundation)
- Gutenberg Project
- NCSTRL (Virginia Tech and Old Dominion University)

This phase will try to link the application of digital libraries, especially, in facilitating teaching and learning processes. The domains of concentration will be science, mathematics, engineering and technology. The types of proposals of interest would be: practical digital library applications for SMET (Science, Mathematics, Engineering and Technology) education, technical studies of digital library capabilities, and general policy studies. One of the major outcomes of second phase is developments such the Open Archived and distributed architecture and theorizing the digital librarymodels.

## 10.6 WHY IS THE DIGITAL LIBRARY INITIATIVE IMPORTANT?

Comprehensive research libraries as we now know them are no longer possible to sustain. The current pricing schemes for both print and electronic information and the exponential growth of knowledge require new models for the development of academic libraries. These new models will increasingly incorporate a growing level of digital content and depend on services enhanced by information technology. They will rely on inter institutional collaboration to develop a tiered approach to information provision: the local, state, regional, national, and international library. And they will influence the development of new pricing and distribution models for print and electronic resourcesthrough experimentation and collective action.

The DLI is a prerequisite for the emerging academic library model. It is the critical infrastructure that will:

- Build a local gateway and archive for scholarly collections and other information resources that will connect digital library gateway to every tier in the new academic library model.
- Create a teaching library that uses web-based and other multimedia modules to help students and others apply critical thinking skills to the identification, filtering, and evaluationof information.
- Offer new kinds of distance and multimedia learning initiatives consistent with university academic standards.
- Foster inter institutional collaboration on joint purchasing of and access to digital informationresources

- Enhance service to the citizens of nation by providing access to databases and other multimedia resources created at information centers.

## 10.7 GLOBAL DL SCENARIO

The early digital library projects and the Digital library initiatives generated interest in the area and several projects were initiated in different countries of the world. The marked difference is in the approach followed by the initiatives. In the U.S. under the digital library initiatives importancewas given to research leading to several breakthroughs in digital library architectures, models and tools. Most digital library projects in other countries however define digital libraries as extension oftraditional library building collections in the electronic forms and providing services.

In Europe there have evolved projects at the European level, national level and local levels. The Telematics for Libraries program of the European Commission (EC) aims to facilitate access to knowledge held in libraries throughout the European Union while reducing disparities between national systems and practices. While it's not exclusively devoted to digital libraries the program covers topics such as networking (OSI–Open Systems Interconnection, Web), cataloging (OPACs), imaging, multimedia, and copy-right, among others many of the 100 or more projects funded by the European Commission do cover issues and activities related to digital libraries (9). In addition there have emerged national digital library initiatives of Russia, Germany, France, Denmark, Sweden, and Spain among others

In UK, the noteworthy efforts in Digital libraries are the ELINOR and the eLib projects (10). The eLib project (http://www.jisc.ac.uk/elib/projects.html) funded by the Joint Information SystemsCommittee (JISC), UK, aims to provide exemplars of good practice and models for well organized, accessible hybrid libraries. The Ariadne magazine (http://www.ariadne.ac.uk/) publishes in depthto the information community at large on progress and developments within the UK Electronic Libraries Programme since its inception.

The Canadian National Library (11) hosts the Canadian Inventory of Digital Initiatives provides descriptions of Canadian information resources created for the Web, including general digital collections, resources centered around a particular theme, and reference sources and databases. In Australia, Australian libraries (at the Federal, State, and University levels) together with commercialand research organizations are supporting a diverse set of Digital Library projects that take on many of technical and related issues. The projects deal both with collection building and servicesand research in terms of metadata. Another main area of focus in digital library retrieval has beenthe subject gateways projects.

In Asia the International conference of Asian Digital Libraries (ICADL, http://www.icadl.org/) is hosted every year by an Asian nation and these provide a forum to publish and contemplate on issues regarding research and developments in the area of digital libraries. In India awarenessof the importance of digital library and electronic information services is spawning, marked by a number of conferences and seminars hosted on the topics. While a national policy on digital library is still pending, a number of individual digital library efforts have emerged. In this context, digital library initiatives in building collections are the nodal agencies working in collaboration with the Carnegie Mellon University, Universal Digital Library project of the US-NSF under and Indo-US Science and Technology Collaboration initiatives and Indian universities participating as membersin the Networked Digital Library of Theses and Dissertations (NDLTD).

In the area of digital library research, Documentation Research and Training, Indian Statistical Institute, researches and implements the technology and methodologies in digital library architecture, multilingual digital information retrieval and related tools and techniques in addition of hosting

workshops to provide training to information professionals in techniques for digital libraries. Other digital library initiatives in Asia are taking shape through the national initiatives such as the Indonesian Digital Library Network (http://idln.lib.itb.ac.id/), the pilot of the Malaysian National Digital library, myLib (http://www.mylib.com.my/), National Digital Library of Korea, (http://www.dlibrary.go. kr/).

## 10.8 DIGITAL LIBRARY RESEARCH TRENDS

The area of digital library has given rise to unprecedented research interest, as the natureof work and functionality of digital library is both complicated and challenging to the academic and research world. Accordingly several research projects have emerged and interesting results have been recorded. Digital libraries are data intensive and hence a major part of the research focuseson software systems that manage the storage and provide access to information. These systems are built by following a typical software engineering lifecycle, with an increasing emphasis on architectural models and components to support the process. One of the important frameworks suggested for naming digital objects and facilitating their access through machine interface was by Kahn and Wilensky (12).

This Repository Access Protocol (RAP) provides an abstract model for the services needed in order to add, modify, or delete records stored in a digital library. Dienst (13) is a distributed digital library based on the RAP model and used initially as the underlying software for the Networked Computer Science Technical Reference Library (NCSTRL). Multiple services are provided as separate modules, communicating using well-defined protocols both within a single system and among remote systems. Other notable pre-packaged systems are E-Prints from the University of Southampton, DSpace from the Massachusetts Institute of Technology & HP alliance team and Greenstone (14) from the University of Waikato. All these provide the ability for users to manage and access collections of digital objects. Software agents and mobile agents have been applied to digital libraries to mediate with one or more systems on behalf of a user, resulting in an analogue toa distributed digital library.

In the University of Michigan Digital Library Project (15), DLs were designed as collections of autonomous agents that use protocol-level negotiation to perform collaborative tasks. The Stanford InfoBus project (16), Stanford University, adopts an approach for interconnecting systems using distinct protocols for each purpose, with CORBA as the transport layer. CORBA also was used asa common layer in the FEDORA project (17), which defined abstract interfaces to structured digitalobjects.

## 10.9 DIGITAL LIBRARY AS A DISCIPLINE

The question then arises by virtue of the depth and width of the work being carried out on digital libraries, is it emerging as a discipline by itself? Information dissemination in a networked environment started with the inception of the Web but organized and processing information services using the Net as a medium is what culminated into digital libraries in the early 1990s. Since then the interest in the area increased to an unprecedented scale and as discussed earlier there have been a rapid proliferation of research and techniques adopted for work at various stages of Digital Libraries. Accordingly projects like the SciDL project proposal of the Virginia Tech digital library research group emphasize the status of digital libraries as an area of research and study and advocates it as a discipline in itself.

It is further necessary to take stock of digital libraries at this juncture as a highly interdisciplinary involving theories and techniques from Computer Sciences, Library science, Artificial Intelligence, Information Systems and Retrieval before defining the boundaries of this

nascent discipline. However it stands that Digital Libraries are kinds of libraries with collections at the core and services and users and the dissemination level. Hence it is essential to study carefully the applicability and theoretical foundation of library and information sciences to the area ofdigital libraries.

## 10.10 SELF ASSESSMENT QUESTIONS

1. Discuss about digital library initiatives.
2. Enumerate the major digital library initiatives.
3. What is the importance of DLI?

## 10.11 REFERENCES


1. Wells, H.G. (1938). World Brain. Garden City, NY: Doubleday.
2. Salton, G. (1968). Automatic information organization and retrieval. New York: McGraw-Hill.
3. Englebart, D.C. (1963). Conceptual framework for the augmentation of man's intellect. In P.W.Howerton & D.C. Weeks (Eds.), Vistas in Information Handling (pp. 1- 20), Washington,D.C: Spartan Books.
4. Madalli, Devika P (2003). Digital Libraries and Digital Library Initiatives. Digital Library: Theoryand Practice, DRTC, Bangalore.
5. http://www.libraries.rutgers.edu/rul/about/long_range_plan.shtml#3
6. Lagoze, C and Van de Sompel. (2001). The Open Archives Initiative: Building a low barrier interoperability framework. In Joint Conference of Digital Libraries. Virginia Tech, Blacksburg.2001.
7. Goncalves, M , et al. (submitted). Scenarios, Streams, Structures, Spaces (5S): A Formal model for digital libraries. Communicated for publication to ACM Transactions for InformationSystems
8. Dienst overview and introduction. http://www.cs.cornell.edu/cdlrg/dienst/DienstOverview.ht m
9. Fox, E. A, et al. (1997). NDLTD: Networked Digital library of Theses and Dissertations. Retrieved November 20th, 2002 from the World Wide Web: http://www.ndltd.org/
10. Manduca, C. A., etal (2001). Pathways to progress: Vision and plans for developing the NSDL. Retrieved on November 16th 2002, from World Wide Web: http://doclib.comm.nsdlib.org/PathwaysToProgress.pdf
11. Kuny, T. (1998) digital Library Projects: European Commission Telematics for LibrariesProgram. At http://www.nlc-bnc.ca/9/1/p1-245-e.html
12. eLib: The Electronic libraries programme at http://www.ukoln.ac.uk/services/elib/
13. The homepage of national library of Canada at http://www.nlc-bnc.ca/index-e.html
14. Kahn, Robert, and Robert Wilensky (1995), A Framework for Distributed Digital ObjectServices. Available http://www.cnri.reston.va.us/k-w.html
15. Lagoze, C., and J. R. Davis (1995), "Dienst - An Architecture for Distributed Document Libraries", in Communications of the ACM, Vol. 38, No. 4, ACM, p. 47
16. Witten, I. H., R. J. McNab, S. J. Boddie, and D. Bainbridge (2000), "Greenstone: A Comprehensive Open-Source Digital Library Software System", in Proceedings of Fifth ACM Conference of Digital Libraries, San Antonio, Texas, USA, 2-7 June 2000, pp. 113-121.
17. Birmingham, W. P. (1995), "An Agent-Based Architecture for Digital Libraries", in D-Lib Magazine, Vol. 1, No. 1, July 1995. Available http://www.dlib.org/dlib/July95/07birmingham.html

LESSON-11

# DIGITAL LIBRARY INITIATIVE IN INDIA

**OBJECTIVE**

After reading this chapter, students will be able to understand:

- Digital Library Initiative in India
- Some of the important digital library initiatives taken by India.

**STRUCTURE**



## 11.1 INTRODUCTION

Information and Communication Technologies (ICTs) have brought significant changes in all round development of the society through transmission of information. Application of information technology to Library and Information Science has provided wider opportunities in archiving and accessing knowledge in the digitized form besides conservation and preservation of the traditionalknowledge. Digitization of materials will provide enhanced access to the electronic information sources and the users can access the digital content irrespective of time and space boundaries.

In India, digital library initiatives were undertaken initially with a view to preserve the art, culture and heritage of the country. Some special libraries are also engaged in digital library initiatives in a limited way. However, initiatives in academic libraries particularly in the Open Distance Learning Libraries (ODL) are yet to venture into the digitization. As such, this paper proposes to:

- Study the digital library initiatives undertaken so far by the Government and other organizations in India.
- Examine the challenges and problems faced in the digital library initiatives, and
- Propose a digital library initiative for the ODL institutions in India.

## 11.2 DIGITAL LIBRARY INITIATIVE IN INDIA

The concept of digital libraries in India began in the mid 1990s with the spread of information technology, the internet and the support of the Central Government. In 1996, this concept was recognized during the Conference on Digital Libraries organized by the Society of Information Science at Bangalore. Though a few libraries have made attempts earlier in this direction, the digital library initiative in India is still at budding stage.

Majority of the Digital library initiatives were largely confined to limited uses such as subscribing to e-journals, scanning documents and installing them on the intranet. But there is every need for rapid change in this scenario of libraries in India to use the Information Technology (IT) and ICTs which are confined so far to the prestigious National institutes such as the Indian Institutes of Technology (IIT), Indian Institutes of Management (IIM), Indian Institutes of Science (IIS) Research Institutes under the control of NISSAT and some special Libraries. Some government agencies and institutions, mostly in the public sector are also engaged in digitization of their libraries in a limited a way. However, it is evident from the initiatives taken so far in this direction that the great potential of ICTs for developing digital libraries has not yet been fully utilized.

***Some of the important digital library initiatives and programmes initiated across the country are given below:***

## 11.3 ARCHIVES OF INDIAN LABOUR

The Archives of Indian Labour was set up in July, 1998 as a collaborative project of V.V. Giri National Labour Institute and the Association of Indian Labour Historians. The core activities of the archive are Digital Archiving, Research, Collection, Public Interface & Dissemination. It was instituted in order to address the urgent need for preservation of rapidly decaying documents and material on labour and to provide for greater public access to the same, as It was felt that documents and data on Indian Labour are being irretrievably lost due to lack of an organized initiative to preservethese documents in the country. The archive, apart from being a repository of documents also builds collections and initiates research in the field of labour history.

## 11.4 DIGITAL LIBRARY OF INDIA

Digital Library of India (DLI) is the biggest national level digital library initiative in India. It is a part of the Universal Digital Library Project, envisaged by Carnegie Mellon University, USA, which has some other international partners such as China and Egypt. DLI is coordinated by Indian Institute of Science, Bangalore and is supported by Ministry of Communications and Information Technology, Government of India. The Mission is to create a portal for the Digital Library of Indiawhich will foster creativity and free access to all human knowledge. As a first step

in realizing this mission, it is proposed to create the Digital Library of one million books, predominantly in Indian languages, available to everyone over the Internet. This portal will also become an aggregator of all the knowledge and digital content created by other digital library initiatives in India.

## 11.5 DIGITAL LIBRARY OF ART MASTERPIECES

HP Labs, (Hewlett Packard's) announced a pilot project with the Centre for Development of Advanced Computing (CDAC) to digitize part of the art collection in the National Gallery of Modern Art (NGMA). NGMA plans to put up images of the paintings on the net, from which customers can order full-sized prints. The museum will make reproductions on demand on Hp design Jet printers and sell them. Similarly, the Indira Gandhi National Centre for the Arts (IGNCA) has taken up multimedia projects for the digitization of traditional artwork and artifacts that will be made availableon the web. The digitization of "Geet Govinda," an important classic of Indian literature, is one oftheir successful ventures.

## 11.6 DOWN THE MEMORY LANE

The National Library of India has initiated in late 1990s a digitization programme, known as Down the Memory Lane', to digitize rare books, manuscripts and other resources from its collection.The English books that were published prior to 1900 and Indian books published before 1920 were taken into consideration. Similarly, the Central Secretariat Library has initiated a programme to digitize government publications like, Gazette of India, Commission & Committee Reports, AnnualReports of the Ministries.

## 11.7 INDIAN NATIONAL DIGITAL LIBRARY IN ENGINEERING SCIENCE& TECHNOLOGY

The Ministry of Human Resource Development (MHRD) has set-up the Indian National Digital Library in Engineering Sciences and Technology (INDEST) Consortium on the recommendation made by the Expert Group. INDEST Consortium is the most ambitious initiativetaken up so far in the country. It welcomes other institutions to join and offers highly discounted rates of subscription and better terms of agreement with the publishers. INDEST Consortium presently include ACM Digital Library, ASCE Journals, ASMe Journals, Capitaline, Euromonitor (GMID), IEL Online, Indian Standards, Nature, ProQuest Science, Sciencedirect, Springerlink andbibliographic databases of Compendex, Inspec and MathSciNet.

### 11.7.1 KALASAMPADA

Indira Gandhi National Centre for Arts (IGNCA), established a Digital Library, known as "Kalasampada", (Digital Library Resource for Indian Cultural Heritage). It includes non-print as well as printed materials. The users will have access to the highly researched publications of the IGNCA from a single window. The integration of multimedia computer technology and software provides anew dimension in the study of the Indian Art and Culture.

### 11.7.2 KHUDA BAKSH ORIENTAL PUBLIC LIBRARY

The Khuda Baksh Oriential Public Library has initiated digitization of Arabic and Persian manuscripts of the medieval India. It is one of the Oriental Libraries having rich collection of Persian, Arabic, Urdu and other languages manuscripts.

### 11.7.3 MOBILE DIGITAL LIBRARY (DWARE DWARE GYAN SAMPADA)

For spreading and promoting literacy among the common citizens, Internet enabled Mobile Digital Library was brought to use. Mobile Van with satellite connection for Internet connectivity isused. The van is also fitted with printer, and binding machine for providing bound books to the enduser from a single point. This is a product from C-DAC-ERDC Noida.

### 11.7.4 MUKHTABODHA DIGITAL LIBRARY AND ARCHIVING PROJECT

The Mukhtabodha Project is an attempt to create a digital library in Indian languages specially the ancient texts on palm leaves, birch tree barks etc.

### 11.7.5 NATIONAL INSTITUTE OF ADVANCED STUDIES (NIAS), BANGALORE

. This institution has started digitization of paintings and the Microfilming of Indian Publication Project (MIPP). The NIAS has also started work on rare manuscript preservation projects for bothmicrofilm and microfiche

### 11.7.6 NATIONAL MISSION FOR MANUSCRIPTS

The Department of Culture, Government of India has launched the 'National Mission for Manuscripts' in 2003 with the main objectives of conservation and preservation of Manuscripts for posterity. India, being the largest repository of Manuscripts, rare books, classics etc., urgently needs digitization to preserve as well as give access to. The National Informatics Centre (NIC) has prepared detailed guidelines for digitization of Manuscripts.

### 11.7.7 NATIONAL RESOURCE CENTRE FOR WOMEN

The National Resource Centre for Women was set up by the Government of India as a Virtual Resource Centre on Women's issues. It serves as a decentralized, participatory and partnership oriented entity aiming at giving access to digital catalogues of different libraries dealing with women's issues, reports of diverse nature, statistics, events, legislation etc using different media to reach the clientele at different levels.

### 11.7.8 PARLIAMENT LIBRARY

In order to cater to the needs of Members of Parliament and officers and staff of Parliament Secretariat a digital library has been set up in the parliament library. A large number of index based databases of information was initially developed by the computer centre. The data stored and available now in PARLIS databases for online retrieval relates to questions, debates, reports, bio-data of present and past members of parliament including photographs and addresses etc.

### 11.7.9 VIDYANIDHI

Vidyanidhi – a Sanskrit word means "Treasure of Knowledge". It is begun as a pilot project to demonstrate the feasibility of electronic Theses and Dissertations. As per the Action Plan of the National Task Force on Information Technology and Software Development, it is mandatory for all universities and deemed universities across the country to host every thesis/dissertation on a designated website. This national policy has provided a policy framework for initiating a digital library of ETDs. 'Vidyanidhi' project was started in the year 2000 at the Department of Library and Information Science, University of Mysore, with the sponsorship of National Information System for Science and Technology, Government of India. The Project's vision is to build and

strengthen theresearch capacities and enhance the quality of doctoral research in India.

