

Building the Open Access Self-Archiving repository for the Bio-Medical Sciences at National Informatics Centre.

Mr. Sukhdev Singh and Ms. Naina Pandita.

Bibliographic Informatics Division,
(Indian Medlars Centre)
National Informatics Centre,
A - Block, CGO Complex, Lodhi Road,
New Delhi-110003. (India).

{sukhi, naina}@nic.in
Phone: 91-11-24362359

Key Words: Open Access; Self-Archiving Repository; Eprints; OpenMED@NIC; Bio-Medical Sciences.

Abstract

Self-Archiving is an important model of the Open Access movement. National Informatics Centre has been providing various services and products to the biomedical community. Building up a Self-archiving repository for Bio-medical and Allied sciences was a natural extension of these activities. To make this repository interoperable with other such repositories Open Access Initiative Protocol for Metadata Harvesting (OAI-PMH) was adopted. The selection of suitable software for the archive was done from OAI-PMH compatible softwares. GNU EPrints was finally selected. A prototype was build for planning of activities, demonstration and checking security aspects. To provide subject-wise browse view to the archive a MeSH based categorization was adopted. A dedicated server was procured and installed in the NIC Network Domain under RedHat Advanced Server Version 3.0. EPrints software was then installed and customized. Making scientists and authors aware of Open Access and its benefits remains a major challenge for any such attempt. However efforts are bearing fruits in the form of Open Self-archiving repository for Bio-medical and Allied Sciences i.e. OpenMED@NIC [<http://openmed.nic.in>].

Introduction:

Creation of scientific knowledge is a systematic process of building upon previous work of others - brick by brick. It a community work and scientific communication plays a vital role in this process. Scholarly journals are the major carrier of scientific communication. Scientific papers published in these journals represent the information that flow within scientific community. Thus scholarly journals are the main communication channel of any scientific community. This communication channel by virtue of its peer review process ensures quality of information and the body of knowledge thus generated. Obviously for the sake of knowledge generation and betterment of human race it is the responsibility of the scientific community to sustain this scientific communication channel. Any barriers in this channel are detrimental to the progress of human race.

Scientists do not write research articles for monetary benefits. They do it to propagate their ideas and results of research work in which they are active. They rather seek appreciation, acknowledge and citations of their work. They desire maximum impact of their work as reflected by citations to them. They are rewarded for their contribution indirectly in form of their carrier advancement. Propagation of ideas is their prime objective for publishing their research articles and other research documents. Scholarly journals are the major source publication, distribution and archival of authors ideas and research results. These attribute research to their authors, provides quality control mechanism through pear-review process. However the traditional business models of these journals introduces certain barriers for wide spread access especially for scientists from developing countries. These are basically of two types: "**price barriers**" and "**permission barriers**". Price barriers are basically high subscription rates for core journals. Libraries, especially from the developing countries, fail to subscribe to journals due to shortage of funds. Online journals are no respite as they may also require licensing fee or pay-per-view fee to be paid to access the full text of articles. Further, journals operating on traditional business models require authors to transfer copyrights in their favour. This practice creates "permission barriers". The subscribers as well as the authors themselves are not permitted to distribute copies of the published papers. The paradox here is that both the consumer and the producer is the scientific community. Yet it has to pay for communication of research results within the community.

Open Access:

Open Access holds promise to remove both price and permission barriers to the scientific communication by using Internet (Suber, Peter. 2004). In fact, 'Open access' (OA) is a step ahead of "Free Access" which removes just the price barriers by providing free access to end users. Open Access removes the permission barrier as well. In other words, under Open Access, the end-user not only has free access to the content but also have the right to further distribute the content. It only requires proper author attribution. Open Access is manifested in two forms – OA Publishing and OA Self-Archiving. OA Publishing is just like any other journal publishing. Like traditional publishing, it involves peer reviewing of submitted articles from authors and publishing. Published content is freely accessible over Internet and the users have right to download, use and further distribute it with proper attribution. The business model is however different here. While in traditional publishing model, it is the "end-user" that pays to access the paper. Here the "author pays" for publication of his/her accepted paper. On the other hand, the OA Self-Archiving model is liberal on peer review. It simply provides persistent digital repository where authors / owners of the content may archive their documents (Pre-referred or post-referred). These repositories normally allow other systems to harvest the metadata associated with the documents of the archive. The exchange of such metadata is in accordance to now well-established "Open Archives Initiative – Protocol for Metadata Harvesting (OAI-PMH)" protocol (Open Archives Initiative. 2002). Open Access Self-Archiving practice has been there as earlier as 1991 within high-energy physics community with help of an archive that is now known as "arXiv".

