Archiving Scientific Literature: An Experience with E-prints Archive Software

By

Madhuresh Singhal
National Centre for Science Information,
Indian Institute of Science,
Bangalore – 560012, Phone: 080-3092511
E-Mail: madhuresh@ncsi.iisc.ernet.in

Francis Jayakanth
National Centre for Science Information,
Indian Institute of Science,
Bangalore – 560012, Phone: 080-3092511
E-Mail: franc@ncsi.iisc.ernet.in

Abstract

The world of academic publishing is undergoing many changes. Paper based publishing is being supplemented by electronic archives. In certain areas, preprint distribution has completely moved away from the paper-based system in to a fully electronic system called e-prints, based on open archives. Arxiv.org, hosted by Los Alamos National Laboratory, is considered the premier example of such e-print archives in the area of physics research. The Open Archive Initiative (OAI) develops and promotes interoperability solutions that facilitate the efficient dissemination of contents amongst the different e-print archives.

The e-prints or electronic pre-prints provide an almost wholly automated and highly efficient organizational framework and distribution mechanism, which is web based. E-print software, developed by the electronic and Computer Science Department at the University of Southampton (http://www.eprints.org) is one such tool, which helps us to build e-prints archive. In this paper, we have discussed the implementation of this software by archiving couple of papers published in the Journal of Indian Institute of Science.

1. Introduction

Online archives on the World Wide Web (WWW) may eventually supersede the original practice of scholars disseminating research in paper journals. In some areas of research, the preliminary stages involved in writing a paper are already making use of the WWW to make research findings public. So, there is no doubt that the WWW has superseded the more traditional methods of distribution.

Indeed, according to **Tenopir -**

"Johannes Gutenberg's printing advances in the 15th century loosened the controls on publishing by broadening and democratising information content and access, and ultimately helped to promote the cultural and scientific upheavals that followed. The Internet- and the World Wide Web in particular- represent another step in this ongoing revolution."

The on-line archives evolved from the practice of authors emailing pre-prints of their papers to peers for informal feedback. With the archives, authors can deposit their pre-refereed work (pre-prints) and published work (post-prints) into an archive for all to see. The cost and sheer number of refereed journals that exist (over 20,000 according to **Bowkers**), it is simply not possible for libraries and institutions to subscribe to more than a small portion of all the journals published every year. For those trying to locate papers, the process can often be a long, laborious and often fruitless one.

According to **Steven Bachrach** et al

"Electronic communication has created new ways to distribute such results and is forcing researchers and publishers to reassess the old procedures and consider new possibilities as we

learn to use the Internet. Now, not only can authors easily disseminate their results, but networked readers can have cheap, fast access to more scientific literature and have it in a form that facilitates its use in their own research".

2. E-Print Origin and Their Impact

The origin of the Eprint lies in increasing interest in alternatives to the traditional scholarly publishing paradigm. In his regard, the development of alternative models for the communication of scholarly results, particularly in the form of online repositories of Eprints, has demonstrated a viable alternative to traditional journal publication.

For this purpose, in October 1999, a meeting in Santa Fe was organized on the belief that the interoperability among these Eprint archives was key to increasing their impact. The Santa Fe convention was the first attempt in which the technical and organizational agreements about OAI were decided.

The following are the examples of few premier eprint archives -

ArXiv (http://www.arxiv.org), hosted by Los Alamos National Laboratory, is considered the premier example of e-print archives. The archive was started in 1991 by Paul Ginsparg, who is internationally recognized as one of the leaders in the area of scholarly publishing alternatives. Over the past decade, the arXiv archive has evolved towards a global repository for non peer-reviewed research papers in a variety of physics research areas. arXiv has also incorporated mathematics, non-linear sciences and computer science.

CogPrints (http://cogprints.soton.ac.uk), hosted by the University of Southampton in the U.K., is modeled on arXiv and focuses mainly on papers in Psychology, Linguistics and Neuroscience.

NCSTRL (Networked Computer Science Technical Reference Library) (http://www.ncstrl.org/) is an international collection of computer science research reports. NCSTRL is based on a distributed model. Documents are stored in distributed archives and are made available through distributed services that communicate via the Dienst protocol.

