Efficient Collection Management in Large Scale Digital Library: A Case Study of National Digital Library of India

Dhara Sharma
Junior Project Officer, NDLI Project IIT Kharagpur, E-mail: dsdharu@gmail.com

Debasis Das
Junior Project Officer, NDLI Project IIT Kharagpur, E-mail: debdas12312@gmail.com

Rajib Kundu
Project Officer, NDLI Project IIT Kharagpur, E-mail: kundu.rajib77@gmail.com

Abstract

Collection development in Digital education plays an important role in our modern education system. In this new era of digitized documents, National Digital Library of India (NDLI) gives a new vision of educational materials to all age population according to their needs. This paper discusses the process of developing various digital collections, outlines the challenges faced and management by giving a special focus on National Digital Library of India (NDLI).

Keywords

Digital Collection, Digital Repository, Collection Management, NDLI

I. Introduction

The digital environment collection has transcended from print environment due to the advancement of information and communication technology, information explosion, and the availability of large number of documents in electronic forms. The selection of quality and quantity of collection is an important and challenging activity of any digital library. A library consists of a numbers of activities related to the development of library collection like determination and coordination of selection policy, identification of user's needs, user studies, selection of information material, planning of resource sharing, collection maintenance and weeding. National Digital Library of India similarly encompasses contents of all educational institutions across India**,** which caters a huge span of students from toddlers to lifelong learners. It is a key driving force for its education, research, innovation and knowledge economy.

II. Collection Development in Large Scale Digital Library – NDLI

The term 'Collection development' refers to the process of systematically building library collections to serve study, teaching, research, recreational, and other needs of library users. The process includes selection of current and retrospective materials, the planning of strategies for continuing acquisition, and evaluation of collections to determine how well they serve user needs. It is one of the important and challenging library management activities.

NDLI has become an online digital treasure-trove of India with approximately 1.26 crore content resources in more than 100 languages, sourced from about 150 institutes / publishers, bringing up a significant change in the domain of online content search for both academics and general readership. The focus of NDLI is to promote universal open access of learning contents. It has promised to co-ordinate digital library development, resource sharing activities among the SAARC countries. It serves all kind of learners including school and college students, teachers, education administrators, researchers, competitive examinations aspirants, professionals/ practitioner's life-long learners or physically disabled. A representative view of the content coverage of NDLI is given below:

