Institutional Repositories for Knowledge Management in Academic and Research Institutions

M.G. Sreekumar, T. Sunitha, P. Sreejaya, R. Biju, Joshy Kuriakose and Prajeeb Kalathil

Center for Development of Digital Libraries (CDDL), Indian Institute of Management Kozhikode (IIMK), Calicut, Kerala, INDIA

{mgsree,sunitha,sreejaya,biju,joshy,prajeeb}@iimk.ac.in

Abstract. Information Retrieval Systems (IRS) provide a simple, Web-based mechanism to researchers to deposit (‘self-archive’) and access their research publications. There are many world renowned open source IR softwares available such as EPrints, DSpace, FEDORA, ARNO, i-TOR, CDSWare etc. Each of this software has a host of features, unique facilities and excellent capabilities, which the users could explore and experiment upon. In this paper we emphasize the need for deploying interoperable open access IRs, and share our experience in creating a state-of-art scholarly Institutional Repository using the DSpace software. The system development and administration illustrated in this paper is on Windows environment as we have observed that a large majority of the library professionals in India are more familiar and comfortable with the Microsoft operating system, though strictly speaking, an open source operating environment for an open source IR software would have been more desirable. The method of developing the IR, which include the institutional open access policies, the installation, configuration and customization of the software along with the related workflow operations such as defining the communities and collections, content development and management, designing and creating standard metadata sets are also illustrated.

1 Introduction
Organizations, especially learning organizations, create a phenomenal amount of scholarly knowledge as part of their academic discourse and research activities. These findings are published through various vehicles of scholarly communication such as journals, conference/symposium proceedings, books, case materials, patents etc. Institutions also produce a great deal of knowledge in the form of ephemeral and unpublished materials such as working papers, technical reports, courseware, classroom presentations (PPTs), lecture notes etc. The world of science and scientific communication are closely knit
entities so much so that one cannot exist without the other. In other words one draws its oxygen, blood and strength from the other. Scholarly communication is a multifaceted and rigorous process involving many stakeholders. Scholarly communication refers to the explicit research findings, formal as well as informal, which the academic and scientific community make known to the world for public consumption. These findings are meticulously brought out research reports, called popularly as ‘articles’ or by the common name ‘papers’, perhaps influenced by the predominant medium used to print them. Authors or contributors of these papers publish them for free unlike the books which form part of the trade. Publishing is the formal system whose key players include researchers, publishers (including scholarly societies), and libraries [CARL]. Researchers submit articles on their findings and publishers add value to these by processing them for peer reviews and necessary editorial corrections. At the other end of the spectrum, libraries subscribe to these publications and facilitate access to primary as well as secondary sources of information for the scientific fraternity. The supply chain of traditional scholarly communication is illustrated in Fig1.

![Figure 1. Scholarly Publication Value Chain](image)

2 The Changing Scenarios in Scholarly Communication
There exists basically three models in scholarly communication which the world has generally accepted, viz., i) the traditional paper based journal publishing process which has a track record of over 300 years, ii) E-publishing on commercial basis iii) the Open Access mode of publishing. The developments in IT and Internet have contributed considerably to shrink the supply chain of information and it is indeed a welcome change. For the information industry, the supply chain extends from the source of information to the point of usage. The traditional journals in paper format took 36-52 weeks for journals to publish. Surprisingly enough, in this total cycle time of 1 year, the value addition (generation, review, correction and printing) takes place in not more than 2-3 weeks, indicating a huge wastage of time and
money contributed mostly by non-value adding links. In this context, the advent of the E-publishing has really been a boost to the scholarly publishing domain, bringing down the publishing time frame to a remarkable 3-4 weeks. A noticeable departure from the traditional systems here is that the printing is delinked from the publishing process. It is important to note here that the much appreciated and respected scholarly value systems are not compromised even by a single degree in relation to quality checks in the new process. In fact it improves and strengthens them with its inherent advantages of being online with respect to processes and procedures. The third category is the growing sets of Open Access Publishing and Scholarly Archive initiatives, which are the offshoots of the novel Open Access movement, catching up globally. Authors are now able to publish their findings at an astoundingly fast pace such as 10-15 minutes or even at a lesser time. The relative features and merits of the three systems are illustrated in the following figure (Fig. 2).

