Digital Conversion and Preservation of old Ph.D. Theses: A Study at Central Library, IIT Kharagpur

M Manivannan *B Sutradhar*

Abstract

The purpose of this paper is to describe how rare library materials and documents are being preserved for long-term usage and archived at the Central Library, Indian Institute of Technology Kharagpur (IIT Kharagpur) and focused on the technical requirements, guidelines for digitization, and digital preservation of old library materials. The paper explains levels of the digitization process and levels of the digital preservation of old Ph.D. theses submitted at the library, also to discuss the digitization infrastructure and software used. The paper shows that the workflow and steps involved in the conducting digitization and digital preservation project at the library. The information contained in the library materials such as rare documents, theses, and other similar old documents can be lost forever unless there are alternative arrangements for digitally converting and or digitizing them. The technological advancement is providing suitable alternatives for digitization and its preservation for the library materials. Before starting up a digitization project, the library should have sufficient funds to acquire infrastructures such as a high-end scanner, computer hardware, other related software, and trained computer-human resources. It can also be noted that for the successful completion of this project, must train the administrator and staff involved in the project. Digital preservation of rare and old documents in an Institutional Digital Repository (IDR) can provide an immediate response to a search, getting information on time, and search enhances the researchers' efficiencies. The IDR supports mechanisms to store, retrieve and manage the digital assets.

Keywords: Preservation, Digitization, Digital Preservation, Rare Documents, Digital Documents, Digital Library, Institutional Digital Repository

1. Introduction

In the modern library, materials like e-books, e-journals, e-databases and all major types of library resources are available in the format of digital contents, and the students, research scholars, faculties, and other academics are accessing online. Simultaneously, the libraries have been migrating from print collection to digital collection ever since digital conversion techniques developed. With availability of resources in digital forms, libraries committing to either procure or create the electronic resources. Preservation and archiving of electronic resources have become a concerned severe of libraries to either subscribed in digital forms or converted.

Radovan (2010), conducted an online web-based survey on digitization and its collection development in public libraries. According to the research findings

the libraries recognized the importance of digital contents as part of the library holdings and for the promotion of their library scientific research in Croatia, and quality of life in the local community. The digitization project in public libraries in Croatia initiated for the preservation and development of digital contents.

Manzuch (2009),focused on objectives of digitization and its budgeting and costs; usage, volume and growth of the digitized collections. The analysis revealed the split between strategic and resource management approaches to digitization, absence of methodology, and problems in developing quantitative measures. Digitization enables extensive use for research, and education and to access great extent. It assists to preserve and to safeguard of fragile materials such as books, newspapers, and other library materials for future uses.

Libraries and librarians should tend to create a human knowledge base and increase its value to create a better society (Choy, 2007, p. 114).The traditional library materials such as books and papers are converting to electronic form in the digitization process, and they can be controlled and preserved by a technology (Witten and Bainbridge, 2003, p. 58). Digitization is one of the most important activities in any libraries are expected to change to global access to information and knowledge. Association for Library Collection & Technical Services (A division of the American Library Association) considered, "Digital preservation combines policies, strategies, and actions that ensure access to digital content over time." The online digital content facilitates the accessibility through the different mediums which are available nowadays as a boon of information technology wherever exists.

1.1 Digital Preservation and Access

Without digital access the digital materials are not helpful to the knowledge society. So, Library should provide access to digital materials converted from print form. Digital preservation has granted in providing access to digital content that would otherwise degrade from repeated use in the hard copy of documents. Digitally converted, born-digital, digitally reformatted, collective resources, data sets, and communication reports are common preservation types of digital resources. Academic libraries like oursare doing preservation of digital resources which rare books, thesis, class notes, question banks, handbooks, practical guides, and other academic-related materials in a systematic manner. The landscape of the digital preservation is one of a multitude of choices that vary widely regarding purpose, scale, cost, and complexity. Publishers and service providers are providing for access to e-journals, e-books, bibliographic databases, and other born digital or print material converted to digital through their websites in subscription mode.This study, mainly focused on - digitization strategies, preservation strategies, digitization methodsand flow, preservation methods and tools, digital access, etc.

