ePrints@IISc: India's First and Fastest Growing Institutional Repository

by

Francis Jayakanth, Filbert Minj, H.S. Usha, & Sandhya J

National Centre for Science Information
Indian Institute of Science
Bangalore – 560012

{franc, filbert, ushashiva, sandhyaj)@ncsi.iisc.ernet.in

Abstract
Case study

Purpose of this paper: To share NCSI's experiences in using GNU EPrints.org software to create and maintain the open-access institutional repository of IISc, ePrints@IISc.

Design/Methodology/Approach: The GNU EPrints.org is the first generic software for creating Open Access Initiative (OAI)-compliant repositories, which enables the researchers to self-archive their research publications thus facilitating open access to their publications. NCSI has been using this software since early 2002. Some amount of customization and value additions to the software are being done periodically by NCSI to meet the local requirements. In this paper, authors are sharing NCSI's experiences in using GNU EPrints.org software.

Findings: GNU EPrints.org software is an excellent tool for creating and maintaining OAI-compliant repositories. The best thing about it is that it can be setup easily even by those who are not too computer-savvy. New features are being introduced on a regular basis. Also, the technical support for the software from the developers and the user community has been very encouraging.

1. Introduction

The Indian Institute of Science (IISc) was established in 1909 through the pioneering vision of J.N. Tata. Since then, it has grown into a premier institution of research and advanced instruction, with more than 2000 active researchers working in almost all frontier areas of science and technology. It is one of the oldest and finest Centres of its kind in India, and has a very high international standing in the academic world as well (http://www.iisc.ernet.in).

In the recent years, IISc has been publishing more than 1500 research articles annually in almost all frontier areas of science and technology. Most of these research articles are published in the subscription-based scholarly journals. Fellow researchers in the country and across the world may or may not have access to these research publications depending on whether their libraries have subscription to the journals in which these

articles are published. The serials pricing crisis is affecting even the wealthiest of the libraries in terms of subscribing to the scholarly journals needed by their researchers. The situation in the developing countries is much more serious. Also, research work reported in the national journals of developing and emerging countries is missing to the international knowledge base because many of the journals from the developing and emerging countries are not indexed by the leading abstracting and indexing journals. Thus, there are both North to South and South to North knowledge gaps, leading to incomplete pictures of global health and other environmental and developmental issues.

The development of open-access (OA) movement offers a light at the end of the tunnel for the information-deprived scientists. The landmark meeting initiated by the Soros Open Society Institute in Budapest in 2001, resulted in the establishment of the Budapest Open Access Initiative and a definition of two primary ways ahead to close the knowledge gaps in science (http://www.soros.org/openaccess). The first OA strategy was called Self Archiving, and subsequently also referred to as Institutional Archives/Institutional Repositories, and proposed that copies of already published search articles should be archived in the author's institutional repository and made available free to all. All such OA archives would be OAI-compliant (i.e. Open Access Initiative Protocol for Metadata Harvesting compliant) and hence interoperable, thus facilitating federated searching (Chan, Kirsop, and Arunachalam, 2006)

In the recent years, the proponents of open-access to scholarly information are strongly urging the institutions to setup interoperable, open-access institutional repositories and impress upon their researchers to self-archive a copy of their research publications in such repositories. Institutional repositories will give better visibility to the research papers that they hold because any person with access to Internet will be able to read and download research papers contained in open-access repositories. After all, any researcher would want their research work to be read and be cited by as many people as possible. There is no better way of realizing this than through the open-access repositories. Studies have shown that research articles that are available in open-access repositories get better citations than the ones that are subscription-based (Lawrence, 2001).

With the open-access movement gaining momentum and its inherent benefits to the developing countries (Chan and Kirsop, 2001), NCSI strongly felt that an Institutional Repository be setup to provide OA to Institute's research publications. Such repositories will, to a great extent, eliminate access barrier to refereed research literature (Harnad). It can also serve as a publications database of an Institute.