Stevan Harnad at the University of Southampton and Paul Ginsparg of Los Alamos National Laboratory are the champion of this movement. The fundamental idea here is that authors would deposit preprints and/or copies of their published versions into such servers, thus providing readers worldwide with a free way of obtaining access to these papers, without needing paid subscription access to the source electronic journals.

3. Advantages of Eprint

In both the pre-refereeing pre-prints and the final published post-prints, the author cites other authors' work that has been used in the research. If a reader wishes to look at a cited piece of work, that paper has to be found elsewhere. This is time-consuming, and often ends with the paper being inaccessible, because the user's institution cannot afford to subscribe to it.

With on-line archives, all papers can be located by anyone quickly and easily and at no cost. Authors can put draft copies and successive updates up for public view, until the final, peer-reviewed (published) version appears. Users can follow the research through all of its successive stages from pre-prints through to the post-prints.

Ultimately we can have one central repository of all the literature. The advantage of central repository will be that anybody can search or browse the literature in their respective area.

4. Eprints Archive Software

Eprints archive software was developed by the Electronics and Computer Science Department of the University of Southampton (http://www.eprints.org). The current version of the software is eprints 1.1.2. This free software can be downloaded from the abovementioned URL by any individual or institution interested in maintaining eprints archive.

This software enables scholars to easily deposit their papers with a reliable and reputable organization that will ensure its availability and accessibility. An Eprint archive is used by scholars to circulate their work quickly and widely. Eprint archives are intended to supplement

traditional mechanisms for the circulation of printed-paper documents. Far greater and more rapid circulation of the document can be achieved by posting it on the archive at no cost to the author. Individual scholars can then be alerted efficiently to the presence of the work by informing them in a brief email of the eprint's unique ID code. Alternatively, scholars may subscribe to receive regular email updates of postings to the archive.

5. Hardware and Software Requirements

- At least Intel Pentium II processor.
- A UNIX operating system. Linux (a very advanced and free UNIX implementation) works just fine, and is in fact the development platform.
- The Apache WWW server, another professional-quality free software product, often included with Linux distributions, such as RedHat.
- The Perl programming language also included with most Linux distributions.
- The mod_perl module for Apache, which significantly increases the performance of Perl scripts.
- The MySql Database, a database system that is free for non-commercial use.
- The **Eprints** software.

The prerequisite software packages that are to be installed are described below in table 1.

Table 1: Prerequisite Software Packages

Package	*Where from*
tar & Gunzip	comes with most UNIX/ RedHat packages
unzip 5.x	`ftp://ftp.freesoftware.com/pub/infozip/`
wget 1.5	`http://www.gnu.org/software/wget/wget.html`
perl 5.005	`http://www.cpan.org/`
apache 1.3.9	`http://www.apache.org/`
mod_perl1.2.1	`http://perl.apache.org/`
Mysql 3.22.x	`http://www.mysql.org/`
sendmail 8.9	comes with most UNIX/ RedHat packages

Many of the software listed in the above table gets installed automatically while installing the Operating System. The Eprints Archive software uses the following Perl modules listed in table 2. These modules are available for free from the CPAN website (http://www.cpan.org). These modules must be installed prior to the installation of Eprint Software. Some of these might already be installed with Perl Distribution.

- CGI 2.6x
- POSIX
- Data-Dumper 2.xxx
- DBI 1.1x
- Msql-MySQL-modules 1.2xxx
- Filesys-Diskspace 0.05
- MIME-Base-64 2.11

- URI-1.06
- XML-Writer-0.4
- ApacheDBI-0.87
- Unicode-String-2.xx

6. Features of Eprints Archive Software

Among the services and functions Eprints archive offers for deposit and management of content are -

- It's easy to setup and install. An installation script automates most of the installation process.
- It can store documents in any format that we (as archive administrator) wish to be accepted. Each individual research paper (or e-Print) can be stored in more than one document format (Fig. 7).
- The archive can use any metadata schema; the administrator decides what metadata fields to held about each e-Print. This is decided in four stages -

1. Decide a maximal set of metadata fields that should be stored (for example, "authors", "title", "journal", "journal volume", etc.) (Fig. 4 & 5)
2. Decide what types of e-Print should be stored (for example, refereed journal article, thesis, conference papers, technical report, unpublished preprint) (Fig. 3)
3. For each e-Print type, decide which metadata fields should be stored for e-Print of that type, and which of those fields are mandatory.
4. Decide how these metadata fields should be projected into the Open Archives world.