Open Access Repository at NIC:

National Informatics Centre (NIC) has been a pioneer in communicating medical research information through the use of Information Technology. It was in 1988 when NIC and Indian Council of Medical Research (ICMR) came together to set up a centre for biomedical information. This center is now emerging as a content creator and aggregator (Singh, Sukhdev; Gaba, Surinder Kumar and Pandita, Naina. 2004). It has developed three major products that are available over Internet - i. UNcat (<http://uncat.nic.in>) - union catalogue of journal holdings of medical libraries of India; ii. IndMED (<http://indmed.nic.in>) - A bibliographic database of Indian biomedical journals and iii. medIND (<http://medind.nic.in>) - full texts of Indian biomedical journals being indexed in IndMED and PubMed. NIC has been providing free access to these databases. Open Access model for communication of biomedical research information thus becomes the natural extension of existing activities.

Today there are about 466 [<http://archives.eprints.org/>] known open access repositories around the world. Out of these, there are very few discipline based self-archiving repositories in the area of bio-medical and allied sciences. Those existing do not provide in-depth subject classification (Bioline International, 2005) for the vast subject area. Thus, a repository was planned, developed and deployed at NIC. It was named as [OpenMED@NIC](#) and made available through website <http://openmed.nic.in> in May 2005. It is a discipline based International Archive. It accepts both published and unpublished documents having relevance to research. Its scope includes Medical and Allied Sciences including Bio-Medical, Medical Informatics, Dental, Nursing and Pharmaceutical Sciences. It accepts various types of document formats like preprints (pre-refereed journal paper), post-prints (refereed journal paper), conference papers, conference posters, presentations, technical reports/departmental working papers, theses etc. It also accepts non-English documents along with metadata in English.

Key Design Considerations:

For working on a suitable design of [OpenMED@NIC](#) archive following considerations were kept in mind:

Interoperability: Authors are interested in maximum exposure to their works. Maximum exposure can only be achieved with some mechanism to expose the metadata worldwide. An interoperable protocol needs to be followed for such a mechanism. Fortunately a program called “Open Archives Initiative” develops and promotes an interoperability standard that aim to facilitate the efficient dissemination of metadata about contents in repositories. This protocol is known as Open Access Initiative Protocol for Metadata Harvesting or simply - OAI-PMH (Open Archives Initiative. 2002). Under this model, metadata is harvested (extracted) from Data Providers (Repositories) by Service Providers (Search Engines). The metadata is exposed by Data Providers as “Tagged Fields” in XML in response to “Service Provider’s” query in accordance to OAI-PMH protocol. The “Service Providers” merge data from various “Data Providers” and store in a back-end DBMS. User queries are run against this aggregated data. OAI-PMH has become almost a standard for Open Archive repositories. [OpenMED@NIC](#) has adopted OAI-PMH Version 2 to expose the Metadata.

Subject Browsing: Categorizing documents in a directory-type structure has always been user friendly. It is easy to navigate especially if user is not looking for a particular document. It also gives context to terms used as keywords. For implementing classification in [OpenMED@NIC](#) archive, the choice was between inventing a new classification scheme and using an

existing one. Inventing a new local classification scheme is time consuming, subjective and expensive. Even with best subject knowledge it is likely to overlook something in the logical subject tree structure, which may be difficult to fit in latter on. Using an existing scheme does not require any such efforts and is interoperable with other systems. Medical Subject Headings [MeSH®] of US National Library of Medicine was a natural choice as it is well suited for Biomedical domain. Moreover using MeSH should make the archive interoperable with databases like PubMed and IndMED in regard subject searching.