IISc comprises of forty and odd departments, Centres and units staffed by high caliber scientists who are engaged in research in frontier areas. The institute has one of the best computing, networking, and equipments for experimental research in the country. It also has access to a large number of online electronic resources, which includes, leading bibliographic and citation databases, data sets, more than 10, 000 electronic journals and other online resources. The main library of the institute, J R D Tata Memorial Library, is one of the best libraries for scientific collections in the country. More than 90% of the main library's budget is spent on subscription towards periodicals. However, many other

universities and other Centres for higher education are not as much privileged as IISc is, especially when it comes to access to scholarly information. This will hamper the scientific progress in the country. For this reason, the universities grants commission of India started the National Centre for Science Information (NCSI) in 1983 at the IISc. The main objective in starting this Centre was to provide information services like current awareness service (CAS) and document delivery service (DDS) to the researchers working in the various universities across the country. The Centre was providing these services in the areas of Biology, Chemistry, Engineering, Geology, Mathematics, and Physics by subscribing to leading abstracting databases and printed journals in the said fields. These and other information services were rendered for almost two decades for both IISc community and also to the researchers working in other universities in the country. With the advent of online electronic journals, consortium-based subscriptions, and availability of several free online resources like the Stanford University's HighWire Press (http://highwire.stanford.edu), PubMed Central (http://www.pubmedcentral.nih.gov/), and host of other online resources, the need for centralized information services like CAS, DDS were no longer needed. Since 2002, the activities of NCSI are confined to the information needs of IISc community. Some of the current activities of the Centre include facilitating easier access to online electronic journals and databases by providing a gateway service (http://e-jis.ncsi.iisc.ernet.in), setting up and maintaining the science information portal of IISc (http://www.iisc.ernet.in), setting up and maintaining institutional repositories for research publications (http://eprints.iisc.ernet.in) and setting up and maintaining the theses and dissertations repository (http://etd.ncsi.iisc.ernet.in).

Realizing the dearth of trained manpower for providing computer-based information services, the Centre had started in the mid 80s, one year training programme called The application of information technologies for providing library and information services. In the recent years, the structure of the course has been re-designed. The Centre now conducts an 18-month training programme called Information and Knowledge Management. The details about the course are available at http://www.ncsi.iisc.ernet.in/training/

One of the students of the IKM training programme had carried out a project entitled "Archiving of Scientific Literature – Experience with EPrints.org software". The student had successfully demonstrated in implementing EPrints.org (http://software.eprints.org) software to create an open-access repository of research publications. The project work was seriously pursued, which resulted in the setting up of country's first and one of the world's earliest institutional repositories, ePrints@IISc (http://eprints.iisc.ernet.in). The open-access movement also gave us the impetus to establish the open-access repository of the institute.

For the ePrints@IISc repository, GNU EPrints.org software (http://eprints.org) is being used right from the beginning and it is serving the needs of an institutional repository very well. Some amount customization to the software has been done locally. In this paper the authors have shared some of their experiences in setting up and maintaining an interoperable institutional repository.

2. About ePrints@IISc

ePrints@IISc (http://eprints.iisc.ernet.in) is India's first interoperable, open-access institutional repository to have been setup. The then director of the Institute, Prof. Govardhan Mehata, had officially launched ePrints@IISc service in Sept 2002. Since its launch, the repository has been growing steadily in terms of number records it holds. To date, the repository holds more than 7000 research publications of the institute. NCSI's aim and desire is to archive all of Institute's publications to date. This is going to be a challenging task but NCSI intends to achieve this feat as early as it can.   This achievement will give a tremendous boost to the open-access institutional repository concept in the country. The repository will also serve as a comprehensive publications database of the institute, which in turn will facilitate better management of research knowledge generated by the Institute. From the researchers' perspective, we are trying to impress upon them the advantages that they stand to gain by depositing their research papers in the open-access, interoperable institutional repository by stressing on the following points:

- Helps in establishing priority for research findings
- Being interoperable, metadata from the repository are available through cross archive service providers like OAIster
- Indexed by Google Scholar, Microsoft's Windows Live Academic Search, Scirus
- Better visibility and wider access
- Better impact and citations
- Rapid communication of research
- Long-term preservation
- Integrated view of IISc research publications
- Value-added services like individual and department-wise publication listings
- Graphical view of usage statistics

The repository is hosted on a recently acquired Terabyte server running the latest version of Red Hat Enterprise Linux. The enterprise edition was opted because the operating system can be kept up-to-date by automatic updates, provided the license for support is active. GNU EPrints.org software version 2.3.13.1 is being used. The Centre's staff has incorporated number of value-added features to the repository software, which includes:

- Alphabetized listing of the author names,
- Browsing of records by the 'keywords' metadata field
- Analysis of access log file
- Customization of repository homepage, metadata fields, and document types

3. NCSI's Experiences so far

3.1 GNU EPrints.org Software

Developed at the University of Southampton, UK, GNU EPrints is meant for creating and maintaining institutional or discipline-based repositories. The current, stable version of

the software is 3.0. Newer versions of GNU EPrints are being released quite regularly. It is very important to update the repository software with the latest stable version so that the newer feature of the software could be made use of. ePrints@IISc is using GNU EPrints version 2.3.13.1. It will be migrating to the latest version in the near future.