## 11.8 SELF ASSESSMENT QUESTIONS

1. Discuss about digital libraries initiatives in India.
2. Enumerate the Indian digital libraries.

## 11.9 REFERENCES

1. Gurram, Sujatha (2008). "Digital library initiatives in India: a proposal for open distancelearning." Proceedings of the IATUL Conferences. p. 25. http://docs.lib.purdue.edu/iatul/2008/papers/25
2. Ali, Amjad (2007). Digital Libraries and Information Networks. – New Delhi: ESS ESS Publications. pp. 170-207.
3. Arms, William Y (2000) Digital Libraries. – Cambridge: M.A: MIT Press.
4. Arms, William Y (2000) Digital Libraries for Distance Education.

Available on http://www.dlib.org/dlib/october00/10editorial.html

5. Das, Anup Kumar, Sen, B.K. and Dutta, Chaitali (2005). Digitization of Scholarly Materials in India for Distance and Open Learners. Available on http://openmed.nic.in/1217/01/ Anup_Kumar_Das_ICDE_Conference_05.pdf
6. Jain, P.K., Jindal, S.C. and Babbar, Parveen(2006). "Digital Libraries in India : initiatives and problems" In International Conference on Digital Libraries 2006 : information managementfor global access. – New Delhi: TERI. pp. 22-31.
7. Ramesha, Karisiddappa, C.R. and Ramesh Babu, B (2008). "Digital Library and Digital Library initiatives in India" In Libraries in Digital Environment: problems and prospects edited by Sunil Kumar Satpathy, Chandrakant Swain and Bijayalaxmi Rautaray – New Delhi: Mahamaya. pp. 1-27.
8. Sree Kumar, M.G and Sreejaya, P (2006). "Digital Library Initiatives and Issues in India: efforts on scholarly knowledge management" In Digital Libraries in Knowledge Management by M.G.Sree Kumar… [et al].– New Delhi: ESS ESS Publications. pp.17-37.

LESSON-12

# OPEN SOURCE INITIATIVES

**OBJECTIVE**

After reading this chapter, students will be able to understand:

- Basic fundamental concepts of open source software
- Open source movement
- Open source initiatives

**STRUCTURE**



## 12.1 INTRODUCTION

Open Source Software (OSS) was born in universities and research labs and today is spreading to the commercial world. Many companies are adopting OSS strategies and licensing models in whole or in part. Some of the most successful OSS initiatives to date include Linux, Eclipse, Apache and Mozilla. For companies looking to define their software strategies for the

coming years, ignoring OSS would be a mistake. In this paper we try to trace the origins of the OSS movement, highlight pros and cons, and discuss alternatives to pure OSS. We talk about different licensing models which exist to serve different segments of the industry and investigate how commercial software company employees can co-exist in the mixed world of proprietary and open source software. Legal issues have had a huge impact on the growth of OSS, so we take a closer look at the legal risks by studying the SCO case. Finally, we end with a case study on open source strategies for Microsoft Windows and other platforms

## 12.2 OPEN SOURCE MOVEMENT

The open-source movement is surfacing more and more often as an undercurrent in the busy flow of discussion swirling around software development in higher education. Most often it comes up for mention as a response to the increasing predomination of commercial, proprietary software in use on campuses. As operating systems, development tools, desktop applications, and enterprise software all have become large, complicated, and expensive, an increasing numberof IT professionals are looking for not just alternative products and sources, but at a different wayto develop and support software. If open-source fulfills its proponents' hopes to even a modest degree, the effect on IT practices in higher education will be substantial.

Open-source can be defined as an approach to software development and intellectual property in which program code is available to all participants and can be modified by any of them. Those modifications are then distributed back to the community of developers working with the software. In this methodology, licensing serves primarily to disclose the identities of all the participants, documenting the development of the code and the originators of changes, enhancements, and derivative off-shoots.

The most widespread and vocal adherents of open-source are the members of the Linux-using community. But projects sponsored by major universities to develop new "open" software are also underway. The most visible of these is the Open Knowledge Initiative, a consortium of American universities led by MIT and Stanford. Their aim is to produce an "architectural specification" for the development of educational software. The Java in Administration Special Interest Group is alarge association of academic and commercial organizations sharing Java code and collaborating in the development of portal, an open-source campus portal product. For the most part, the opensource technologies and products existing or under development today are not primarily unique orgroundbreaking in functionality. Instead, they are alternatives to commercially well-established software, distinguished more by the way they are owned, operated, and further developed. A collegeor university buying a commercial portal or operating system agrees to license terms and conditionsthat almost always prohibit any modifying of the software. The software itself comes only in compiledform and so is not amenable to being changed in any event. Frustrations with those constraints are the basis for interest in open source.

## 12.3 OPEN SOURCE INITIATIVE (OSI)

In 1998, Netscape Communications, Inc. related an open source version of its browser in connection with the Mozilla project. As part of the project, Bruce Perens developed the initial draft of the open source Definition that led to the establishment of the Open Source Initiative. The OSI has been responsible for the continuing evolution of the Open Source Definition and certification of vendor licenses as compliant with the definition.

The OSI prefers to use the term "Open Source" instead of the FSF's "Free Software". In addition to a change in terminology, the OSI has been avoided some of the strong political positions taken by the FSF and has gravitated toward an approach to open source software more palatableto

commercial enterprises that want to develop open source products and related services. This transition has not been without substantial controversy.

The OSI Open Source Definition requires license to include the following rights and obligations" No fee or royalty may be imposed on redistribution.

- The source code must be made available.
- The licensee must have the right to create derivative works and modifications.
- The license may require modifications to be distributed as the original version plus patchescontaining the modifications.
- The licensor cannot discriminate against any user or group.
- All rights granted in the original license must be granted in any redistribution of the code.
- The license applies to the software as a whole and each of its components.
- The license may not restrict other software that is distributed with the licensed software.

## 12.4 DEFINITION OF OPEN SOURCE

Open source does not just mean access to the source code. The distribution terms of open-source software must comply with the following criteria:

### *12.4.1 FREE REDISTRIBUTION*

The license shall not restrict any party from selling or giving away the software as a component of an aggregate software distribution containing programs from several different sources. The license shall not require a royalty or other fee for such sale.

### *12.4.2 SOURCE CODE*

The program must include source code, and must allow distribution in source code as well as compiled form. Where some form of a product is not distributed with source code, there must be a well-publicized means of obtaining the source code for no more than a reasonable reproduction cost preferably, downloading via the Internet without charge. The source code must be the preferred form in which a programmer would modify the program. Deliberately obfuscated source code is not allowed. Intermediate forms such as the output of a preprocessor or translator are not allowed.

### *12.4.3 DERIVED WORKS*

The license must allow modifications and derived works, and must allow them to be distributed under the same terms as the license of the original software.

### *12.4.4 INTEGRITY OF THE AUTHOR'S SOURCE CODE*

The license may restrict source-code from being distributed in modified form *only* if the license allows the distribution of "patch files" with the source code for the purpose of modifying the program at build time. The license must explicitly permit distribution of software built from modifiedsource code. The license may require derived works to carry a different name or version numberfrom the original software.

### *12.4.5 NO DISCRIMINATION AGAINST PERSONS OR GROUPS*

The license must not discriminate against any person or group of persons.

### *12.4.6 NO DISCRIMINATION AGAINST FIELDS OF ENDEAVOR*

The license must not restrict anyone from making use of the program in a specific field of endeavor. For example, it may not restrict the program from being used in a business, or from being used for genetic research.

### *12.4.7 DISTRIBUTION OF LICENSE*

The rights attached to the program must apply to all to whom the program is redistributed without the need for execution of an additional license by those parties.

### *12.4.8 LICENSE MUST NOT BE SPECIFIC TO A PRODUCT*

The rights attached to the program must not depend on the program's being part of a particular software distribution. If the program is extracted from that distribution and used or distributed within the terms of the program's license, all parties to whom the program is redistributed should have the same rights as those that are granted in conjunction with the original software distribution.

### *12.4.9 LICENSE MUST NOT RESTRICT OTHER SOFTWARE*

The license must not place restrictions on other software that is distributed along with the licensed software. For example, the license must not insist that all other programs distributed on the same medium must be open-source software.

### *12.4.10 LICENSE MUST BE TECHNOLOGY-NEUTRAL*

No provision of the license may be predicated on any individual technology or style of interface.

## 12.5 OPEN SOURCE IS NOT PUBLIC DOMAIN SOFTWARE, SHAREWARE, OR FREEWARE

It is important to distinguish open source software from shareware, freeware or software that is dedicated to the public domain. Public domain software is generally provided without a license, with the developer relinquishing all copyrights that it may have in the programming. Users can copy, modify and distribute the software freely and generally have no obligation to give any attribution to the original developer. In contrast, open source software is always provided under a license, although licensees although licensees receive far more rights than a traditional software license. "Shareware" or "freeware" is software provided under a traditional license that prohibits access to source code or the creation of derivative works. Shareware is typically provided free of charge during a trial period. If the licensee desires to continue to use the shareware following the trial period, it must pay a license fee. "Freeware" does not require the payment of any license fee for use of the software. (Freeware should not be confused with the Free Software Concept used by the Free Software Foundation).

## 12.6 OPEN SOURCE LICENSES

Open source licenses are licenses that comply with the Open Source Definition in brief, they allow software to be freely used, modified, and shared. To be approved by the Open Source

Initiative (also known as the OSI), a license must go through the Open Source Initiative's license review process.

### 12.6.1 POPULAR LICENSES

The following OSI-approved licenses are popular, widely used, or have strongcommunities (as defined in the 2006 Proliferation Report):

- Apache License 2.0
- BSD 3-Clause "New" or "Revised" license
- BSD 2-Clause "Simplified" or "FreeBSD" license
- GNU General Public License (GPL)
- GNU Library or "Lesser" General Public License (LGPL)
- MIT license
- Mozilla Public License 2.0
- Common Development and Distribution License
- Eclipse Public License

### 12.6.2 ALL APPROVED LICENSES

Many other licenses are also OSI-approved, but fall into other categories, such as special-purpose licenses, superseded licenses, or retired licenses. Complete lists that include all approved licenses are available:

- sorted by name (alphabetical)
- sorted by category

Open Standards Requirement for Software

***THE REQUIREMENT***

An "open standard" must not prohibit conforming implementations in open source software.

***THE CRITERIA***

To comply with the Open Standards Requirement, an "open standard" must satisfy the following criteria. If an "open standard" does not meet these criteria, it will be discriminating againstopen source developers.

1) **No Intentional Secrets:** The standard MUST NOT withhold any detail necessary for interoperable implementation. As flaws are inevitable, the standard MUST define a process for fixing flaws identified during implementation and interoperability testing and to incorporatesaid changes into a revised version or superseding version of the standard to be releasedunder terms that do not violate the OSR.

2) **Availability:** The standard MUST be freely and publicly available (e.g., from a stable web site) under royalty-free terms at reasonable and non-discriminatory cost.

3) **Patents:** All patents essential to implementation of the standard MUST:
   a. be licensed under royalty-free terms for unrestricted use, or
   b. be covered by a promise of non-assertion when practiced by open source software

4) **No Agreements:** There MUST NOT be any requirement for execution of a license agreement, NDA, grant, click-through, or any other form of paperwork to deploy conformingimplementations of the standard.

5) **No OSR-Incompatible Dependencies:** Implementation of the standard MUST NOT require any other technology that fails to meet the criteria of this Requirement.

## 12.7 OPEN SOURCE EDUCATION

The Open Source Initiative (OSI) actively promotes open source software by educating developers, decision makers, and users about the advantages of open source software and collaboration techniques. OSI's members are active in the core open source development communities as well as in government, academic and industry circles, helping to educate people about open source. As part of its mandate on education, OSI members deliver presentations about open source technologies, collaboration techniques, and community building at conferences and seminars across the world. OSI Board members and individual members have also conducted workshops as well as short and long courses on open source concepts, techniques, and technologies.

OSI's Education Committee focuses on the use and teaching of open source software in the educational context, from high school through graduate and post-graduate levels. The Education Committee is responsible for developing, arranging, and conducting educational conferences, programs, courses of instruction, and online educational seminars covering open source software, licensing, and communities.

The committee chairperson currently is Prof. Tony Wasserman. Additionally, the members of osi-edu-discuss at members.opensource.org also participate in the discussions around education.

## 12.8 SELF ASSESSMENT QUESTIONS

1. What is open source?
2. Discuss about open source initiative.
3. Write definition of OS?

## 12.9 REFERENCES

1. Overly, Michael (2003). The Open Source Handbook. Silver Spring, MD; Pick & Fischer,p.5-6.
2. Warger, Thomas (2002). The Open Source Movement. Educause Quarterly. http:// skat.ihmc.us/rid=1071396961281_699700851_2705/ope%20source%20movement.pdf
3. http://opensource.org/osd

LESSON-16

# OPEN ARCHIVE INITIATIVE (OAI)

## OBJECTIVE

After reading this chapter, students will be able to understand:

1. Basic concepts of Open Archives Initiative
2. The technical framework of the Open Archives Initiative

## STRUCTURE



## 13.1 INTRODUCTION

The name Open Archives Initiative reflects the origins of the OAI in the E-Prints community where the term archive is generally accepted as a synonym for a repository of scholarly papers. Members of the archiving profession have justifiably noted the strict definition of an "archive" withintheir domain; with implications for preservation of long-term value, statutory authorization and institutional policy. The OAI uses the term "archive" in a broader sense: as a repository for stored information. Language and terms are never unambiguous and uncontroversial and the OAI respectfully requests the indulgence of the archiving community with this less constrained use of"archive".

Some explanation of the use of the term "Open" in OAI is also due. Our intention is "open" from the architectural perspective defining and promoting machine interfaces that facilitate the availability of content from a variety of providers. Openness does not mean "free" or "unlimited" access to the information repositories that conform to the OAI technical framework. Such terms are often used too casually and ignore the fact that monetary cost is not the only type of restrictionon use of information any advocate of "free" information recognize that it is eminently reasonableto restrict denial of service attacks or defamatory misuse of information.

## 13.2 OPEN ARCHIVE INITIATIVE

The Open Archives Initiative develops and promotes interoperability standards that aim to facilitate the efficient dissemination of content. The Open Archives Initiative has its roots in an effort to enhance access to e-print archives as a means of increasing the availability of scholarly communication. Continued support of this work remains a cornerstone of the Open Archives program. The fundamental technological framework and standards that are developing to support this work are, however, independent of the both the type of content offered and the economic mechanisms surrounding that content, and promise to have much broader relevance in openingup access to a range of digital materials. As a result, the Open Archives Initiative is currently an organization and an effort explicitly in transition, and is committed to exploring and enabling this new and broader range of applications. As we gain greater knowledge of the scope of applicability of the underlying technology and standards being developed, and begin to understand the structure and culture of the various adopter communities, we expect that we will have to make continued evolutionary changes to both the mission and organization of the Open Archives Initiative.

## 13.3 E-PRINT ORIGINS

The origins of the OAI lie in increasing interest in alternatives to the traditional scholarly publishing paradigm. While there may be disagreements about the nature of what changes needto take place, there is widespread consensus that change, perhaps radical change, is inevitable.There are numerous motivating factors for this change. An increasing number of scholarly disciplines, especially those in the so-called “hard sciences” (e.g., physics, computer science, life sciences), are producing results at an increasingly rapid pace. This velocity of change demands mechanisms for reporting results with lower latency times than the ones experienced in the established journal system. The ubiquity of high-speed networks and personal computing has created further consumer demand for use of the Web for delivery of research results. Finally, the economic model of scholarly publishing has been severely strained by rapidly rising subscription prices and relatively stagnant research library budgets.

In some scholarly fields, the development of alternative models for the communication of scholarly results – many in the form of on-line repositories of EPrints has demonstrated a viable alternative to traditional journal publication. Perhaps the best known of these is the Physics archive1 run by Paul Ginsparg [2] at Los Alamos National Laboratory. There are, however, a number of other established efforts (CogPrints2, NCSTRL3, RePEC4), which collectively demonstrate the growing interest of scholars in using the Internet and the Web as vehicles for immediate dissemination of research findings. Stevan Harnad, among the most outspoken advocate of change, views such solutions as the first step in radical transformation of scholarly publishing whereby authors reclaim control over their intellectual property and the publishing process [3].

## 13.4 TECHNICAL FRAMEWORK

The technical framework of the Open Archives Initiative is intended to provide a low-barrier approach to interoperability. The membership of the OAI recognizes that there are functional limitations to such a low-barrier framework and that other interoperability standard, for example Z39.50, addresses a number of issues in a more complete manner. However, as noted by Bill. At the root of the technical agreement lies a distinction between two classes of participants:

- Data Providers adopt the OAI technical framework as a means of exposing metadata about their content.

- Service Providers harvest metadata from data providers using the OAI protocol and use the metadata as the basis for value-added services.

The remainder of this section describes the components of this technical framework. A theme carried through the framework and the section is the attempt to define a common denominator for interoperability among multiple communities while providing enough hooks for individual communities to address individual needs (without interfering with interoperability). More details on the technical framework are available in the Open Archives Metadata Protocol specification availableat the OAI web site.8

## 13.5 METADATA

The OAI technical framework addresses two well-known metadata requirements: interoperability and extensibility (or community specificity). These issues have been a subject of considerable discussion in the metadata community [12, 13] the OAI attempts to answer this in a simple and deployable manner. The requirement for metadata interoperability is addressed by requiring that all OAI data providers supply metadata in a common format – the Dublin Core Metadata Element Set [14]. The decision to mandate a common element set and to use the Dublin Core was the subject of considerable discussion in the OAI. One approach, which has been investigated in the research literature [15], is to place the burden on the consumer of metadata rather than the provider, tolerating export of heterogeneous metadata and relying on services to map amongst the representations. OAI, however, is purposely outside the domain of strict research, and in the interest of easy deployment and usability it was decided that a common metadata format was the prudent decision.

The decision to use the Dublin Core was also the result of some deliberation. The original Santa Fe convention took a different course – defining a metadata set, the Open Archives Metadata Set, with some functionality tailored for the E-Print community. The broadening of the focus of the OAI, however, forced reconsideration of this decision and the alternative of leveraging the well- known and active work of the DC community in formulating a cross-domain set for resource discoverywas chosen.

Those familiar with the Dublin Core will note that all fields in DC are optional. The OAI discussed requiring a number of DC elements in OAI records. While such requirement might be preferable from the perspective of interoperability, the spirit of experimentation in the OAI persuaded the committee to keep all elements optional.

## 13.6 RECORDS, REPOSITORIES, AND IDENTIFIERS

The OAI technical framework defines a record, which is an XML-encoded byte stream that serves as a packaging mechanism for harvested metadata. A record has three parts:

- Header – containing information that is common to all records (it is independent of the metadata format disseminated in the record) and that is necessary for the harvesting process. The information defined in the header is the unique identifier for the record (described below), and a datestamp indicating the date of creation, deletion, or latest dateof change in the metadata in the record.