- Eprints can be placed in a configurable, extendible subject hierarchy. This hierarchy can be used to view and search the archive. (Fig. 1)
- Submission of all e-Prints is via a simple but extremely powerful WWW-based interface. Papers can be uploaded just as files, in a compressed bundled file (such as a .zip file), or automatically mirrored from an existing website by specifying a URL. (Fig. 8)
- Authors can have associated metadata. Again, the site administrator can decide what metadata fields will be stored with each author's record.
- Users can subscribe either as authors or readers, via a web form or an automatically processed e-mail account.
- Submitted papers (if administrators desire) go through a moderation process. Submitted papers are placed in a buffer, where they can be approved by a moderator, rejected outright, or returned to the author for amendment. (The extent of this moderation is, of course, up to the individual institution running the archive.
- Moderation process is also performed using a WWW-based interface. In fact, once the site is up and running, it will require very little maintenance.
- Automatic notification of newly submitted content in selected subject areas.

7. Installing and Configuring the Software

Much of the installation process has been automated by the use of some installation scripts. It is also possible to install the software manually, allowing the maximum amount of flexibility when installing the software on machines that are running other services. Here is an overview of the steps that one can perform automatically or manually.

1. Install the code, setting relevant path information.
2. Configure the code to work with the rest of the system.
3. Create the MySQL database for Eprints.

4. Configure Apache to execute Eprints scripts and serve Eprints documents.
5. Install a crontab, which periodically executes various Eprint scripts.

These steps will complete the installation and configuration of Eprint Archive Software.

8. Implementation

The e-prints archive software offers two kinds of services.

1. Archive viewing service
2. Registration based service.

1. Archive viewing service

Everybody can access archive-viewing service. The viewing service allows us to **browse** the archive by subject and **search** the archive. Fig. 1 shows the screen shot of browsing the archive by the subject. We can configure the fields to be displayed for searching. Every e-print, which is archived, is automatically assigned document id code. The archive can also be viewed using this id code.

2. Registration based service

Registration based service includes subscription service and deposit items service. In order to access registration based services, one has to get a login id/password by sending an email to the archive administrator account.

a. Subscription Service

This feature allows us to instruct the archive to automatically e-mail the details of new eprints appearing in the archive. One can choose the subject areas in which s/he is interested, and how often s/he wishes to be informed of updates: Every day, week or month.

b. Deposit Item service

When this option is selected for depositing an item, a pop up screen for logging in to the server comes up. After successfully logging in, the screen shot shown in Fig 2 will come up. From the subsequent screen shown in Fig 3, one has to select appropriate document type for deposit from the list. The list items are configurable. Upon selecting the document type, the bibliographic details about the document have to be provided in the subsequent screens. These screen shots are shown in Fig 4 and 5 respectively. These fields (Metadata) are configurable by the archive administrator. The fields with asterisks are mandatory fields. After providing the details of the mandatory fields, the next screen will be of the subject category of the item to be deposited (Fig. 6). Item can be assigned as many subject categories as required. The administrator can add new subject categories, whenever such a request comes from the author. In the next screen as shown in the Fig 7, the format of the document to be uploaded is specified. By default the archive accepts the following file formats: HTML, PDF, PS, ASCII text. The administrator can add new formats or can remove any existing format/s.

In the next screen as shown in Fig. 8, one has to select the file upload method. The file to be uploaded can be a plain text file, from an existing website, zip archive or compressed tar archive. After selecting the appropriate file, one can also verify the file by viewing it before final submission, and then finally submit it for archiving. Once an e-print is deposited, it doesn't get archived immediately. The deposit has to be approved by the site administrator. In this way, to a certain extent quality control can be ensured.