Apart from being the first generic repository software, some of the very useful features of GNU EPrints.org are

- Very straight forward to install, customize, and maintain
- It is Unicode-compliant, so the repositories' content and user interface can be in any of the language that has a script
- Depending on the OS on which GNU EPrints is being installed, most of the dependent software can be installed along with the operating system.
- Easy to upgrade
- Repositories can have customizable subject classification scheme. For example, ePrints@IISc is using the faculty divisions and the corresponding departments as the subject classification
- The document types and their corresponding metadata elements can be easily configured, if required.
- Browse views can be generated for any of the metadata fields.
- The look and feel of the repository site is easily customizable as required.
- If required, Access to full texts can be restricted. At times to comply with the publishers' policy the repository administrator has to enforce this restriction. The restriction can be revoked anytime. In the latest version of the software, it is possible to send an email request to the authors to get a copy of such restricted articles.
- Local customizations can be retained when the repository is upgraded to newer versions of the software.
- Porting the contents of a repository to a physically different server can be easily done.
- Multiple repositories could be setup on a single instance of GNU EPrints.org
- Batch inputting of records can be easily done using the 'import_eprints' script
- Rendering of mathematical and chemical equations and symbols in their actual form can be achieved by encoding the equations and symbols in Latex
- Version 3 has a range of new features (http://www.ariadne.ac.uk/issue50/eprints-v3-rpt/):

Apart from the above-mentioned features, the discussion lists are very active. Most the queries are answered very quickly, many a times by the development team members themselves. Hence the solutions provided to the queries raised on the discussion lists are very reliable. The wiki and the mailing list archive are the other useful resources for trouble-shooting and other issues related to GNU EPrints.org

3.2 Strategies for filling the Repository

Since the time the repository was officially launched, not much of publicity campaign has been made to popularize the institutional repository concept and its inherent benefits to

the researchers. This is evident from the fact that self-archiving by the researchers is very sporadic. The NCSI staff has been doing the bulk of the deposits.

The SHERPA/RoMEO project site (http://www.sherpa.ac.uk/romeo.php) provides the policies of many of the publishers of scholarly journals towards self-archiving of published articles in an institutional repository. It is evident from the site that more than 92% of the publishers allow some form of self-archiving or the other. While some publishers allow only preprints, others allow both preprints and post-prints. Some publishers even allow / insist that their versions of PDFs be put in the institutional repositories..

From the SHERPA/RoMEO site, we first identified all such publishers who allow self-archiving of publishers' version of PDFs. Some of the publishers who come under this category include BioMed Central, IEEE, and Cambridge University Press. IISc being part of the INDEST consortium (http://paniit.iitd.ac.in/indest/), which is being funded by the ministry of human resources development, Govt. of India, has access to many of the leading electronic full text journals like IEEE/IEE Electronic Library Online, Emerald Fulltext, Elsevier ScienceDirect, ACM Digital Library, Springer Link etc. By doing an affiliation search, NCSI downloaded all the records with the full texts published by IISc from the publishers who allow their versions of PDFs to be put in the repository and ingested them into the repository. Some publishers, however, do not allow their versions of full texts in the repository. In such cases, we do make an attempt to get the authors' version of the post-prints by sending an email to the concerned authors. But in many of the instances either the authors do not have a copy of the post-prints or there is no response to the emails. It is also likely that the authors in question have left the institute or are away from the institute at that instance of time. In all such cases, we put the publishers' version of the full text and restrict the access to the IISc community only or just have the bibliographic details of such records.

NCSI also carries out searches on the other databases like the Web of Science, EiVillage, SciFinder Scholar, etc. to identify IISc publications. IISc has access to these and other resources either through the INDEST consortium or through the subscription of main library.. The downloaded records are converted to XML and are ingested into the submission buffer of the repository using the 'import_eprints' script, which is bundled with GNU EPrints.org software. These records will later be associated with the corresponding full texts depending on their availability.

Some of the faculty members send us hard copies of some of their publications to be deposited in the repository. If such publications are not available online, NCSI digitizes such articles and ingest them into the repository. Researchers can also send their publications to a specific email id, which will be attended to by one of the NCSI staff.