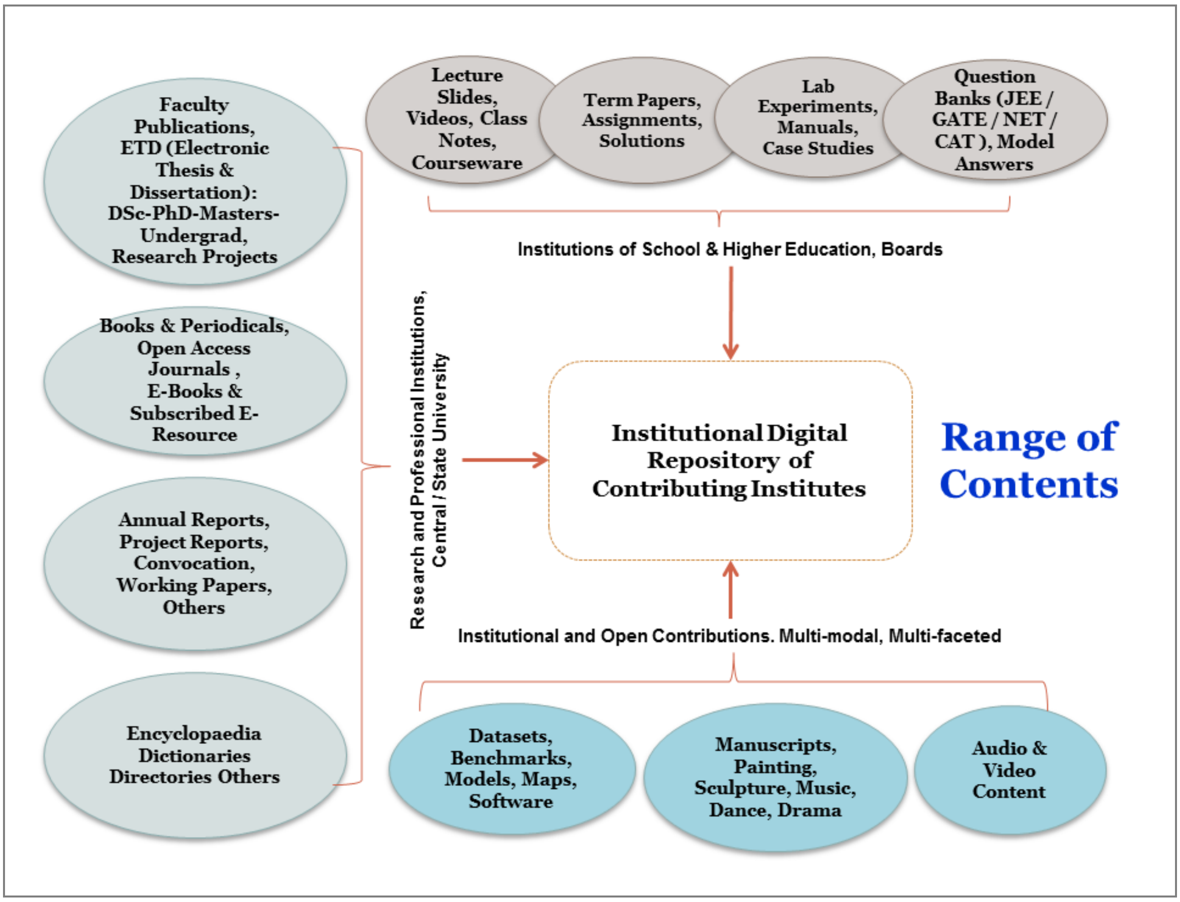
Lecture Slides, Videos, Class Notes, Courseware
Term Papers, Assignments, Solutions
Lab Experiments, Manuals, Case Studies
Question Banks (JEE / GATE / NET / CAT), Model Answers
Institutions of School & Higher Education, Boards
Faculty Publications, ETD (Electronic Thesis & Dissertation): DSc-PhD-Masters-Undergrad, Research Projects
Books & Periodicals, Open Access Journals , E-Books & Subscribed E-Resource
Annual Reports, Project Reports, Convocation, Working Papers, Others
Encyclopaedia Dictionaries Directories Others
Research and Professional Institutions, Central / State University
Institutional Digital Repository of Contributing Institutes
Range of Contents
Institutional and Open Contributions. Multi-modal, Multi-faceted
Datasets, Benchmarks, Models, Maps, Software
Manuscripts, Painting, Sculpture, Music, Dance, Drama
Audio & Video Content

Figure I: Range of Contents in NDLI

III. Challenges in Collection Management

3.1 Diversity of Resources

NDLI covers all domains such as School Board, Competitive Examination, Science and Engineering, Medical, Legal, Culture and Heritage, Humanities and Arts. It's not only restricted in books and journals but also audio and video, software's, animations, simulations, lab materials, question banks, model answers, etc. NDLI facilitates students to prepare for entrance and competitive examinations.

3.2 Schema

Digital library system rests on 'Metadata' which is the basis of document exploration. NDLI follows integrated metadata standards from national and international origin, such as Dublin core, IEEE-LOM, MARC, Shodhganga. The aggregation of NDLI contents and metadata spans through varied sources, content type and categories and domain. For NDLI metadata are collected, curated and assigned to each document. The challenge in this process is to remove noise, manage, and maintain the same.