9. Open Archive Interoperability

A vital component of the archiving of research literature on-line is the interoperability of different eprint archives. Otherwise, in order to find relevant material, one would have to visit many archives in turn and search each individually.

Interoperability is a broad term, touching many diverse aspects of archive initiatives, including their metadata formats, their underlying architecture, their openness to the creation of third-party digital library services, their integration with the established mechanism of scholarly

communication, their usability in a cross-disciplinary context, their ability to contribute to a collective metrics system for usage and citation, etc.

Interoperability among archives offers substantial benefits to the scholars that use them. An important attribute of the traditional research library as an information provider is its role as a common entry point for a variety of information resources, not necessarily divided along disciplinary or institutional boundaries. The move from physical to digital sources should not be accompanied by the breakup of this entry point into a collection of fragmented archives. An increasing number of scholars move fluidly in their research across domain boundaries; the technology for delivering digital information should facilitate rather than hinder such fluidity. Mechanisms for interoperability offer the potential for discovery tools and virtual collections that extend across the contents of multiple archives. Authors also benefit from such archive spanning tools, since their works will be accessible by a wider audience.

Interoperability is also beneficial to the archive and service provider. Rather than having to provide an entire suite of services for its users, individual archives can instead establish a well-defined interface on which external providers can build enhanced services. A variety of such services can be envisioned, including those that facilitate discovery, linking, and reviewing. An intriguing and essential set of services would be those that provide metrics to assist in the evaluation of the impact of certain scholarship and aid in tenure review and promotion decisions.

In brief we can say that because of Interoperability, the papers in all Eprints Archives can be harvested and searched by Open Archive Services. One such Cross Archive Searching Service http://arc.cs.odu.edu/ is providing seamless access to all the eprints, across all the Eprint Archives, as if they were all in one global, virtual archive.

10. Conclusion

The Open Archives initiative (OAI) promotes and encourages the development of author self-archiving solutions (also commonly called e-print systems) through the development of technical mechanisms and organizational structures to support interoperability of e-print archives. Such interoperability can stimulate the transition of e-print systems into genuine building blocks of a transformed scholarly communication model. There is now well-established system within the scholarly publishing world, to enhance public access to scholarly journal articles through the use of e-prints server.

References

1. **Eprints Archive Software Installation Document**: http://www.eprints.org/docs/eprints-install.html