- Metadata – containing metadata in a single format. As noted in section 4.1, all OAI data providers must be capable of emitting records containing unqualified DC metadata. Othermetadata formats are optional.

- About – an optional container to hold data about the metadata part of the record. Typically, the "about" container could be used to hold rights information about the metadata, terms and conditions for usage of the metadata, etc. The internal structure of the "about" containeris not defined by the protocol. It is left to individual communities to decide on its syntax andsemantics through the definition of a schema.

## 13.7 SELECTIVE HARVESTING

A protocol that only enabled consumers of metadata to gather all metadata from a data provider would be cumbersome. Imagine the transactions with large research libraries that expose the metadata in their entire catalog through such a protocol. Thus, some provision for selective harvesting, which makes it possible in the protocol to specify a subset of records to be harvested,is desirable. Selection, however, has a broad range of functionality. More expressive protocols include provisions for the specification of reasonably complete predicates (in the manner of database requests) on the information requested. The OAI decided that such high functionality was not appropriate for a low-barrier protocol and instead opted for two relatively simple criteria for selectiveharvesting.

Date-based – "the date of creation, deletion, or latest date of modification of an item, the effect of which is a change in the metadata of a record disseminated from that item". Harvesting requests may correspondingly contain a date range for harvesting, which may be total (between two dates) or partial (either only a lower bound or an upper bound). This date-based harvesting provides the means for incremental harvesting. For example, a client may have a weekly schedule for harvesting records from a repository, and use the date-based selectivity to only harvest records added or modified since the last harvesting time.

## 13.8 OPEN ARCHIVES METADATA HARVESTING PROTOCOL

The initial protocol that came out of the Santa Fe meeting was a subset of the Dienst protocol. While that subset protocol was functionally useful for metadata harvesting, aspects of its legacy context presented barriers to simple implementation. The current technical framework is built around a more focused and easier to implement protocol – the Open Archives Metadata Harvesting Protocol.

The Open Archives Metadata Harvesting Protocol consists of six requests or verbs. The protocol is carried within HTTP POST or GET methods. The intention is to make it simple for data providersto configure OAI conformant repositories by using readily available Web tools such as libwww- perl9. OAI requests all have the following structure:

- Base-URL – the Internet host and port of the HTTP server acting as a repository, with an optional path specified by the respective HTTP server as the handler for OAI protocol requests.

- Keyword Arguments – consisting of a list of key-value pairs. At a minimum, each OAI protocolrequest has one key value pair that specifies the name of the OAI protocol request.

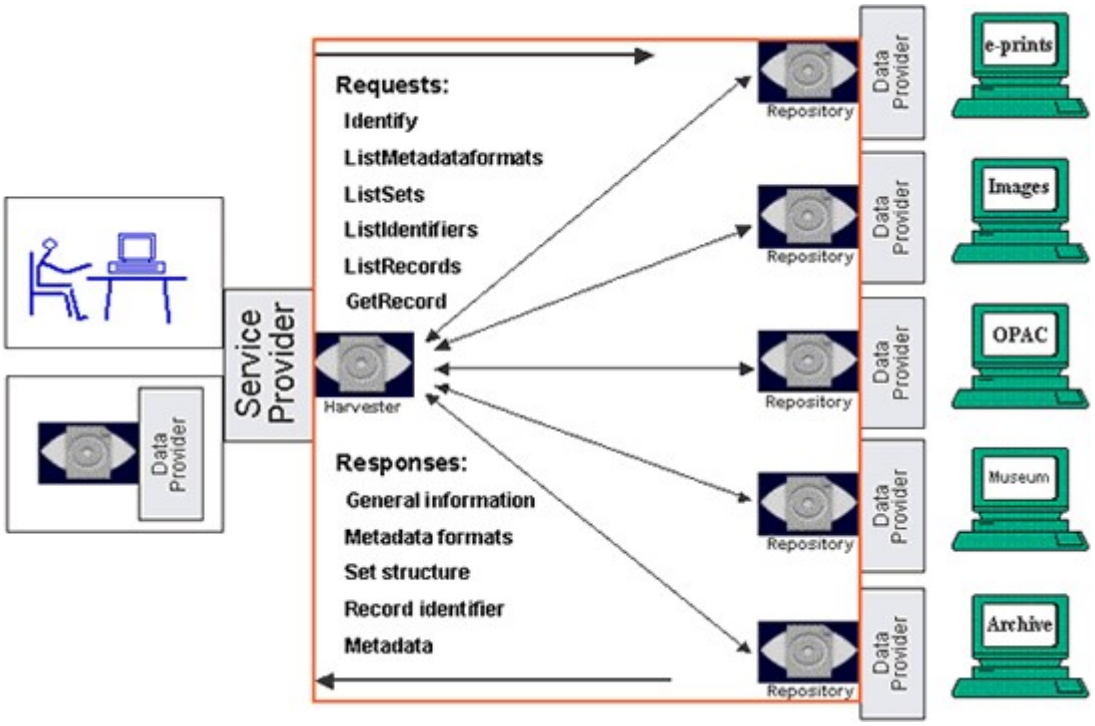


Figure 4 - Sample OAI Request Encoding

## 13.9 DATA PROVIDER CONFORMANCE AND REGISTRATION

The OAI expects that data providers will fall into three layers of participation, each higher layer implying the preceding layer(s);

1) OAI-conformant – These are data providers who support the protocol definition. As stated earlier, conformance is testable since there are XML-schemas to validate all responses. No doubt, the OAI will not be able to track every provider using the protocol since use of it does not require any licensing or registration procedure.

2) OAI-registered – These are data providers who register in an OAI-maintained database, which will be available through the OAI web site. Registration will entail that the data provider gives a BASE-URL, which the registration software will then use to test compliance.

3) OAI-namespace-registered – These are data providers who choose to name their records in conformance with an OAI naming convention for identifiers. Names that follow this convention have the following three components:

a) oai – A fixed string indicating that the name is in the OAI namespace.
b) <repoID> - An identifier for the repository that is unique within the OAI namespace.
c) <localID> - An identifier unique within the respective repository.

An example of a name that uses this naming scheme is: oai: arXiv: hep-th01

## 13.10 SELF ASSESSMENT QUESTIONS

1. What is metadata harvesting?
2. What are the technical frameworks of open archive Initiative (OAI)?

## 13.11 REFERENCES

1 http://www.arxiv.org.
2 http://cogprints.soton.ac.uk.
3 http://www.ncstrl.org.
4 http://netec.mcc.ac.uk/RePEc/.
5 The Santa Fe meeting was sponsored by the Council on Library and Information Resources (CLIR), the Digital Library Federation (DLF), the Scholarly Publishing & Academic Resources Coalition (SPARC), the Association of Research Libraries (ARL) and the Los Alamos National Laboratory (LANL).
6. Van de Sompel, H., T. Krichel, and M.L. Nelson, The UPS Prototype: an experimental end-user service across e-print archives, in D-Lib Magazine. 2000.
7. Lagoze, C. and J.R. Davis, Dienst - An Architecture for Distributed Document Libraries. Communications of the ACM, 1995. 38(4): 47. 8. A New Approach to Finding Research Materials on the Web, 2000, Digital Library Federation. http://www.clir.org/diglib/architectures/vision.htm.
9. Anderson, K.M., et al., ACM 2000 digital libraries : proceedings of the fifth ACM Conference on Digital Libraries, June 2-7, 2000, San Antonio, Texas. 2000, New York: Association for ComputingMachinery. xiii, 293.
10. Borbinha, J. and T. Baker, Research and advanced technology for digital libraries: 4th European Conference, ECDL 2000, Lisbon, Portugal, September 18-20, 2000: proceedings. Lecture notes in computer science 1923. 2000, Berlin ; New York: Springer. xvii, 513.
11. Arms, W.Y., Digital libraries. Digital libraries and electronic publishing. 2000, Cambridge, Ma.:MIT Press. 12. Lagoze, C. Accommodating Simplicity and Complexity in Metadata: Lessons from the Dublin Core Experience. in Seminar on Metadata. 2000. Archiefschool, Netherlands Institutefor Archival Education and Research, The Hague.
13. Weibel, S., The Dublin Core: A simple content description format for electronic resources. NFAIS Newsletter, 1998. 40(7): p. 117-119.
14. Chang, C.-C.K. and H. Garcia-Molina. Mind your vocabulary: query mapping across heterogeneous information sources. In International Conference on Management of Data and Symposium on Principles of Database Systems. 1999. Philadelphia: ACM: 335-346.
15. Lagoze, C. and T. Baker, Keeping Dublin Core Simple. D-Lib Magazine, 2001. 7(1).
16. Fallside (ed.), D.C., XML Schema Part 0: Primer. 2000, World Wide Web Consortium. http://www.w3.org/TR/xmlschema-0/.
17. Thompson, H.S., et al., XML Schema Part 1: Structures. 2000, World Wide Web Consortium. http://www.w3.org/TR/xmlschema-1/.
18. Biron, P.V. and A. Malhotra, XML Schema Part 2: Datatypes. 2000, World Wide Consortium.http://www.w3.org/TR/xmlschema-2/.
19. Van de Sompel, H. and P. Hochstenbach, Reference Linking in a Hybrid Library Environment , Part 3: Generalizing the SFX solution in the "SFX@Ghent & SFX@LANL" experiment. D-Lib Magazine, 1999. 5(10).http://www.dlib.org/dlib/october99/van_de_som pel/10van_de_sompel.html.
20. Van de Sompel, H. and P. Hochstenbach, Reference Linking in a Hybrid Library Envronment:, Part 1: Frameworks for Linking. D-Lib Magazine, 1999. 5(4).http://www.dlib.org/dlib/april99/ van_de_sompel/04van_de_sompel-pt1.html.
21. Van de Sompel, H. and P. Hochstenbach, Reference Linking in a Hybrid Library Environment, Part 2: SFX, a Generic Linking Solution, in D-Lib Magazine. 1999.
22. Paepcke, A., et al., Search Middleware and the Simple Digital Library Interoperability Protocol. D-Lib Magazine, 2000. 5(3).http://www.dlib.org/dlib/march00/paepcke/03pae pcke.html.

**LESSON-14**

# BUILDING THE DIGITAL LIBRARY: DIGITIZATION-PROCESS AND METHODS, PLANNING FOR DIGITIZATION

**OBJECTIVE**

After reading this chapter, students will be able to understand:

- Basic fundamental concepts of digital library protocols
- Different protocols of DL
- Basic fundamental concepts of digitization
- Need of digitization
- Planning & implementation of digitization

**STRUCTURE**

## 14.1 INTRODUCTION

Libraries have always been a community's 'portal' to information, knowledge and leisure. Beyond their shelves, libraries are a community's gateway to information from many sources nationally and internationally. Libraries provide professionals trained to distinguish and verify content,build collections and provide a reference and information service. Today more libraries rely on electronic sources for collecting, organizing and distributing information. The information age has created unprecedented opportunities to acquire electronic content from many sources including existing digital content in many different types of libraries. The concept of a "world library for the blind" rests on the ability of digital libraries to share and coordinate collection-building resources and to use digital technology to share content. It requires designing these systems and services with interoperability in mind and using common standards. It begins with a

shared understanding that technology does not fundamentally change library service, but rather the way in which it is organized and delivered. Therefore, guidelines for the development of the digital library must begin with the assumption that the library remains a collection of organized content reflecting works of imagination and information necessary to facilitate life long learning, career development and an informed citizenry. Its digitization is a means of ensuring that its collections are preserved and accessible to all regardless of disability.

Digital Libraries are being created today for diverse communities and in different fields e.g. education, science, culture, development, health, governance and so on. With the availability of several free digital Library software packages at the recent time, the creation and sharing of information through the digital library collections has become an attractive and feasible proposition for library and information professionals around the world. Library automation has helped to provide easy access to collections through the use of computerized library catalogue such as On-line Public Access Catalog (OPAC). Digital libraries differ significantly from the traditional libraries because they allow users to gain an on-line access to and work with the electronic versions of full text documents and their associated images. Many digital libraries also provide an access to other multi-media content like audio and video.

## 14.2 DIGITIZATION

Witten and David (2003) defined Digitization as the process of taking traditional library materials that are in form of books and papers and converting them to the electronic form where they can be stored and manipulated by a computer. Ding, Choo Ming (2000) has elaborated the works of Getz (1997), Line (1996) and McKinley (1997) on the advantages of digitization. They maintained that:

- Digitization means no new buildings are required; information sharing can be enhanced and redundancy of collections reduced.
- Digitization leads to the development of Internet in digitalized based libraries. As Internet is now the preferred form of publication and dissemination.
- Digital materials can be sorted, transmitted and retrieved easily and quickly.
- Access to electronic information is cheaper than its print counterpart when all the files are stored in an electronic warehouse with compatible facilities and equipment.
- Digital texts can be linked, thus made interactive; besides, it enhances the retrieval of more information.

In the light of the following advantages, it is natural today to find more information being digitizedand uploaded into the Internet or Compact-Disc Read Only Memory (CD-ROM) in order to be made correspondingly accessible globally.

## 14.3 WHY DIGITIZATION?

There are three main needs for digitization; two or all the three of them may apply to your digital library project.

- To preserve the Documents: That is to allow people to read older or unique documents without damage to the originals.
- To make the documents more accessible: This is to serve the existing users better; e.g. to

allow the users to search the full text of the documents or to serve more users than envisagedin remote locations, example, more than one person at a time.

- To reuse the documents. It means to convert documents into different formats; for exampleto use images in a slideshow and to adopt the content for a different purpose.

Digitizing documents can take a lot of time, effort and money. Smith (2001), narrated the following reasons that should be considered before going into digitization.

## 14.4 REASONS TO BE CONSIDERED

### *14.4.1 Is it worth digitizing?*

Do the documents contain the information that is valuable enough to warrant the costs of digitization? There is no point digitizing the documents that are already out of date, no matter how bulky they, but it is worthy to digitize the old, unique documents that can be easily damaged so that the people can be allowed to use them without handling the originals. These unique documents are sometimes called the heritage documents.

### *14.4.2 Who is your audience?*

If there are only few users, or maybe there are large numbers of potential users, but they do not have computers to access the digital library, they can be served by sending them photocopies. It may be difficult to judge the demand for documents. It is, however; wise to get other people's opinions. Ask the potential users of the documents what they see as their priorities.

### *14.4.3 Do the documents form a collection?*

It is important to verify if the documents form a collection. In fact, the documents in a digitallibrary should have something in common like a common subject focus

### *14.4.4 How easy is it to digitize documents?*

Another important factor to take into account is how easy it will be to digitize the documents. Not all the hard copy documents can be easily converted to electronic format. There is the need to check the physical characteristics of the documents to understand how easy it will be to digitize them. If you have a lot of documents that are hard to digitize, you might choose not to include them in the digital library. It is advisable to put them in the image files, rather than in the searchable text document. According to Maxine (2000), creating a digital library collection involves the following steps: planning, implementation and promotion. These are essential if the finished product is to successfully meet the user's needs and conform to the accepted quality standards.

## 14.5 PLANNING

Planning mainly involves identifying various tasks related to creating a digital library collection, developing strategies for handling these tasks, identifying required resources and formulating a timeline for accomplishing these tasks. If there is a need to have a large digital project, you may consider conducting a feasibility study to assess the viability of the project before detailed planning. The outcome of the feasibility study could be a formal proposal for obtaining management approvalor grant for the project.

a) The first step in planning a digital library collection development project is to specify the need for creating the digital library collection, its purpose and target user community. You should indicate if management, the users or others have expressed this need and defined what this need is. The purpose could be improving preservation of some rare or delicate materials, improving access to and the visibility of certain material or facilitating re-use of documents. It is important to identify the target user community for a digital library collectionand their profile.
b) There is the need to define the source material that constitutes the digital library collections and the key attributes of this source material. Examples of source material include project reports, staff publications, working papers, theses, dissertation, audio and video lectures, songs and musical scores etc. There is also the need to specify what portion of the materialis to be digitized and if all the material or only a sub-set will be covered in the digital collection. Remember to assess copyright restrictions.
c) Define the key features of the digital library collection you plan to build. Identify the nature of the collection e.g. static or dynamic. Indicate the type of usages you would allow the users to adhere to and the kind of service delivery they should expect from you e.g. CDROM oron-line or both. Define metadata, search and retrieval requirements.
d) The important task in creating a digital library collection is the conversion of the source materials available in hardcopy into a digital format. There should be a clear cut statement about the related requirements and their processes, namely:
    - How to convert the source material into required digital format.
    - What are the digitization requirements?
    - The workflow involved in digitizing the source material.
e) Identify the resources and money required for creating and maintaining digital collections. There is a need to identify:
    - What type of information technology (IT) infrastructure is required for establishingand maintaining the digital collections?
    - What are the personnel requirements and
    - What are the financial requirements involve for setting up and maintaining the collection.
f) Finally, there is the need to define how the project is going to be implemented and what the major milestones and time requirements are?

## 14.6 IMPLEMENTATION

Planning is followed by implementation. That is getting down to the actual steps required to set up the collection. This means that there must be a need to obtain the management approval for the plan and the required resources before proceeding with the implementation. There is a need to identify and designate a project manager to lead the implementation of the digital project. For large digital library projects, it is essential to have a full time project manager for the project period. The Implementation of a digital library project involves the following activities.

- Establish the project team
- Set up the Information Technology (IT) infrastructure
- Procure and install digital library software
- Finalize policies and specifications
- Complete arrangement of workflow for digitization
- Set up the digital library collection site in case of Internet distribution
- Obtain copyright permissions and
- Release the digital library collection for use.

## 14.6 PROMOTION AND PROVISION OF SERVICES

The digital library collection created should be visible, and it should provide an easy access for users. One-way of achieving this is to include links to the collection site in the appropriate pages of the library website and other related on-line services in the organization. In addition to, or in the absence of remote on-line access to the digital collection, there is the need to explore other modesof providing access to the digital collection. These may include:

- Setting up local public access computers on the library Local Area Network.
- Provision of e-mail based services and
- CD-ROM based distribution of the collection.

## 14.8 DIFFERENT STAGES IN DIGITIZING DOCUMENTS

Cornell University Library/Research Departments (2000), provides six stages in digitizing documents for a digital library: Registering, Scanning, Optical Character Recognition, Proofreading and formatting and producing the Final Version.

### *14.8.1 Registering*

Before scanning large number of documents, there is the need to first register them and use a filing system to keep their track. If not, you risk misplacing hardcopies, losing files, skipping steps in the process or duplicating work, perhaps without realizing it. There is also the risk of losing electronic versions of files because they have been misnamed or saved in the wrong subdirectory. Moreover, a good filing system is vital, so everyone in the digitizing team knows what he is supposed to do, and he can fill in for another person in case of absence.

### *14.8.2 Scanning documents*

It is necessary to clean and dust off the documents to be scanned; make sure that all the pages are present and in the right order. If the document is in poor condition, try to find a fresh copy. If it is a sheet fed scanner, cut the document open to get individual sheets to feed through the scanner. If necessary, you can rebind the documents later. If you do not want to damage the documents, you can photocopy each page and feed in the photocopy through the scanner, though this uses a lot of paper and reduces the quality of the scan. To scan a document on a flatbed scanner, place it face down on the scanner platen or put the pages into the sheet feeder. Then, inthe software, choose a setting, resolution and colour and scan each page of the document at the settings you have chosen.