3.3 Mapping of Data

After harvesting the source data, the harvested data is being mapped to the NDLI metadata schema, curated through the computer programs or by the human curators. The usual first choice of file formats is either comma delimited text files, since these are easy to dump from many databases, or JSON format, often used for event data or data arriving from a rest API. Text files are readable by humans and therefore easy to generate, debug and troubleshoot. But there are also significant drawbacks to this approach, and often these drawbacks only become apparent over time, when it can be challenging to modify the file formats across the entire system. But the worst problem by far is the fact that with CSV and JSON data, the data has a schema, but the schema isn't stored with the data. Since the schema is stored in the file, programs don't need to know about the schema in order to process the data. Humans who encounter the file can also easily extract the schema and better understand the data they have. The problem of managing schemas across diverse teams in a large organization was mostly solved when a single relational database contained all the data and enforced the schema on all users.

3.4 Heterogeneous Data

Heterogeneous data structures are those data structures that contain a variety or dissimilar type of data. NDLI aggregates heterogeneous data from various sources which of the most challenging issues to map all types of data in to a single data standard.

3.5 Technical Architecture

In order to provide digital contents in an accessible way technical challenge is a big hurdle**.** Typical problems includes source site is not opening; resources not harvestable etc. Within a coordinated NDLI scheme, some common standards have been needed to allow other digital libraries to interoperate and share resources. The problem, however, is that across multiple digital libraries, there is a wide diversity of different data structures, search engines, interfaces, controlled vocabularies, document formats**,** etc. Because of this diversity, federating all digital libraries nationally or internationally would an impossible effort.

3.6 Copy right issues

Copyright has been called the barrier to digital library development. NDLI systematically collect digital information on a larger scale, the provision of effective access could be questionable. NDLI, simply caretakers of contents they don't own the copyright of the material they hold. It will ever be able to freely digitize and provide access to the copyrighted materials in their collections. Providing the copyright status of each digital object, and the restrictions on its use or the fees associated with it.

IV. Collection Management Workflow

NDLI collaborates with different sources (institutions, publishers, digital content owners, etc) to get permission of incorporating their metadata and contents in NDLI. It does not store contents, NDLI only ingests metadata for Search & Browse, Content (Full-text) is delivered from Source, following approaches to acquire metadata.

For IDRs' the metadata is harvested through communication link, mapped to NDL metadata schema and curated, through computer programs followed by human review. Sources which are making available metadata of their contents in XML, MARC or CSV format are also mapped to NDL metadata schema and curated via programmatic and manual interventions. Contents available in structured educational websites are being crawled through computer programs to extract metadata from the website as per NDL metadata. Some amount of manual curation may need to be done. Manual annotation of metadata, directly from contents, is also being applied though to a limited scale. Software tools have been developed to automatically extract metadata from contents. The extracted metadata then goes through a round of manual checking and curation / validation.

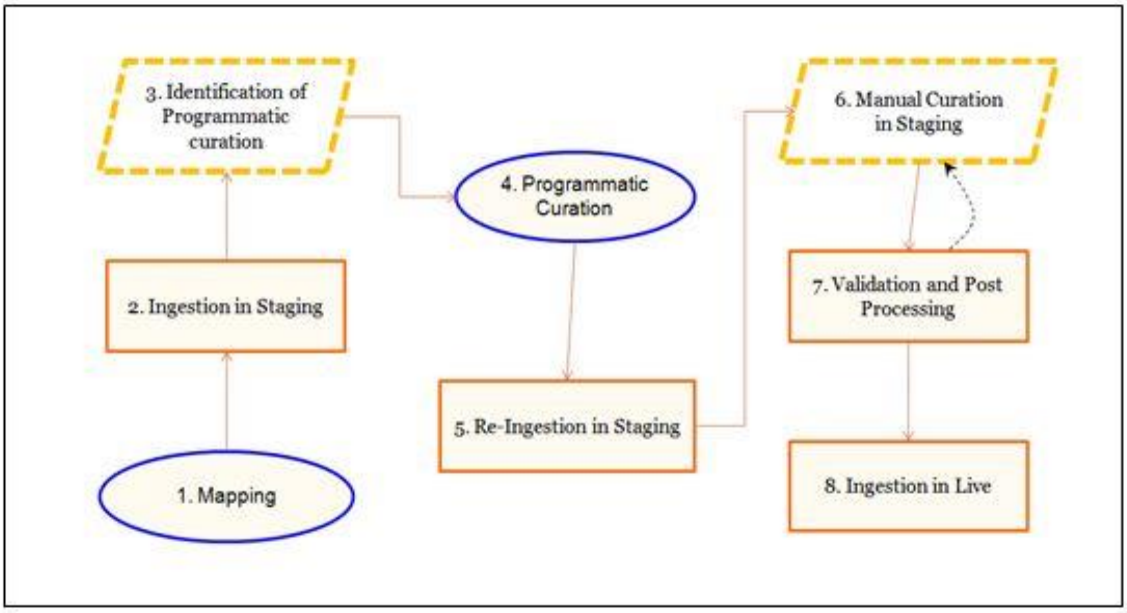
3. Identification of Programmatic curation
4. Programmatic Curation
6. Manual Curation in Staging
2. Ingestion in Staging
7. Validation and Post Processing
5. Re-Ingestion in Staging
1. Mapping
8. Ingestion in Live