2. http://www.eprints.org/results/report.html

3. http://www.cogsci.soton.ac.uk/~harnad/Tp/resolution.htm

4. http://opcit.eprints.org/dl00/dl00.html

5. http://library.cern.ch/HEPLW/4/papers/3/

6. http://htpprints.yorku.ca/faq.html

7. http://lib-www.lanl.gov/libinfo/preprints.htm

8. http://www.nature.com/nature/debates/e-access/Articles/harnad.html

9. http://www.openarchives.org/

10.http://www.arl.org/sparc/core/index.asp?page=g20#6

11.http://www.arl.org/newsltr/217/mhp.html

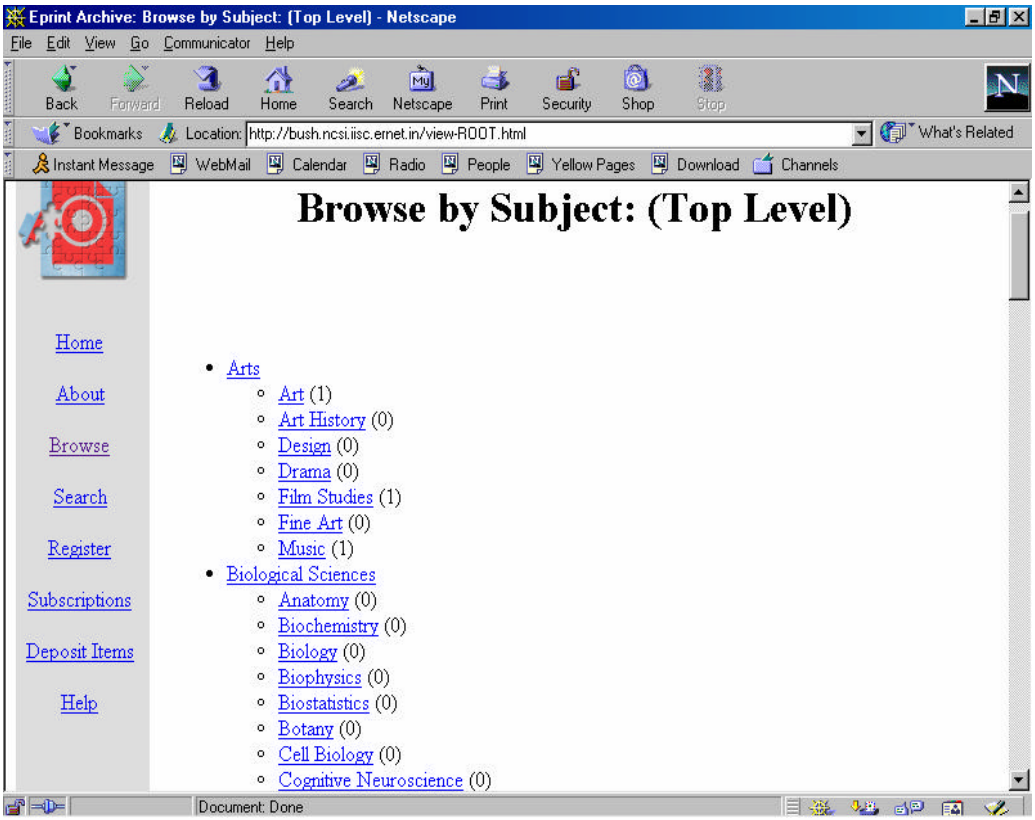
Eprint Archive: Browse by Subject: (Top Level) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://bush.ncsi.iisc.ernet.in/view-ROOT.html What's Related
Instant Message WebMail Calendar Radio People Yellow Pages Download Channels
Browse by Subject: (Top Level)
Home
About
Browse
Search
Register
Subscriptions
Deposit Items
Help
Arts
Art (1)
Art History (0)
Design (0)
Drama (0)
Film Studies (1)
Fine Art (0)
Music (1)
Biological Sciences
Anatomy (0)
Biochemistry (0)
Biology (0)
Biophysics (0)
Biostatistics (0)
Botany (0)
Cell Biology (0)
Cognitive Neuroscience (0)
Document: Done

Fig. 1

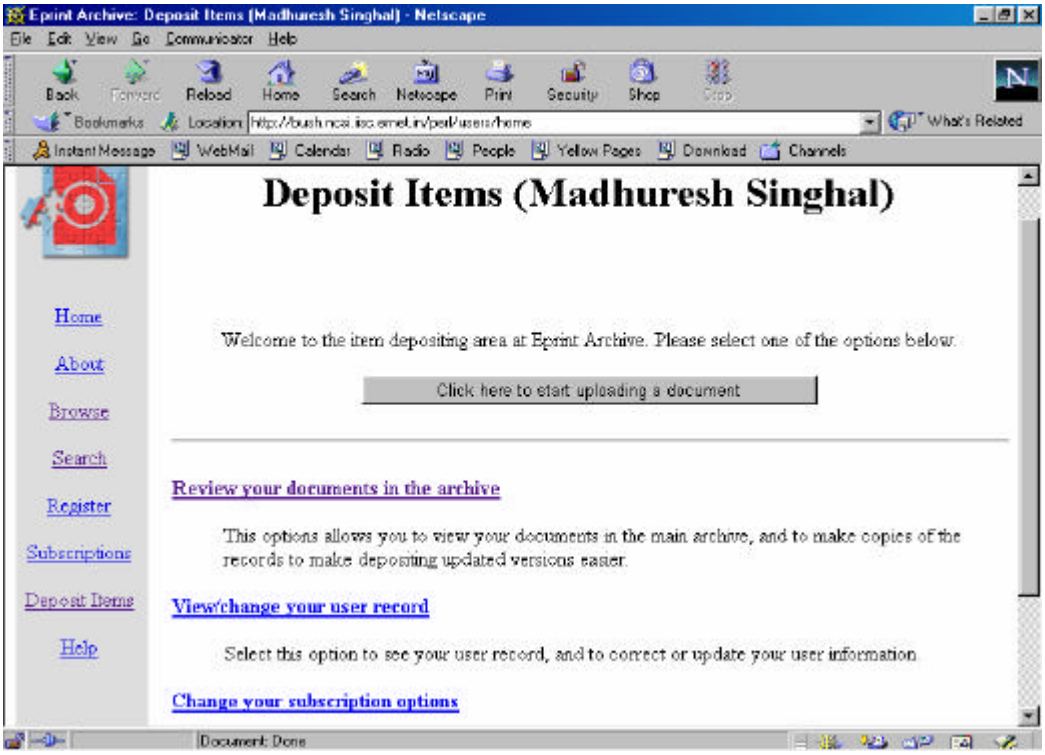
Eprint Archive: Deposit Items (Madhuresh Singhal) - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://bush.ncsi.iisc.ernet.in/perl/users/home What's Related
Instant Message WebMail Calendar Radio People Yellow Pages Download Channels
Deposit Items (Madhuresh Singhal)
Home
About
Browse
Search
Register
Subscriptions
Deposit Items
Help
Welcome to the item depositing area at Eprint Archive. Please select one of the options below.
Click here to start uploading a document
Review your documents in the archive
This options allows you to view your documents in the main archive, and to make copies of the records to make depositing updated versions easier.
View/change your user record
Select this option to see your user record, and to correct or update your user information
Change your subscription options
Document: Done