2. Central Library of IIT Kharagpur

In support to the teaching, learning, and research mission of the Indian Institute of Technology Kharagpur which well known in the name of IIT Kharagpur, the Central Library is regarded as the hub of the institute to fulfill the informational needs of the institute; mainly towards to completion of its academic programmes and the research activities. At present, the Central Library is catering to the needs of more than eleven thousand students consists of undergraduates, postgraduates, and research scholars; and seven hundred faculty members, and more than one thousand staff members

of the institute. It is a matter of prestige; the Central Library has certified with ISO 9001:2008 since 2014 and it transitioned to ISO 9001:2015 since July 2017.

Central Library of IIT Kharagpur receives copies of all Ph.D. theses awarded at this institute.

At present, Central Library has a collection of more than 5132 volumes of individual titles of Ph.D. theses which research conducted and submitted by research scholars of IIT Kharagpur. These are all presented since 1956 at this library. Authority coined a policy that the research scholars must submit a soft copy of the thesis in a CD-ROM along with a print copy. This policy was taken in the year 2009. Therefore, the Central Library is now receiving a softcopy of thesis and one print copy, and these softcopies are born-digital materials. These submitted documents have value further for both research and study. The Central Library has taken an initiative to digitize all old Ph.D. theses as well as rare documents and to ensure continued existence and management of digital materials. DSpace open source software is used to providing access these digital documents to the library users.

3. Digitization Process

IFLA - International Federation of Library Associations and Institutions deliberated and given a "Guidelines for Planning the Digitization of Rare Book and Manuscript Collections". IFLA clearly pointed out the following strategies and planning for digitization of rare documents

- ❖ Designing the digitization project;
- ❖ Selection of documents for digitization;
- ❖ Workflow for creating digital collection;
- ❖ Metadata for created digital collection;
- ❖ Display the created digital collection to the users;
- ❖ Dissemination, promotion, and reuse;
- ❖ Evaluation; and
- ❖ Long-term preservation of the digital collection.

The center of interest of these guidelines are planning and collaboration with users to reach the expected outcomes and viable results, and help professionals for developing sustainable digital collections. The collections will continue to have research and re-usable in the future.

A successful project is expecting the right answers to the following questions raised in the guidelines for digitization given by IFLA.

i. What is the vision for the project? What are the goals and objectives? Who will use it?

ii. How will they use it?

iii. Who should be involved in the planning?

iv. What level of complexity is desired? What level of complexity can be achieved?

v. What do you want to digitize, and why?

vi. Are there any copyright issues regarding the materials?

vii. Do you have the space, money, and equipment, and expertise?

viii. What is the final format of the project? Do you have the means to achieve it?

ix. How will you incorporate quality management into all stages of the project?

The initial objective of the digitization project is to convert about 3197 old Ph.D. theses in to digital format and improving accessibility, preservation, and spreading knowledge to institute community and other research community.