3.3 Repository Workflow

To upload an article into the repository one needs to first register with the repository to obtain a username and password. Once the registration process is successfully completed,

a user can login to the repository to deposit an article. For ePrints@IISc repository, the registration is restricted to the IISc community only.

Every submission that is made to the repository resides in the intermediary buffer called the submission buffer. The repository administrator will then scrutinize each of the submissions before they are approved for archiving. During the process of approval, the administrator ensures that the submissions adhere to the policies of the repository with respect to

- At least, one of the authors is affiliated to IISc
- Author names appearing in the original publications will be taken as authority
- The titles of the articles should be entered as in the original publications. This implies that formulae, equations, mathematical, chemical, and other symbols should appear as in the original publications.
- Correlation between metadata and full text
- The link to the publishers' version should be correct
- The administrator will also ensure that all the other metadata details are entered correctly, especially copyright information. Some publishers do insist that their copyright information be included.

3.4 Value Additions

3.4.1 Homepage Customization

The look and feel of the repository has been customized considerably as compared to the default installation of GNU EPrints.org software. Figure 1 is the screen shot of ePrints@IISc homepage. The homepage gives the total count of the deposits it holds, a listing of titles of a few recent publications, which have been archived, a link to the abridged version of publishers' policy, which is based on the SHERPA/Romeo project, and links to various other components of the repository.

3.4.2 Customization of Browse Views

By default, the browse views are generated for author, subject, and year metadata fields. The software can be easily configured to include any other metadata fields as needed for the browse view. ePrint@IISc has additionally included 'keywords' and 'document type' for browse views. For the author field, by default, the system generates one long, alphabetical listing of names. As the repository grows, this listing will also grow proportionately and eventually the navigation will become unwieldy. NCSI has developed a script, which generates hyperlinked, alphabetized listing, which makes navigation easier and convenient. Figure 2 gives the screen shot of hyperlinked, alphabetized listing of author names. NCSI has also developed a script to render the browse view of 'keywords' by splitting the multiple keywords associated with eprints into individual keywords and then sorting the individual keywords. These and other scripts that NCSI has developed are made available for free downloading from NCSI's wiki pages.
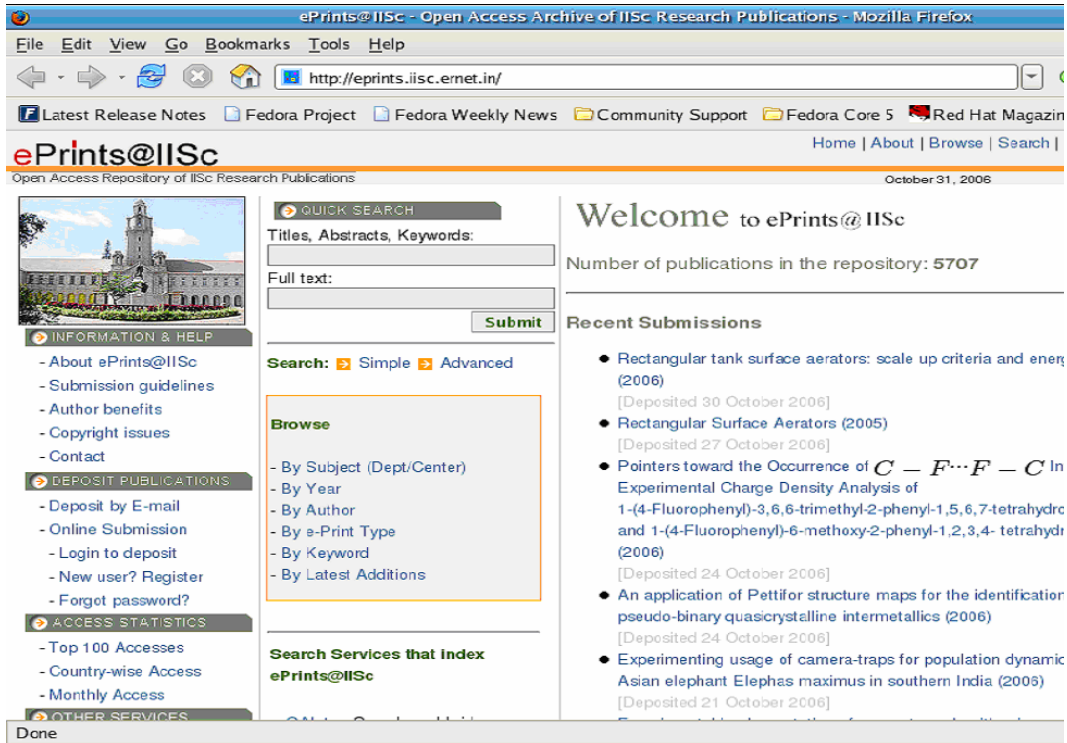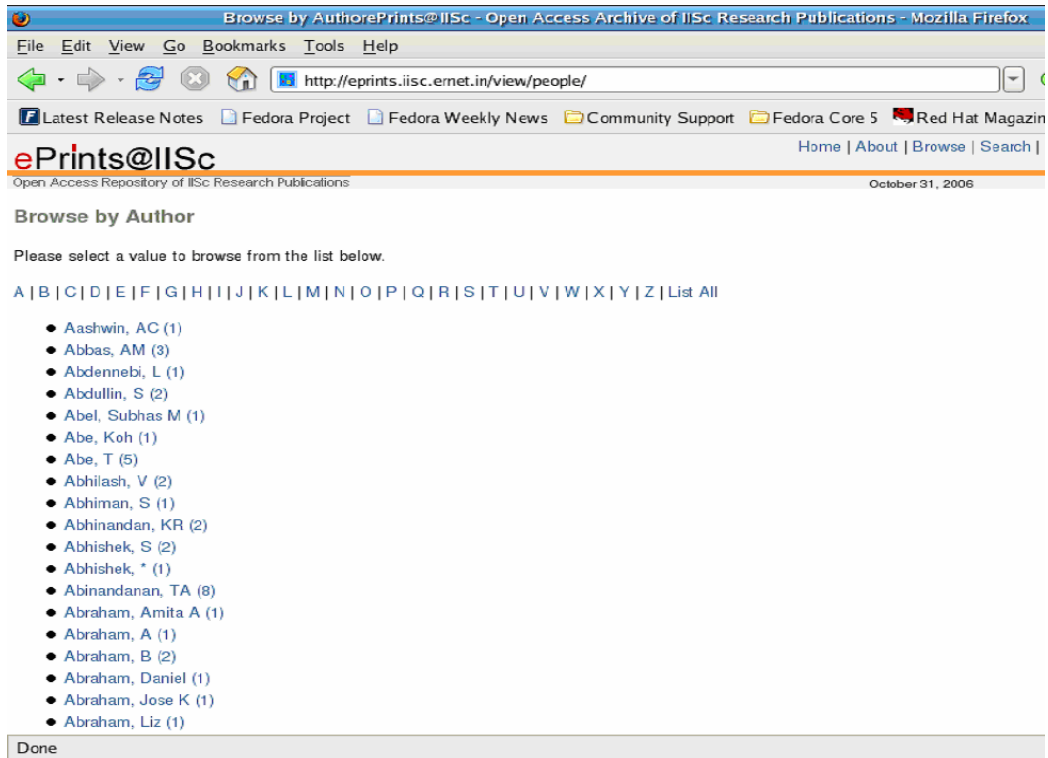
Figure 1 – ePrints@IISc Homepage