Fig. 2

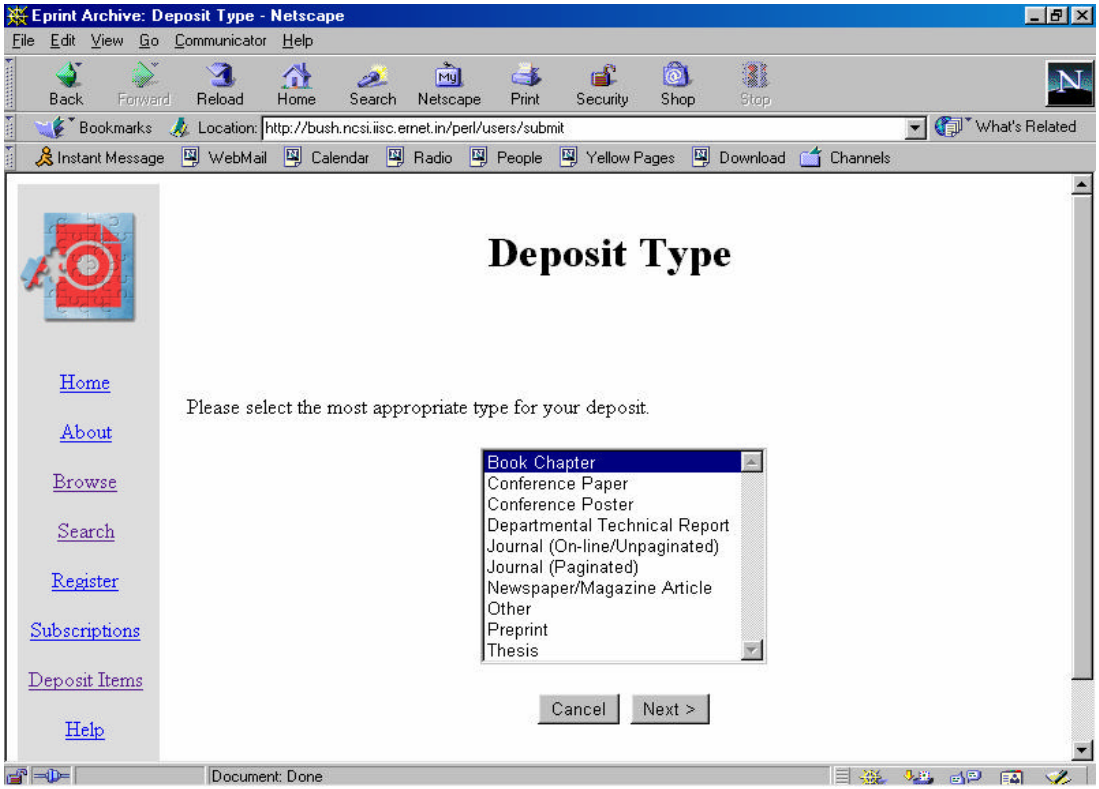
Eprint Archive: Deposit Type - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://bush.ncsi.iisc.ernet.in/perl/users/submit What's Related
Instant Message WebMail Calendar Radio People Yellow Pages Download Channels
Deposit Type
Home
About
Browse
Search
Register
Subscriptions
Deposit Items
Help
Please select the most appropriate type for your deposit.
Book Chapter
Conference Paper
Conference Poster
Departmental Technical Report
Journal (On-line/Unpaginated)
Journal (Paginated)
Newspaper/Magazine Article
Other
Preprint
Thesis
Cancel
Next >
Document: Done