### *14.8.3 Optical Character Recognition (OCR)*

Optical Character Recognition (OCR) software converts a scanned image into a text file that a word processor can read. To do this, it must first recognize where the text is on the page. The software breaks the text blocks down into lines or into an individual character. It tries to match the image of each letter against patterns it recognizes as an “a”, “b”, etc. There is a problem to encounter with languages that use Latin scripts with accented characters. As a solution, you should use the OCR software that is specific for language.

### *14.8.4 Proofreading*

This is the act of making corrections to the document text and layout. This is done in two ways:

a. Comparing the scanned text on the screen with the hardcopy and entering the corrections directly into the computer. The word processor's spellchecker will help in spelling errors quickly.

b. Printing out the scanned text and comparing it with the original copy. Mark any corrections on the printout, and then enter them into the computer. This is a slower method, but may be the best option if there are no enough computers for each proofreader.

### *14.8.5 Reformatting*

The Optical Character Recognition (OCR) software may produce a document that consists of straight text, no columns, no headers and footers. There is the need to reinsert these by hand or correct where they appear on the page. There may be also need to change the typeface, heading styles and so on, to make the document more attractive and readable. Alternatively, you may be able to adjust the settings of your OCR program to preserve the layout of the page.

### *14.8.6 Final Version*

For many documents, there is a need to add some information to the text so that readers can identify it easily. As for a book you must make sure that the book title, the author or the editor, the publisher and the publication date are all included. As for chapter in a book, you should include the title and the author of that chapter and the original page numbers in the printed version of the book. As for the journal articles you should include the journal title, the date, the volume and the issue number, the article title and the authors and the page numbers in the original printed journal. In other words there is the need to add Metadata to describe each document.

## 14.9 TECHNOLOGY INFRASTRUCTURE AND PERSONNEL

Several resources are required for the creation of digital library collections, their maintenance and provision of services. The two major resources needed are technology infrastructure and personnel.

### 14.9.1 Infrastructure

Access to a digital library collection can be provided on-line or off-line. The On-line access today typically means that the client uses a web browser on a desktop computer or laptop and access the collection by connecting to the digital library website over the Internet. The On-line access requires a connection to the Internet or to an internal network (Intranet). In Off-line access, the digital library is not accessible over a network. One way of providing an Off-line access to a digital library collection is to receive and respond to the user queries over e-mail. Another way is to distribute the digital library collection on a CD-ROM. A digital library project would typically require the following equipment. Server computer, Desktop computers, digitization equipment, Network connectivity and other equipment.

Another aspect is the software to be used in digital library. The Digital library software works with the web server in providing various digital library functionalities including creation, organization, maintenance, indexing, search and retrieval. In choosing the software, some features

should be taken into consideration. These include: Support for different document types, Support for customized metadata, Collection administration, Support for standards like Dublin core metadata standard, Search and retrieval and Multi-lingual support.

Several free digital library software packages are now available which could facilitate the easy creation and sharing of information through digital library collections. Examples of open source free digital library software include: Greenstone Digital Library software by New Zealand Digital Library; Academic Research in the Netherlands On-line (ARND); Tilburg University, The Netherlands; CDSware; CERN Document server software, Geneva, Switzerland; D-space; MIT Libraries, Cambridge, MA USA. etc.

## 14.10 PERSONNEL

Personnel are most important digital library's resource, not only during its initial creation and set up, but also for its operation, maintenance and provision of services. Since the access to the digital library is easy, compared to a physical library, more users are likely to access it. If the digital library does not meet the expectations of the users in terms of currency and quality of content, they will lose confidence, and it is likely for them not to visit the digital library again.
It is therefore important to assign the personnel with the right skills and attitude to handle the various tasks associated with the digital library project.

Broadly speaking, the personnel will be required for the following tasks:

- Project management.
- Selection and preparation of source material
- Digitization and conversion
- Cataloguing and metadata assignment
- Quality assessment
- System administration and maintenance of digital library server and website.
- System analysis/programming for digital library application/interface development
- Promotion and provisions of services.

Moreover, the rapid changes in the digital library technologies require constant re-training and re-positioning of staff for an effective practice in technological application.

Digitization has opened up new audiences and services for libraries, and it needs to be integrated into the plans and policies of any institution to maximize its effectiveness. Digitization is a complex process with many crucial dependencies between different stages over time. Utilizing a holistic life-cycle approach for digitization initiatives will help develop sustainable and successful project. It is hoped that the approach of the issues outlined, the software mentioned in this paper and the references to more detailed source and past project will contribute to the future success of initiating digitization of library resources.

Digitization as a complete process that broadly includes: selection, assessment, prioritization, project management and tracking, preparation of originals for digitization, metadata collection and creation, digitizing, quality management, data collection and management, submission of digital resources to delivery systems and into a repository environment, and assessment and evaluation of

the digitization effort. This document divides the processes involved in a digitization workflow into four main phases:

- Project planning
- Processes occurring prior to digitization
- Digital conversion
- Post-digitization work

Project planning and management, data collection, and quality management are consideredto be some of the ongoing activities throughout all four phases of the digitization workflow. The activities described within each phase address library/archival issues, imaging and conversion work, and IT infrastructure issues. Library and archival issues include preparation of originals for digitization, indexing, collection and creation of metadata of all types, and quality control of the digital versions, indexing data, and other metadata. Imaging/conversion work includes digitization, creation of derivative versions for access, quality control, and metadata creation. IT infrastructure issues include: collection and transfer of data to other systems, networked and Web services, databases, and managed storage and backup. Additional IT infrastructure issues include: short-term/intermediate data storage, backup of digital resources for disaster recover, and safeguards and checks to protect against data loss and to ensure data integrity.

*The Archives New Zealand S-6 Digitization Standard* provides a similar framework for digitization projects as this document, and lists a set of mandatory requirements for digitization processes. Apartial list of those requirements, appropriate for cultural heritage institutions, is listed below some of the requirements have been reworded to be more generic. We have listed them here as they dovetail nicely with concepts outlined in the four phases of digitization workflow as described earlier.

- All digitization and digitization processes must be planned, scoped and documented
- An appropriate digitization approach must be selected, documented and implemented
- Technical specifications aligned to the digitization requirements must be selected, documented and implemented
- Equipment and software aligned to the digitization requirements must be implemented
- Systems to support management of the digital output of digitization must be in place
- Guidelines for the preparation of original collections/records must be documented and implemented
- All digital objects created must be assigned metadata to document digitizing processes and to support ongoing business processes.
- Quality assurance and quality control procedures must be defined, documented and implemented
- Digital storage and disaster recovery procedures for digital objects and metadata must be defined, documented and implemented.
- Systems for the long-term management of digital objects and metadata must be documented and implemented.
- Preservation strategies and processes for digital objects and metadata must be defined,

documented and implemented

A large component of post-digitization activities includes content management and preservation within a digital repository environment. This document does not detail those processes, but instead leaves off after submission of digital objects into a managed repository environment. In Open Archival Information System (OAIS) parlance, the document addresses "pre-ingestion" activities, or those activities that take place prior to submission of digital resources to a digital repository for long-term management. This document does not address the processes, procedures, and actions surrounding management of Archival Information Packages (AIP) and subsequent creation and dissemination of Digital Information Packages (DIP). Creation of Submission Information Packages (SIP) is only discussed generically; as it is assumed that different repositories will have different ingest requirements or procedures. In many organizations, digitization processes may create official "record" copies. If this is the case, policies, workflow, and infrastructure shouldbe designed to ensure the integrity and authenticity of record copies is maintained during the digitization process and through submission of digital copies into various management and accesssystems and repositories. ARMA International has developed a set of "Generally Accepted Recordkeeping Principles" that suggests records should be created, managed, and maintained according to these principles. These *Generally Accepted Recordkeeping Principles*, http:// www.arma.org/garp/, include:

- Principle of accountability
- Principle of integrity
- Principle of protection
- Principle of compliance
- Principle of availability
- Principle of retention
- Principle of disposition
- Principle of transparency

Also, we believe factors developed around the sustainability of file formats can inform the conceptualization of digitization activities, related workflow, and supporting IT infrastructure. Severalorganizations have developed these criteria, including the Library of Congress; the National Library of the Netherlands; the State and University Library, Arhus, and the Royal Library, Denmark; and the National Archives UK. These sustainability concepts may be applied to the digitization processas a whole, not just to the creation and management of data formats. Factors to consider may include:

- Library of Congress - (Sustainability of Digital Formats, Planning for Library of CongressCollections - http://www.digitalpreservation.gov/formats/index.shtml)
- ✓ Disclosure
- ✓ Adoption
- ✓ Transparency
- ✓ Self-documentation
- ✓ External dependencies
- ✓ Impact of patents
- ✓ Technical protection mechanisms

- KB – (National Library of the Netherlands, *Evaluating File Formats for Long-term Preservation* http:// www. kb. nl/ hrd/ dd/ dd_links_en_publicaties/ publicaties/KB_file_format_evaluation_method_27022008.pdf)
  - ✓ Openness
  - ✓ Adoption
  - ✓ Complexity
  - ✓ Technical Protection Mechanism (DRM)
  - ✓ Self-documentation
  - ✓ Robustness
  - ✓ Dependencies
- KB – Denmark (Denmark, the State and University Library and the Royal Library, *Handling File Formats*, 2004 - http://netarchive.dk/publikationer/FileFormats-2004.pdf)
  - o Openness
  - o Portability
  - o Quality
  - o Monitoring obsolescence
- National Archives UK - (National Archives, UK, *Selecting File Formats for Long-TermPreservation-*
- http://www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf)
  - o Open standards
  - o Ubiquity
  - o Stability
  - o Metadata Support
  - o Feature Set
  - o Interoperability
  - o Viability

o The following additional criteria should be considered for migration:
- ☐ Authenticity
- ☐ Processability
- ☐ Presentation

## 14.12 ASSUMPTIONS

- Cultural heritage organizations digitize materials for many purposes:
  - o To facilitate access
  - o For preservation
- To prevent loss of information due to:
- Obsolescence
- Deterioration
- Handling
- Theft or destruction
  - o For exhibits, publications, and web use
  - o For researcher/patron requests
  - o To support current business processes

- There is not a single approach to digitization and metadata creation/management for all projects.

  - Type of original, media type, quality and condition of originals, nature of information, preservation risk level, etc., will determine the complement of approaches for digitization and metadata requirements that are matched to the originals.

  - The specific work process should be tailored to each individual digitization effort

- Due to differences in the nature of and the media type of the originals and existing copies
- Due to the extent and nature of existing descriptive or bibliographic information
- For maximum efficiency and cost-effectiveness

  - If there is pre-existing descriptive or bibliographic information (metadata), in either hardcopy
    or electronic form, that is related to candidate materials, it should be collected and come into the digitization work process as early as possible, ideally in the pre-digitization phase.
    - More descriptive and technical information (metadata) will be created during the conversion(digitizing) process.
    - Descriptive work will need to be done post-digitization but prior to the completion of the project,if the metadata does not exist prior to conversion.
  - Approach to digital object identifiers may vary by class of digital object

- These activities encompass all types of originals (manuscripts, books, still photography, maps, plans, artifacts, audio, video, motion pictures, aerial photography, etc.)
- Long-term preservation functionality for digitized versions of originals should be provided by amanaged repository environment.

  - Accordingly, digitization activities should be aligned with a managed repository and its workprocesses and requirements.

- Digital resources will require metadata at the appropriate level - the level and completeness of the descriptive and technical metadata will vary depending on the original and media type. The level and extent of metadata provided should be determined prior to conversion and should be created according to defined criteria/standards, whenever possible.

## 14.13 GENERAL POLICY ISSUES

- Policies and procedures relating to digitization activities and the management of digitization projects

- Define policies and procedures
- Implement and ensure compliance with policies and procedures
- Different procedures selected depending on class or category of project

- Determine complement of digital objects, file types, and file formats for both preservation and access digitization projects

- Recommendations for sustainable formats for digital archival records and digital master copies

o Digital object characteristics (i.e., nature of raster image files, digital audio and video files, etc.)

o Digital conversion parameters (i.e., technical specifications followed for the creation of digitalobjects)

- Creation and management of metadata schema(s)

o Define the complement of metadata needed for preservation of digital objects
o Define appropriate complement of various metadata to ensure management of assets for desired retention time period (short or long term)
o Determine when in workflow what metadata is added or created
o Determine where metadata will be stored (embedded or in external system or both)
o Determine in what formats metadata will exist
o Determine relationship to identifiers of digital objects

- Definition of essential characteristics (significant properties) of the original– curatorial/archivaland technical.
- Determination of appropriate approach and quality levels for digitization

o Approaches to digitization
o Conversion specifications for preservation reformatting
o Establish quality management approach
o Establish metrological approach for scanner and digital conversion equipment performance

- Ensure authenticity of digital copies

o Verification procedures for digital copies
o Comparison and review of digital copy to original record from the library/archival/curatorial perspective to ensure digital copy satisfies requirements for authentic digital versions.
o Document chain of custody – both original records during digitization and the digital copies
o Audit-trail – history of actions on the digital copies and related metadata, from creation through final submission to digital repository.
o Verification of fixity information such as checksums and digital signatures
o Relationship to identifiers of digital objects

- If applicable, ensure appropriate records management of the digital resources to be created

o Define records management issues related to the digital copy
o Define class/status of new copies (e.g., records, copies, non-records, master files, access/ distribution/derivative files) in order to determine digitization approaches and management of file types and content over time.
o Identify party/parties responsible for managing the various digital versions and ensure appropriate records management depending on status of digital copies.
o Define procedures and methods for accessing the digital copies for most or all use requests

## 14.14 GENERAL COMPONENTS FOR DIGITIZATION TECHNICAL REQUIREMENTS

### *14.14.1 Technical Guidelines*

Requirements for capture/conversion of original materials, production workflow, editing and processing operations, use of targets, equipment characterization and performance, quality control, etc.

o Types of originals addressed – textual, microfilm, still photographs, aerial photographs, maps, architectural and engineering plans, and objects and artifacts.
o Digitization information capture levels-
- Recommended requirements – preferred minimum requirements for preservation reformatting

o For textual records – generally considered by the digital library community to be equivalent of 35mm microfilm
- Alternative minimum requirements – appropriate for reference quality scanning

o Equivalent to the Transfer Guidance for Federal Agencies
o Specification used for NARA's pilot Electronic Access Project

### *14.14.2 Definition of Digital Object Classes and Data Formats*

o Recommendations for digital object types for different types of originals and the corresponding data formats for the digital objects.

### *14.14.3 Classification of Digital Copies Produced from Physical Records*

o Define the copy or record status of digital objects, the purpose and type of digital files, and the generation of digital file (master or derivative). Copy or record status is a characteristic of a digital file, and the copy or record status can change over time. The purpose and type of digital file relates to the initial intended use and related information capture levels for digitization. The combination of copy/record status, purpose and type of file, and generation determines both retention as well as management requirements to ensure data integrity, authenticity, disaster recovery, preservation, and access over time.

### *14.14.4 Identification of Work Processes, Responsibilities, Roles, Resources*

o Identification of units and staff that will perform different digitization related activities and
o Identification of current and future IT systems/applications that may be used for different digitization related activities, including storage

## 14.15 METADATA FOR DIGITIZATION

o Defines the recommended complement of metadata that could be considered for a digitization project, and identifies a list of minimum metadata elements for digitization projects, the category of metadata to which they belong, and the level at which they might apply. In general, these elements encompass metadata that documents: the original record, the digital resource, the process of creating the digital resource, changes made to the digital resource over time, and the content.

o Identifies methods, tools, systems for implementing metadata in the digitization workflow

o Defines standards, formats, and schemas followed

## 14.16 IDENTIFIERS

o Determination of approach to identifiers, whether identifiers are system-assigned, actionable, descriptive, or all of the above; at what level identifiers are applied, what is their role in production workflow, metadata creation, and fixity information; how are they cross walked across systems or structures; are they local or standardized; etc.

## 14.17 QUALITY MANAGEMENT

o Defines all activities that determine quality policies, objectives, and responsibilities, as well as the implementation of these activities by processes such as quality planning, quality control, quality assurance, and quality improvement. See ISO 8402, *Quality Management and Quality Assurance*
.

## 14.18 DIGITIZATION ACTIVITIES AND PHASES

***Activities***

The processes necessary to appropriately manage digitization projects include the following high-level activities, which can be grouped into a sequence of general phases (see below):

- Selection, Assessment, and Prioritization (to determine what materials will be selected, what projects will be approved, and how they will be prioritized).

- Determination of access/use restrictions or copyright, condition of records, copy status of digital resources, and approach to digitization and metadata

- Review and approval process

o Of digitization projects
o Of technical and metadata approaches

- Project Planning, Management and Tracking
- Determine what resources are required and available for projects
- Communication about and coordination of digitization projects
- Tracking of records throughout the process
- Digital Copy Status and Records Management
- Review reasons for digitization and evaluate status of originals
- Determine copy and record status for digital objects and for metadata to be created
- Manage and document process appropriately to ensure authenticity of digital copies
- Finalize status of digital copies and related metadata
- Update status of original records if needed
- Preparation of Originals for Digitization
- Bibliographic or archival preparation, preservation preparation, etc.
- Metadata
- Collection, creation, management, and reuse in other systems of all types of metadata (not just descriptive)
- Quality assurance and quality control of metadata
- Validation and verification of metadata (both technical and curatorial)
- Identifiers
- Determination of format, use in workflow and systems, standardization
- Digitization

- Digital reformatting
- Quality management, quality assurance, and quality control of digital copies
- Metrological assurance and device conformance
- Validation and verification - curatorial verification of the digital copies; technical verification ofdigital objects to technical approach
- Submission of Digital Resources to
- Access and delivery systems
- To digital repository

o May include creation and ingestion of submission information package

- Data Collection and Management
- Entry, collection, import, export, etc. of digital copies and metadata
- Links to all appropriate IT systems

Manage and make available digital copies and related metadata

o May include management of archival information package (content preservation) and provision of access to dissemination information package to end-users

- Assessment and Evaluation
- Per project
- Assessment of impact of digitization on other activities –such as business processes
- Data collection

- Project assessment, evaluation, and reporting

***Phases***

Digitization can be broken out into a sequence of project phases. In all phases, activities (above) can be grouped into management, operational, and program assessment categories. The phases follow a general sequence of steps, and can be grouped into:

- Project planning activities
- Pre-digitization activities (selection, assessment, and prioritization; preparation, and metadata collection and creation).
- Digitization activities
- Post-digitization activities (submission to delivery and repository systems, data collection and management; making digital copies and associated metadata available; assessment and evaluation). Post digitization activities also include creation and ingestion of a submission information package into a repository, management and preservation of the archival information package, and provision of access to the dissemination information package to end users; however, these activities are not specifically addressed here.

Activities such as project management and tracking, quality management, process improvement, as well as metadata management/collection (of all types of metadata) are ongoing processes that will continue throughout the entire digitization project.

