Union Catalogue of Digitised Documents in India

Sanjo Jose
Librarian,
St. Thomas' College, Thrissur -1
sanjojose@gmail.com

Abstract

A lot of digitisation projects are in progress in different parts of India. But many of these efforts are individual and lack a holistic view. The article stresses the need to have coordinated digitisation activities and a union catalogue of the digitised contents in India.

Keywords: Digitisation, India, Coordination, Union catalogue, archival collection

1. Introduction

Libraries are considered to be the store house of knowledge from the past. Most of the collections of the libraries in the past centuries are in print form. Many old libraries houses manuscripts in leaves, handmade paper, cloth etc. Past two decades saw the emergence of World Wide Web (WWW) and proliferation of digital documents, which changed the role of libraries from store houses of books and other documents to institutions those provide access to information.

2. Electronic information sources

With the introduction of information and storage technologies, libraries have started to use them in providing access to information in electronic databases through different media in the forms of magnetic disks like floppy disks and optical disks like CD-ROMs. Although the amount of information stored in these media were large compared to the conventional sources like print journals, the pace at which they were distributed was same as that of printed articles. The electronic disks were to be dispatched through air or surface mail to the user institutions or libraries just like the print media. Along with the development of communication technologies, the mode of delivery of electronic information was revolutionized by Internet and WWW. The electronic databases and journals were able to be accessed at the end users desktop personal computers from the places of their own origin through high end servers.

3. Online information sources

Most of the information sources like books, journals, periodicals, theses, dissertations etc. are composed in the electronic form with the aid of computers today. Since they are born digital, it is easy to distribute them in the same form online through WWW. Sometimes they are enhanced with multimedia for an improved user experience. They get printed if they are intended to be distributed in that format.

4. Microreproduction

Before the advent of the above mentioned technologies, libraries store a vast variety of documents which were handwritten or printed and distributed as such. Many of them are precious with immense value to the mankind in terms of their archeological value and the information they possess. They have to be preserved in their present form without much wear and tear while using them. In addition to that the information in them should be accessible to be those interested in their perusal for studies and research. Microreproduction or microphotographic reproduction was introduced to reproduce these materials for the use of information in them. Microforms like microfilms and microfiches were used to convert the materials in 1920s. But the conversion to these media require much cost. These media required space to store and special readers to read them. The media themselves were prone to corrosion very soon.

5. Digitisation

Digitisation was suggested to overcome the disadvantages of the above media. Digitisation involved the scanning of these materials with digital scanners and conversion of the scanned images into user accessible formats.

As briefly explained above, most of the modern information sources are born digital, they can be hosted and accessed through communication networks without much additional tasks and costs involved. But when it comes to archival materials like manuscripts, old and rare books etc. they have to be converted to digital form using much manual efforts, time and cost.

Many organizations in different parts of the world like libraries, archives, universities, cultural institutions etc. have started digital libraries populated with the digitised documents. India is also not an exception. Many digitisation projects are going on in different parts of this large country.

6. Major digitisation projects

Some of the major digitisation projects are listed with brief descriptions. The descriptions of the projects are taken from the official sites of the respective projects which are mentioned in the paratheses.

6.1 Million Book Project
The Million Book Project, was a book digitisation project, led by Carnegie Mellon University School of Computer Science and University Libraries. Scanning of a million books was completed in 2007 and they are available online at the Universal Library (http://www.ulib.org/). The project is also associated with Digital Library of India.

6.2 Google Books

Google Books (https://books.google.com) is a service from Google Inc. that searches the full text of books and magazines that Google has scanned, converted to text using optical character recognition, and stored in its digital database.

6.3 Internet Archive

Open Content Alliance (http://www.opencontentalliance.org) is a consortium of organizations like Yahoo!, Internet Archive, University of California, University of Toronto and others contributing to a permanent, publicly accessible archive of digitised texts. Scanning for the Open Content Alliance is administered by the Internet Archive, which also provides permanent storage and access through its website at https://archive.org/details/texts.

7. Scope

The digitisation projects in India which involve conversion of documents are the subjects of discussion in this article. It excludes the digital library projects which involve only born digital documents and digitisation of theses and dissertations. Theses and dissertations are usually confined to the institutions like colleges and universities and the availability of their multiple copies are generally scarce, which in turn reduces the possibility of the digitization of multiple copies at multiple locations.

8. Digitisation projects in India

8.1 Digital Library of India

Digital Library of India (http://www.dli.ernet.in/) is a result of collaboration of Million Book Project with Ministry of Communications and Information Technology for the Government of India and research partners in India like Indian Institute of Science, Bangalore, International Institute of Information Technology, Hyderabad, ERNET and 22 major participating centres. The project scanned books in many languages, using optical character recognition to enable full text searching, and providing free-to-read access to the books on the web. The library is migrating to structured database based on DSpace software. At present, the contents of the digital library are hosted in Tiff image format which is less user-friendly on web browsers. The quality of metadata is very poor, which is evident in the new home page of DLI hosted in DSpace platform. For example some author names in DLI are XXXX, xxxx, N.A., N. A. and titles include UNDERSANDING TAXES etc.

8.2 Traditional Knowledge Digital Library

Traditional Knowledge Digital Library (http://www.tkdl.res.in/)
TKDL was set up in 2001 as a collaboration between the Council of Scientific and Industrial Research (CSIR) and the Department of Ayurveda, Yoga and Naturopathy, Unani, Siddha and Homoeopathy (Department of AYUSH), Ministry of Health & Family Welfare, Government of India. The objective of the library is to protect the ancient and traditional knowledge of the country from exploitation through biopiracy and unethical patents, by documenting it electronically and classifying it as per international patent

classification systems. TKDL is a representative database of 1200 Ayurvedic, Unani and Siddha formulations. TKDL provides access to 2.50 lakhs medicinal formulations to patent offices only under TKDL Access Agreement.