Figure 2 – Hyperlinked, Alphabetized Listing of Author Names

3.4.3 Wiki Pages

The concept of maintaining service-based wiki pages is catching up. We have setup wiki pages for sharing our experiences, scripts, and tools related to institutional repositories and OAI-compliance. It is accessible at http://ncsi-net.iisc.ernet.in/pmwiki. These pages provide details and references to the new features and functionalities that NCSI has been facilitating for ePrints@IISc repository.

3.4.4 Research Publications Report System (RPRS)

ePrints@IISc uses the organizational hierarchy (faculty divisions and departments) for subject browsing instead of a traditional classification scheme (Rajashekar and Jayakanth, 2004). RPRS facilitates in generating division or department-wise break up of different document types that have been deposited for particular year(s) in the repository. Figures 3 and 4 are the screenshots of RPRS for the years 2000-2006 for biological sciences division.
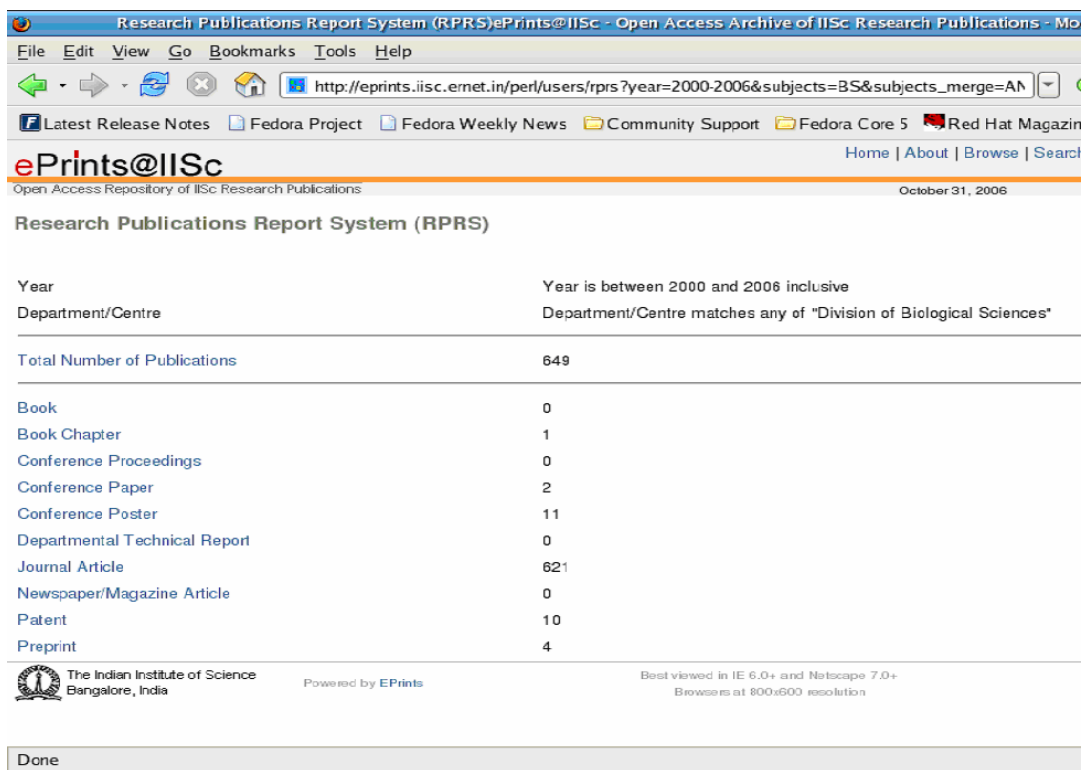

Figure 3 – RPRS for ePRints@IISc

Figure 4 – RPRS for Biological Sciences Division

3.4.4 Usage Statistics

Analysis of log files to generate usage statistics is an important task for any web-based service. NCSI has developed in-house software for this purpose. It is still in the testing phase. Figures 5, 6, and 7 shows screen shots of some usage statistics generated using the in-house software. At present, generation of usage reposts and statistics using the in-house software is not very straightforward. NCSI is also evaluating other open-source software for this purpose.

