Indian Digital Repositories: A State of the Art Report

Nanaji Shewale[1] and Sunita Barve[2]

[1]Gokhale Institute of Politics and Economics (GIPE), Pune 411004, India
nanamani@gmail.com

[2]National Centre for Radio Astrophysics (NCRA), Pune 411007, India
sunitab@ncra.tifr.res.in

Abstract. In the 21st century, libraries are handling major part of their collections in digital format. Today, large number of digital collections is accessible through digital repositories available around the world. There are over 2700 Institutional Repositories (IR) registered on Registry of Open Access Repositories (ROAR), and around 2186 IRs on Directory of Open Access Repositories (DOAR). Apart from these, there are several closed and open access repositories which are not registered either with ROAR or DOAR. From India, there are more than 73 digital archives with the oldest one is "Librarians' Digital Library (LDL)" hosted by Documentation Research and Training Center (DRTC), Bangalore which was initiated during January 2004. All these repositories together contribute to more than 2,00,000 digital documents. Lot of efforts are going on to make these collections available either on Internet and/or Intranet. The professional involvement of developing digital archives is generally done by the affiliated library professionals from respective institutions/organizations. These repositories, deliver digital contents from different discipline. Due to continuous changes in hardware and software, sustainability poses an important challenge for any repository in the long term. In the present paper, a survey has been carried out and data is collected from these repositories to find out sustainability issues handled by these repositories from future point of view.

Keywords: Digital repositories, Sustainability, India, Report

1. Introduction

In India, many research and development (R&D) organizations as well as college and university libraries are building Digital Repositories (DR) / Institutional Repositories (IR). The first digital repository was built by Indian Institute of Science (IISc) using EPrints digital

repository software during 2002. Since then many institutions have taken active role in establishing digital repositories by making use of Open Source Digital Repository Software (OSS-DL). A large amount of data is being made available in Internet during the last decade due to ease of use of OSS-DL. While these repositories are available online, one of the challenge which remains with these repositories is, sustainability of these repositories for future retrieval of information for next generations. In this paper, we have looked into sustainability issue mainly and accordingly survey is conducted and results are presented.

2. Sustainability

In the digital library context, "sustainability is the ability to generate or gain access to the resources needed to protect and increase the value of the content or service for those who use it.

2.1 Types of Sustainability

Organizational, economical, technical, collaborative, collections, integral are few major aspects of sustainability in the context of digital libraries. Many organizations initiate projects such as digital repositories. Members who are involved in the development of digital repository play an important role for the development of digital repositories. This issue is addressed as the concept of organizational sustainability. Organizations require financial resources to continue their digital repository activities. This concept is called economic sustainability. Economic sustainability refers to the revenues and investments necessary to support digital libraries. Collections sustainability refers to strategies for ensuring that the information added in digital repositories persists for a long time with quality of information. Digital repositories initiated are mainly based on the backend technologies used. The technology used for building digital repositories is constantly changing hence main aspect in digital repositories is sustainability of digital repositories from technological point of view. This issue is known as technology sustainability. Economic and technological aspects are the key parts in sustainability of digital libraries. Overall, it is not possible to preserve digital information without a sustainable organizational, economic, social, structural, and technical infrastructure, nor is it sensible to preserve material without sustained value.

2.2 Major Issues in Sustainability

While managing digital collections by making use of either proprietary or commercial software tools, it is imperative that the software tool should continue to provide their support over the years. Practically, nothing is sustainable forever. So, instead of asking such as question about sustainability, one should ask how long we can be confident to take care of the data added in the digital repository along with the software and hardware for the future access. Since most digital library projects are long-term efforts, they require commitment of long-term resources. In sustainability, one of the areas is digital preservation where we need to preserve all the software, hardware and operating systems, etc. More money has to be spent on sustaining any of these digital repository projects which requires continuous funding from the respective organization to sustain those repositories over the years. Considering all these aspects, a survey was conducted from Indian digital repository stakeholders.

3. Scope of the study

To identify digital repositories from India and to further study sustainability issues of these repositories DOAR and ROAR registries were used for initial work. As per the DOAR and ROAR repository registry, a large number of repositories are initiated all over the world. For practical purposes, to study the sustainability issues of digital repositories, it was felt that the study may be restricted to repositories developed within geographic span of India and the ones registered either on DOAR or ROAR. Secondly, the study was further restricted to repositories which are using Open Source Software tools. Some of the IRs excluded from the present study were the ones which practice the closed access and were built using the commercial software. While studying the sustainability aspect of Indian IRs, authors also noticed a cluster of CSIR Labs IRs in India.