Specific work activities within these phases may take place in a single phase or in more than one phase. As an example, activities like collection/creation of descriptive information may take place at different and/or multiple points during the metadata collection and creation process, the digitization process, and/or the data collection and management process. Therefore, many activities may take place at different points in the chronology and/or repeat at different stages of the workflow depending on the:

- Type of original, media type, and physical copies available to be digitized
- Condition and usability of the originals and/or copies to be digitized
- Nature of the digitization effort, see approaches listed in the table on page 4
- State and extent of processing done for the originals being digitized
- Nature and extent of descriptive information available in hardcopy and/or electronic form

many activities are also likely to occur concurrently, rather than sequentially.

Planned digitization projects will likely start at the beginning of the sequence of phases, while other efforts (like exhibits, fee requests, reference requests, etc.) will probably start in the middle of the sequence of phases. For example, exhibits, fee requests, and reference requests may start directly with digitization.

### 14.18.1 PROJECT PLANNING

Selection, assessment, and prioritization

- Selection, assessment, and prioritization of digitization projects/candidate materials

o Includes both external partnerships and internal access projects
o Nominations for proposals/priorities for digitization informed by agency/institutional priorities and researcher/public interest

- Priorities to provide enhanced access to high value and/or high use collections/records

- Scale, scope, comprehensiveness may be factors in selection and drive priorities for external/ partnership digitization

o Collect information, evaluate needs, and analyze collections/records selected to determine status/ extent of description, cataloging, and processing, access/use restrictions and permissions, best format for digitization (images, full text, etc.), physical characteristics (bound/disbound, foldouts, etc.), physical condition, restrictions and copyright, etc.

o Will the whole series/collection, a selected segment, or a cross section be digitized? Will access be provided via bibliographic records, a finding aid, or some other means? If the collection is treated in selective fashion, will the access tool also be selective or will it describe the entire collection or content body?

o Digitization done in-house, by a partner, or by a contractor
-Development of contract, RFP, procurement, etc.

o Preservation reformatting
o Priorities for digitization informed by formal risk assessments
o Classified Records Review (for federal agencies)
o Priorities for digitization informed by formal risk assessments
o Exhibits and Publications
o Priorities may be driven by Exhibits schedule
o Evaluation of whether digitizing more than is required for an exhibit would benefit an existing digital collection or future digitization projects.
o Fee and Reference Requests (Keep digital copies produced across institutions? May be a decision to be made by staff based on importance of collections/records or illustrates a type of original?)
o Conservation Treatment Documentation and Object Inventory

- Digitization of Current Business Documents

- Approve projects and work based on defined decision-making criteria, including available resource and capacity analysis

## 14.19 PROJECT MANAGEMENT AND TRACKING

- Project management
- Communication and coordination
- Coordinate digitization activities across institution
- Define organizational roles and responsibilities
- Plan and establish staff resources
- Identify any constraints and challenges - relating to technical, staffing, financial, and schedulingissues
- Acceptance and review of project proposals
- Access driven projects
- Risk assessments for preservation reformatting projects
- Identification and definition of project parameters and approaches to digitization and metadata(consult existing metadata schema if possible)
- Determination of identifiers and/or file naming approaches
- Identification of any IT issues: databases, software, storage needed, Web issues, etc.
- Identification of available bibliographic records, finding aids, and existing descriptive metadata,plan for processing/cataloging
- Review/approval of procedures for conversion/reformatting, metadata, digital copies, etc.
- Ensure compliance with specification/guidelines for digital resource creation and metadata, andwith pre-defined templates and profiles
- Creation and management of appropriate metadata schema
- Manage workflow for all activities
- Management of project documentation, including compliance with industry standards
- Identification of available resources – staff, supplies, equipment, etc.
- Determine if any staff training is needed
- Address issues of conversion site if applicable (particularly with partner projects) in terms oftransfer of originals, security, etc.
- Establish project timeline
- Project timeline may be influenced by other requirements; if so, resources required to complete the digitization effort may need to be changed to meet specific deadlines
- Project tracking
- Track, manage, and document activities and inventory projects
- Document all procedures and processes

## 14.20 COPY STATUS AND RECORDS MANAGEMENT

- Review reasons for digitization and evaluate status of original records/originals
- Preliminary determination of copy and/or record status for digital objects and for metadata to be created

### 14.20.1 PRE-DIGITIZATION

- Project management continues
- Project tracking continues
- Tracking location of originals during pre-digitization processes
- Data assessment and aggregation
- Establish access to any existing metadata, documentation, cataloging, or archival description to be used to facilitate the digitization process and the intellectual organization of the digital resources
- Identification of bibliographic records, finding aids, indices, folder lists, inventories, etc. in bothhard copy and electronic format
- Identification of electronic metadata held in management systems and access systems
- Quality assurance, quality control, verification and validation
- Preparation
- Curatorial/archival preparation of physical originals/records
- Analysis of originals (formats, organization, condition, copies, size, etc.)
- Physical and intellectual organization
- Collect and record a more detailed level of descriptive metadata during the course of curatorial/ archival preparation work to enhance description in existing systems
- Create, assign, and record appropriate records management/administrative metadata for new digital resources
- Batch records for conversion
- Preservation preparation
- Evaluation of physical condition and readiness for scanning
- Holdings maintenance, if needed
- Conservation prep, if needed
- Batch records for conversion
- Requirements review
- Define metadata requirements for different collections/groupings/classes of resources and determine minimal level of appropriate metadata to provide adequate access to and long-term preservation of digital copies.
- Identification of appropriate metadata schema or templates
- Identification of appropriate minimum complement of metadata
- Descriptive, administrative, technical, and structural metadata
- Indexing– if done before scanning
- Records management metadata, if applicable
- Determination of identifiers and/or file naming approach
- In registry
- Assigned by system or repository
- Actionable at file level, resolved by system
- Identifiers used by/in descriptive systems
- Identifiers used in production workflow
- Original identifiers
- Role of identifiers in fixity, authenticity auditing and reporting

- Application of identifiers at what level (collection, series, item; digital object, file)
- Cross walking of identifiers across systems
- Standardized identifiers or local identifiers
- Definition of essential characteristics – curatorial/archival and technical
- Define legal admissibility/authenticity of digital copies of records, if applicable
- Determination of appropriate approach and quality levels for digitization
- Approaches to digitization (including image capture specifications, testing and evaluation, workflow, header information, image processing, compression, file naming, file directory structure,file formats for archiving and for presentation, etc.)
- Conversion specifications for preservation reformatting
- Determine any special production needs
- Versions - determination of types and number of digital versions to be created during project
- Define copy type/record status for resources being created (such as preservation master, production master, derivative files, etc.)
- Define levels of access and storage for copy types
- Determination of naming and directory structure schemes
- Determination of file storage needs
- Establish QC/QA Procedures
- Define user interface and digital resource delivery requirements, if necessary
- Copyright and privacy issues
- Review and identify use and access restrictions on collections/items
- Resolve and restrictions and permissions issues
- Digital copy status and records management
- Finalize recommendations for copy status of digital objects and status of metadata
- Determine and assign responsibility for managing the digital objects and metadata
- May vary depending on copy/file types for versions of the same resources

### 14.20.2 DIGITIZATION

- Project management continues
- Project tracking continues
- Monitor status and products of all activities
- Check in and checkout of items to production unit
- Data entry
- Record any pre-existing metadata needed to begin conversion (may include job trackinginformation, descriptive metadata, etc.)
- Digital Conversion
- Capture done according to specifications in-house, by partners, and/or by contractors
- Image target use for performance verification
- Device conformance testing and calibration
- Initial and on-going testing of digital image quality and equipment based on established benchmarksand specifications
- Digitization of existing documentation, if not already in electronic form
- Digitization of descriptive information, finding aids, indices, folder lists, inventories, etc. if not inelectronic format

- Perform any correction/editing/processing to digital files
- Image evaluation – objective and subjective
- Create and track production metadata.
- OCR and text conversion/mark-up, rekeying, etc.
- Technical, structural, administrative, and descriptive metadata creation and collection
- Define requirements for and record metadata for different collections/groupings/classes ofresources at different levels.
- Create and record/embed metadata into appropriate systems/headers
- Auto characterization and manual and automated collection of technical and other metadata tocarry forward as files are moved into other systems
- Indexing – minimal intellectual organization of digital objects to match the appropriate level within the archival descriptive hierarchy or to match the intellectual organization of the collection. Indexing is primarily geared towards describing and organizing large groups of digital versions of physical records. Indexing provides a level of association and organization of digital resources sothey can be effectively searched and retrieved.
- Role of identifiers
- Quality management - quality assurance and quality control of digital copies and metadata toensure conformance to guidelines
- As with any manufacturing process, exceptions or defects can consume an inordinate amount ofresources; the further downstream the error detection, the greater the resource use to correct
- Defect identification and inspection and verification of files
- Automated quality assurance/quality control for both digital objects and for related metadata (alltypes of metadata – including technical, administrative, descriptive, etc.)
- Follow up by staff on problems identified by automated checks
- Statistically valid sampling checks by staff, automated identification of resources to be checked
- Rework for error identification
- Ensure compliance with templates/profiles
- Follow established metrology protocols and document certifications, or correct and replace asrequired
- Documentation of quality assurance/quality control process
- Create and record QC/QA metadata
- Data entry/import
- Import technical, structural, descriptive, production, administrative, rights, QC/QA metadata intoappropriate systems on local level
- Import assets into appropriate systems on local level
- Collect and manage new data in central and local systems
- Version control
- Define and record relationship between types of files (such as preservation master, productionmaster, derivative files, etc.)
- Automate production of derivative files and versions
- Automation of metadata into and out of header tags and files (such as XMP, IPTC, etc.)
- Perform inspection and verification of derivative files and versions

- ❖ Create and apply checksums to appropriate versions
- ❖ Create batches
- ❖ Aggregate multiple versions, files, and metadata files into a "package" for submission/deliveryinto storage.
- ❖ Role for identifiers
- ❖ Copy status and records management
- ❖ Manage and document process appropriately to ensure authenticity of digital copies

### 14.20.3 POST-DIGITIZATION

- ✓ Project management continues
- ✓ Project tracking continues
- ✓ Copy status and records management
- ✓ Finalize status of digital copies and related metadata
- ✓ Update status of original records/originals if needed
- ✓ Complete bibliographic/archival description and collection and creation of any additional appropriate metadata (descriptive, structural, administrative, technical) not collected in earlier processes

- Manage metadata

## 14.21 FINALIZE THE COMPLEMENT OF METADATA NEEDED

- ✓ Appropriate complement of various metadata to ensure management of assets for desiredretention time period (short or long term).
- ✓ Quality assurance and quality control of metadata and digital objects
- ✓ Conformance to standards, data types, templates/profiles
- ✓ Accuracy
- ✓ Defect identification and error correction
- ✓ Automated quality assurance/quality control for both digital objects and for related metadata (alltypes of metadata – including technical, administrative, descriptive, etc.).
- ✓ Follow up by staff on problems identified by automated checks
- ✓ Statistically valid visual checks by staff (i.e., color and tone accuracy), automated identificationof resources to be checked.
- ✓ Record actual rework/defect correction efforts
- ✓ Documentation of quality assurance/quality control process
- ✓ Curatorial/archival validation and verification of digital versions in comparison to originals from curatorial/archival perspective to ensure digital copies satisfy requirements for authentic digital versions
- ✓ Technical validation to industry specifications for well-formed digital objects and data formats;and assessment of digital objects to verify they meet local profile and requirements
- ✓ Make digital objects and metadata available to staff and researchers
- ✓ Deliver digital objects via web-based/delivery systems for research
- ✓ Deliver high-quality digital products via the web and via optical media
- ✓ Aggregate and associate digital objects and metadata files for packaging and transfer

- ✓ Create and associate multiple low resolution derivative files
- ✓ Assign checksums
- ✓ Export – flexible packaging of both digital objects and metadata for delivery into other systemsusing different metadata schema
- ✓ Submit resources to access/delivery systems and make resources available online
- ✓ Submit resources to digital repository
- ✓ Export metadata in different formats to other systems
- ✓ Export digital files to other systems
- ✓ Acceptance/confirmation of export/submission of digital objects and metadata into other systems
- ✓ Update metadata in other management and access systems as needed to synchronize or replace with new metadata generated during digitization projects
- ✓ Linking of metadata between systems
- ✓ Provide routine reference to digitized records via on-line systems
- ✓ Track and associate new digital/analog versions to the physical originals
- ✓ Manage digital resources in appropriate actively managed storage environment
- ✓ After submission of completed digital objects and related metadata to long term digital repository
- ✓ Ensure provenance and authenticity of digital resources
- ✓ Ensure data integrity
- ✓ Ensure disaster recovery
- ✓ Project assessment, reporting and evaluation
- ✓ Project Assessment
- ✓ Web, Image File, and Database Usage Analyses
- ✓ Cost-Benefit Analyses
- ✓ Assessment of impact on other activities
- ✓ Assess effects of digitization on traditional reference activities (e.g., online access, in-person access, send all source analog content offsite?) and researcher requests, and update procedures.
- ✓ Identify and correct problems and errors relating to both digital objects and related metadata
- ✓ Correct problems/deficiencies on a routine basis for all categories of digitization
- ✓ Lessons learned
- ✓ Unexpected results, scoping errors, etc.
- ✓ Process improvement – as needed update workflows, tools, procedures, policies, etc.

## 14.222 IT INFRASTRUCTURE NEEDS

### *14.22.1 Access:*

- Access to both digital files and metadata will be needed by both the public (online) and by internalstaff for the purposes of research, exhibits, publications, sale, etc.
- Provide a centralized IT workspace so copies are accessible to all staff during the work processin order to complete description, quality control work, etc.
- Versions will be moved into other systems for access, display, presentation, etc.
- Identify new digital versions in management systems to reduce duplication of digitization

effortsand to minimize handling of records.

- Metadata will be moved into other systems for access, display, presentation, etc.
- System to assign/register identifiers
- System to resolve identifiers, if actionable

### *14.22.2 Managed Storage:*

Infrastructure to store, manages, and provides access to digital copies

- Data storage - master files, access files, metadata, and data migration issues
- Need to ensure the- data integrity, disaster recovery, and authenticity of the digital resources created
- Provide minimal bit preservation activity
- Accept packages of digital files, versions, metadata, and information about submission process (verification and validation information, etc.).

o Ensure viability of data and maintenance of essential characteristics
o Incorporate checksums, validation, and verification functionality
o Monitoring
o Track change history to digital objects
o Authenticity/provenance chain

- Perform backups and redundancy to appropriate levels to ensure data integrity and disaster recovery
- Define server requirements - develop configuration management plan
- Network security issues
- Systems documentation
- Plan and budget for systems upgrades
- Site licenses and hardware/software maintenance contracts
- Move objects into one or more long-term destinations
- Ability to transfer digital objects and metadata into other systems for access purposes
- Ensure appropriate intellectual control of digital resources

o Synchronization of metadata and digital objects
o Updating of metadata and digital objects
o Manage relationships and associations between versions/multiple components, parent-child relationships, etc.

- At some appropriate point, digital resources and metadata move into digital repository

## 14.23 SELF ASSESSMENT QUESTIONS

1. Explain in brief what digitization is.
2. How to plan for digitization
3. What are the technical infrastructures needs for digitization?

## 14.24 REFERENCES

1. Cornell University Library/Research Departments. (2000), Moving theory into practice: digital Image for libraries and archives. Research Libraries Group. Available at http:// www.library.cornell.edu/preservation/tutorial
2. Digital Library Federation. (2001), Registry of Digitized Books and Serial Publication, Available at http://www.digilib.org/collections
3. Ding, Choo Ming. (2000), Access to Digital Information: Some Breakthrough and Obstacles, Journal of Librarianship and Information Science, Vol.32 No.1

4. Greenstone Training Workshop Material.(2002), Available at http://www.greenstone.org
5. Ian, H. Witten & David, Brainbridge. (2003), How to Build a Digital Library, London: MorganKaufman Publishers
6. Sitts, Maxine K. (2000), Handbook for Digital Projects: A Management Tool for Preservation and Access. Northeast Document Conservation Center, Andover, Massachusetts. USA.
http://www.nedcc.org/digital/dman.pdf
7. Smith, Abbey (2001), Strategies for Building Digitized Collection. Washington, D.C. Digital Library Federation, Council on Library and Information Resources. Available at http:// www.clir.org
8. Federal Agencies Digitization Guidelines Initiative (FADGI)http://digitizationguidelines.gov
9. FADGI Glossary
http://www.digitizationguidelines.gov/glossary.php
10. Library of Congress, *NDLP Project Planning Checklist.*
http://memory.loc.gov/ammem/prjplan.html
11. *NC Echo Guidelines for Digitization 2007 Revised Edition*
http://www.ncecho.org/dig/digguidelines.shtml
12. Washington State Library, *Digital Best Practices*
http://digitalwa.statelib.wa.gov/newsite/best.htm
13. Northeast Document Conservation Center, *Handbook for Digital Projects: A Management Tool for Preservation and Access,* 2000.
14. http://www.nedcc.org/resources/digitalhandbook/dman.pdf
15. Bibliographic Center for Research, Collaborative Digitization Project, *Best Practices and Publications*
16. http://www.bcr.org/dps/cdp/best/index.html
17. *NINCH Guide to Good Practice in the Digital Representation and Management of CulturalHeritage Materials,* 2002.
http://www.nyu.edu/its/humanities//ninchguide/
18. *The Archives New Zealand S-6 Digitisation Standard*
http://continuum.archives.govt.nz/files/file/standards/s6.pdf
19. ARMA International, *Generally Accepted Recordkeeping Principles*
http://www.arma.org/garp/
20. Library of Congress, *Sustainability of Digital Formats, Planning for Library of CongressCollections*
http://www.digitalpreservation.gov/formats/index.shtml
21. National Library of the Netherlands (KB), *Evaluating File Formats for Long-term Preservation* . http:// www. kb. nl/ hrd/ dd/ dd_links_en_publicaties/ publicaties/KB_file_format_evaluation_method_27022008.pdf
22. The State and University Library, Arhus, Denmark, and the Royal Library, Copenhagen, Denmark, *Handling File Formats* (2004)http://netarchive.dk/publikationer/FileFormats-2004.pdf
23. National Archives, UK, *Selecting File Formats for Long-Term Preservation*.http:// www.nationalarchives.gov.uk/documents/selecting_file_formats.pdf
24. National Archives and Records Administration, *Technical Guidelines for the Digitization of Archival Materials for Electronic Access – Creation of Production Master Files – Raster Images,* June 2004.http://www.archives.gov/preservation/technical/guidelines.pdf
25. Puglia, Steven. *The Cost of Digital Imaging Projects*. RLG DigiNews, Volume 3, Number 5,October 15, 1999.
http://digitalarchive.oclc.org/da/ViewObjectMain.jsp?fileid=0000070511:000006278991&reqid=7049#feature.