Conclusions

With the availability of several open-sources, professional quality digital library software meant for creating and maintaining institutional repositories, institutions can easily setup a repository in matter of weeks. However, sustaining the repositories for perpetuity is going to be a challenge, for new versions of the OS and the repository software will be emerging constantly. The repository team should ensure that the hardware and software components related to the repository are up-to-date. Depending on the magnitude of publications being handled, a dedicated team is a necessity. Maintaining consistency in the rendering of author names is very critical. Right now, there seem to be no easy solution to tackle this problem
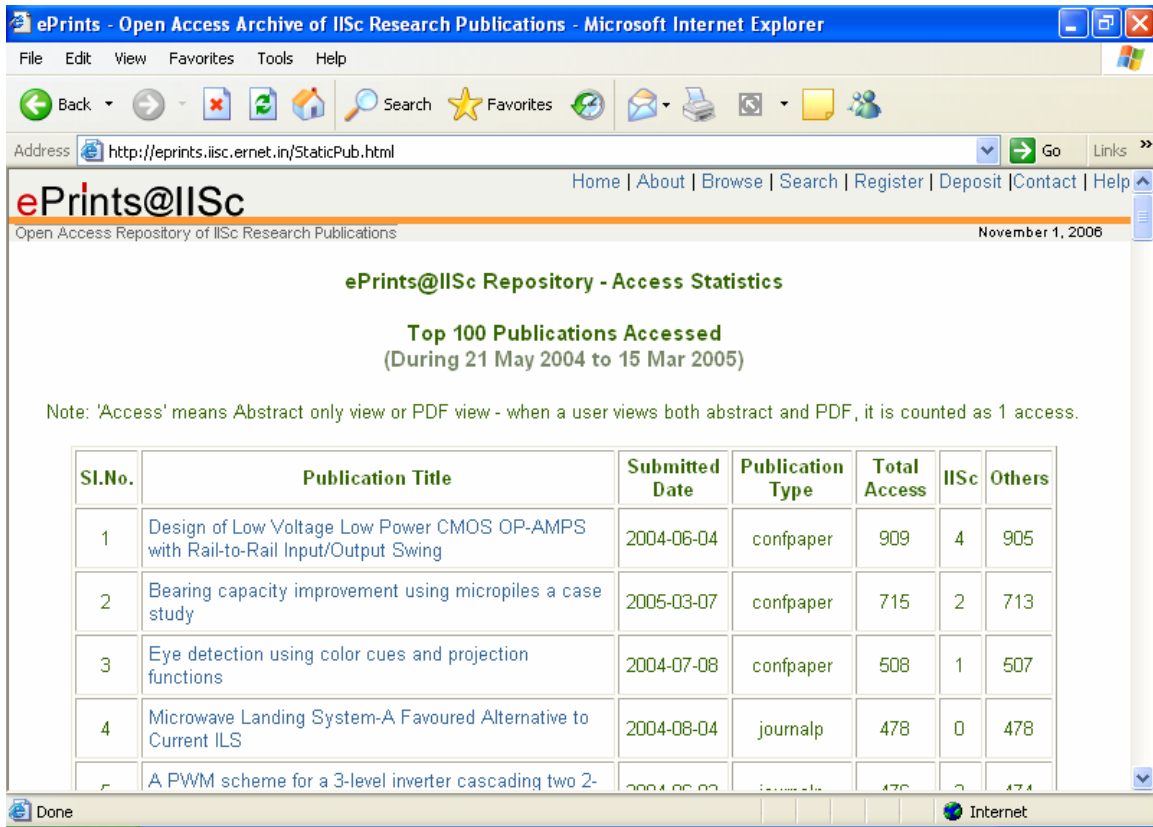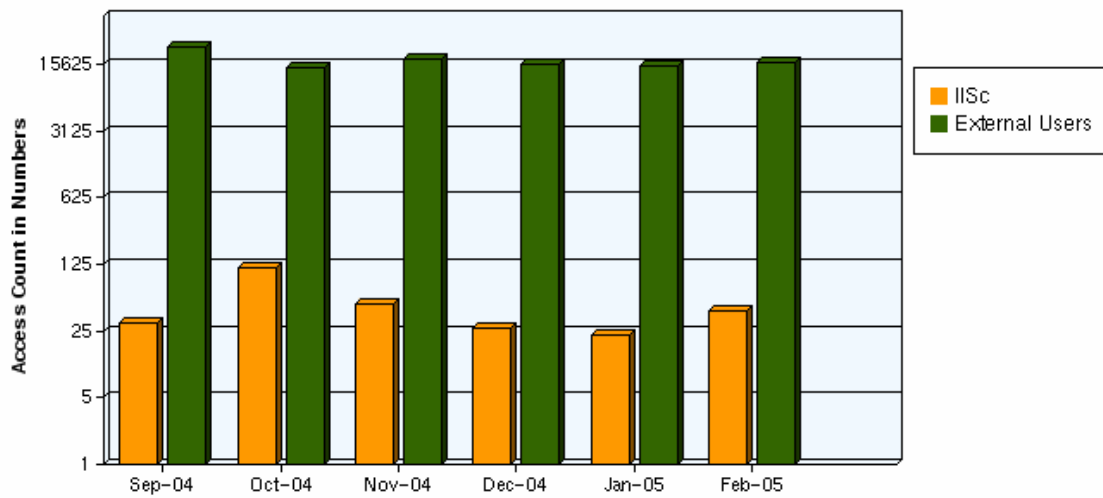
Figure 5 – Partial Listing of Top 100 Accesses

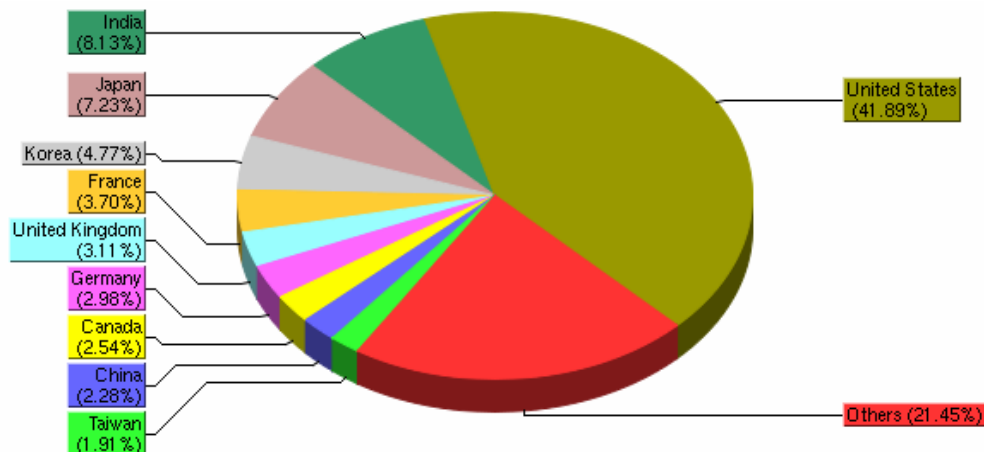Figure 6 – Internal and External Access Counts