4. Indian Scenario

During 2002, first digital repository ePrints@IISc was initiated by Indian Institute of Science (IISc) using EPrints open source software. Currently the repository holds 32876 documents. After 2002, many institutions and R&D organizations from India took initiatives in creating digital repositories and as a result, large digital contents are

being made available to end users across the world. These collections mainly include annual reports, technical reports, publications (Preprints /Post prints), digitized documents, talks, etc. In open access repositories, the major parts of the contents are available from Indian stakeholders. As mentioned earlier, list of digital repositories from India is found out from ROAR/DOAR sites. DOAR or OpenDOAR, is an authoritative directory of academic open access repositories. ROAR is a searchable International Registry of Open Access Repositories indexing the creation location and growth of open access institutional repositories across world along with details about the contents. ROAR was created by EPrints at University of Southampton in 2003.

Today, many countries have initiated either IR or DR and uploading large volumes of data. There are more than 2700 open access repositories registered on either DOAR or ROAR. A variety of software like DSpace, EPrints, Greenstone, OPUS, FEDORA, etc. are being used to maintain digital repositories. It is worth mentioning that DSpace and EPrints are the most popular software used across the world for building digital repositories. Table 1 shows the percentage of use of DSpace and EPrints across world and India.

Registering Agency	World Scenario			Indian Scenario		
	IRs	DSpace	EPrints	IRs	DSpace	EPrints
ROAR	2914	1171 (40%)	458 (16%)	91	55 (58%)	28 (30%)
DOAR	2165	882 (41%)	315 (15%)	54	37 (70%)	13(25%)

Table 1. Repository Software used in Indian IR

4.1 Cluster of CSIR Labs IRs

Apart from the above registered DOAR/ROAR, IRs, there is a cluster of Indian IRs initiated by Council of Scientific and Industrial Research (CSIR) India. Established in 1942, CSIR is one of the largest R & D organizations in India with 39 laboratories, with a collective staff of more than 17000. The initiative of CSIR is a welcome step towards the open access movement. The mandate of CSIR research is "all research papers published from all CSIR laboratories and supported by a grant from CSIR will be made open access by depositing the full-text and the metadata of each paper in an institutional repository". Considering this mandate, the CSIR labs have started, cluster of their IRs that is

centrally maintained by CSIR/URDIP. Currently the site contains more than 48,000 records across all CSIR labs.

5. Analysis of the Survey

During last decade, a large number of repositories are being initiated and are registered under DOAR or ROAR. From India, there are about 80 IRs listed on DOAR / ROAR but only 53 IRs are found to be live and accessible after carrying out the survey. This indicates the need for present study of sustainability of these repositories. Prima facie, some of the reasons for non-accessibility of these IRS could be 1) repository address has changed, 2) some over enthusiastic IR managers registered on ROAR/DOAR site, but never updated the details; 3) the IR agencies do not keep track of live or active repositories. Leaving apart some of these issues; a survey was conducted from these 53 live / active Indian IRs to assess the sustainability aspect. The questions framed in the questionnaire were around four aspects of the sustainability, i. e. Economical, Social, Technical and Organizational. While collecting the data, most of the information was directly collected from the IR sites whereas remaining part of the information was collected by sending questionnaire to the IR managers of these repositories. The survey questions were framed on various aspects like software tools used, hardware infrastructure available, manpower available, financial and technical support available, organizational support available, etc.

5.1 Establishment of Repositories

Data was collected to find out the establishment year of each repository from DOAR. DOAR provides date of registration instead of year of establishment. From the DOAR site, it was observed that the first two IRs traced back to 2004 are Librarians Digital Library (LDL) and ePrints@IISc. Majority of the Indian IRs are developed between 2006 and 2011, whereas two Repositories have been added in 2012. Secondly, the cluster of CSIR Labs IRs came into existence from late 2012.

5.2 Type of contents of Repositories

Repositories which are initiated from India cover contents in variety of documents with major input in the form of articles. Other type of contents include in IR include, digitized books, conference

proceedings, learning objects, multimedia, patents, theses, unpublished literature, etc.

5.3 Subject coverage of Repositories

Wide range of subjects covered in Indian IRs that range from subject areas like pure science, applied sciences to social science. The Table 2 below shows number of repositories from different subject areas. It is observed that there is not a single repository that is dedicated to a language and literature. Again, the subject categories are very broad ones which may not reflect exact subject of the IR. Lastly, many IRs show the subject as a multidisciplinary, which is bit misleading.