Workflow of the Digitization Project

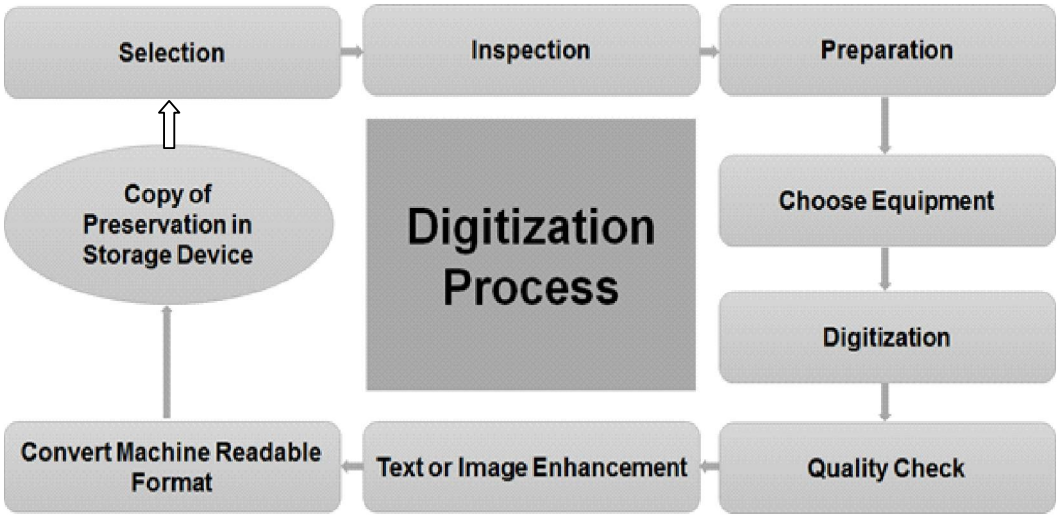

Figure 1: Workflow Diagram of the Digitization Project

4. Document Selection, Inspection, and Preparation

The documents which are to be digitized determined as old thesis and at least one print copy already submitted at the library. The print copies are arranged by subject-wise (by DDC call number) in a separate area. Ph.D. thesis which submitted before 2009 is to be digitized, and it is picked up from the collection and transferred to the Digital Library section. Initially, the duplicate copies are extracted from the total collection to check either the selected print copy is already converted into digital format or not. If it is already digitized, a visible mark 'Digitized' is put on the print copy of the thesis. Sometimes, the document selection exists as an urgent requirement for users. So, it is prioritized on the demand of the users.

Physical review of the document is included in the document selection process. The staff who is selecting the document is also inspecting possibility of digital conversion with the existing infrastructure; along with size and type of the document. If the selected document is feasible for digitization, then it is transferred to the preparation area. The digitized documents are replaced in its original place after completion of the digitization process. Movements of the documents in the digitization process are being registered wherever it is necessary.

5. Digitization

A high-end overhead book scanner BookEye® 4 is used to scan the documents. Using this scanner originals up to A2 in sizes such as books, magazines, posters, folders or bound documents of all types, can be digitized at high speed and a maximum resolution of 600 dpi. It is a high-performance production scanner is complemented by a robust capturing software BCS-2. The V-shaped book cradle was optimized to reduce the risk of damage to

book spines and binding during digitization. The book cradle holds the open book at an angle of 120 degrees; enough to effectively scan the contents of a page, while gently preserving the original subject matter.

5.1 Original Copy

Scanning of a document is a batch process in our practice. The batch process started scanning from the title page and continue to end page of the document, and the contents scanned from the pages as exist. The provided software along with the scanner is takes care of the batch scanning process. Each process considered as one document or thesis and the process saves the file as accession number provided in the document or thesis. In the scanning batch process includes digitization process of manuscripts, texts, images, photographs, graphs, tables, maps, and artworks.

Single side printed documents have been scanned by Flat-mode of the scanner, and each scanned image of the single page stored in a short sequence. Double-side printed documents were scanned by V-mode of the scanner, and scanned images of both sides stored in a short sequence; then it is automatically split into two pages by user settings on the scanning software. The scanning software is doing the page separation process for the entire complete batch scanning process.

As default setup 300 dpi, black & white mode, and 1-bit color depth was chosen for manuscripts, texts, tables, and graphs. Despite choosing the default setup, the dpi may increase, and the color mode may change wherever it is necessary. 400 dpi, color mode, and 24-bit color depth are chosen for images, photographs, maps, and artworks. Resolution may vary depending upon the quality of the document. Maximum 600 dpi resolution can be used.

5.2 Enhancement of Scanned Pages

After completion of scanning of one complete document, some enhancements on the scanned pages were carried out like removal of dot marks in pages, black lines in page edges; and despeckle filtering. Some of the processes are done automatically and some of the processes are done manually depending upon its quality. The final output quality of the document is to consider which includes color saturation, image brightness, image integrity, and other optical flaws; and it is varying in different documents. Inbuilt tools of software are helping to reach the quality of the document digitization.