**LESSON-15**

# INSTITUTIONAL REPOSITORIES

**OBJECTIVE**

After reading this chapter, students will be able to understand:

- Basic fundamental concepts of Institutional Repository.
- Definitions of Institutional Repository.
- Benefits of an IRs
- Challenges of an IRs

**STRUCTURE**

## 15.1 INTRODUCTION

Information and Communications Technology (ICT) continues to transform the scholarly environment and management of higher education institutions. For example, ICTs are core resources required for digital publishing and online teaching and learning. ICT has created platforms and opportunities for scholars to work collaboratively through extensive infrastructures, with access to resources and knowledge services in borderless environments. The rapid growth of digital assets creates challenges in the use, management, archiving and application of digital information and datasets. That is why digital scholarship is the fastest growing academic phenomena today. Digital Scholarship (DS) has been popularly defined as "any element of knowledge or art that is created, produced, analyzed, distributed, published, and/or displayed in a digital medium, for the purpose of research or teaching" (Kirsten Foot, Assistant Professor, Department of Communication, University of Washington). However, this definition does not seem to capture the totality of digital activities in any tertiary institution. Therefore DS can be defined as "an integrated collaborative blended environment which embraces cutting-edge technologies in learning, teaching, research, professional and administrative services" (Task Group on UB and Digital Scholarship, 2008).

Recently academic institutions have been grappling with how to manage the digital intellectual output they produce including journal articles, conference papers, reports, theses & dissertation, teaching materials, artwork, research notes, and research data. Clearly, technology has made it easy to create, store and access digital material. Paradoxically however, while there is potential for instantaneous access, all too often many materials are not usually made accessibleto many users and they remain marooned in the authors' computers. About 80-85% of digital intellectual output of universities is never made accessible to the public (The Open Citation Project, 2004). Also the escalating costs of online journals prohibit subscription and it is becoming more unrealistic and challenging for libraries to subscribe to all, or even most of the online academic journals (Warren, 2003).

In response to the above mentioned conditions, Massachusetts Institute of Technology (MIT) announced a research project titled DSpace "to build a stable and sustainable long-term digital storage repository that provides an opportunity to explore issues surrounding access control, rights management, versioning, retrieval, community feedback, and flexible publishing capabilities"(DSpace Project, 2000).

## 15.2 INSTITUTIONAL REPOSITORIES

### *15.2.1 The Purpose of Institutional Repositories*

The institutions of higher education all over the world are experiencing the necessity of managing their education, research and resources in a more effective and open way. By making the research and scientiûc output easily available, they will support the development of new

relationships between the academicians and both national and international research centres. This will facilitate: The institutions of higher education will have to assume that the learning improvement is the key for the success of the Information Society.

### *15.2.2 What do we call an Institutional Repository?*

In the simplest sense of the term, an Institutional Repository is an electronic archive of the scientiûc and scholarly output of an institution, stored in digital format, where search and recovery are allowed for its subsequent national or international use. A repository contains mechanisms to import, identify, store, preserve, recover and export a set of digital objects, usually from a web portal.

Those objects are described by labels ('metadata') that facilitate their recovery. From a more conceptual point of view, the IR forms an authentic management system of contents, given that, apart from the documents themselves, the repository offers to the academic community a set of services for the management of that output. The IR is a means of scientiûc communication,but it cannot be understood as a publication channel; it must be understood as a complement to the process of scientiûc publication formalised with peer review. The intellectual collections includethe research output (articles, theses, communications, etc.), teaching materials, and administrative documents as well as those documents generated by the institution, all in various formats like texts, presentations, audio-visual records and e-learning objects.

Until now, interest in IRs has mainly concentrated on the research output, since this constitutes an indicator of the institutions' performance at the time of obtaining funding. On occasions, research grants require as a condition that publications are open. In the teaching ûeld, the IR which is usually integrated in the e-learning system– facilitates a change of paradigm in teaching and learning, contributing a pedagogical environment rich in information.

## 15.3` DEFINITIONS OF INSTITUTIONAL REPOSITORY

There are a number of definitions for "institutional repository" (IR). Here are a few key ones:

***Clifford Lynch, "Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age"***

In my view, a university-based institutional repository is a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution.

***Mark Ware, Pathfinder Research on Web-based Repositories***

An institutional repository (IR) is defined to be a web-based database (repository) of scholarly material which is institutionally defined (as opposed to a subject-based repository); cumulative and perpetual (a collection of record); open and interoperable (e.g. using OAI-compliant software); and thus collects, stores and disseminates (is part of the process of scholarly communication). In addition, most would include long-term preservation of digital materials as a key function of IRs.

*Raym Crow, the Case for Institutional Repositories: A SPARC Position Paper*

Institutional repositories . . . digital collections capturing and preserving the intellectual output of a single or multi-university community provide a compelling response to two strategic issues facing academic institutions. Such repositories:

Provide a critical component in reforming the system of scholarly communication a component that expands access to research, reasserts control over scholarship by the academy, increases competition and reduces the monopoly power of journals, and brings economic relief and heightened relevance to the institutions and libraries that support them; and

Have the potential to serve as tangible indicators of a university's quality and to demonstrate the scientific, societal, and economic relevance of its research activities, thus increasing the institution's visibility, status, and public value.

## 14.4 IR AS A MAJOR BENCHMARK OF DIGITAL SCHOLARSHIP

IR proponent Lynch (2003) regards institutional repositories as essential infrastructure for modern scholarship. He argues that "the development of institutional repositories emerged as a new strategy that allows universities to apply serious, systematic leverage to accelerate changes taking place in scholarship and scholarly communication". Also, Markey et al (2007) remarked that, "a considerable portion of the scholarly record is born digital, and some scholarship is produced in digital formats that have no physical, in-the-hand counterparts. The proliferation of digital scholarship raises serious and pressing issues about how to organize, access, and preserve it in perpetuity. The response of academic institutions has been to build and deploy institutional repositories (IRs) to manage the digital scholarship their learning communities produce".

All the above definitions of an IR agree that it is an enabling component of digital scholarship and this paper works on the premise that IR is currently a cornerstone of DS and an important tool to manage institutional intellectual output. This seems to suggest that the success of DS can at least partially, if not fully depend on and be measured by the success of IRs.

## 14.5 IR CONCEPTS AND DIGITAL SCHOLARSHIP

There are specific concepts integral to digital scholarship that an IR does or potentially can characterize, both in concept and functionality. These are interoperability, research management and new forms of publishing. They further support the assertion that an IR is currently a major benchmark of digital scholarship.

## 14.6 INTEROPERABILITY

The Open Archives Initiative (OAI) designed a shared code for metadata tags (e.g., 'date', 'author', 'title', 'journal' etc.). So full-text documents may be in different formats and locations, butif they use the same metadata tags they become interoperable. It means that their metadata can be harvested and all the documents can then be jointly searched and retrieved as if they were in one global collection, accessible to everyone (Open Archives Initiative 2002). Signing up to OAI compliance (by choosing for example open access, open source platforms) and committing to interoperability with other repositories indicates an ability and openness on behalf of the institution to contribute to global scholarship. Opening the repository to cross repository searching tools (OAI and other harvesters as well as the generic indexing in tools like Google Scholar) exposes scholarshipin a new way and puts it in an international context (Westell 2006). The institutional repository maybe the only vehicle to attract an international audience to the institution.

Another type of interoperability is integration of the IR with course management tools. There is potential to extend the objective and content of an IR beyond the 'purist' content, viz: copies of peer reviewed research articles, to embrace classroom, and distance teaching and learning materials and offer open educational resources. Arguably these developments have evolved from the open access culture proselytized first by digital research repositories.

## 15.7 LINKAGE WITH DIGITAL RESEARCH MANAGEMENT SYSTEMS

Some developers see potential for developing IRs beyond their current capacity, even believing that the current repository concept is flawed (Stuart, 2008). They posit that institutions should provide a system for 'work in progress' complete with descriptive data to associate it with the appropriate research grants, researchers, departments and funders. The current conceptual 'institutional repository' should be a slice in this data-corpus, as a logical front end for a system which manages the entire research process, rather than an independent initiative. In effect IR functionality would be part of a current research information system (CRIS), bringing together the information that underpins the complete process of research, from grant application up to and including peer-reviewed publications.

## 15.8 NEW PUBLISHING AND COMMUNICATIONS MODELS

Discussions concerning IRs as a possible alternative publishing model split IR adherents into different camps, between innovators and purists. One model frames deposit in a repository as an adjunct, and complementary to, the traditional publication process (Hunter, 2007). Lynch (2003) for example firmly believes "it underestimates the importance of institutional repositories to characterize them as instruments for restructuring the current economics of scholarly publishing rather than as vehicles to advance, support, and legitimize a much broader spectrum of new scholarly communications". The other group sees repositories as the beginning of new forms of academic publishing (Hunter, 2007). For example, in an interview about the University of California (UC) eScholarship repository, Candee (2006), the director of publishing and strategic initiatives inthe Office of Scholarly Communication, agreed that the eScholarship Repository had effectively become a publishing platform for the University of California Press. She further added, "I think it is unfortunate that the term institutional repository has come to mean something narrower.., ultimatelyI envision a very different arrangement between universities and publishers than we have now". The UC vision is that universities will be in control of the publishing process, with all content hosted and managed by universities themselves. It is possible that in future IRs will be regarded as the principal route for the dissemination of research papers.

## 15.9 BENEFITS OF AN IR

The benefits of repositories to institutions and individuals are numerous and can be grouped into the following categories (Pickton & Barwick (2006)):

## 15.10 SPECIFIC TO THE UNIVERSITY, AN IR OFFERS

1. Increasing visibility and prestige. A high profile IR may be used to support marketing activitiesto attract high quality staff, students and funding.
2. Centralization and storage of all types of institutional output, including unpublished literature.
3. Support for learning and teaching. Links may be made with the virtual teaching environment and library catalogues.

4. Standardisation of institutional records. The compilation of an 'Institutional CV' and individualonline dossiers linked to the full text of articles becomes possible.
5. Ability to keep track of and analyse research performance.
6. Breaking down of publishers' costs and permissions barriers.
7. Alleviation of requirement to trust publishers to maintain information in the long term, withoutany commercial benefit for the authors.
8. Promotion of a philosophy of wider communication.

## 15.11 SPECIFIC TO AUTHORS, AN IR ENHANCES:

- Dissemination and impact of scholarship. Some studies have estimated that open access articles are cited 50% to 250% more than non open access articles. In some disciplines, online files receive on average 300% more citations than materials available only in paper format (The Open Citation Project, 2004). Also, Google Scholar gives preferential treatment to materials in IRs; a paper picked up from an IR would appear higher up on the Google results list (Ashworth 2006).
- Storage and access to a wide range of materials. Many authors lack time, resources, or expertise to ensure preservation of their scholarly work. Through an organizationally based IR strategy, long-term accessibility and greater security of work is assured. Research items get a permanent URL compared to a personal or departmental web site.
- Feedback and commentary from users. Authors are able to receive and respond to commentary on 'pre-prints'.
- Added value services; such as hit counts on papers, personalised publication lists and citation analyses.
- A central archive of a researcher's work.
- A researcher's profile.
- Benefits to researchers and their institutions in terms of prestige, prizes and grant revenue.

## 15.12 SPECIFIC TO SOCIETY/COMMUNITY

As scholarship is shared, society at large benefits. Maximising public access to research findings online, in turn maximises its visibility, usage and impact. It also maximises its benefits to research itself (and hence to the society that funds it) in terms of research dissemination, application and growth, research productivity and progress. (ePrints website FAQ, n.d.). Also more sponsors of funded research now have mandates for authors to deposit their articles and other research outputs as a condition for funding.

Some policies promote Open Access for funded research. These requirements are intendedto increase readership, re-use and dissemination of research outputs. The message to researchersis that research is incomplete until the output is widely disseminated.

To summarise, the potential uses of an IR are: scholarly communication; management and storage of learning materials, electronic publications and research collections; preservation of digital research work; building university profile by showcasing academic research work; providing an institutional leadership role for the library; research assessment; encouraging open access; and housing digitised collections (Barton and Waters, 2004).

## 15.13 CHALLENGES OF AN IR

Despite the numerous benefits of an IR, there are implications and potential barriers to itssuccess as summarized below (Pickton & Barwick, 2006):

### 15.13.1 COST:

The initial financial cost for an open source software adopted by most institutions for creating IRs is not high but the recurrent costs, especially staff costs (e.g. time spent drafting policies, developing guidelines, publicising, training, supporting users and creating metadata, specialist IT consultancy) may be significant. This is further discussed below.

### 15.13.2 DIFFICULTIES IN GENERATING CONTENT

A successful IR depends on the willingness of authors to deposit their work voluntarily and there may be local barriers and hindrances to be overcome. There are acknowledged difficulties in generating content, especially at the beginning. Unless the value of an IR can be demonstrated quickly, the organization's long-term commitment to the project may begin to wane. The best way to prove the enduring value of the IR and to ensure its long-term survival is to quickly populate it (Gibbons, 2004).

### 15.13.3 SUSTAINING SUPPORT AND COMMITMENT

Far too often, it is difficult to sustain continuous support and commitment from the management and academic staff. Lynch (2003) has succinctly described this obstacle: "Stewardshipis easy and inexpensive to claim; it is expensive and difficult to honour, and perhaps it will prove tobe all too easy to later abdicate". There is a need for institutions to think seriously before launching institutional repository program as it may disintegrate rapidly if not properly managed.

### 15.13.4 RIGHTS MANAGEMENT ISSUES

Sometimes researchers are apprehensive about infringing publishers' copyright and lack adequate awareness about their own intellectual property rights. They may be uncertain about making their work available online before it is published by a traditional publisher.

### 15.13.5 POLICY ISSUES

Experiences suggest that an IR will only function to its capacity when a mandate is in place to populate it but clearly researchers can react negatively to any suggestion of compulsion. Lynch (2003) has cautioned that an IR should not become a tool for enforcing administrative control over academic work.

### 15.13.6 LACK OF INCENTIVES

In the absence of any incentive academics feel reluctant to provide even bibliographic details of their scholarly output especially when they know that incentives are available in other institutions.

## 15.14 ROLE OF LIBRARIANS IN AN IR

Pro-activity and responsibilities relating to IRs are assumed by different people in various institutions. Largely they will be undertaken collaboratively by officers within the library in

partnership principally with research and development, and information technology sections. Stimulating engagement for buy-in is crucial in the early stages of an IR when efforts are made to build a critical mass of material. Nixon (2002) rightly observed that "Reference librarians are a library's eyes and ears. They understand users needs and perceptions. They know what's working and what's not. When they act as subject selectors, they are the library's primary liaison with faculty intheir subject areas and its most visible representatives. They know how to help, inform, persuade, and teach users. For an IR to succeed, it is essential that they be involved in its planning, implementation, and operation." So librarians have critical roles to play in both establishing and maintaining an IR through:

### 15.14.1 ADVOCACY:

Librarians need to know all about the IR, its principles, benefits and operational processes in order to promote it and act as 'IR evangelists' (Ashworth 2006). Librarians will need to develop advocacy programs, publicise IR through institutional news media and respond to questions by the stakeholders.

### 15.14.2 BUILDING CONTENT:

Librarians can employ advocacy and marketing strategies to promote engagement with faculty members and help to generate content. They can also assist by proactively searching for content independently.

### 15.14.3 COLLECTION ADMINISTRATORS AND METADATA SPECIALISTS:

Librarians have potential roles as collection administrators and metadata specialists. For effective implementation of IR, libraries will need to recruit or train librarians with digital collectionmanagement and provide a mediated deposit service for reluctant 'self-archives''.

### 15.14.4 TRAINING:

Librarians should be able to train staff and students to use the IR and help them prepare their digital products.

## 15.15 SOFTWARE ARE USED FOR INSTITUTIONAL REPOSITORIES

A variety of systems are in use. IR software may be supported in various ways (e.g., locally supported, centrally supported by a consortium of institutions, or supported for a fee by a vendor). Four commonly used systems are:

- **Digital Commons**, commercial software, http://www.bepress.com/ir/
- **DSpace**, free open source software, http://www.dspace.org/
- **EPrints**, free open source software, http://www.eprints.org/
- **Fedora**, free open source software, http://www.fedora-commons.org/

## 15.16 SELF ASSESSMENT QUESTIONS

1. What is an Institutional Repository?
2. Explain the benefits & Challenges of IRs.

## 15.17 REFERENCES


1. American Council of Learned Societies. (2006). Our Cultural Commonwealth: The report of the American Council of Learned Societies Commission on Cyber infrastructure for the Humanities and Social Sciences. Retrieved April 4, 2008 from http://www.acls.org/programs/Default.aspx?id=644
2. Ashworth, S. (2006). Role of Librarians in the Development of Institutional Repositories Retrieved on March 15th 2008 from http://www.pfsl.poznan.pl/oa/ppt/2.ppt
3. Barton, M. R. & Waters, M.M. (2004). Creating an Institutional Repository: LEADIRS Workbook. MIT Libraries. Retrieved April 6, 2008 from http://www.dspace.org/implement/leadirs.pdf
4. Candee, C. (2006). Changing the paradigm, An Interview. Retrieved May 1, 2008 from http://poynder.blogspot.com/2006/01/changing-paradigm.html
5. Chan,D.L.H., Kwok, C.S.Y. & Yip, S.K.F. (2005). Changing roles of reference librarians: thecase of the HKUST Institutional Repository. Reference Services Review, 33(3), 268-282, Retrieved on March 15th2008 from http://repository.ust.hk/dspace/bitstream/1783.1/2039/2/p268.pdf
6. Crow, R. (2002). The Case for Institutional Repositories: a SPARC Position Paper, Retrieved April 26,2008 from http://www.arl.org/sparc/bm~doc/ir_final_release_102.pdf
7. DSpace Project. (2000). MIT Faculty Newsletter, XII (4), Retrieved April 27, 2008 fromhttp://www.dspace.org//news/articles/dspace-project.html
8. ePrints website FAQ. (n.d.). What is the purpose of self-archiving? What is the Open Archives Initiative (OAI)? Self-archiving FAQ. Retrieved April 24, 2008 from http://www.eprints.org/openaccess/self-faq/
9. Gibbons, S. (2004).Benefits of an institutional repository. Library Technology Reports, 40(4),11-16.
10. Hunter, D. (2007). Repository: publication or archive? JISC_Repositories listserve discussion thread, Retrieved April 6, 2008 from www.jiscmail.ac.uk/archives/jisc-repositories.html
11. Johnson, R. (2002). Institutional repositories: partnering with faculty to enhance scholarly communication. D-Lib Magazine, 8(11), retrieved May 8th from http://www.dlib.org/dlib/ november02/johnson/11johnson.html
12. Lynch, C. (2003). Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age. ARL, 226, Retrieved April 13, 2008 from http://www.arl.org/bm~doc/br226ir.pdf
13. Nixon, W. (2002). The evolution of an institutional e-prints archive at the University of Glasgow. Ariadne, Retrieved April 24, 2008 from http://www.ariadne.ac.uk/issue32/eprint-archives/
14. Open Archives Initiative. (2002). The Open Archives Initiative Protocol for Metadata Harvesting version Retrieved April 24, 2008 from http://www.openarchives.org/OAI/ openarchivesprotocol.htm

OpenDOAR. (2008). The Directory of Open Access Repositories. Retrieved April 4, 2008 from http://www.opendoar.org/index.html

**LESSON-16**

# OPEN SOURCE SOFTWARE FOR DIGITAL LIBRARIES (GSDL, DSPACE & E-PRINT)

**OBJECTIVE**

After reading this chapter, students will be able to understand:

- Basic fundamental concepts of Open Source Movement
- Advantages of Open Source Software (OSS)
- Selection criteria of open source software
- Overview of Open Source Softwares (GSDL, Dspace & Eprint)

**STRUCTURE**

## 16.1 INTRODUCTION

Open source software is computer software whose source code is available under a license (or arrangement such as the public domain) that permits users to study, change, and improve the software, and to redistribute it in modified or unmodified firm. It is often developed in a public, collaborative manner. It is the most prominent example of open source development and often compared to user generated content. For many libraries, organizing their books and other media can be daunting task, especially as the library grows with more material. Years ago we had crude card catalogue systems (remember the Dewey Decimal System) that kept things organized, but were difficult to maintain. With today's computing technology, organizing our libraries has never been easier or more efficient. Gone is the card catalogue and in some libraries, it's much easier to locate a book through and internet connection and picking it up upon your arrival, rather than wasting the time scouring the aisles looking for your next read. Now just because the world has been blessed with wonderful software solutions that make everything easier to do, doesn't mean that every library in the universe is using these solutions.

