Designing Institutional Digital Repository for the University of Burdwan: A FLOSS Based Prototype

A Synopsis

Thesis submitted to the University of Burdwan in partial fulfillment of the requirements for the Award of the Degree of Doctor of Philosophy (PhD) in Arts (Library and Information Science)

Mr. Bijan Kumar Roy

Supervisors

Professor Subal Chandra Biswas
and
Dr. Parthasarathi Mukhopadhyay

Department of Library and Information Science
The University of Burdwan
Burdwan, West Bengal
2013
1.0 Introduction

Utilizing the World Wide Web (WWW), science has created many new tools for communication that is increasing used commercially. After this technological revolution, the present scholarly communication system has changed. People were dissatisfied with the present publishing system and in the face of rising costs, flat budgets, and restricted access to information, as well as rapid changes in technology, traditional publishing model are being changed to electronic publishing system. New publishing model has developed and due to the open access to knowledge movement, two open access model has became popular throughout the world. IR has emerged as a new OA publishing tool that supports free online access to public funded research. The term ‘IR’ is not new a concept but became popular relatively late with the first discipline based repositories being implemented in the early ‘90s. It has become a hot topic in academic world in the past few years. In these circumstances, very recently an academic movement called the ‘OAM’ was organized as an alternative to the conventional scholarly communication. Two contemporary developments in particular have helped shape the nature of today’s institutional repositories (IRs): the emerging knowledge management movement; and the maturing, but still rapidly advancing, technology of content or asset management in the digital information system. The knowledge management movement of the 1990’s influenced the development of IRs. The major driving force behind the development of IRs has been the dramatic shift in scholarly communication especially within the past five to ten years. The open source software, open access and open standards movement has become much popular and got importance in academic world. This is typically reflected in the success story of open access digital repositories worldwide which mostly use open source software and provide open access to public funded research outputs.

1.1 Background of the Study

While there is a progressive growth of digital libraries and IDRs in several developed countries like USA, UK, Germany, Australia and many countries in Europe, India is witnessing moderate growth. Indian universities and other academic and research institutes are lagging behind in this direction. There are near about 2500 repositories in the world. But this repository movement is neglected in India and only eighty (80) repositories have so far been developed in the country. The concept ‘IR’ is not new but it became popular relatively late with compare to other developed countries. Many research institutes, universities, but not all, have repositories for academic papers. Many are well beyond the conceptual stage in their implementation, and many are operating with a limited scope. Some elite educational and research institutes (such as Indian Statistical Institute, some CSIR Laboratories, IITs, IIMs etc.) have already started their initiatives in building institutional repositories including a few Universities (such as University of Hyderabad, University of Delhi.). Indeed, many ETD initiatives are from institutes of national importance. University Grants
Commission has already developed a policy document on building university level Institutional Digital Repository (http://www.ugc.ac.in) in India. National Knowledge Commission have strongly recommended that all research articles published by Indian authors receiving any government or public funding must be made available under OA and should be archived in the standard OA repository. But there is as such no mandate to dictate that publicly funded research should be reported in OA journals and repositories. All the initiatives are experimental in nature (except a few such as Librarian’s Digital Library (LDL) of DRTC, ISI) and are not based on open standards as far as policy issues.

1.2 Objectives of the Study

The main objective of the study is to develop a standard IDR model and formulate a policy for the University of Burdwan that will support information needs of the university stakeholders. The other objectives are mentioned below:

- Formulating policies in the line of global recommendations and best practice guidelines;
- Applying open standards in the design and development of the IDR system as far as information organization, retrieval and harvesting are concerned;
- Developing Unicode-compliant Bengali language based User Interface to support integrated searching and browsing of regional languages based resources;
- Developing federated search mechanism in order to harvest metadata from OAI-PMH compliant repositories;
- Design and integration of interactive communication tools for scholarly interaction among community members; and
- Incorporation of Indic script based Web-enabled subject access system in Bengali script.

