DESIDOC Journal of Library & Information Technology, Vol. 32, No. 5, September 2012, pp. 417-420

Building Digital Library using DSpace: Case Study of GIPE's Dhananjayarao Gadgil Digital Library

Nanaji Shewale

Gokhale Institute of Politics and Economics (GIPE), Pune-411 004
E-mail: nanamani@gmail.com

ABSTRACT

Dhananjayarao Gadgil Library of Gokhale Institute of Politics and Economics (GIPE) has started building up the digital library collections by scanning its valuable rare books during July 2011. The GIPE's digital library uses open source software DSpace for building its digital repository. Currently, repository holds more than 7500 digital documents. These documents are available to all the scholars across the world. The present paper discusses the methodology adapted for its digital library building. The various processes involved were: Identification of the vendor, identifying the documents to be scanned or digitised, overall work flow, preparation of the metadata in excel, converting the data into Dublin Core XML format and then finally uploading the metadata into DSpace server. Unique feature of this digital library is the bulk upload of the metadata along with digital data files of all the digital documents.

Keywords: Open source software, DSpace, digital library, digitisation, metadata, Dublin Core metadata, XML, Gokhale Institute of Politics and Economics, Dhananjayarao Gadgil library

1. INTRODUCTION

Dhananjayarao Gadgil Library was established in 1905 as a public library under the aegis of Servants of India Society with great Indian nationalist leader late Gopal Krishna Gokhale as one of its founding members. With the establishment of the Gokhale Institute of Politics and Economics (GIPE) in 1930, the library has started to act as the library of the Institute, and since then it has been mainly responsible for building its collection. It has the largest and most important collection of social science documents in India which includes publications on economics and allied social science subjects. Presently, the library successfully caters to the needs of scholars, academicians, policy makers and others working in social science subjects in India and abroad. Though attached to GIPE, the library extends membership to public and thereby, performs the dual function of an academic or research and a public library.

At present the library possesses more than 2.75 lakh documents. Other than books, it has bound volumes of around 1873 journal titles, around 400 unpublished theses and dissertations and some microfilms, floppies, CDs, and maps. Indian economic history in general and land legislation in particular are most comprehensive and unique collections of this library. Majority of its collection includes books on agricultural economics, Indian economy and reports of various committees and commissions appointed by Government of Maharashtra and Government of India from time to time.

The GIPE library has supported various research projects in the areas of rural development, co-operation, poverty alleviation, employment guarantee, crop insurance, planning, federal finance, input-output economics, agricultural development, irrigation and several other studies in the social and economic problems of India. According to Haider[1], "Libraries are central organs in a university and play a significant role in achieving the due objectives of higher education". Similarly, GIPE library too has played an important role of supporting higher education and research for more than a century.

2. DIGITISATION AND DIGITAL LIBRARIES

According to Wikipedia[2], a digital library is a library in which collections are stored in digital formats (as opposed to print, microform, or other media) and accessible by computers. The digital content may be stored locally, or accessed remotely via computer

Received 14 May 2012, accepted 14 May 2012, online published 7 September 2012

networks. A digital library is a type of information retrieval system Deegan & Tanner[3], describe a digital library as:

(a) A digital library is a managed collection of digital objects.

(b) The digital objects are created or collected according to the principles of collection development.

(c) The digital objects are made available in a cohesive manner, supported by services necessary to allow users to retrieve and exploit the resources just as in the case of other library materials.

(d) The digital objects are treated as long term stable resources and appropriate processes are applied to them to ensure their quality and survivability.

The purposes of digitising and building the GIPE's digital library can be summed up as:

(a) Preservation of most of the rare and valuable documents by reducing wear and tear of the books which are frequently used by users for reference and photocopying.

(b) It was observed that most of this collection is used by researchers from all over the world and it may not be practical for every user to visit the library and to access the same.

(c) To avoid the direct access to rare and valuable books which are in a wear and tear condition and need physical preservation? So digitisation will give an access to the materials that can no longer be accessed in their original format.

(d) Enhance customer service as per their expectations from existing resources at their convenience and may not be at any extra cost.

3. NEED FOR DIGITISATION

Hundreds of the libraries in the world have launched digitisation projects to digitise their collections and place them on the web. Usually, digitisation of the old and rare materials helps to present information in new and important way. Dhananjayarao Gadgil library houses books that are published right from 1680 onwards. There are thousands of the books which are now free from copyright. The main purpose of digitising is to enhance access and improve preservation. By digitising these collections, GIPE library can make information accessible that was previously only available to a selected group of researchers that could walk into the library. The idea of digitisation is that, it will allow users to search collections rapidly and comprehensively from anywhere at any time. Table 1 shows the type of documents available in GIPE along with its number which required immediate action to digitise for preservation and providing wider access to the research community.