Many Libraries do not have enough amounts of money to burn, and any that they do get usually goes to purchasing additional resources. Because of this need for software (and the installation and training costs associated with any), and the lack of money available to spend on it, many libraries are left to fend for themselves when it comes to staying up to date with the latest technology. Unless, of course, they embrace the open source movement and use some of the countless software solutions available to help out. Most software that we all use every day is known as "proprietary", which in a nutshell means that it costs money and that the actual code of the software is restricted, in that the code of the software cannot be modified, copied, or changed from its original construction. The code is "unreadable" and pretty much is what it is. *Open source software*, on the other hand, is quite the opposite. The open source mentality revolves around sharing and collaboration, and these two important elements describe open source software perfectly.

First and foremost, open source software is free for anyone to have; more importantly, not only is the software free, but it is also free for anyone to copy, hack, modify, etc. This increases the possibilities of a software program's potential because of this free-thinking model. Many large groups of programmers have customized basic open source programs into whatever they deemed necessary, and have in turn given these modifications back to the open source community for free where others can continue to build on their work. There are many different kinds of open source software solutions out there today that could be embraced by the library. There's basic operating system, document processing programs, Library Management Software (LMS) and Digital Library software.

## 19.2 OPEN SOURCE MOVEMENT

In 1998, a group of individuals advocated that the term free software be replaced by open source software (OSS) as an expression which is less ambiguous and more comfortable for the corporate world. Software developers may want to publish their software with an open source software license, so that anybody may also develop the same software or understand how it works. Open source software generally allows anybody to make a new version of the software, port it to new operating systems and processor architectures, share it with others or market it. The aim of open source is to let the product be more understandable, modifiable, duplicatable, reliable or simply accessible, while it is still marketable.

The Open Source Definition, notably, presents an open-source philosophy, and further defines a boundary on the usage, modification and redistribution of open-source software. Software licenses grant rights to users which would otherwise be prohibited by copyright. These include rights on usage, modification and redistribution. Several open-source software licenses have qualified within the boundary of the Open Source Definition. The most prominent example is the popular GNU General Public License (GPL). While open source presents a way to broadly make the sources of a product publicly accessible, the open-source licenses allow the authors to fine tune such access. The "open source" label came out of a strategy session held in Palo Alto in reaction to Netscape's January 1998 announcement of a source code release for Navigator (as Mozilla). A group of individuals at the session included Todd Anderson, Larry Augustin, John Hall, Sam Ockman, Christine Peterson and Eric S. Raymond. They used the opportunity before the release of Navigator's source code to clarify a potential confusion caused by the ambiguity of theword "free" in English.

The 'open source' movement is generally thought to have begun with this strategy session. Many people, nevertheless, claimed that the birth of the Internet, since 1969, started the open source movement, while others do not distinguish between open source and free software movements. The Free Software Foundation (FSF), started in 1985, intended the word 'free' to mean "free as in free speech" and not "free as in free beer." Since a great deal of free software already was (and still is) free of charge, such free software became associated with zero cost, which seemed anti-commercial.

## 16.3 ADVANTAGES OF OPEN SOURCE SOFTWARE

**16.3.1 Lower software costs:** Open source solutions generally require no licensing fees. The logical extension is no maintenance fees. The only expenditures are for media, documentation, and support, if required.

**16.3.2 Simplified license management:** Obtain the software once and install it as many times and in as many locations as you need. There's no need to count, track, or monitor for license compliance.

**16.3.3 Lower hardware costs:** In general, Linux and open source solutions are elegantly compactand portable, and as a result require less hardware power to accomplish the same tasks as on conventional servers (Windows, Solaris) or workstations. The result is you can get by with less expensive or older hardware.

**16.3.4 Scaling/consolidation potential:** Again, Linux and open source applications and services can often scale considerably. Multiple options for load balancing, clustering, and open source applications, such as database and email, give organizations the ability to scale up for new

A growthor consolidate to do more with less.

**16.3.5 Support:** Support is available for open source often superior to proprietary solutions. First,open source support is freely available and accessible through the online community via the Internet. And second, many tech companies are now supporting open source with free online and multiplelevels of paid support. For example Liblime.

**16.3.6 Escape vendor lock-in:** Frustration with vendor lock-in is a reality for all IT managers. In addition to ongoing license fees, there is lack of portability and the inability to customize software to meet specific needs. Open source exists as a declaration of freedom of choice.

**16.3.7 Unified management:** Specific open source technologies such as CIM (Common Information Model) and WBEM (Web Based Enterprise Management) provide the capability to integrate or consolidate server, service, application, and workstation management for powerful administration.

**16.3.8 Quality software:** Evidence and research indicate that open source software is good stuff.The peer review process and community standards, plus the fact that source code is out there forthe world to see, tend to drive excellence in design and efficiency in coding.

## 16.4 SELECTION CRITERIA OF OPEN SOURCE SOFTWARE

Evaluation of open source software is different from proprietary programs. A key difference for evaluation is that the information available for open source programs is usually different than for proprietary programs; source code, analysis by others of the program design, discussion between users and developers on how well it is working, and so on. Often proprietary programs always hide all information from users and only allow running the software. Following criteria's can be adopted for open source software selection:

### 16.4.1 OPEN SOURCE SOFTWARES ON THE WWW

Most convenient option to identify particular software for your library need is to ask professional friends who have experience in using open source softwares. You can directly contact other libraries in your locality or post a message in any popular email discussion forum of librarians. Certain open source softwares are highly popular among librarians community, for example Greenstone digital library software is a favorite candidate for the libraries who make use it for the collection and organization of digital materials. Librarians can select the software without much effort, if more popular software's are available for various library purposes. Websites which provide detailed listing of open source software are:

- Free Software Foundations software directory (www.fsf.org)
- UNESCO Free & Open Source Software Portal (www.unesco.org)
- Source Forge (http://sourceforge.net/)

### 16.4.2 OPEN SOURCE LICENSES

Open source licenses are assure users freedom to use, copy, improve and distribution of software. GPL is the most popular license for free and open source software and provides feasible terms of use. Using GPL license, a user can modify the software without the permission of its creator. At the same time BSD license impose certain restrictions on modification of software without the permission of its developer. If one has decided to choose the software with non General

Public License, check the license if it contains any un-acceptable clauses.

### 16.4.3 FUNCTIONAL MODULES

Certain features or modules essential for day to day work may not be available with the initial development stages of open source softwares. In such cases, libraries have to purchase additional modules from open source service providers or make use the in-house expertise to build the required features. Functional modules essential for library management systems (ILS) are cataloguing, circulation, OPAC, serial control and acquisition. It is essential to read release notes of latest version and software roadmap to know which features are already available and are expected in future. Ensure the availability of standards like MARC, Z39.50, and Dublin Core whichare essential for exchange of bibliographic information in library softwares.

### 16.4.5 STABLE RELEASES

Stable release of open source software shows its developer's ability to fix and correct bugs along with new features. Version history of open source software is often available from project websites or any other project repositories like Source Forge (www.sourceforge.net), Savannah (savannah.net) and Free Software Foundations software directory (www.fsf.org). These services help users to check the information regarding software origin, releasing history, version numbering scheme, developers details etc. Actively maintained open source projects mention even the releasing dates of forthcoming versions.

### 16.4.5 DEVELOPERS AND USER COMMUNITY

The development and maintenance of open source software is a social collaborative activity. Open source software is actively developed on a 24-hour basis by a large number of programmers from all over the world. Depending on the success of a certain open source software project, this results in a development process that out paces that of many competitors. Another aspect of open source software is that, many different people and organisations look at the software from a different perspective. This leads to invaluable discussions on what direction the development should be taken. Many IT experts claim that, it is this multi-cultural and multi-organisational influence that, combined with the global spreading and fast development pace, makes open source software more innovative than closed software. Active projects usually have regularly updated web pages and busy development email lists. They usually encourage the participation of those who use the software in its further development. If everything is quiet on the development front, it might be thatwork has been suspended or even stopped.

### 16.4.6 USER INTERFACE

Most of the open source library softwares are available with web interface. Software with web interface is easier to learn and use. Graphical templates of open source softwares are possibleto customize and users can add new design. Through redesigning the templates and style sheetsopen source software can easily integrate with library/institutional websites. Separate administrative and user interface is essential for remote access and maintaining security.

### 16.4.7 DOCUMENTATION

So users are mainly responsible for the deployment of open source software; detailed and up-to-date documentation is a prerequisite for successful installation and maintenance. Open source software documentation is available through project websites, wikis, blogs and email lists.

They give information of software installation in various operating systems, software architecture,database structure, history of bug fixes, changes in new release, road map(wish list) of future releases etc. Installation details and information for users are also available with installation package.Individual documentation for developers, administrator and user is another advantage of open source software documentation. Software community incessantly updates the online documentation and it is better to make use the online wiki or email lists for error fixing and clearing doubts.

So, it seems that there are some very powerful solutions available today that could be usedto create a much more resourceful library. By using open source software in the library, money that otherwise would be spent on software solutions can be used for other important resources, suchas purchasing additional media resources (books, journals, etc.), or can be used to hire educated, technical support that provides patrons with the know how to better use already existing resources. In addition, this free software is constantly being updated, changed, and customized to meet the library's needs. While all of this is fine and dandy, and sounds like the win-win solution for your library, there are still pitfalls and hurdles we'll need to overcome. Hopefully this article provides some introductory information as to how to wean your library off of traditional computing productsand dive into the pool of open source resources available today.

## 16.5 OVERVIEW OF OPEN SOURCE SOFTWARES (GSDL, DSPACE &EPRINT)

### 16.5.1 GSDL

The Greenstone Digital Library Software (GSDL) offer exciting was to build and distributedigital document collections. It helps us to publish digital collections on the Internet or on CD-ROM. Within a few minutes time, one can build full-text search indexes and browsing classifiers for anycollection of digital documents. Once initiated, the collection building process will take place mechanically, running into several hours or days for a very large collection.

Downloading digital documents from World Wide Web, organizing them into focused collections and making the materials accessible to others can be a prime application area of digital libraries.

### 16.5.2 SPECIAL FEATURES

The GSDL software is an open source software available from the New Zealand Digital Library (www.nzdl.org) under the terms of GNU, general public library license. Greenstone CD-ROMs have been published by the United Nations and other agencies for distribution in developing countries. Some of the features of the software are:

- ✓ It suits both Windows (3.1/3.11, 95/98/ NT/2000) and UNIX (Linux Sun OS) any of these systems can be used as a web server.
- ✓ The administration function build in it enables the items to authorize new users to build collection, protect documents so that they can only be accessed by registered users on presentation of password.
- ✓ It build collection with effective full-text searching and metadata-based browsing facilities.Collection containing millions of documents, up to several gigabytes can be built. Full-textsearching is fast because compression is used to reduce the size of the indexes and textusers can browse the list of authors, titles, date, class no., etc.

- ✓ Plug Ins can be written to accommodate new document types; the collection can contain pictures, music, audio, video clips, etc. It also supports multilingual documents.
- ✓ Collection can be updated and new one brought online any time without bringing down thesystem.

The Z39.50 protocol is supported for accessing external servers and for presenting Greenstone collection to external clients.

### 16.5.3 SOFTWARE REQUIREMENTS

- OS Windows/Linux
- Apache web server/IIS
- PERL
- Java 2 Runtime Environment 'version 1.4.2_03'
- Web browsers—Netscape Navigator or
- Internet Explorer
- GSDL 2.41

### 16.5.4 STEPS FOR INSTALLATION OF SOFTWARE

The following steps are needed for installation:

- Install the web server IIS/Apache
- Install the Java 2 Runtime Environment from the internet (latest version)
- After installing J2RE, Go to GSDL Folder "gsdl-2.41-win32" (Setup file) from the Internet"MyComputer-GSDL-"gsdl-2.41-win32. exe"
- Choose Setup Language. English [United States] is the default
- The InstallShield Wizard will begin the installation of GSDL software. Click<next>
- Accept all the terms of license agreement by clicking on <yes> button
- Choose the type of installation you need and choose the collection/s that you want to be installed.
- Set the admin password
- (The above step will install web library edition of GSDL and any other sample collection/s and/or GSDL documentation, CD exporting function depending on what was checked or unchecked).
- Check the Greenstone Directory Structure:D: gsdl/Collect
- Cgi-binMicrosGli etc.
- We should create virtual directories for GSDL home and CGI executable directory.

## 16.6 DSPACE

Dspace is community-based, open source software platforms you can download free of charge and use to create your own digital repository. Organizations and institutions can more easily share and preserve their scholarly collections with an archiving system that stores digital representations of books, theses, 3-D digital scans of objects, photographs film, video, and research data sets and other forms of content.

### 16.6.1 WHAT IS DSPACE?

DSpace is a platform that allows you to capture items in any format – in text, video, audio, and data. It distributes it over the web. It indexes your work, so users can search and retrieve your items. It preserves your digital work over the long term. DSpace provides a way to manage your research materials and publications in a professionally maintained repository to give them greater visibility and accessibility over time.

DSpace is typically used as an institutional repository. It has three main roles:

- Facilitate the capture and ingest of materials, including metadata about the materials
- Facilitate easy access to the materials, both by listing and searching
- Facilitate the long term preservation of the materials

### 16.6.2 WHAT ARE THE BENEFITS OF USING DSPACE?

- ✓ Getting your research results out quickly, to a worldwide audience
- ✓ Reaching a worldwide audience through exposure to search engines such as Google
- ✓ Storing reusable teaching materials that you can use with course management systems
- ✓ Archiving and distributing material you would currently put on your personal website
- ✓ Storing examples of students' projects (with the students' permission)
- ✓ Showcasing students' theses (again with permission)
- ✓ Keeping track of your own publications/bibliography
- ✓ Having a persistent network identifier for your work, that never changes or breaks
- ✓ No more page charges for images. You can point to your images' persistent identifiers inyour published articles.

### 16.6.3 WHAT CAN DSPACE BE USED FOR?

DSpace can be used to store any type of digital medium. Examples include:

- Journal papers
- Data sets
- Electronic theses
- Reports
- Conference posters
- Videos
- Images

### 16.6.4 WHAT DOES DSPACE LOOK LIKE?

At a very high level, DSpace looks like this:

- ♦ Web-based interface makes it easy for a submitter to create an archival item by depositing files. DSpace was designed to handle any format from simple text documents to datasetsand digital video.
- ♦ Data files, also called bit streams, are organized together into related sets. Each bit

stream has a technical format and other technical information. This technical information is kept with bit streams to assist with preservation over time.

- An item is an "archival atom" consisting of grouped, related content and associated descriptions (metadata). An item's exposed metadata is indexed for browsing and searching.Items are organized into collections of logically-related material.
- A community is the highest level of the DSpace content hierarchy. They correspond to parts of the organization such as departments, labs, research centers or schools.
- DSpace's modular architecture allows for creation of large, multi-disciplinary repositories that ultimately can be expanded across institutional boundaries.
- DSpace is committed to going beyond reliable file preservation to offer functional preservation where files are kept accessible as technology formats, media, and paradigms evolve overtime for as many types of files as possible.
- The end-user interface supports browsing and searching the archives. Once an item is located, Web-native formatted files can be displayed in a Web browser while other formatscan be downloaded and opened with a suitable application program.

### 16.6.5 SOFTWARE RELEASES

Releases of the DSpace software have taken places as follows:

DSpace version 1.0 - 8th November 2002
DSpace version 1.1 - 8th May 2003
DSpace version 1.2 – 13th August 2004
DSpace version 1.3 – 3rd August 2005
DSpace version 1.4 – 26th July 2006
DSpace version 1.5 – 25th March 2008

## 16.7 E-PRINT

The origin of the Eprint lies in increasing interest in alternatives to the traditional scholarly publishing paradigm. In his regard, the development of alternative models for the communicationof scholarly results, particularly in the form of online repositories of Eprints, has demonstrated a viable alternative to traditional journal publication.

For this purpose, in October 1999, a meeting in Santa Fe was organized on the belief that the interoperability among these Eprint archives was key to increasing their impact. The Santa Fe convention was the first attempt in which the technical and organizational agreements about OAI were decided

**Advantages of E-Print**

In both the pre-refereeing pre-prints and the final published post-prints, the author cites other authors' work that has been used in the research. If a reader wishes to look at a cited pieceof work, that paper has to be found elsewhere. This is time-consuming, and often ends with the paper being inaccessible, because the user's institution cannot afford to subscribe to it. Users can follow the research through all of its successive stages from pre-prints through to the post- prints.

A Ultimately we can have one central repository of all the literature. The advantage of central repository will be that anybody can search or browse the literature in their respective area.

**Eprints archive software**

Eprints archive software was developed by the Electronics and Computer Science Department of the University of Southampton (http://www.eprints.org). The current version of thesoftware is eprints 1.1.2. This free software can be downloaded from the abovementioned URL byany individual or institution interested in maintaining eprints archive.

**Hardware and software requirements**

- At least Intel Pentium II processor.
- A UNIX operating system. Linux (a very advanced and free UNIX implementation) worksjust fine, and is in fact the development platform.
- The Apache WWW server, another professional-quality free software product, often includedwith Linux distributions, such as RedHat.
- The Perl programming language also included with most Linux distributions.
- The mod_perl module for Apache, which significantly increases the performance of Perlscripts.
- The MySql Database, a database system that is free for non-commercial use.
- The Eprints software.
- The prerequisite software pack

**Installing and configuring the software**

Much of the installation process has been automated by the use of some installation scripts.It is also possible to install the software manually, allowing the maximum amount of flexibility when installing the software on machines that are running other services. Here is an overview of the steps that one can perform automatically or manually.

1. Install the code, setting relevant path information.
2. Configure the code to work with the rest of the system.
3. Create the MySQL database for Eprints.
4. Configure Apache to execute Eprints scripts and serve Eprints documents.
5. Install a crontab, which periodically executes various Eprint scripts.