Figure 2. Upcoming Models of Scholarly Communication Value Chain
3 Background
The focus of this paper is on Open Access Archives (OAA), primarily engaged in the promotion, propulsion and instant dissemination of the research findings, intellectual discourse and scholarship in an academic setting. Hence the thrust is on Institutional Repositories and specifically, in the creation, configuration and customization of an IR using DSpace software on a Microsoft platform. Documentation on the installation, configuration and customization of DSpace on Windows is available at DSpace Wiki [DSpace Wiki].

4 Open Access Movement towards Scholarly Content
OA springs from the interesting discipline of Critical Information Studies (CIS). CIS and OA are emerging scholarly fields that sit at the intersection of many important areas of study. CIS considers the ways in which culture and information are regulated by their relationship to commerce, creativity, and other human affairs. CIS promises to let scholars contribute to the greater good while ensuring effective peer review and distribution of work. By all indications, CIS has succeeded in changing the terms of the conversation about issues such as copyright, cultural policy, and the relationships among democracy, culture, and technology. More practically, CIS has helped generate the ‘open content’ and ‘open journals’ movement, which allows authors and artists to retain more control over the ways that publishers exploit their work and enables authors to ‘lock content open’ [Siva, 2006]. The ambitious Open Archives Initiative (OAI) movement has two major arms – the Golden route of ‘Open Access publishing’ and the Green route of ‘Open Access Archives’.

Open Access or the popular abbreviation OA refers to the innovative and the most ambitious concept of universal free access to scholarly information [Budapest Open Access Initiative (BOAI)]. The purpose of OA is to release the scholarly content, through various novel interventions and in phases, from the clutches of the monopolizing commercial publishers and make it available free to the scientific fraternity worldwide. The rationale is that the scientific community publishes their research findings without any expectation of payment. The tangible benefits derived out of the publications are career prospects, recognition and visibility. The findings are given free for consumption by the scientific world. In other words, OA is a rephrasing of the historic dream concept- the universal availability of publications that the library and the academic fraternity have been pondering over for the past several decades.

A pressing problem which had led to this coordinated and concerted movement worldwide, was the exorbitant cost of scholarly journals which have been sky rocketing, and the shrinking library budgets, which had rendered libraries and the academic community to resort to alternative models
of access to their own content. The paradox was that the universities and research institutes who contribute to the scholarly content through their researchers had apparently no access to the content when it went into the hands of the publishers. In the name of the ‘copyright’ which they take away from the authors at the time of submission of the manuscript, publishers have literally locked up the content for good.

The OA process is all about assuring that the scholarly content is made available online. This presupposes that the content, especially scholarly content, is to be born in digital format and this is the prevailing mode of scholarly content authoring. In other words, Open Access is technology’s answer to the scholarly journals crisis the world faces today.

The Budapest Open Access Initiative (BOAI) launched by the Open Society Institute (OSI) gives a candid explanatory note on OA [OSI]. It says “Open Access is the ultimate convergence of an age old tradition and the new technology, to make possible an unprecedented public good. The old tradition is the willingness of scientists and scholars to publish the fruits of their research in scholarly journals without payment, for the sake of inquiry and knowledge. The new technology is the internet. The public good they make possible is the world-wide electronic distribution of the peer-reviewed journal literature and completely free and unrestricted access to it by all scientists, scholars, teachers, students, and other curious minds. Removing access barriers to this literature will accelerate research, enrich education, share the learning of the rich with the poor and the poor with the rich, make this literature as useful as it can be, and lay the foundation for uniting humanity in a common intellectual conversation and quest for knowledge”.

BOAI, by ‘open access’ to literature, means its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself.

The ultimate goal of the Open Access movement is open access to peer-reviewed scholarly journal literature. Presently BOAI suggests two ways to attain this goal:

1. Open Access Publishing (OAP), famous as ‘golden’ road to Open Access. In this model, as against the tradition, journal/s are published as open access publications and they provide instant online open access to articles upon publication.