1.3 Statement of Research Problems

The research problem of this study is represented by a descriptive statement as below:

“DESIGNING INSTITUTIONAL DIGITAL REPOSITORY FOR THE UNIVERSITY OF BURDWAN: A FLOSS BASED PROTOTYPE”.

A close examination of this descriptive statement reveals that it is amenable to be converted into a set of five interlinked questions and these are as follows.
How to design a theoretical model for University-specific IDR? What should be the standard components and parameters for it? How to link the parameters with the requirements of users and administrators of IDR?

How to convert the theoretical model into a software framework? How to apply open source software and open standards in designing the IDR software framework? What should be the selection criteria for IDR software and related standards?

What are the procedures for making the IDR software framework Unicode-compliant entity? How to develop mechanisms for processing, retrieval and display of Indic script based information objects? What are the ways to develop and incorporate Indic script based subject access system?

How to incorporate into the IDR software framework for the University of Burdwan utilities like federated searching and interactive communication tools for scholar?

How to implement and integrate the prototype IDR in the existing information access system of the University of Burdwan? How to design and apply multilingual user interfaces at different levels and different points of utilization? How to make the prototype IDR interoperable and open for harvesting? And how to make distribution copy of the software framework in the form of ready off-the-shelf product for the benefit for research and academic bodies?

1.4 Research plan

This research study seeks to understand what university members believe about IDRs including their perceived benefits, their ease of use, and their role within greater system of scholarly publication. It aims to support successful implementation and maintenance of the university’s IDR. However, in response to the research problems, as formulated in the foregoing section, this research study centres on the following hypotheses:

- It is possible to design a theoretical model with carefully crafted techno-organizational components on the basis of global recommendations, best practice guidelines, and taking into account specific requirements for the University of Burdwan;

- The theoretical model is amenable to be converted into an accommodative software framework through the application of related open standards and open source IDR software;
It is quite feasible to make the software framework Unicode-compliant through the application of an array of FLOSS based multilingual tools for storing, processing and retrieving of digital knowledge objects available in Indic scripts in general and Bengali script in particular;

SKOS (Simple Knowledge Organization System), a W3C standard can be utilized for converting any standard classification into Web-enabled access format for its seamless integration with IDR framework;

The Unicode-compliant and FLOSS based software framework can act as IDR system for the University of Burdwan in web-enabled environment and can be integrated seamlessly with the existing information access system of the University of Burdwan; and

The entire model framework with all of its complexities can be converted into ISO image for utilization by other institutes as an off-the-shelf product.

1.5 Scope

The following scope notes define the focus of the document, but in each case it will be important to specify the boundary relations between that which is in scope and that which is out of scope for the roadmap. Near about two thousand four hundreds repositories (2400) from OpenDOAR and ROAR databases have been analyzed and ROARMAP database has been consulted and this provided the information used in this report. The scope of this study is described in details:

- **Geographical scope:**
  - In scope: The University of Burdwan, West Bengal, India
  - Out of scope: Any type of organizations/or institutes throughout the world

- **Object types:**
  - In scope: Public funded research outputs (text, data, projects, learning materials, articles, dissertations and theses etc.
  - Out of scope: Administrative records and other legal matters

- **Media types:**
  - In scope: Potentially all (text, image, sound, moving image, simulation, etc)
  - Out of scope: None
• **Object provenance:**
  - In scope: Objects created, owned, contributed and shared by members of the BURA community
  - Out of scope: objects made available to BURA on a commercial basis

• **Documents consulted:**
  - In scope: Published articles, books, technical reports, conference papers and unpublished e-prints (very selective) published in English language
  - Out of scope: digital media works (such as MP3 files), editorials, e-mail messages, interviews, letters to the editor, news articles, presentation slides or transcripts, questionnaire or web log postings

• **Language Covered:**
  - In scope: English and Bengali
  - Out of scope: None

1.6 **Methodology**

The methodology followed for this research work has been divided into three conceptual areas -

- **Part - I: Development of the theoretical model:**
- **B. Part - II: Selection of standards and software and designing a prototype system**
- **C. Part -III: Implementation of the proposed model.**