Subjects Covered	Number
Computers and IT	2
Ecology and Environment	2
Journal Publication	2
Library and Information Science	2
Social Sciences General	3
Mathematics and Statistics	3
Agriculture, Food and Veterinary	4
Biology and Biochemistry	4
Business, Economics, Management	4
Science General	5
Physics, Astronomy and Astrophysics	5
Engineering	5
Health and Medicine	5
Chemistry and Allied Sciences	6
Technology General	6
Multidisciplinary	20

Table 2. Subject Coverage

5.4 IR administrators

The qualification, experience and skills of an IR administrator are important aspects for the successful implementation and maintenance

of IRs. This information too was collected in this survey. Of the total responses, it was found that 60% of the IR administrators are having doctorate degrees whereas 24% hold master's degree and the remaining are either Masters of Philosophy or engineering degree holders. It is also found that 80% of the IR administrators are having an experience between 5 to 12 years, whereas only 20% of the IR administrators have the experience of less than 3 years. An attempt has also been made to find out the quantum of manpower available for these repository maintenance. After the analysis of survey, it was found that 50% of the IRs are manned by single personnel, while 25% of IRs are manned by 2 personnel and the rest are manned by more than 3 personnel. Librarians have always played an important role in the implementation and initialization of the Indian IRs. The survey analysis shows that 79% of the librarians are maintaining their IRs while 13% of the IRs maintained by system administrators. Teachers of library and information science are also playing an important role in maintaining the IRs which is 8%, and an interesting factor to note in the Indian IRs.

While conducting a survey, it was also asked that whether the initialization of IR are part of the organization or of an individual's interest and it was found that 87% of the IRs are an initiation of the individuals and the rest 13 % IRs are the initiatives by an organization. The reason behind this could be many library professionals in India are trying to adopt and learn open source technologies and especially DSpace and EPrints software.

5.5 Funding for Repositories

While analyzing the data, it was observed that there are only three repositories having received funding for initializing the IR project. It was also observed that directly or indirectly, the hardware and manpower is provided by the parent organization. Software required for the IRs is open source that is available freely. Manpower required for IR is generally pooled from the existing library staff strength available in every organization and hence there is no separate manpower recruitment for maintaining IRs. Currently, IRs demand funding for retrospective conversion of data from print format to digital format. A large number of digitization projects are being initiated in India to digitize rare collections available in the libraries.

5.6 Technological Sustainability

5.6.1 Hardware Configuration

While managing IR backend, hardware and software support needs to be with good configuration. After the survey analysis, it was found out that 32% of the IRs use IBM Systems, 32% use Dell systems, 27% use HP systems and 9% use Sun systems. It is commonly known that all these companies are branded companies and provide robust support for hardware requirements and the downtime of all these brands is very rare. This shows that from the hardware point of view these repositories will be sustainable in future.

5.6.2 CPU Configuration

Majority of the IRs use only PIV processors. Higher configuration processors (such as Dual Core, Triple core or Quad Core processors, hexa-core, octa-core, etc.) are yet not used by many of the respondents. To handle large volumes of data in future, it would be necessary that these IRs use higher processor configuration machines.

5.6.3 Storage Capacity

Digital documents added in the Indian IRs contain mainly the textual data. The textual data contains articles and digitized thesis / books / reports etc. The textual data does not need large file storage capacity. It is observed that present IRs use hard disks capacities which range from 20 GB to 4 TB. It will be necessary for the IR managers to upgrade their data storage devices to accommodate large volumes of data may be in the tune of few hundred terabytes. IR users may start uploading data with multimedia formats where each audio / video file size can be in gigabytes. To manage this large volume of data, repositories will need to upgrade their existing storage capacities to accommodate large volume of text, audio/video file formats.

5.6.4 RAID Support

Wikipedia defines RAID as redundant array of independent disks, is a storage technology that combines multiple disk drive components into a logical unit. Data is distributed across the drives in one of the several ways called "RAID levels" depending on what level of redundancy and performance is required. RAID is used to save operating system and data used in live configuration to save disk failures and keep data

accessible without any interruption. RAID increases data reliability as well as increase input/output performance of the system used. Present study indicates that 76% of the IRs support RAID configuration whereas 24% have no idea of what is RAID and how it is used to save data from failure.