5.3 File Format

The master copy of the images are created in TIFF (Tag Image File Format). which is one of the most common graphic image formats.

Each page of the scanned document is saved in TIFF with separate sequence file number generated by the software. Finally, these are saved in a folder name with the assigned accession number of the document. Entire collections of images of one document are included in the folder. Final copy of one document is converted into a single PDF file.

5.4 Optical Character Recognition (OCR)

OCR is a technology that converts into searchable data of different types of documents such as scanned documents, images, and PDF files captured in a scanner. Adobe ABBYY® 6.0 OCR conversion software is used to convert into a machine-readable format of the scanned documents at the same time it

is used to create into a single PDF file. If conversion of machine-readable form is not possible for any scanned pages, OCR software keepsthe same in image format and to make PDFs. Now, the final version of the scanned document is full-text searchable, machine-readable, and with low file size.

5.5. Preservation in Storage Device

Digital Preservation is a continuation of the digitization process of the print materials. It is an original digital copy of a print document captured at the best possible quality/resolution, for long-term usage, and production of a range of dissemination. Typically, master copies of scanned images are stored in an off-line mode on a hard disc or CD-ROM and are accessed only for the production of a derivative. The saved folder of entire collections of TIFF images of one document and the full-text searchable PDF file is stored in a CD-ROM/DVD-ROM as well as in a master computer. After satisfactory preservation, the scanned image files are removed from the scanning computer.

Quality of scanning the document is considered based on the document quality, size, and type. The Digitization process mentioned above are minimum standards and ensuring that the captured digital images are to be accessible and usable in the best quality.

5.6 Few difficulties While Scanning

- ❖ Some Civil and Architectural drawing size has exceeded than the scanning area. When the document pages are oversized, need to scan two or more parts, then these are joined together.
- ❖ Manuscript contents are in-depth in the spine area, in this case, dismantled the document binding, scanned all the separated pages, and then the document is rebound as it is.
- ❖ Fragile documents are handled carefully and taken extra care; such documents are taking more time to complete the digitization.

6. Preservation of Digital Documents in Dspace

Even though many software available in the market, we decided to go with DSpace to set up an IDR (Institutional Digital Repository). The open source repository software DSpace developed by MIT & HP chosen above Fedora, EPrints, and Greenstone Digital Library because it has a good Web UI, structure, and the ability to manage various file formats. DSpace is emphasizing flexibility in capturing a wide range of content while EPrints focused more on publication aspects (Burns et al., 2013).

The DSpace software installed on a separate high-end server and the software is customized in such a way to satisfy administrator and user requirements.

The next step in the digitization and preservation process is the creation of metadata on the digitally created documents. The metadata creation includes three main categories: Descriptive, Structural, and Administrative metadata. DSpace using the Dublin Core metadata standard. But some of the metadata fields required for thesis document are not focused in the DSpace data structure, i.e., Dublin Core. So that, it was decided therefore to customize the metadata fields for digitized thesis available in our library. DSpace's metadata standard can be modified to suit particular document types like thesis, articles, etc. Even though the DSpace metadata is customizable, we did not input the metadata about the structure of the document.

The next step of the preservation process is data ingestion or content submission. Through the DSpace Web UI, metadata ingestion, uploading of a full-text thesis and it's abstract, which are done by a system administrator or a staff authorized by the administrator. All item creation and management are done by the administrator so that minimal descriptive metadata of thesis are considered to preserve in the DSpace.