*The major steps included in Part - I are discussed*

- Step 1: Understanding IDR, its application and implementation;
- Step 2: Study of existing IDR initiatives at national and international level;
- Step 3: Identification of requirements for BURA on the basis of recommendations made at national agencies such as UGC, NKC etc.;
- Step 4: Identification of parameters for developing IDR;
- Step 5: Development of the model for BURA on the basis of identified parameters;
The major steps included in Part - II are discussed

- Step 6: Development of criteria for different facets of BURA;
- Step 7: Identification and selection of metadata standard for each document type against the selected criteria;
- Step 8: Determination of technical specifications of BURA;
- Step 9: Identification and development of FLOSS based IDR software selection criteria for BURA;
- Step 10: Installation of the selected IDR software along with necessary FLOSS based companion and/or dependent software;
- Step 11: Incorporation of selected metadata schemas into the prototype BURA; and
- Step 12: Development of Unicode-compliant environment for processing Indic script based documents.
- Step 16: Development of Indic script based subject access system;
- Step 17: Inclusion of federated search mechanism; and
- Step 18: Design and inclusion of communication support system for scholars.

The major steps included in Part - III are discussed in details

- Step 13: Implementation of crosswalks and interoperability standards and development of OAI-PMH (Open Archive Initiative-Protocol for Metadata Harvesting) compliant IDR;
- Step 14: Designing of user interface in two levels (Submission interface & End-User access);
- Step 15: Incorporation of vocabulary control devices (existing open standards tools) in two levels (searching as well as indexing);
- Step 16: Uploading of records through distributed processing i.e. remote submission process;
- Step 23: Design of single window embedded search interface for multiple IDR.
Step 24: Development of auto alerting services and platform for scholarly interaction through blogs and discussion forum;

Step 17: Beta testing of BURA and final release through Web-hosting.

Step 26: Generating ISO image from BURA software framework for its ready implementation in other institutes.

1.7 Significance of the Study

In India for several years, research organizations, universities have been generating valuable knowledge resources in the form of theses, dissertations, project reports, courseware, pre-prints, etc through research and development activities. But there is no mechanism for holding such intellectual output of the University for scholarly communication. Unfortunately these resources are not available and inaccessible due to the absence of appropriate proper mechanisms. There is increasing awareness that universities and research institutions lose valuable digital and print material due to difficulties in accessing them and lack of good preservation practices.

In this situation a repository model based on global standards and best practice guidelines is required for Indian universities that will organize and preserve university’s intellectual resources for future use and will provide global access to Indian research outputs in order to serve the following purposes:

- as a tool for preserving university’s research output for perpetual access;
- act as a guiding tool to assist universities to its digital publishing ambitions;
- act as a tool to enable universities offer digital courseware and online learning; and
- provides an ISO image of the BURA software framework so that other universities/or institutes can utilize it in designing IDR system for their own.

1.8 Findings

The document is a first pass at formulating a roadmap for university-specific repository system. There are many unknowns in this area, so the roadmap is aspirational and, to some extent, speculative. This is the first iteration; the intention is to seek further input based on feedback to this study. It is likely that versions of the roadmap will be produced in future as supporting material for various organizations. The barriers are various and derived from the different stakeholders such as organization, funding body, publishers, users, academicians, administrators, and the governmental policies. The results show that only a few elite institutes have established IDRs in India. In practice the ranges of resources is limited and are mainly text based. The problems are three folds: technical (hardware and software), non-technical (administrative, policy issue) and cultural (social acceptance, advocacy etc).
Synopsis

- **Problems related to Policy**

Majority of the repositories do not have well defined regarding organization and management of resources, document types, archiving, submission, workflow, legal contributor, access to contents, quality of contents, versioning, publicity and branding, copy right and licensing, embargo period, preservation and data backup, file formats, metadata, removal or withdrawal of contents, etc.