Table 1. Number of documents need to be digitised on priority basis

Type of documents	No. of documents
Rare books published from 1680s	01,700
Old and books free from copyright	50,000
GIPE publications (books, reports, etc.)	01,000
(quarterly journal)	
GIPE Publications - *Artha Vijnana*	00,200
Maps	00,300
Servants of India Society's annual reports,	00,150
newsletters	
Thesis, dissertations and other	00,200
unpublished materials	
Miscellaneous documents	01,000

4.1 Software Identification

The GIPE library has preferred to use open source software both for the operating system as well as the digital repository software. Some of the reasons for using open source software are: (1) available free of cost, one can just download from the internet site like http://sourceforge.net. (2) Open source software like DSpace allows customisation to suite the institute's websites look and feel, (3) It does not have any vendor lock or copyright, (4) Open standards used by this software allow easy sharing of information with vast community of implementers and developers around the world, (5) Software is updated frequently, and (6) Online manuals, help and technical support is available either through mailing lists or internet relay chats (IRCs).

According to Directory of Open Access Repositories (OpenDOAR)[4] and Registry of Open Access Repositories (ROAR)[5], more than 80 known software are used for building the digital repositories. The ROAR site indicates that 1089 repositories from its 2730 registered repositories use DSpace alone. At the same time, it is noticed that 859 repositories from 2191 The DOAR's registered repositories use DSpace. The DOAR[6] also gives a detailed graph showing the usage pattern of open source software for digital repository.

Due to some of these reasons, GIPE adapted DSpace for content management and Ubuntu 10.04 Long Term Stable Version as an operating system. Installation of the Ubuntu, DSpace and other components is done by using LibLiveCD[7] which is developed by Prof ARD Prasad (DRTC, Bangalore) and Dr Sunita Barve (NCRA, Pune). One can download the LibLiveCD from sourceforge.net. Other components installed with DSpace from LibLiveCD are: Apache Tomcat server, Exim-4 mail server, Squirrel mail–web-based mail client, PosgreSQL/MySQL–database server, Open SSH server, PHP5, Z39.5 server, etc. According to Shewale & Barve[8], some of the advantages of using DSpace are:

- Largest community of users and developers. More

than 800 digital repositories in the world are using DSpace which is highest than any other software of this type

- Manage and preserve all types of digital documents
- Supports handle system for every document that is added into DSpace repository (http:.//handle.net)
- Support excellent workflow for all operations within DSpace
- Completely customisable to fit ones needs in terms of users' interfaces, metadata, OAI-PMH compatible, database configurations, choice of language, etc.
- Used by number of educational, government, private and commercial institutions

4.2 Identifying the Vendor

After receiving substantial funds for starting the digitisation project it was decided that rare and old but valuable books from the library be digitised. Accordingly, quotes were invited from various vendors who are involved in the process of digitisation. Following criterion were taken into consideration to get the best suppliers/vendors.

(a) Vendor had do the scanning, cleaning text or images, skew correction, de-speckle, cropping, providing equal margins, etc. and make searchable PDF for each book alongwith chapter link, insert watermark (if any) and upload the data on DSpace or any other software, etc.

(b) Vendor was to be responsible for maintaining confidentiality about work, follow the data protection acts.

4.3 Selection of Books

After identifying the vendor, the actual digitisation work started in July 2011. First step involved in this project was the selection of books. Even though there are more than 50,000 books that need digitisation/preservation, priority is given to the rare books. Other criteria to identify the books for digitisation is the publication year, i.e., book must be published before 1940, means free from copyright but at the same time the book or title should be most sought by the library users. Till March 2012, the library has successfully completed the digitisation of more than 6700 books and uploaded almost 5900 titles on the DSpace server.

4.4 Preparation of Excel Data-sheets

Once the books are identified, the bibliographic data of the document is downloaded from the Library Management System used by GIPE library and converted it in the Excel format. This data includes the bibliographic fields like accession number, title, subtitle (if any), author(s), edition, publisher, place of publication, year of publication, class number, keywords (if any), and abstract (if any).

4.5 Quality Check of Digitisation work

After digitisation, the vendor is asked to submit two searchable PDF files: (1) preliminary pages alongwith the content pages of the document and (2) entire book, i.e. from cover page to the last page. For ease of access and identifying the documents, the accession number was used as file name like GIPE-000019-Contents.pdf and GIPE-000019.pdf. The content file helps the end-user to judge whether to download the full book or not, thereby saving the bandwidth and the traffic on the net. First of all, it is checked that the number of documents returned by the vendor are same both in terms of the number and the titles. Secondly the physical condition of the books is in good condition or not after the books are returned back by the vendor. Later the PDF files of both the content pages as well the entire book are checked.

4.6 Preparing DC XML Metadata

In the excel sheet other fields such as type of document, number of pages, content page file, and digitised book PDF file, etc. were also included. As DSpace accepts the data in XML format with Dublin Core standard format, this excel data is converted to Dublin Core metadata format using some utilities. Presently, in GIPE the data is converted by developing a directory from the Microsoft Mail Merge as the data is already available in the excel format. There are also a number of utilities available on the internet to convert the excel data to Dublin Core XML format. Table 2, shows the sample record in XML format to be uploaded on DSpace server.