Fig. 3

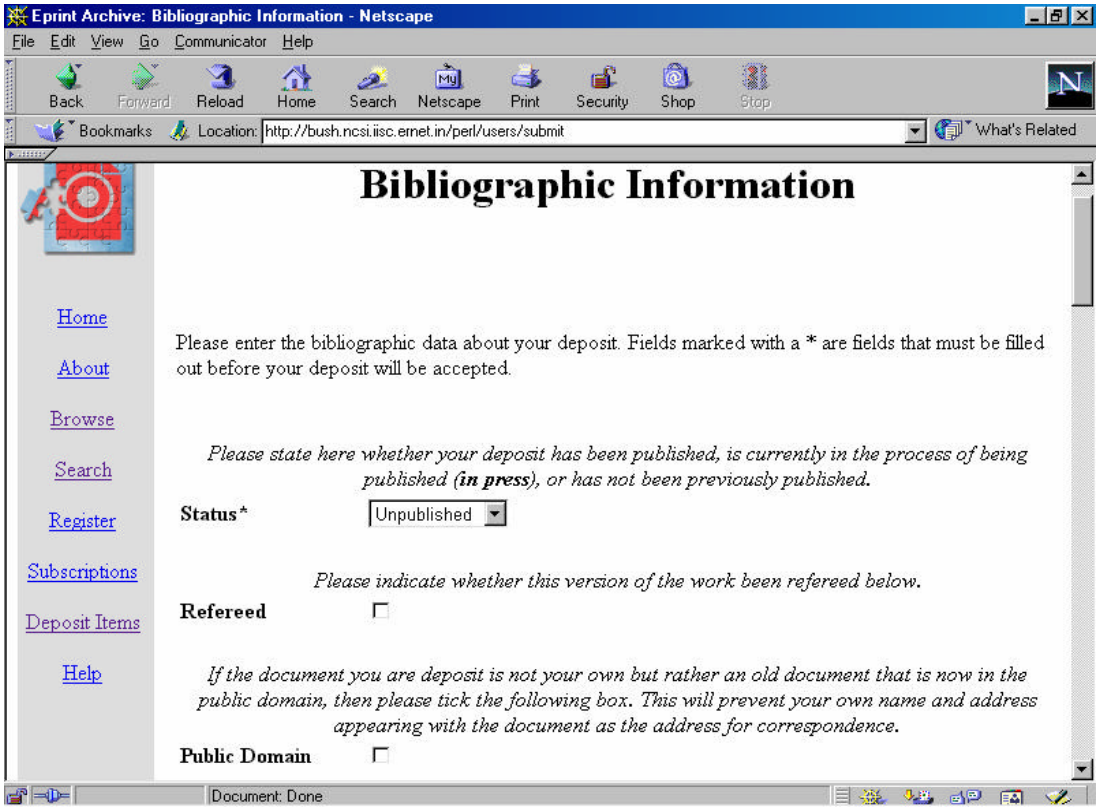
Eprint Archive: Bibliographic Information - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://bush.ncsi.iisc.ernet.in/perl/users/submit What's Related
Bibliographic Information
Home
About
Browse
Search
Register
Subscriptions
Deposit Items
Help
Please enter the bibliographic data about your deposit. Fields marked with a * are fields that must be filled out before your deposit will be accepted.
Please state here whether your deposit has been published, is currently in the process of being published (in press), or has not been previously published.
Status*
Unpublished
Please indicate whether this version of the work been refereed below.
Refereed
If the document you are deposit is not your own but rather an old document that is now in the public domain, then please tick the following box. This will prevent your own name and address appearing with the document as the address for correspondence.
Public Domain
Document: Done