Figure II: Curation Process

V. Statistics in NDLI Collection

NDL is structured to cover all domains of knowledge i.e Technology, Natural Science and Mathematics, Social Science, Computer Science, Literature, History & Geography, Philosophy, Pscychology etc. All types of learning contens- not just books and journals but also video, audio materials, software, animations, stimulations, web-courses, hands on, lab materials, question banks, model answers etc.

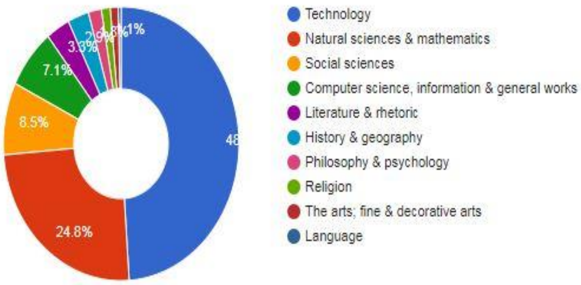

Figure III: Subject-wise Content Distribution

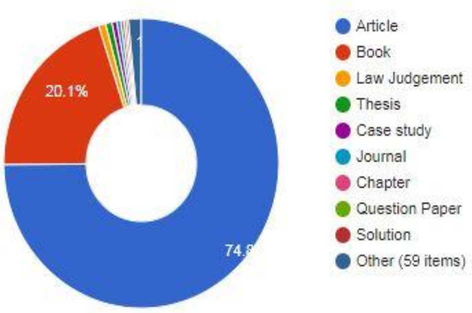

Figure IV: Learning Resources Type Distribution

VI. Conclusion

The collection development is a planning process by which the library selects and manages its collection of digital resources. With facing various challenges i. e. Effective crowd- sourcing mechanism for contents, National licensing, Copyright policies, Heterogeneous data structure, NDLI collection development has become the most important and affecting at great extent. This digital library has taken an initiative of providing right resources and services to its users. It is an educational ecosystem platform where all stakeholders can participate, contribute and benefit.

VII. Acknowledgment

This work would not have been possible without support of Prof. Partha Pratim Das, Joint P.I of NDLI Project, Prof. Plaban Kumar Bhowmik, Co P.I of NDLI Project and Chief Technical Officer Mr. Nanda Gopal Chattopadhyay to write this article.

VIII. References

1. Chakrabarti, P. P., Das, P. P., Bhowmick, P., et.al (2016). National Digital Library: Building a National Asset. Yojana, 60, 19-23.

2. Das, P. P., Bhowmick, P., Sarkar, Sudeshna, et.al. (2016). National Digital Library: A Platform for Paradigm Shift in Education & Research in India. Science and Culture, 82 (1-2), 4-11.

3. Fordham, A. E. (1999). The collection development planning process. Special Libraries Management Handbook: The Basics. University of South Carolina College of Library and Information Science, 2004.

4. Wang, Lidong (2017). Heterogeneous Data and Big Data Analytics. Automatic Control and Information Sciences, 3(1), 8-15.