- **Lack of Institutional Mandate at National Level**

The main problem is undoubtedly the absence of a national mandate on OA publishing. There are now statements of support from different groups both in public and private sector, but no real mandate to deposit copies of papers in an institutional repository yet exists. At national level, UGC, NKC are still to issue a policy statement on the dissemination of research outputs. But there is as such no mandate to dictate that publicly funded research should be reported in OA journals and repositories. There is no initiative to setting up Registry of Indian repositories in line with ROAR and OpenDOAR registries.

- **Problems related to Stakeholders**

The success of any repository system depends on cooperation and active participation of different stakeholders namely researchers, faculties, administrator or government etc. But till now they are not aware of the system and hesitates to submit article in repository.

- **Access related problems**

The growth in usage of repository has been very slow and the number of papers deposited remains small. There are many problems related to the non-availability of the facility to download full text OA articles. Repositories do not provide full text access to their resources and only metadata is searchable.

- **Problem related to Multilinguality**

India is multilingual country having twenty-two (22) constitutionally recognized Languages and less than 5% people can read and write English. But repositories are not Unicode compliant and major portion of their collections are based on only English language. There is no such provision for multilingual information processing, searching and retrieval.

- **Technological problems**

Majority of the IDR initially faces the problems regarding software such as installation, up-gradation, customization; maintenance and back up databases. The basic technology is in place for ingest and harvesting/access by third parties but
seamless technical links between repositories and authoring systems or research information systems do not yet exist.

➢ Problems related to Standard

Most of the IDRs are not working well and are not up to the global standard. Majority of IDRs do not follow global standards and specifications in the following areas like network, data format, software, encoding, interoperability etc.

1.9 Solution achieved

The findings of the study suggest that the needs of IDRs in India are pressure of work, and the need to offer new services and systems that are slow, inefficient and cumbersome. The majority of the repositories, however, are still struggling with their manual operations and systems which were reported to be unsatisfactory in the sense that they are slow and inefficient.

The study suggests that while the current technical infrastructure in the India is in need of some development, it is primarily in the areas of policy (both national and institutional), culture and working practices that changes need to be made. So, it is necessary for India to have a network of National Digital Library centers to spearhead the process of this movement. The government, research councils and other funding bodies need to mandate that all scholarly publications generated by publicly-funded research are made available on an open access basis. The ‘reward structures’ and ‘professional development’ infrastructure within the academic community need to recognize open access as a valuable and important part of the profession. The community needs to find ways to encourage academics to share and re-use publications, research data and learning resources as openly as possible.

1.10 Novelties of research

This doctoral dissertation has developed a software framework using open standards and open source software taking into considerations of global recommendations and best practice guidelines. This software framework has developed Unicode-compliant Bengali script based interfaces (Fig. 1.10) that supports not only browsing and searching IDR resources but also allows repository administrator to perform various operations like creating community/collection, user registration, password management etc. This research study has also incorporated an ontology driven Web-enabled KOS (DDC 22nd edition – up to 3rd summary) in BURA software framework that supports hierarchical browsing and searching of subject categories in English as well as in Bengali script (Fig. 1.10.1). In addition, this study has designed a single window federated searching mechanisms for harvesting metadata from multiple OAI-PMH compliant repositories globally (Fig. 1.10.2). Finally the entire model
framework with all of its complexities has been presented as a ready to use off-the-shelf product for the benefit of research and academic communities.
Conclusion

IDRs throughout the world are at a critical point in their development. This research study is an approach towards the development of open access to knowledge movement by establishing IDRs following global recommendations and best practices as suggested in this doctoral dissertation. In order to provide global access to the public funded research outputs generated by Indian universities and research institutes, this research study has developed a standard IDR software framework using open standards and open source software keeping in mind the requirements of the organization. This software framework may be used as a guiding tool for other organizations in designing IDR for their own and in managing IDR resources in local languages other than default English language. To achieve this success, this software framework has the provision of storing, processing and retrieving Indic-script based resources including integration of Web-enabled Indic script based subject access system. In this context this research study may be viewed as an action in hope rather than hoping for action.

1.

2.

Signature of Supervisors

Signature of Candidate
References