6.1 Submission Process

A series of steps followed to submit an item in the "DSpace Submission" process, and it may include one or more web UI pages. By default, the DSpace Submission process includes the following steps, in this following order that is creation/selection of collection for which content is to be ingested, describing the content metadata, upload the full-text and other files, review before final submitting, and agreeing to license text before completing the process

6.2 Browse and Search

Users expect that the interaction to the preserved digital documents should not be complicated, and to be open and free access with a standard search engine. DSpace provides a simple web-based user interface to browse and to search for their required information. The essential component of discovery in DSpace is search, and the attributes Communities & Collections, By Issue Date, Authors, Titles and Subject wise can be browsed and searched by a user, and also limited to items within a particular community or collection. The browse and search facility is given only to institute's community through an intranet.

The IDR contains 3944 e-version of theses in various subjects of research outputs. Out of these, 2063 converted digitally with the help of the above-discussed digitization project. About 354 are ready to upload in the IDR. About 834 theses are yet to be digitally converted and uploaded in the IDR.

6.3 Re-Use

An interoperable protocol or standard is being adopted to expose metadata to an external repository or system, such a protocol facilitates the efficient dissemination of repository metadata. "Open Archives Initiative" has developed the protocol for metadata dissemination or harvest. Open Access Initiative Protocol for Metadata Harvesting called OAI-PMH is known protocol to disseminating metadata.Under this model, metadata is harvested (extracted) from Data Providers (Repositories) by Service Providers (Search Engines) (Singh et al., 2008). The whole DSpace metadata of our IDR is harvested by the National Digital Library of India (NDLI) project with the use of OAI-PMH model.

6.4 Copyright Issues

Copyright law is the central issue of converting physical form to a digital form of library materials or copyrighted materials. Before starting the digital conversion process, the librarians or administrators have to consider whether the digital conversion is violating the copyright and intellectual property law. Minow (2002), a library law consultant suggested in his title "Library Digitization Projects and Copyright", Copyrighted materials published prior to 1922 may be digitized without the need to obtain the author's permission. If an item is still under

copyright, it may be digitized for the purpose of education and research or in-house use only. Even if a work is copyrighted, under certain conditions libraries have the right to create digital copies on works published and must not make for commercial purpose. A notice of copyright must be included in any converted digital material. The default copyright notice will be issued to our users when they are downloading the full-text thesis. Copyright statements are varying from institution to institution, but most of the institution allow to access provided this is for educational or research purposes only.

Our current scenario is that the research scholars of the institute submitted their thesis in a printed as well as a soft copy for archiving in the Central Library of the institute. Also, a 'Transfer of Copyright' is received from the research scholars like i) agree to allow the Institute to place the electronic version of thesis on a private intranet maintained by the Institute for its academic community; ii) agree to allow the Institute to upload the electronic version of my thesis on a public access website of the internet, should it so desire.

7. Conclusions

In the contemporary technology development, the technology and infrastructure should be used to provide service on information and knowledge dissemination. Conversion of digital format or digitization is the most suitable alternative for dissemination, preservation, and archive and to provide ease of access to print materials. The digitization and preservation project (Old Ph.D. Theses) taken by Central Library is to fulfill the needs of our user community. Latest overhead scanner, scanning software, OCR conversion software, and DSpace open source software are used to complete and to achieve the goal.

The IIT Kharagpur, one of the most important technological institute in India, has been reaching many significant achievements in researches and has been releasing significant research outputs. This digitization project's goal is to convert all old Ph.D. thesis into digital format and make it available online to ease of search and access. Almost 80% of old theses have converted into digital format, and the same have uploaded in the Institutional Digital Repository (IDR), users can search and access abstracts of the e-version of theses. Dissemination service of full-text of theses is provided on request by the user community.

This project will enhance the use and online access to the digital copy of old theses, and create a system to preserve and archive the intellectual output of the research scholars and old print materials of this institute.