Trust of Users: Self-archiving repositories are based on premise that authors will deposit their research papers. In an institutional repository, an institutional regulatory policy ensures that institutional output is archived in the repository. However a cross-institutional or discipline based repository does not enjoy such a regulatory policy. Authors deposit their documents in such repositories on voluntary basis. They expect repositories to expose their works to their peers. Repositories must win trust of the depositors by providing persistent URLs associated with their works. Also their works should be retrievable at any time from anywhere. Reliability of links encourages other authors as well to cite works along with associated URLs. Trust of authors and end users on a repository is linked to its ability to attract documents. [OpenMED@NIC](#), being a discipline and cross-institutional repository, has to be reliable and persistent in order to win the trust of depositors and end-users.

Application Software: Open Source license and OAI-PMH compliance were the prime criteria for selecting application software for [OpenMED@NIC](#). There are number of such software that could have been used – Archimede, ARNO, CDSware, DSpace, Eprints, Fedora i-Tor, MyCoRe, and OPUS (Open Society Institute. 2004). However in terms of installation base, EPrints and DSpace were the forerunners among these. While EPrints is based on with time tested mature technologies, DSpace offers feature like persistent “handles”. Choosing one of these was really a difficult task. After carefully studying pros and cons, EPrints was selected. It is comparatively mature and easier to install, customize and maintain.

User Interface: [OpenMED@NIC](#) is meant to be self-archiving in nature. Much of the required efforts are to be performed by submitters. The archive staff requires minimum efforts. This requires intuitive and user-friendly web interface.

Prototype Development:

A prototype of the [OpenMED@NIC](#) archive was developed before the final system. While building the prototype, the above-mentioned considerations were kept in mind. An old unused P-II was selected for developing the prototype. It was formatted and RedHat Linux 9 was installed. However, P-II offered limited hardware resources for smooth running of the system. It was reformatted and downgraded to RedHat Version 7.3. Other major softwares installed were Apache 1.3.31, Mod_Perl 1.25 and MySQL 3.23.49. Finally EPrints 2.3.4 was installed after number of attempts of matching and installing correct versions of various required modules.

In EPrints subject categorization scheme can be provided at the time of installation. This can be done with the help of "import_subjects" command. It uses a colon separated ASCII text file "**subjects**" for this purpose. Each line of this file is in the following format:

subjectid:name:parents:deposable

Where -

subjectid is a unique ID for this subject.

name is the name of this subject.

parents is a comma separated list of the parents of this subject.

deposable is a boolean value indicating if this may have documents under it.

Development of subject categorization based on MeSH was the major task in the prototype development. MeSH has around 23,000 terms. Building such a huge categorization was not justified to start with as the repository was expected to grow at a much slower phase. A representative broader categorization was decided to start with. Which can be enhanced later on by the administrator. Even for such representative broader categorization scheme, manually creating EPrints "**subjects**" file was difficult. So, a PERL script was written to extract a representative scheme from computer readable MeSH file. The script created the "**subjects**" file based on statistical sampling of subject tree depths. The "**subjects**" file thus generated was imported in the prototype.

Once the prototype was ready, it was used to demonstrate the core features and functions of the proposed archive. Security aspects of the system were also checked. Experience with creating the prototype provided insight in determining the exact technical requirements. It also provided insight in planning for activities required for final system.

Deployment of OpenMED@NIC:

Once, the prototype was in place it was tested with depositing dummy documents. Once satisfied, the development of final repository system was initiated. Changes were made in the look and feel of the web based user interface of the system. Additional document types like PPT and PPS were added to allow depositing of "Power Point" slides.

EPrints software takes total control of "Apache" and runs with a privileged mode. This requirement introduces conflicts if it shares web server with other applications. It was therefore decided to deploy it on a separate independent server. This also reduced the security risks associated with allowing it to write on file system. A **Rack Mount Server** – RS2 (1 U) with 4 GB RAM and dual processors was procured. This was loaded with RedHat Advanced Server (AS 3). For sake of taking periodic back-ups it was deployed under "Storage Area Network (SAN)" in NIC. Some of the EPrints installation procedures used in prototype were required to be changed due the change in version of webserver from apache 1.3 to 2.0. SMTP gateway for sending emails was setup that was not done in prototype. Cron procedures were set up to perform routine functions automatically. Relevant DNS entry and firewall rules were added in NIC Network to make it available through Internet.