2. Open Access Archiving (OAA) or open access self-archiving, also called the ‘green’ road to OA. In this model, as a new trend, authors
submit their pre-peer reviewed versions of papers (pre-prints) or post peer-reviewed published article/s (post-prints) into the archive for open access.

Open-access (OA) literature is digital, online, free of charge, and free of most copyright and licensing restrictions [Suber]. It is all about democratizing the scientific intellectual capital, which often draws its energy from publicly or privately funded research. It is not against proprietary or peer reviewed scholarly journals.

5 Open Access Archives (OAA) and Institutional Repositories (IR)

An archive is a generally accepted synonym for a repository. A repository is a network accessible server that holds scholarly digital content or eprints. Scholarly Archives or Institutional Repositories are established medium to communicate peer reviewed (post-prints) and non-peer reviewed scholarly literature (pre-prints). There are basically three types of scholarly archives in vogue, viz., author archives, institutional archives and subject archives. Subject archives are also called as central archives. According to Stevan Harnad, open archiving is just self-archiving the articles the author has published in (peer-reviewed) non-OA journals. Hence it neither bypasses nor replaces peer-review. It has nothing to do with changing the peer reviewing process. Self-archiving is a way of supplementing non-OA journal access with an OA version for those would-be users whose institutions cannot afford the non-OA journal.

There are numerous advantages that OA boasts while they campaign worldwide. Authors as well as Institutions can derive a number of benefits out of Archives. For authors, instant dissemination of the fruits of their long years of rigorous research to a global audience is the first and foremost. OA papers get increased visibility through novel models of harvesting done by search engines such as the Google, Citeseer etc. and the interoperability among similar archives achieved through the Open Archives Initiative (OAI) Protocol for Metadata Harvesting (PMH) are unparallel value additions to OA Archives. While more visibility leads to more citations, one’s research impact naturally gets scaled up. Authors are therefore attracted to come to OA Archives. Additional benefits to self archiving include the assurance of the long term preservation of their articles and the facility to have a proper control as well as meticulous monitoring of one’s own publications.

For institutions too, a long list of advantages and benefits invite them to OA. Firstly, the institute’s archive, popularly known as Institutional Repository helps in pooling the organization’s Intellectual Capital, which is otherwise scattered, distributed and unnoticed, in one central place. The archive therefore serves as a one-stop-source or a single access point for the research output of an institution. It provides ample scope for introspection as to
whether the institute is going in the right direction in its research activities. Necessary strategies and meticulously designed action plans could be charted out based on the feedback. Institutional repositories facilitate instant generation of research reports and thereby save a valuable amount of time. Most importantly the archives ensures long term preservation of its scholarly materials with the help of Open Source softwares and Open Standards of data models and data structures.

6 Open Archive Directories and Search Engines
There are many value added services which index OA archives spread globally, as well as harvest metadata records for search and retrieval. OpenDOAR, the Directory of Open Access Repositories lists 502 OA archives situated worldwide [OpenDOAR]. OpenDOAR is a joint effort led by the Open Society Institute (OSI), along with the Joint Information Systems Committee (JISC), the Consortium of Research Libraries (CURL) and SPARCEurope [JISC][CURL]. DMOZ, the largest open directory of the Web, lists 59 free access online archives [DMOZ]. The Registry of Open Access Repositories (ROAR) hosted by Eprints.Org lists 607 plus open access archives [ROAR]. OAister, one of world’s outstanding OA repository registry services offered by the University of Michigan, indexes over 663 OAI-compliant open repositories worldwide with an overwhelming 8,593,164 records [OAister]. Arc, developed by the Old Dominion University, is among the early federated search services based on OAI-PMH protocol [Arc].

7 Institutional Repository (IR) Softwares
There are many world renowned free open source Institutional Repository (IR) softwares available such as EPrints, DSpace, FEDORA, ARNO, i-TOR, CDSware etc. They are issued either under GNU public license or the BSD license and can be downloaded from their own sites or open source software directories such as SourceForge [Sourceforge]. Each of the software has a host of features, unique facilities and excellent capabilities, which the users could explore and experiment.