8.3 National Mission for Manuscripts

The National Mission for Manuscripts, NAMAMI (http://www.namami.org/database.htm) was established in February 2003, by the Ministry of Tourism and Culture, Government of India. NAMAMI seeks to unearth and preserve the vast manuscript wealth of India. India possesses an estimate of five million manuscripts, probably the largest collection in the world. These cover a variety of themes, textures and aesthetics, scripts, languages, calligraphies, illuminations and illustrations. Together, they constitute the 'memory' of India's history, heritage and thought. These manuscripts lie scattered across the country and beyond, in numerous institutions as well as private collections, often unattended and undocumented. NAMAMI aims to locate, document, preserve these manuscripts and digitise them to make them accessible to prospective users.

8.4 Indira Gandhi National Centre for the Arts

Indira Gandhi National Centre for the Arts (http://ignca.nic.in/rare.htm) prepared a descriptive catalogue of the illustrated rare books at its library. Kalasampada: Digital Library- Resource for Indian Cultural Heritage (DL-RICH, http://www.ignca.nic.in/dgt_0001.htm) is a project by IGNCA. But both the projects lack structured database and search facility.

8.5 Panjab Digital Library

Panjab Digital Library (http://www.panjabdigilib.org) is envisaged to locate, digitise, preserve, collect and make accessible the accumulated wisdom of the Panjab region, without distinction as to script, language, religion, nationality, or other physical condition. The full text of the documents in the library is available online.

8.6 National Library of India

National Library (http://www.nationallibrary.gov.in/SearchIndexDigital.php) is under the Ministry of Culture, Government of India. But the full text of any of the documents is not available online in the official website of the Library.

9. Digitisation in Kerala

Four major libraries in Thiruvananthapuram, i.e., Kerala Legislature Library, State Central Library, University of Kerala Library, and Oriental Research Institute and Manuscripts Library have their own digitisation projects.

9.1 Digital Archive of Kerala Legislative Assembly

The Digital Archive of Kerala Legislative Assembly (http://klaproceedings.niyamasabha.org/) is mainly intended to make the Assembly documents accessible online to the general public. The archive encompasses the full texts of deliberations of various Legislative Bodies of Kerala since 1888, Bills Passed, Reports

of Assembly Committees and other Committee/ Commissions and Bulletins Part I & II. The documents can be searched and accessed through different options like name of legislative body, number of session, date, name of business transacted, name of member and subject. Search options are available in both Malayalam and English.

9.2 Kerala State Central Library
Kerala State Central Library (http://statelibrary.kerala.gov.in/digital_archives.html) digitise the rare collection of old and precious documents that are not available anywhere else. The digital archive of the library contains Travancore/Travancore-Cochin/Kerala gazette from 1903 in addition to other rare documents. The archive has a structured database and search facility which is available in both Malayalam and English.

9.3 University of Kerala Library
University of Kerala Library is also involved digitisation. But no information except a single sentence, "The library is currently in the process of digitising its rare collections." in its homepage at http://www.kulib.in/ is available online about its efforts.

9.4 Oriental Research Institute and Manuscripts Library
Digitisation at Oriental Research Institute and Manuscripts Library in University of Kerala is done as part of National Manuscript Mission.

9.5 Kerala Sahitya Akademi
Online Library of Kerala Sahitya Akademi of digitized books and periodicals in Malayalam and English (http://www.keralasahityaakademi.org/online_library/index.html) is available. But the catalogue lacks a structured database and search facility. Moreover the links given in the website directing to the domain <http://www.dli.ernet.in/> are all dead.

10. Suggestions for better digitisation efforts

Process of digitisation involves a lot of manual labour and time for scanning the pages. The costs involved in these processes are high. The creation of documents for a digital library those are born digital involves very less efforts, time and cost.

10.1 Catalogue of digitised documents
Considering this very important factor, the libraries and institutions which are involved in digitisation activities must provide a catalogue of the digitised documents they posses with a minimum metadata like the title, author, date of creation/publication, publisher, place of publication, subject or contents, date of scanning etc. This will help other libraries in the possession of same document repeating the scanning process, which will save a lot of time and effort.

10.2 Coordinated efforts

A coordinated effort should be taken by the libraries or institutions which are digitising their collections to avoid duplication of efforts by different libraries digitising same documents. A nodal agency should coordinate the whole process, just as Indian Institute of Science coordinated the creation of Digital Library of India.

10.3 Union catalogue

A union catalogue of the documents digitised and hosted in libraries and institutions in different parts of India, will be helpful for users and other institutions involved in digitisation to locate the digitised documents. Search in vernacular languages in which the documents are written should be included with the catalogue. Search through transliteration also should be integrated. A model which can be made as a reference is National Union Catalog of Manuscript Collections (http://www.loc.gov/coll/nucmc/) maintained by Library of Congress, United States of America. NUCMC catalogs the archival and manuscript collections held by eligible repositories located throughout the USA and its territories.

10.4 Full text accessibility

The digitised documents should be available online to make use of them to the maximum. Usually there will be no copyright impediments to make these documents online in full text, as most of them might have completed the copyright period.

10.5 Full text search

Full text searching within the digitised documents will be a desirable quality for future research. For providing this facility the documents should have a reasonable quality of optically character recognised text.

11. Conclusion

Coordinated efforts and cooperation among the institutions and libraries will always help the users to locate their document easily. By making the availability of the documentary heritage of the country over a large network, the next generation will be more aware of the depth and breadth of the knowledge our ancestors possessed.