5.6.5 File Formats

As mentioned earlier, majority of the contents uploaded by Indian repositories is based on textual data which is saved in PDF format. While analyzing data, IR administrators mentioned that their data is preferably uploaded in PDF form. For audio/video files mp3/mp4 formats is used by 10% of the IRs. Few IRs have used Microsoft Office Documentation format (.doc) as well as TIFF/JPEG file formats for uploading pictures. PDF cannot be used as an archival format, hence long term solutions are needed to keep digital PDF records accessible for a long time length. The PDF/A format have been expressly introduced for the purpose. From the digital preservation point of view PDF/A attempts to achieve the objectives of device independence, self-containment, and self-documentation. It was observed that all the repositories are using PDF as a standard file format for uploading textual data. PDF is not an archival format hence these repositories will have to convert their data into PDF/A format for future sustainability of these files over the network.

5.6.6 OAI-PMH Support

Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) provides an application-independent interoperability framework based on metadata harvesting. Metadata harvesting is a framework for the exchange of metadata in distributed and decentralized electronic information environment. OAI-PMH protocol enables automatic exchange of metadata over the network. The survey result shows that only 50% of the Indian IRs have implemented OAI-PMH protocol and are harvesting metadata from other repositories. It was also found that around 70% IRs support both RSS 1.0 and RSS 2.0. With the growing knowledge of this concept one can hope that in future these IRs will be OAI-PMH/RSS/Atom compliant.

5.6.7 Handle System Support

Digital resources should be identified by unique identifiers on Web. For this purpose some mechanisms have been developed which assigns persistent identifiers to digital objects irrespective of their location, just like ISBN for books / monographs. In the beginning, URLs were popularly used as the identifier to a digital resource but many times URLs don't work and end users get error as "Page not found". To overcome this error in digital libraries persistent identifier schemes were introduced. A persistent identifier is a name for a resource (digital) that will remain the same regardless of where the resource is located. To provide unique identifier numbers to each digital document added in the digital repository, handle.net system provides persistent identification numbers. Handle System is a distributed information system designed to provide an efficient, extensible, and secured global name service for use on networks such as the Internet. The Handle System includes an open protocol, a namespace, and a reference implementation of the protocol. The protocol enables a distributed computer system to store names, or handles, of digital resources and resolve those handles into the information necessary to locate, access, and otherwise make use of the resources. After analyzing the data of Indian repositories, it was found that only 52% respondents have used handle.net system for unique identification of every document that is added in their repositories. It is necessary that these repositories register with handle.net system to provide unique identification number for every document added. One of the reasons for not registering with handle.net could be the registration costs.

5.6.8 Backup of Repositories

A backup or the process of backing up is making copies of data which may be used to restore the original after a data loss event. The primary purpose of backup is to recover data after its loss, be it by data deletion or corruption. As backup plays a vital role in the IRs. A question was asked to find out how backup of data is handled by individual IR administrators. The findings of the same revealed that administrators give more preference to the data backup, because 63% of them take the data backup whereas 33 % of them take full backup i.e. system as well as data backups. Surprisingly, there are 4% administrators who have never taken backup of their IRs at all. Regarding the frequency of the backup, only 50% of the IR administrators have their scheduled backups, i.e. daily/weekly/monthly. Remaining IR administrators take

the backup on ad-hoc basis, i.e. as and when the data is added or updated. One of the important aspects of sustainability of IR is, the backup should be available for future generation. If regular backups of the systems are not available it is difficult to recover data.

5.6.9 Security of Repositories

Security of data uploaded on the IR is an important aspect. While analyzing data it was observed that 90% of Indian IRs run on Linux Operating System and hence none of them have any antivirus installed on their systems. The security adopted by these IRs is mainly based on firewall supported by Linux Operating System except that of Librarians Digital Library (LDL) initiated by DRTC, Bangalore which has its own self-written firewall policies. Out of 54 IRs only 2 IRs run on Windows Operating System and use commercial antivirus software for data security.

5.6.10 Network Connectivity

Network connectivity is one of the important aspects while providing data over network. Many of the digital repositories share bandwidth made available from parent organization. Overall Internet bandwidth available within these organizations ranges from 2 MBPS to 1 GBPS.

6. Conclusion

From the survey analysis, it can be concluded that, sustainability of these repositories from Economic, Social, Technical & Organizational needs improvements. Following are some of the suggestions from our survey analysis. Economical sustainability of an IR depends on the financial support required for the manpower and IT infrastructure. In the present study, it was observed that manpower and IT infrastructure is provided by Institute. Secondly, majority of IRs seek financial support for retrospective conversion of conventional documents.