7.1 DSpace
DSpace is a digital asset management software jointly developed by Hewlett-Packard and MIT Libraries, and it is arguably one of the appreciated open source software deployed worldwide for building digital institutional repositories that captures, stores, indexes, preserves, and redistributes content in digital formats. DSpace facilitates the institutions and universities operate an open access and interoperable institutional repository at the local level. It is also intended to serve as a repository back up for future development to address long term preservation and remote/online access issues. The system was launched during late 2002 as a live service hosted by MIT Libraries, and the source code made publicly available according to the terms of the BSD
open source license, with the intention of encouraging the formation of an open source community around DSpace [DSpace Wiki].

7.1.1 Features & Functions
DSpace is 100% open source software and is freely available for download from the open source software directory SourceForge (http://sourceforge.net). The software has been built on a strong architecture supported by state-of-art digital library technologies and embracing almost all latest trends in information sciences. It provides the users, especially the librarians and system administrators, every freedom for building, managing, customizing, administering and Internet publishing world class institutional repositories and digital libraries. Its major features include the ability to accept all forms of digital materials including text, images, video, and audio files. Possible content includes scholarly articles and preprints, technical reports, working papers, conference papers, books, e-theses, multimedia publications, Datasets: statistical, geospatial, matlab, etc. Images: visual, scientific, etc.; audio files, video files, learning objects, bibliographic datasets, reformatted digital library, collections, Web pages etc.

For enhancing the resource discovery features, DSpace supports Dublin Core metadata unqualified element set as well as provisions for the qualified Dublin Core metadata registry. The software allows the communities and users to publish their articles remotely on the archives. It has CNRI ‘Handles’ support for Persistent URLs (PURL) which assigns and resolves persistent identifiers for each digital item. Interoperability is another salient feature of DSpace, and it supports the Open Archives Initiative’s Protocol for metadata harvesting (OAI-PMH) V2.0 as a data provider. OAI support was implemented using OCLC’s OAICat open-source software to make DSpace item records available for harvesting. DSpace uses the versatile Lucene search engine for full text searches. Lucene search engine is a part of Apache Jakarta project, and brings along laudable search features like ‘fielded’, ‘boolean’, ‘exact term’, ‘proximity’, ‘wild cards’, ‘fuzzy’, ‘range’, ‘boosting terms’ etc. DSpace supports unlimited exporting/importing of digital content, along with its metadata in a simple XML-encoded file format. The database management system used is PostgreSQL which supports transactions between Oracle as well as MySQL. DSpace enjoys international acceptance across the world and it provides a customizable Web interface. The workflow process for content submission, the decentralized submission process, the remote publishing facility are regarded as the unique features of DSpace. Most importantly it is Open URL compliant also [DSpace Wiki].
7.1.2  **DSpace Installation**
Mainly six prerequisite softwares are essential for running a DSpace server, viz., i). Java SDK, ii). Apache, iii). Tomcat, iv). Apache Ant v). PostgreSQL and vi). the DSpace software itself. These softwares are to be installed in sequence also. It is very important to setup corresponding HOME variables and modify the PATH variables after the installation of Java SDK and Apache Ant respectively. After the installation of PostgreSQL, we need to create a database named ‘dspace’ owned by user ‘dspace’ with UNICODE encoding. To load DSpace you have to start the three services, namely, Apache, Apache Tomcat and the PostgreSQL Database server.

7.1.3  **DSpace Configuration**
Primarily configuration of DSpace Software is done by editing the file ‘dspace.cfg’ located at ‘dspace\config’, which contains basic information about a DSpace installation, including system path information, network host information, SMTP mail server address and other things like site name etc. We configured ‘dspace.url’ line in the ‘dspace.cfg’ file with the desired DSpace site address, i.e., the URL ‘dspace.iimk.ac.in’.

For mailing purposes we need to modify the ‘mail.server’ configuration item in ‘dspace.cfg’ file on a case to case basis. There could be two instances here – i). the Dspace server itself has got a mail server configured and running (say, sendmail) or ii). the mail server is running elsewhere. You need to furnish the SMTP mail server address as per the situation, in the ‘dspace.cfg’ file, located at c:\dspace\config\. If the Institute has a separate mail server, this server will have to relay the DSpace server IP.