References

1. Burns, C. S., Lana, A., & Budd, J. M. (2013). Institutional repositories: exploration of costs and value. D-Lib Magazine, 19(1-2).
2. Choy, F. C. (2007). Libraries and librarians – what next?.LibraryManagement, 28(3), 112-124.
3. Manzuch, Z. (2009). Monitoring digitisation: lessons from previous experiences. Journal ofDocumentation 65(5), pp. 768-796.
4. Minow, M. Library digitization projects and copyright. Available athttp://fultonhistory.com/

L L R X . c o m % 2 0 - %20Library%20Digitization%20Projects% 20and%20Copyright.pdf (Accessed on 18/02/ 2018)

5. Radovan, V. (2010). Public libraries in Croatia and the digitization challenge. Library Review, 59(5), 325-340.
6. Singh, S., Pandita, N., & Dash, S. S. (2008). Opportunities and challenges of establishing open access repositories: a case study of OpenMED@NIC. Paper presented at the Trends and Strategic Issues for Librarians in Global Information Society: ICCSR Sponsored Seminar, Chandigarh, Chandigarh.
7. Witten, I., Bainbridge, D., & Nichols, D. (2003). How to Build a Digital Library: Morgan Kaufmann.

Further Reading

1. ALA. (2007). Definitions of Digital Preservation. Paper presented at the ALA Annual Conference, Washington, D.C.Available at http:// www.ala.org/alcts/resources/preserv/ defdigpres0408 (Accessed on 11/01/2018)
2. Barbara, C., & Agnieszka, W. (2013). Providing access to historical documents through digitization. Library Management, 34(4/5), 324-334.
3. DSpace. Available athttp://dspace.org/sites/ dspace.org/files/archive/1_5_2Documentation/ ch02.html (Accessed on 11/01/2018)
4. Duraspace. Available athttps:// wiki.duraspace.org/display/DSDOC4x/ Submission+User+Interface (Accessed on 11/ 01/2018)
5. FileFormat.Info. File Format. Available athttp:// www.fileformat.info/format/tiff/egff.htm (Accessed on 08/02/2018)
6. IFLA. (2014). Guidelines for Planning the Digitization of Rare Book and Manuscript Collections. International Federation of Library Associations and Institutions.
7. ImageAccess. Scanners BE4-SGS-V2 Professional. Available athttp:// w w w . i m a g e a c c e s s . d e / index.php?lang=en&page=ScannersBE4-SGS-V2Professional (Accessed on 15/02/2018)
8. Joanna, B. (2007). Building an institutional repository at Loughborough University: some experiences. Program, 41(2), 113-123.
9. Londhe, N. L., Desale, S. K., & Patil, S. K. (2011). Development of a digital library of manuscripts: A case study at the University of Pune, India. Program, 45(2), 135-148.
10. Lynch, C. (2003). Institutional repositories: essential infrastructure for scholarship in the digital age. ARL Bimonthly Report(226), 1-7.
11. Mary, W. (2006). Institutional repositories: proposed indicators of success. Library Hi Tech, 24(2), 211-226.
12. Sadanand, B. (2008). Creation of Digital Library of Manuscripts at Shivaji University, India. Library Hi Tech News, 25(1), 13-15.

13. Shampa, P., andSashi, P. S. (2014). Digitization initiatives and special libraries in India. The Electronic Library, 32(2), 221-238.

14. Sutradhar, B. (2006). Design and development of an institutional repository at the Indian Institute of Technology Kharagpur. Program, 40(3), 244-255.

15. Yan, Q. L. (2004). Best practices, standards and techniques for digitizing library materials: a snapshot of library digitization practices in the USA. Online Information Review, 28(5), 338-345.

16. Zinaida, M. (2009). Monitoring digitisation: lessons from previous experiences. Journal of Documentation, 65(5), 768-796.

About Auhors

Mr. M. Manivannan, Assistant Librarian, Central Library, IIT Kharagpur, Kharagpur
Email:vannan_mm@library.iitkgp.ac.in

Dr. B. Sutradhar, Librarian, Central Library, IIT Kharagpur, Kharagpur.
Email: bsutra@library.iitkgp.ac.in

Note:

Online version of this paper, associated data, files and other supplementary materials are available on Institutional Repository of INFLIBNET Centre. It can be accessed online by scanning QR Code or using following URI: http://ir.inflibnet.ac.in/handle/1944/2297