These steps will complete the installation and configuration of Eprint Archive Software

## 16.8 SELF ASSESSMENT QUESTIONS

1. What is open source software (OSS)? Explain
2. Enumerate the digital library software.
3. Discuss the criteria of selection of digital library software.

## 16.9 REFERENCES

1. Altman, Micah (2001). Open Source Software for Libraries: from Greenstone to the Virtual Data Center and Beyond. *IASSIST Quarterly*, Winter 2001, 5-11. Retrieved

January 17, 2008, from Web site:
http://iassistdata.org/publications/iq/iq25/iqvol254altman.pdf
2. Bailey, Charles W., Jr. (2006). Open Access and Libraries. Retrieved January 15, 2008, from Web site:
http://www.digital-scholarship.com/cwb/OALibraries2.pdf
3. Balas, Janet L. (2004). Considering open source software. *Computers in Libraries*. 24 (8), 36-39.Retrieved February 10, 2008, from Web site: http://www.infotoday.com/cilmag/sep04/ balas.shtml
4. Bretthauer, David (2002). Open Source Software: A History. *ITAL: Information Technologyand Libraries*. 21(1), 3-11. Retrieved January 21, 2008, from Web site:
http://www.ala.org/ala/lita/litapublications/ital/2101bretthauer.cfm
5. Corrado, Edward M. (2005). The Importance of Open Access, Open Source, and Open Standards for Libraries. *Issues in Science & Technology Librarianship*. 42. Retrieved February 3, 2008, from Web site:
http://www.istl.org/05-spring/article2.html
6. Ferraro, Joshua. (2006). Why Your Library Needs Open Source. Retrieved February 9, 2008, from Web site: http://liblime.com/c/welcome.html
7. Free Software Foundations software directory. http://www.fsf.org/

A

8. Hebert, Eric. How Open Source Software Can Improve Our Library. Retrieved January 15, 2008, from Web site: http://www.degreetutor.com/library/managing-expenses/open-source-library
9. Kumar, Vimal (2007). Selection and Management of Open Source Software in Libraries. In Kumar, Manoj K., Eds. *Proceedings CALIBER 2007: 5th International Convention on Automation of Libraries in Education and Research Institutions*, 1-5.
10. Mackenzie, Adarian (2001). Open Source Software: When is a Tool? What is a Commodity? *Science as Culture,* 10(4), 541-552.
11. Morgan, Eric Lease (2002). Possibilities for Open Source Software in Libraries. *ITAL: Information Technology and Libraries.* 21(1), 12-15. Retrieved January 19, 2008, from Web site:http://www.ala.org/ala/lita/litapublications/ital/2101morgan.cfm
12. Open source software. Wikipedia. Retrieved February 5, 2008, from Web site: http://en.wikipedia.org/
13. Source Forge. http://sourceforge.net/
14. UNESCO Free & Open Source Software Portal. http://www.unesco.org/
15. http://greenstonesupport.iimk.ac.in/Documents/GSDL%20Begineers%20Guide.pdf
16. http://dspace.org/sites/dspace.org/files/dspace%20brochure_v4.pdf
17. http://eprints.rclis.org/5702/1/Archiving.pdf

## LESSON - 20

# FUTURE DIGITAL LIBRARIES

**OBJECTIVE**

After reading this chapter, students will be able to understand:

- Basic concepts of future digital library
- Digital library creation & management

**STRUCTURE**



## 17.1 INTRODUCTION

Research on digital library (DL) systems started in Europe in the mid-nineties. At that time DLs were seen essentially as repositories of digital texts accessible through a search service which was operating by indexing information stored in a centralized metadata catalogue. The construction of a DL was very resource consuming since, for each new DL, both the content and the software providing the DL functionality were built from scratch. As a result of this development approach, only powerful user communities [1] or user communities with in-house computer science technical skills (Leiner, 1998) could afford the building up of DLs. These DLs were created to serve end-users only as consumers of information. They did not provide any functionality for submittingthe documents.

The submission was usually performed either by the author or by a librarian operator by means of specific procedures residing outside the DL. Today, the requirements imposed on DLs are very different from that early time. Anovel notion of DLs, also referred to as “knowledge commons”(Ioannidis, 2005), has recently emerged, whose fulfillment requires new technologies and new organizational models. This paper focuses on such new DLs by first discussing the motivations for their introduction, then presenting an innovative DL technology, called DILIGENT, and, finally,illustrating the role that libraries can play in this new scenario.

## 17.2 DIGITAL LIBRARIES OF THE FUTURE

According to the most recent understanding, the DLs of the future will be able to operate over a large variety of information object types - far wider than those maintained today in physical libraries and archives. These information objects will be composed of several multi-type and multimedia components aggregated in an unlimited number of formats. These, for example, can mix text, tables of scientific data and images obtained by processing earth observation data, or they can integrate 3D images, annotations and videos. These new information objects will offer innovative and more powerful means to researchers for sharing and discussing the results of their work. In order to be able to support these objects, the DL functionality has to be appropriately extended far beyond that required to manipulate the simple digital surrogates of the physical objects. In order to support these objects the DL may need considerable resources. For example, the creation and handling of the new documents may require access to many different, large, heterogeneous information sources, the use of specialised services that process the objects storedin these sources for producing new information, and the exploitation of large processing capabilitiesfor performing this task.

New DLs are also required to offer a much richer set of services to their users than in the past. In particular, they must support the activities of their users by providing functionalities that may range from general utilities, like annotation, summarization or co-operative work support, to very audience-specific functions, like map processing, semantic analysis of images, or simulation. The availability of this new DL functionality can, in principle, change the way in which research isconducted. By exploiting such types of DL, for example, a scientist can annotate the article of a colleague with a programme that extracts useful information from a large amount of data collected by a specific scientific observatory. This programme, executed on demand when the annotation is accessed, can complement the content of the paper with continuously refreshed information. In the new DLs users are not only consumers but also producers of information. By elaborating information gathered through the DL they can create new information objects that are published in the DL, thus enriching its content. The new DLs are thus required to offer services that support the authoring of these new objects and the workflows that lead to their publication.

In parallel with the above evolution of the role of DL systems, we are now observing a large expansion in the demand for DLs. Research today is often a collaborative effort carried out by groups belonging to different organizations spread worldwide. Motivated by a common goal and funding opportunities, these groups dynamically aggregate into virtual research organizations that share their resources, e.g. knowledge, experimentation results, or instruments, for the duration of their collaboration, creating new and more powerful virtual research environments. These virtual research organizations, set up by individuals that do not necessarily have great economic poweror technical expertise, more and more frequently require DLs as tools for accelerating the achievement of their research results. This new potential audience demands less expensive and more dynamic DL development models. They want to be able to set up new DLs that serve their needs for the duration of their collaborations in an acceptable timeframe and with an acceptable cost. The current DL development model cannot satisfy this large demand; a radical change is needed if we want to be able to address these new emerging requirements. A great contribution towards the satisfaction of all the above mentioned requirements can certainly come from the introduction of mechanisms that support a controlled sharing of resources among different organizations. Sharing in this context is not only applied to repositories of content, as is usually meant today, but can be extended to any type of resource needed to build a DL, i.e. language and ontology resources, applications, computers and even staff with the necessary skills for supporting the DL development, deployment and maintenance.

Supporting this type of sharing requires the introduction of appropriate solutions at both the technological and organizational levels. These two levels are not independent; instead they strongly influence each other. In fact, the availability of a good technological solution favors the creation of an appropriate organization, and vice-versa, a successful organization stimulates the developmentof new supporting technologies. In the next section we present the DILIGENT infrastructure as an example of a technological solution for these new DLs. The organizational aspects stimulated by the introduction of this technology are briefly discussed afterwards.

## 17.3 DILIGENT

DILIGENT (Digital Library Infrastructure on Grid Enabled Technology) [2] is a three-year Integrated Project (2004-2007) funded by the European Commission under the 6th Framework Programme for Research and Technological Development. The objective of this project is to developa Digital Library Infrastructure that will enable members of dynamic virtual research organizations to create on-demand transient digital libraries that exploit shared resources. Resources in this context are multimedia and multi-type content repositories, applications, and computing and storageelements. Following the understanding of DLs expressed in Borgman *et al.* (2002), this project focuses on the development of DLs that "are not ends in themselves; rather they are enabling technologies for digital asset management, electronic commerce, electronic publishing, teachingand learning, and other activities" (p. 7).

From an abstract point of view, the DILIGENT infrastructure can be understood as a broker serving DL resource providers and consumers. The providers are the individuals and the organizations that decide to publish their resources under the supervision of the broker, according to certain access and use policies. The consumers are the user communities that want to build their own DLs. The resources managed by this broker are content sources (i.e. repositories of information searchable and accessible through a single "entrance"), services (i.e. software tools that implement a specific functionality and whose descriptions, interfaces and bindings are defined and publicly available) and hosting nodes (i.e. networked entities that offer computing and storage capabilities and supply an environment for hosting content sources and services). Providers register their resources and give a description of them by exploiting appropriate mechanisms provided by the infrastructure. The infrastructure also automatically derives other properties of the resources that are used to enrich the explicit description. The infrastructure manages the registered resourcesby supporting their discovery, monitoring and usage, and by implementing a number of other functionalities that aim at realizing the required controlled sharing and quality of service.

A user community can create one or more DLs by specifying a set of requirements. These requirements specify conditions for the information space (e.g. publishing institutions, subject of the content, document types), for the operations that manipulate the information space (e.g. type of search, tool for data analysis), for the services for supporting the work of the users (e.g. type of personalized dissemination, type of collaboration), for the quality of service (e.g. configuration, availability, response time) and for many other aspects, like the maximum cost, or lifetime. The broker satisfies the community's requirements by selecting, and in many cases also deploying, a number of resources among those accessible to the community, gluing them appropriately and, finally, making the new DL application accessible through a portal. The composition of a DL is dynamic since the DL broker continuously monitors the status of the DL resources and, if necessary, changes them in order to offer the best quality of service. By relying on the shared resources many DLs, serving different communities, can be created and modified on-the-fly, without big investments and changes in the organizations that set them up.

In order to support the transactions between the providers and the consumers, the DILIGENT infrastructure exploits the virtual organizations (VOs) mechanism that has been introduced in the

Grid research area (Foster *et al.*, 2001). These mechanism models sets of users and resources aggregated together by highly controlled sharing rules, usually based on an authentication framework. VOs have a limited lifetime, are dynamically created and satisfy specific needs by allocating and providing resources on demand. Through the VOs mechanism the DILIGENT infrastructure glues together the users and the resources of a DL.

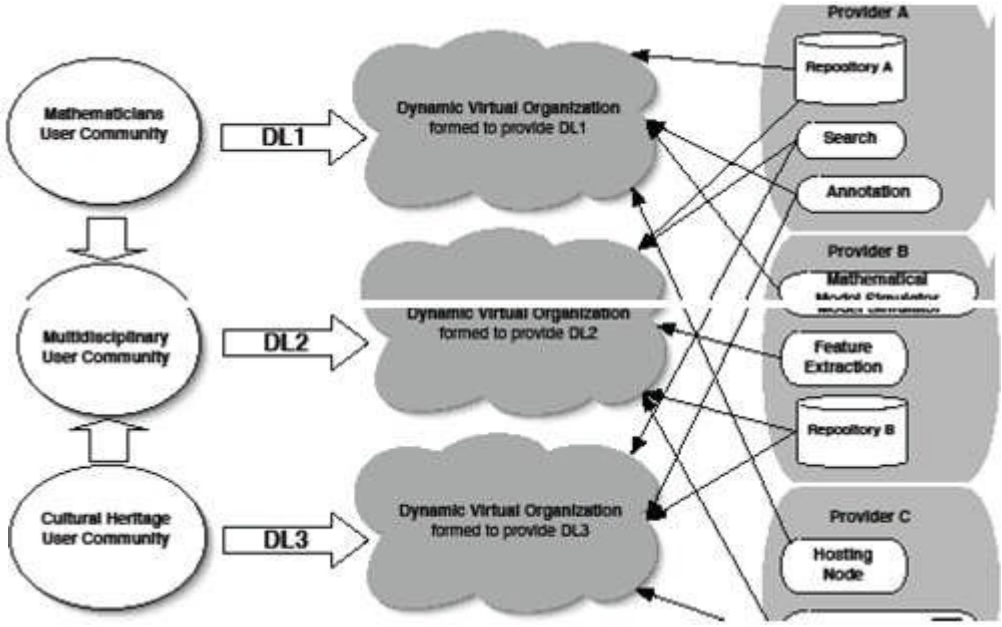


**Figure 1: The role of virtual organizations in DILIGENT**

Figure 1 graphically illustrates the role of VOs in supporting the brokerage model. The consumers, i.e. the user communities that require DLs to support their needs, are on the left of the figure. The providers, each of which makes a number of resources available, are on the right. The infrastructure acts as a mediator by maintaining a framework where multiple virtual organizations, active on the same shared resources, can co-exist.

The DL development model proposed by DILIGENT is radically new. Within the described framework each DL consumes the required resources only for the time it needs them. This opens a lot of new opportunities for the creation of the functionalities required by the new "knowledge commons" environments. In particular, the exploitation of more effective, but also very computationally expensive algorithms becomes viable at an acceptable cost for many communities. For example, thanks to sharing, the use of the high process consuming algorithms that automatically extracts features from multi-media objects can be exploited in a large number of DLs. Moreover, in the framework established by the new development model, the user communities can easily, and in a timely manner, create and maintain their own DLs with limited resources since the management of the DL is automatically and transparently carried out by the infrastructure.

The system that implements the functionality of the DILIGENT infrastructure is being built by integrating DL and Grid technologies (Foster and Kesselman, 2004). The motivation for this design choice relies on the similarity between many of the problems encountered through our new notion of DLs and the issues addressed by the most recent research in the Grid domain.

## 17.4 THE DILIGENT SYSTEM IS DIVIDED INTO FIVE FUNCTIONALITY CLUSTERS

From the functional point of view, the DILIGENT system is divided into five functionality clusters aregiven below:

### *17.4.1 DL Creation & Management*

Is responsible for the dynamic construction and maintenance of the transient DLs and for the controlled sharing and management of the resources that are used to implement them. The functionalities offered by this cluster allow users to express the requirements that the DL must fulfill. Moreover, they automatically identify and arrange the pool of resources needed to satisfy these needs.

### *17.4.2 Content & Metadata Management*

Implements the handling of DL content and related metadata, the consistent and distributed management of annotations, and the integration of external content and metadata sources.

### *17.4.3 Process Management*

Manages the creation of user processes composed of existing services, the validation of their correctness, the automatic optimization of their definition according to the resources available and the service characteristics, and their reliable execution. Thanks to this feature, the DILIGENT system can easily be enriched with additional operational workflows to meet new user requirements.

### *17.4.4 Index & Search Management*

Is responsible for enabling cost-efficient search and retrieval of information in DLs, while satisfying the level of quality required for the overall data retrieval and delivery operations.

### *17.4.5 Application Specific Functionality*

Provides the functionality needed to support user-specific scenarios, like portals, document visualization, or features extraction.

## 17.5 THE ROLE OF LIBRARIES IN FUTURE DLS

In the framework envisaged by DILIGENT, libraries play an important role at the organizationallevel. In particular:

- As providers of resources, they can help to enhance the amount of available resources by making stakeholders aware of the importance of sharing. In particular, as far as the sharingof content is concerned, they can operate by promoting digitization campaigns and the Open Access approach. These actions may result in a vast amount of new digital information accessible online which can be exploited by advanced services.

- Also within a digital framework, libraries are certainly the best candidates for carrying out content description, maintenance and preservation of resources. By exploiting their large experience acquired in the past, they can contribute to the long-term availability and to the quality of the resources disseminated by the DLs.

- Long-term availability also requires the implementation of models able to support the sustainability of the resources provided. Libraries, either alone or as members of library consortia, can also act as the organizations deputed to define and put in place these models.

- As main resource providers, libraries can work jointly on the definition of common policies and standards. An agreement on these aspects would strongly contribute towards facilitatingthe design and development of the new complex services required to fulfill the emerging user needs.

- In the future envisaged by DILIGENT, libraries can also play an important role as mediators between the infrastructure and the user communities. In particular, they can proactively promote and facilitate the creation of DLs that respond to the needs of the user communities. They can also assist users by providing, if necessary, the skills required to select, update and exploit the DL content and services.

## 17.6 SELF ASSESSMENT QUESTIONS

1. Discuss about future digital libraries.
2. What are the roles of future digital libraries?

## 17.7 REFERENCES

1. Borgman, C., Sølvberg, I. and Kovács, L. (Eds.) (2002), *Proceedings of the Fourth DELOS Workshop, Evaluation of Digital Libraries: Testbeds, Measurements, and Metrics*, Budapest, 6-7 June 2002, available at: www.sztaki.hu/conferences/deval/presentations/DELOSWorkshop4OnEval_report.pdf (accessed 4 May 2006).

2. Foster, I. and Kesselman, C. (Eds.) (2004), *The Grid: Blueprint for a Future Computing Infrastructure*, 2nd ed., Kaufmann, Amsterdam.

3. Foster, I., Kesselman, C. and Tuecke, S. (2001), "The anatomy of the Grid: enabling scalable virtual organizations", *The International Journal of High Performance Computing Applications*, Vol. 15, No. 3, pp. 200–222.

4. Foster, I., Frey, J., Graham, S. and Tuecke, S. (Eds.) (2004) *Modeling Stateful Resources with Web Services*, Version 1.1, 3 May 2004, Whitepaper, available at: www-128.ibm.com/ developerworks/library/ws-resource/wsmodelingresources. pdf (accessed 4 May 2006).

5. Ioannidis, Y. (2005), **"**Digital libraries from the perspective of the DELOS Network of Excellence", in *Proceedings of the IEEE-CS International Symposium: Global Data Interoperability - Challenges and Technologies*, June 20th-24th, 2005, Sardinia, Italy, pp. 51-55, available at: http://globalstor.org/(accessed 4 May 2006).

6. Leiner, B.M. (1998), "The NCSTRL approach to open architecture for the Confederated Digital Library", in *D-Lib Magazine*, Vol. 4, No. 11, available at: www.dlib.org/dlib/december98/leiner/12leiner.html (accessed 4 May2006).

(**202 ML21)**

**MODEL QUESTION PAPER**

**ACHARYA NAGARJUNA UNIVERSITY: CENTER FOR DISTANCE EDUCATION**

**M.L.I.Sc.**

**SEMESTER - II**

**Digital Libraries**

Time: 3 Hours Maximum Marks: 70

**Answer any five questions.**
**All questions carry equal marks.**

1. What is Digital Library? Discuss the fundamentals of Digital Library.
2. Describe the advantages and limitation of Digital Libraries.
3. Write an account on design of Digital Libraries.
4. State the Protocols of Digital Libraries
5. Write an account on Digital Archiving.
6. Describe various types of Digital Resources
7. Write an account on open archive initiatives.
8. Write an account on Digital Library Software.
9. Write an account on open source software of digital libraries.
10. Describe various stages in Digitations.