7.1.4  **Communities and Collections**
In DSpace, a digital repository is organized in terms of communities, sub-communities and collections. In other words, communities, sub-communities and collections can be arranged hierarchically. In our repository we followed the subject approach for creating communities and publication type classification for creating collections within each community.

7.1.5  **Collection Building**
Adding content to DSpace is quite easy and straight forward. As DSpace has a workflow based remote publishing facility, authorized users can submit their items from their own client machines. You will need to be logged in to DSpace before you can submit. Most collections will also require specific authorization for you before you can submit items to it.

7.1.6  **Workflow**
After starting a submission, you will be led through a seven-step workflow process. These include some basic metadata descriptions about the materials first, then several screens where you describe the details, then file uploads, a
verification screen, a license granting screen and finally a submission complete screen. The following figure (Fig. 3) shows the workflow process of item submission in DSpace.

![Submit: Submission Complete!](image)

**Figure 3. DSpace Collection Submission Workflow**

After submission is complete, the submitted item will go through some formalities like review, edit, or approve according to the collection's policies. This means that the submitted item might not go directly into the main archive, before the validation process.

### 7.1.7 DSpace Customization

DSpace is implemented by using Java Servlets and JavaServer Pages which produce the HTML pages for DSpace. As JSP coding is similar to HTML, changing the look and feel of DSpace page is very easy. You can make your own header image by replacing the existing ‘dspace-blue.gif’ located at `\tomcat5\webapps\dspace\image\` with customized image. We have edited the news-top.html file located at `\dspace\config\` for giving an introductory note about the repository in the DSpace home page. By editing the news-side.html located at `\dspace\config\`, we added information about our library related programmes/seminars/workshops etc. We also configured the item count against communities and collections by setting the ‘webui.strengths.show’ configuration item’s value to ‘true’ in the ‘dspace.cfg’ file. DSpace uses a ‘styles.css.jsp’ file which is located at `\Tomcat5\webapps\dspace\`, which we modified for altering font type, size and colours from default style. The home page of the IIM Kozhikode’s DSpace repository is shown in Figure 4.
7.1.8 DSpace Administration

In DSpace administration the administrator has to do a wide range of tasks for the successful maintenance of a digital repository. When we first configure a digital repository using DSpace, we begin with creating Communities and Collections. Policy setting is the most important step in the administration of a digital repository. DSpace administration provides a number of archive control features. We can literally control the access, usage and preferences of each and every collection as well as user(s) or community through this versatile tool. We have to take a decision with regard to whom or which group can submit digital items to each collection. In addition, we also have to take a decision as to who or which groups of members (E-people) are authorized to review, approve and modify metadata while submission / collection building.

7.1.9 E-People

Collection items can be accessed by everyone (anonymous group), but users must be authenticated to perform functions such as submission, email alert or administration. DSpace calls its registered members as e-people. DSpace holds the details about each e-person such as their E-mail address, first and
last names etc. E-people can be members of 'groups' to make administrator's tasks easier when manipulating authorization policies.

7.2 Publishing the Archive on the Internet

This is one of the simplest yet most interesting step in the development of your IR. You need to identify a suitable domain name, configure the same, and give the value at 'dspace.url' in the ‘dspace.cfg’ file, say ‘dspace.url = https://dspace.iimk.ac.in:8080/’. You will then need to register the domain with a public IP. By default, DSpace is running on 8080 port, i.e anybody accessing 'dspace.mydomain' will have to give http://dspace.mydomain:8080. If you want to avoid 8080, you have to configure port forwarding 8080 to 80 (the default port) using a firewall. We can now access the archive as http://dspace.mydomain.