Populating OpenMED@NIC:

Deploying and maintaining a repository is challenging task. However, populating the repository is even more challenging. It takes lot of efforts in making the content owners and authors aware of Open Access and virtues of self-archiving. They are reluctant to deposit their documents in Open Access repositories (Westrienen, Gerard van and Lynch, Clifford A. 2005). The prime reasons for this could be following:

- Confusion, Uncertainty and Fear on Copyright Issues.
- Doubts regarding how the material would be used and by whom.
- Doubts on getting proper attribution, impact and scholarly credit.
- Myth of low quality material in institutional repositories.
- Unfriendly submission procedures.
- Lack of mandatory provisions to deposit.
- Lack of Internet Access facilities.

Open Access Movement is striving to overcome these barriers. A series of attempts were made at NIC to spread awareness about Open Access among the bio-medical community. These included writing letters and emails to scientists working in all ICMR Institutions and other eminent scientists. Number of emails was sent to various discussion groups. Open Access topic was introduced to participants of various training programmes. Online

tutorials (Naina, Pandita and Singh, Sukhdev. 2005) were prepared and archived in the [OpenMED@NIC](#) itself. These efforts has resulted in more than 335 registered users of [OpenMED@NIC](#) with around 660 documents since its inception in May 2005.

Role of Librarians:

We expect all Indian librarians of medical and allied sciences institutes / organizations to play an important role in populating the [OpenMED@NIC](#) repository. They can conduct training / seminars in their institutes for awareness about [OpenMED@NIC](#). They can train and help their library users in self-archiving. To start with they can also offer doing "proxy" self-archiving, on behalf of their institutional authors (Budapest Open Access Initiative. 2005). They can build their institutional repository by registering with an institutional account in [OpenMED@NIC](#). If required, NIC can train librarians wanting to play an active role in submissions to the archive.

Conclusion:

Building up a self-archiving repository is a challenging task. It requires meticulous planning of various processes and resources. It requires dedicated hardware, software and competent human resources. Internet connectivity should be of high bandwidth and available around the clock. Once the repository is established, winning trust of content owners and end users for populating repository is another major challenge. It requires spreading awareness among scientific community about various benefits of open access self-archiving. Collaborative efforts are needed to promote and populate [OpenMED@NIC](#). Librarians should take a lead and collaborate in this effort.

References:

Bioline International (2005). Bioline Eprints Archive. <http://bioline.utsc.utoronto.ca/>

Budapest Open Access Initiative (2005). Self-Archiving FAQ. <http://www.eprints.org/openaccess/self-faq/#libraries-do>

Kwasik, Hanna (2005) Open Access and Scholarly Communication -- A Selection of Key Web Sites. Issues in Science and Technology Librarianship. Summer 2005. <http://www.istl.org/05-summer/internet.html>

Open Archives Initiative (2002). The Open Archives Initiative Protocol for Metadata Harvesting. <http://www.openarchives.org/OAI/2.0/openarchivesprotocol.htm>

Open Society Institute (2004). A Guide to Institutional Repository Software v 3.0, 3rd Edition. New York, Open Society Institute.
<http://www.soros.org/openaccess/software/>

Pandita, Naina and Singh, Sukhdev (2005). OpenMED self help tutorial. IMC.
<http://openmed.nic.in/135/>

Singh, Sukhdev; Gaba, Surinder Kumar and Pandita, Naina (2004). Architecture and building of medical digital library at National Informatics Centre: what exists and what is required for MeDLib@NIC? In: International Conference on Digital Libraries, 24-27 Feb 2004, New Delhi. <http://openmed.nic.in/12/>

Suber, Peter (2004). Open Access Overview.
<http://www.earlham.edu/~peters/fos/overview.htm>

Westrienen, Gerard van and Lynch, Clifford A. (2005). Academic Institutional Repositories: Deployment Status in 13 Nations as of Mid 2005. D-Lib Magazine, 11(9), <http://www.dlib.org/dlib/september05/westrienen/09westrienen.html>