Social sustainability of an IR depends on the awareness and willingness of the authors for IR input and administrators for the maintenance. From the present study, it was noticed that there is a need to increase awareness of Open Access Repositories. Open Access Repositories need to get an equal importance as that of the commercial publishers when it comes to the evaluation of individuals in terms of their

performance when publishing their research. If authors have published their research in open access repositories it should have given more weightage than commercial publications. At any moment of time, it will always be difficult to say that the IRs are technologically sustainable. This is because of the frequent updates in the hardware and software technologies. To keep an IR technologically sustainable, the administrators have to keep their hardware and software in current configuration and with current released versions from time to time. It was observed that the organizations support development of IRs either directly or indirectly in terms of manpower, technology and financial aids. After sustainability, it is important to study the overall usage of the documents uploaded in these repositories from the usability point of view and actual downloads from each of these repositories. These studies will reveal the worthiness of these repositories and their future needs.

References

ARD Prasad, Nabonita Guha, 'Interoperability and the OAI-PMH', DRTC-HP International Workshop on Building Digital Libraries using DSpace, (Bangalore, 2005), <http://http://drtc.isibang.ac.in/xmlui /handle/1849/245 >

ARD Prasad, Nabonita Guha, 'Persistent identifiers for digital resources', DRTC-HP International Workshop on Building Digital Libraries using DSpace, (Bangalore, 2005) <URL:http://http://drtc.isibang.ac.in/xmlui/handle/1849/246 >

CSIR CentralOpen Archive Repositories Harvester. <URL: http://oa.csirexplorations.com>.

Directory of Open Access Repositories, <URL:http://www.opendoar.org/>

Handle System Overview. <URL: http://http://www.ietf.org/rfc/rfc3650.txt/>

Kevin Bradley, 'Defining digital sustainability', Library Trends, Summer (2007), <http://findarticles.com/p/articles/mi_m1387/is_1_56/ai_n21092805/?tag=content;col1>

Nancy L. Maron, K. Kirby Smith and M. Loy, 'Sustaining digital resources: an on-the-ground view of projects today; Ithaka case studies in sustainability, 40 p. (July 2009), <http://www.ithaka.org /ithaka-s-r/ strategy/ ithaka-case-studies-in-sustainability/report/

SCA_Ithaka_ SustainingDigitalResources_Report.pdf>[1 March 2012].

Registry of Open Access Repositories, <URL:http://roar.eprints.org/ >

Sunita Barve, 'File Formats in Digital Preservation', Proceedings of the International Conference on Semantic Web & Digital Libraries (ICSD-2007), (India, 2007), 239-248 <URL:http://drtc.isibang.ac.in/xmlui/handle/1849/312/>

Tylor O. Walters. 'Digital sustainability: weaving a tapestry of interdependency to advance digital library programs', in Katherine Skinner, ed., Strategies for sustaining digital libraries (Emory University, 2008), 22-40.

Institutional Repository (IR): An Assessment of Trends in India

H. C. Nagarathna[1] and K. S. Chudamani[2]

[1,2]JRD Tata Memorial Library
Indian Institute of Science, Bangalore - 12
{hcn, ksc}@library.iisc.ernet.in

Abstract. Nowadays, the trend in all library and information centers is Institutional Repositories (IR). The increasing amount of digital content produces new difficulties for long-term preservation and accessibility. Many of the institutions have their repositories, which they have built on various open source software. Institutional repository increases the institutions visibility, status and public value. Institutional repository is becoming one of the most popular tools for self-archival and dissemination of an organization's intellectual or scholarly output. IR that supports a broad institutional effort and offers direct and immediate benefits to each institution are important. IR is currently the focus of intense study among libraries supporting scholarly research and higher education. Large academic libraries were the early adopters of institutional repositories. Many digital repositories are being created on a small scale. For example specific special library collections have implemented repositories. The mission of an IR is to be Institutionally defend, scholarly, cumulative and perpetual open and interoperable. IR have the potential to bring significant benefits to educational and research institutions. This study aims to find out whether long-term preservation is part of the mission of institutional repositories and if so, what plans IRs have to provide long-term preservation of their content.

Keywords: Institutional repository, Digital archives, preservation, India, Copyright.

1. Introduction

Institutional repositories are useful for libraries. The digital content may be stored locally, or accessed remotely via computer networks. A digital library is a type of information retrieval system. Digital collections in IR capture and preserve the intellectual output of university communities. Have a potential to serve as tangible indicators of institutional quality and to demonstrate the scientific societal, and economic relevance of its research activities. Many academic libraries are actively involved in building institutional repositories of the