Figure 7 – Top 10 Countries Accessing ePrints@IISc Repository

.
Acknowledgments

This paper would not be complete without acknowledging the contributions of late Dr. T.B. Rajashekar in setting up the ePrints@IISc repository. He was the team leader, which did all the ground works in setting up the repository. His immense knowledge in the areas of Library and Information Science, and in the emerging field of digital library, was effectively put to use in some of the innovative and unique features of ePrints@IISc. He was also an ardent advocate of open-access in the country. He had organized several national and international level workshops on GNU EPrints.org and other digital library software. At the time of his untimely demise, he was the principle investigator of two sponsored research projects. Both the projects were about providing open-access to research publications. The main objective of one of the projects was in setting up a national level cross archive search facility for the institutional / discipline-based repositories in the country. This objective has since been realized, when NCSI setup CASSIR (cross archive search services for Indian repositories) and has harvested more than seventeen thousand metadata records from seventeen repositories in the country (http://cassir.ncsi.iisc.ernet.in)

NCSI would also like to thank its former employees, Mr. Aneesh Joy, Ms. Jyoti Kittur, Ms. Kavitha Shree, Mr. Madhuresh Singhal, Ms. Kaini, Ms. Anitha, present and the previous batches trainees and many of its alumni who all have contributed immensely to the growth of ePrints@IISc repository.

NCSI also takes this opportunity to thank department of scientific and industrial research, ministry of science and technology, govt. of India, for sponsoring the project 'Development of OAI-based Institutional Research Repository Services in India', being carried by NCSI. As part of this project, NCSI has already setup a prototype cross-repository service for the OAI-compliant repositories in the country (http://cassir.iisc.ernet.in), developing tools for making legacy databases OAI-compliant,

tools to facilitate importing of records into the repository, and handling author names in the repository, which are rendered differently by different publishers

References

Lawrence, S. (2001). Free online availability substantially increases a paper's impact. *Nature Webdebates*. Available at:
http://www.nature.com/nature/debates/e-access/Articles/lawrence.html

Leslie, Chan and Barbara Krisop. (2001). Open Archiving Opportunities for Developing Countries: toward equitable distribution of global knowledge. Ariadne, 30. Available at:
http://www.ariadne.ac.uk/issue30/oai-chan/intro.html

Leslie, Chan, Barbara, Krisop, and Subbaiah, Arunachalam. (2006). Open Access Archiving: the Fast Track to Building Research Capacity in Developing Countries. Available at:
http://www.scidev.net/open_access/files/Open%20Access%20Archiving.pdf

Rajashekar, T.B. and Francis, Jayakanth. (2004). INASP Newsletter, June 2004, pp 6-7 Available at: http://www.inasp.org.uk/newslet/jun04.html#4

Stevan, Harnad. For Whom the Gate Tolls? How and Why to Free the Refereed research Literature Online Through Author/Institution Self-Archiving, Now. Available at http://www.cogsci.soton.ac.uk/~harnad/Tp/resolution.htm

Brief Bio-data of the authors

Mr. Francis Jayakanth holds a master's degree in library and information science. Apart from managing the ePrints@IISc repository, he offers courses on digital libraries and web-based information systems in the 18-month IKM training programme.

Mr. Filbert Minj holds a M.Tech. degree in computer science. Apart from managing the etd@IISc theses repository, he offers courses on 'C' programming language and database management system in the 18-month IKM training programme. He is also involved in the systems and network administrations

Ms. Usha Shiva holds a master's degree in library and information science. She is been involved in depositing articles, editorial reviewing, and other repository related work.

Ms. Shandhya Jagirdar holds a master's degree in library and information science. She is been involved in depositing articles, editorial reviewing, and other repository related work.

.