7.3 DSpace Lifeline

DSpace’s mailing lists (eLists) are very useful and powerful, and there are three active listservs maintained by DSpace which shares and clarifies user experiences and stories dealing with real life DSpace situations. The ‘dspace-general’ list is reachable through ‘dspace-general@mit.edu’, or subscribed to by visiting ‘http://mailman.mit.edu/ mailman/listinfo/dspace-general’ and following the instructions to subscribe. Beginners are advised to first check the ‘FAQ’ (http://wiki.dspace.org/EndUserFaq) or the ‘archive’ (http://mailman.mit.edu/ pipermail/dspace-general/) to see if the question in hand has been answered before. For systems professionals and developers there are two more Lists, viz., the ‘DSpace-Tech’ (technology discussion list) and the ‘DSpace-Devel’ (developers’ list), and both of them are very informative and supportive for posting questions or contribute one’s expertise to other developers working with DSpace, and to share ideas and discuss code changes to the open source platform. To subscribe to these, you may visit ‘http://lists.sourceforge.net/lists/listinfo/dspace-tech’ and ‘http://lists.sourceforge.net/lists/listinfo/dspace-devel’ respectively. Also, there are separate archives for these two lists for getting access to the old postings.

From India, the eList managed by DRTC is very active and useful. You can subscribe to the List via the Web at http://drtc.isibang.ac.in/mailman/listinfo/dlrg or by sending a message with subject or body 'help' to dlrg-request@drtc.isibang.ac.in.

DSpace Wiki (http://wiki.dspace.org/) is also a very useful reference tool which is worth looking at before we send out messages to eLists, as there could be possible solutions already provided in the Wiki.
8 Institutional Policy on Open Access
Setting up an institutional repository is not a big deal now a days. But arriving at a suitable and feasible open access policy at the institutional level is a Herculean task and this need the active participation of the information professionals and the scholarly community of the institution. We need to do a bit of scouting, and if necessary, little lobbying also towards this. A reasonable amount of guidance on this and also on submission guidelines, author benefits, copyright issues etc. could be well seen at the IR at the Indian Institute of Science (IISc) set up by NCSI (http://eprints.iisc.ernet.in).

9 Conclusion
This paper has touched only the tip of the iceberg on OA as well as that of DSpace. ‘Open Access’, as rightly told, is an unprecedented ‘public good’ made possible with the convergence of the good old academic and scientific tradition of sharing scholarly knowledge and the great potential of latest technologies. Ensuring free and unrestricted peer-reviewed scholarly literature accessible worldwide shall definitely accelerate research and enrich education and thereby bridge the widening digital divide between the developed and the underdeveloped nations. It is heartening to note the enormous amount of initiatives that have taken place already and further taking place world over, towards this innovative and novel goal. The selfless contributions pumped in by the open source software (OSS) front, the compliments from the Free Software Foundation (FSF) and the generous funding received for the cause of science are absolutely praiseworthy [OSS],[FSF].

Both Open Access Publishing (OAP) and Open Access Archiving (OAA) are central and crucial to India as far as its academic and scientific agenda for the coming years are concerned. Thanks to the selfless efforts of a few OA experts, there are laudable OA efforts and initiatives on training and workshops. Institutions such as DRTC, INFLIBNET, NCSI, INDEST, MSSRF, UNESCO, DSIR, TERI, IIMK and many others are imparting and organizing workshops and training programmes on OA softwares in India. However, given the magnitude of India’s wealth of scholarly knowledge base, IT expertise and the strength of the academic/scientific community, it is strongly felt that they are not just good enough unless the policy makers' whole hearted support is ensured for the successful implementation of repositories in the respective institutes from where the professionals were rigorously trained. Similarly the country's major journal publishing organizations should come up with open access models of publishing. CSIR, ICAR, ICMR, ICSSR, DRDO are just a few examples. What is more important is the involvement of the University system in the Open Access movement, which is unfortunately not making much headway. UGC should take the lead and involve Institutions like the INFLIBNET in capturing the invaluable intellectual capital being untapped in our universities and put them
together into Open Access Archives. Similarly the national level research councils, viz., CSIR, ICAR, ICMR, DRDO, ICSSR etc. could consider collecting and archiving their research papers into a central archive which could be accessed by all, while individual institutional repositories which are interoperable, could be attempted simultaneously. Ultimately, it is not the progress of technology, but lack of long range vision and proactive policies which pull us back from success in most cases.

References

[8] JISC. http://www.jisc.ac.uk/
[9] CURL. http://www.curl.ac.uk/
[10] DMOZ. http://dmoz.org/Science/Publications/Archives/Free_Access_Online_Archives/
[18] OSS. http://www.opensource.org