

Digitization Of Library Resources And The Formation Of Digital Libraries

V.Thirumalar¹, P.Sunitha²

¹(MCA M.Phil, B. Asst. Professor Department of computer science Pioneer College of Arts and Science)
malarvenu2001@gmail.com

²(Asst. Professor Department of computer science Pioneer College of Arts and Science)

Abstract: *This paper discusses the new activities, methods and technology used in digitization and formation of digital libraries. It set out some key points involved and the detailed plans required in the process, offers pieces of advice and guidance for the practicing Librarians and Information scientists. 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. The paper ends with a call to integrate digitization into the plans and policies of any institution to maximize its effectiveness.*

I. Introduction

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.

II. What Is Digital Libraries?

A digital library is a collection of digital documents or objects. This definition is the dominant perception of many people of today. Nevertheless, Smith (2001) defined a digital library as an organized and focused collection of digital objects, including text, images, video and audio, with the methods of access and retrieval and for the selection, creation, organization, maintenance and sharing of collection.

Though the focus of this definition is on the document collection, it stresses the fact that the digital libraries are much more than a random assembly of digital objects. They retain the several qualities of traditional libraries such as a defined community of users, focused collections, long-term availability, the possibility of selecting, organizing, preserving and sharing resources.

The digital libraries are sometimes perceived as institutions, though this is not as dominant as the previous definition. The following definition given by the Digital Library Federation (DLF) brings out the essence of this perception.

“Digital Libraries are organization 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.” (DLF 2001)

The point in this definition is on the digital library as a dynamic, growing organism. As digital libraries evolve and become the predominant mode of access to knowledge and learning, institutionalization of digital libraries appears to be on the increase.

III. Benefits of Digital Libraries

Digital libraries bring significant benefits to the users through the following features:

i. Improved access

Digital libraries are typically accessed through the Internet and Compact Disc-Read Only Memory (CD-ROM). They can be accessed virtually from anywhere and at anytime. They are not tied to the physical location and operating hours of traditional library.

ii. Wider access

A digital library can meet simultaneous access requests for a document by easily creating multiple instances or copies of the requested document. It can also meet the requirements of a larger population of users easily.

i. Improved information sharing.

Through the appropriate metadata and information exchange protocols, the digital libraries can easily share information with other similar digital libraries and provide enhanced access to users.

ii. Improved preservation.

Since the electronic documents are not prone to physical wear and tear, their exact copies can easily be made, the digital libraries facilitate preservation of special and rare documents and artifacts by providing access to digital versions of these entities.

IV. Functional Components Of Digital Library

Most digital libraries share common functional components. These include:

i. Selection and acquisition

The typical processes covered in this component include the selection of documents to be added, the subscription of database and the digitization or conversion of documents to an appropriate digital form.

ii. Organization

The key process involved in this component is the assignment of the metadata (bibliographic information) to each document being added to the collection.

iii. Indexing and storage

This component carries out the indexing and storage of documents and metadata for efficient search and retrieval.

iv. Search and retrieval

This is the digital library interface used by the end users to browse, search, retrieve and view the contents of the digital library. It is typically presented to the users as Hyper-Text Mark-up Language (HTML) page.

These mentioned components are the important characteristic of digital library, which differ it from others collections of online information.

V. 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:

i. Digitization means no new buildings are required; information sharing can be enhanced and redundancy of collections reduced.

i. Digitization leads to the development of Internet in digitalized based libraries. As Internet is now the preferred form of publication and dissemination.

ii. Digital materials can be sorted, transmitted and retrieved easily and quickly.

iii. 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.

iv. 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 digitized and uploaded into the Internet or Compact-Disc Read Only Memory (CD-ROM) in order to be made correspondingly accessible globally.

VI. Why Digitization?

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

i. To preserve the Documents: That is to allow people to read older or unique documents without damage to the originals.

ii. 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 envisaged in remote locations, example, more than one person at a time.

- iii. To reuse the documents. It means to convert documents into different formats; for example to 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.

VII. Reasons To Be Considered

i. 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.

ii. Who is your audience?

If there are only few users, or may be 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.

i. Do the documents form a collection?

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

ii. 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 with the accepted quality standards.

VIII. 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 approval or 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 collection and 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 material is 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.
- a. 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. CD-ROM or on-line or both. Define metadata, search and retrieval requirements.
- b. 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:
 - i. How to convert the source material into required digital format.
 - ii. What are the digitization requirements?

- iii. The workflow involved in digitizing the source material.
- c. Identify the resources and money required for creating and maintaining digital collections. There is a need to identify:
 - i. What type of information technology (IT) infrastructure is required for establishing and maintaining the digital collections?
 - ii. What are the personnel requirements and
 - iii. What are the financial requirements involve for setting up and maintaining the collection.
- d. Finally, there is the need to define how the project is going to be implemented and what the major milestones and time requirements are?

IX. 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.

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

X. 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 modes of providing access to the digital collection. These may include:

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

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.

i. 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.

ii. 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, in the software, choose a setting, resolution and colour and scan each page of the document at the settings you have chosen.

iii. 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.

iv. 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 not enough computers for each proofreader.

v. 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.

vi. 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.

XI. Technology Infrastructure And Personal

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.

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.

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:

- i. Project management.
- ii. Selection and preparation of source material
- iii. Digitization and conversion
- iv. Cataloguing and metadata assignment
- v. Quality assessment
- vi. System administration and maintenance of digital library server and website.
- vii. System analysis/programming for digital library application/interface development
- viii. 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.

XII. Greenstone Digital Library Software

Greenstone is a freely available suite of software for building and distributing digital library collections. It provides a new way of organizing information and publishing it on the Internet or on the CD-ROM.

The Greenstone is open source software, issued under the terms of the GNU General Public License. The aim of the software is to empower the users, particularly in the Universities, Libraries and other public service institutions, to build digital libraries. The software has the following features such as multi-platform availability for windows, linux, access and distributed through the Internet, Intranet and CD-ROM, powerful indexing from full-text and creation of indexes for various metadata, powerful search and browse, support different file formats (html, pdf, doc rtf, ppt etc), extensibility by allowing customization and configuration. Greenstone also allows the building of non-textual multimedia such as audio, video and pictures accompanied by textual description to allow for searching and browsing.

XII. Conclusion

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.

References

- [1] Digital Library Federation. (2001), Registry of Digitized Books and Serial Publication,
- [2] Ding, Choo Ming. (2000), Access to Digital Information: Some Breakthrough and Obstacles, Journal of Librarianship and Information Science, Vol.32 No.1
- [3] Greenstone Training Workshop Material.(2002).
- [4] Ian, H. Witten & David, Brainbridge. (2003), *How to Build a Digital Library*, London: Morgan Kaufman Publishers
- [5] 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>
- [6] Smith, Abbey (2001), Strategies for Building Digitized Collection. Washington, D.C. Digital Library Federation, Council on Library and Information Resources.<http://www.clir.org>