4.7 Bulk Uploading on DSpace Server

In July 2011, the records were uploaded one by one at an average of 10 minutes per record as one has to input various fields and upload the data files. It was observed that the process was very time consuming and there was a possibility of typographical mistakes while entering data. After exploring the Internet, it was found that there are some utilities to bulk upload data on DSpace Server in XML format. But these utilities were limited to uploading only the metadata and not the data files called the bit-streams. Due to the limitation of these utilities on internet, Prof ARD Prasad, has developed some utilities in Perl for uploading and importing bulk metadata records alongwith respective bit-streams. The advantage of these utilities is that one can upload any number of records into DSpace in a very short span of time. Usually the time taken for uploading the records depends on the number of records, the size of bit-streams and the internet bandwidth. Till 8 August 2011, 95 books (37791 pages) were digitised and uploaded. During 13 March 2012 to April 2012, 350 books

Table 2. Sample record in XML format

DC Code Start	<?xml version=”1.0" encoding=”utf-8" standalone=”no”?><dublin_core schema=”dc”>
Title	<dcvalue element=”title” qualifier=”none”>History of Hindu civilization during British rule Vol.III </dcvalue>
Author	<dcvalue element=”contributor” qualifier=”author”>Bose, Pramatha Nath</dcvalue>
Class Number	<dcvalue element=”subject” qualifier=”cc”>Y73Q2:1.2.M/C6.3</dcvalue>
Type of Document	<dcvalue element=”type” qualifier=”none”>Book</dcvalue>
Contents (.pdf)	<dcvalue element=”identifier” qualifier=”uri”>GIPE-000019-Contents.pdf</dcvalue>
Full Text (.pdf)	<dcvalue element=”identifier” qualifier=”uri”>GIPE-000019.pdf</dcvalue>
No of Pages	<dcvalue element=”identifier” qualifier=”citation”>228p</dcvalue>
Accession No.	<dcvalue element=”identifier” qualifier=”accno”>GIPE000019</dcvalue>
Publisher, Place	<dcvalue element=”publisher” qualifier=”none”>W. Newman and Company, Calcutta</dcvalue>
Year	<dcvalue element=”date” qualifier=”issued”>1896</dcvalue>
DC Code End	</dublin_core>

(92575 pages) were digitised and uploaded. Total 6700 books (2092246 pages) have been scanned.

The GIPE is in the process of acquiring new IP addresses with new bandwidth and hence present site address/domain name of repository (http://library.gipe.ac.in) is likely to change soon. In such circumstance, one should google the keywords like ‘GIPE’, ‘Pune’, ‘Digital’ and ‘Library’ to locate repository on internet.

4.8 Statistical View

Bragan[9] quotes, ‘Say you were standing with one foot in the oven and one foot in an ice bucket. According to the percentage (statistics) people, you should be perfectly comfortable’. Table 3 shows some of the statistics (facts) generated by DSpace from Dhananjayarao Gadgil digital library.

Table 3. Statistics of Dhananjayarao Gadgil digital library

Type of statistics	Number
Bit-stream views	3,99,830
Item views	2,42,007
Community views	2,52,132
Collection views	2,18,010
Searches performed	0,18,224
OAI requests	0,02,180
User logins	0,00,528

5. CONCLUSIONS

With the invention of the printing technology almost four centuries ago, huge literature is available in printed format. With the limited life of the print material, some of the old and rare books are now being found in an acutely precarious condition afflicted by termite and time rote decay**.** As these are priceless and most valuable to researchers, it is imperative that these books be retrieved and saved from extinction by the use of technology visa–a-vies the process of digitisation. Secondly, there is a huge demand from the scholars to access this good old literature, and hence it’s high time that the LIS professional start digitising, uploading and making these collections on the web for the academic community.

ACKNOWLEDGEMENTS

Author acknowledges the contribution of Prof A.R.D. Prasad, DRTC, Bangalore for developing Perl utilities for bulk upload of metadata with digital bit-streams onto DSpace server; encouragement by Prof Rajas Parchure, Director, GIPE for initiating and successful implementation of GIPE’s digital library; technical support from Dr Sunita Barve, SIRC-in-Charge, NCRA; and Mr Pramod Joshi, Information Scientist, GIPE.

REFERENCES

1. Haider, S.J. Coping with change: Issues facing university libraries in Pakistan. *J. Acad. Lib.*, 2004, **30**(3), 229-36. http://en.wikipedia.org/wiki/Digital_library, http://www.opendoar.org/index.html
2. Digital Library. Wikipedia–The Free Encyclopedia.
3. Deegan, M. & Tanner, S. Digital futures: Strategies for the information age. Library Association Publishing, London, 2002. 22 p.
4. OpenDOAR: Directory of Open Access Repositories.
5. ROAR: Registry of Open Access Repositories. http://roar.eprints.org/
6. OpenDOAR Chart: Usage of open access repository software–Worldwide.
7. Prasad, A.R.D. & Barve, Sunita. LibLiveCD. http://liblivecd.sourceforge.net/
8. Shewale, Nanaji & Barve, Sunita. LibLive CD for novice users. *In* Advances in digital libraries development, edited by Antony Jose. Macmillan, Delhi, 2012. pp. 185-92.
9. Bragan, Bobby. Statistics quotes and sayings. 1963. http://www.quotegarden.com/statistics.html