Fig. 4

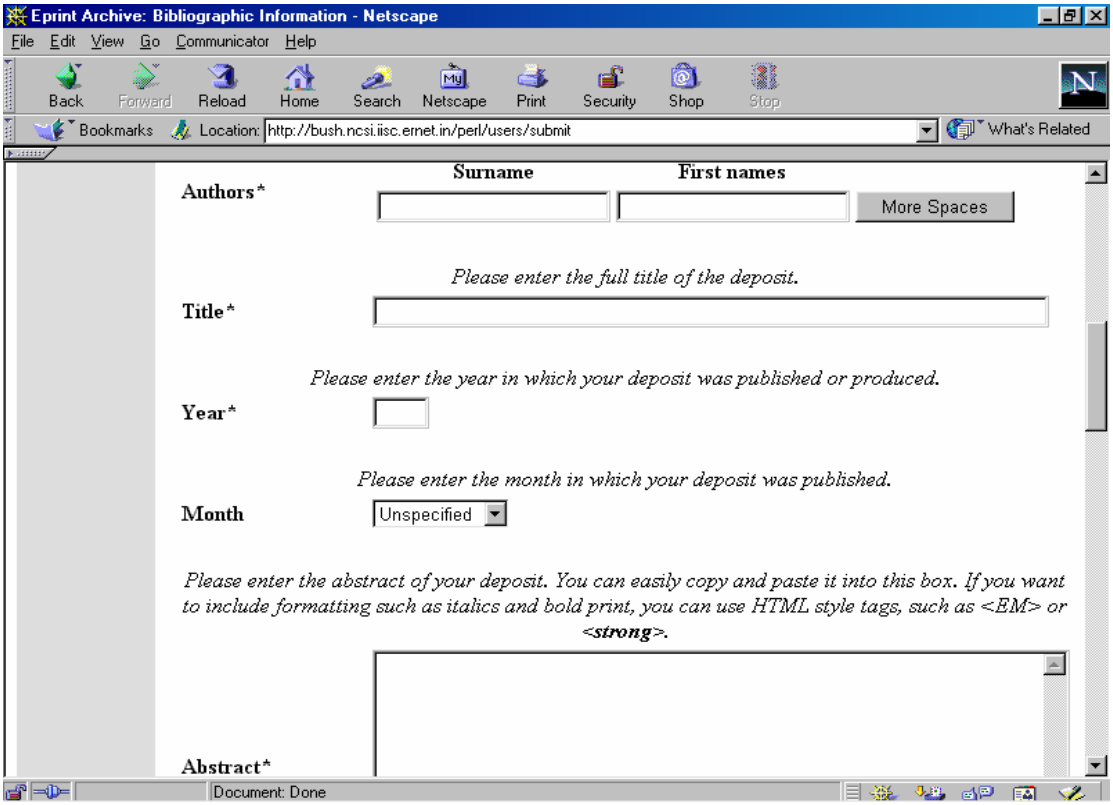
Eprint Archive: Bibliographic Information - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://bush.ncsi.iisc.ernet.in/perl/users/submit
What's Related
Surname
First names
Authors*
More Spaces
Please enter the full title of the deposit.
Title*
Please enter the year in which your deposit was published or produced.
Year*
Please enter the month in which your deposit was published.
Month
Unspecified
Please enter the abstract of your deposit. You can easily copy and paste it into this box. If you want to include formatting such as italics and bold print, you can use HTML style tags, such as <EM> or .
Abstract*
Document: Done

Fig. 5

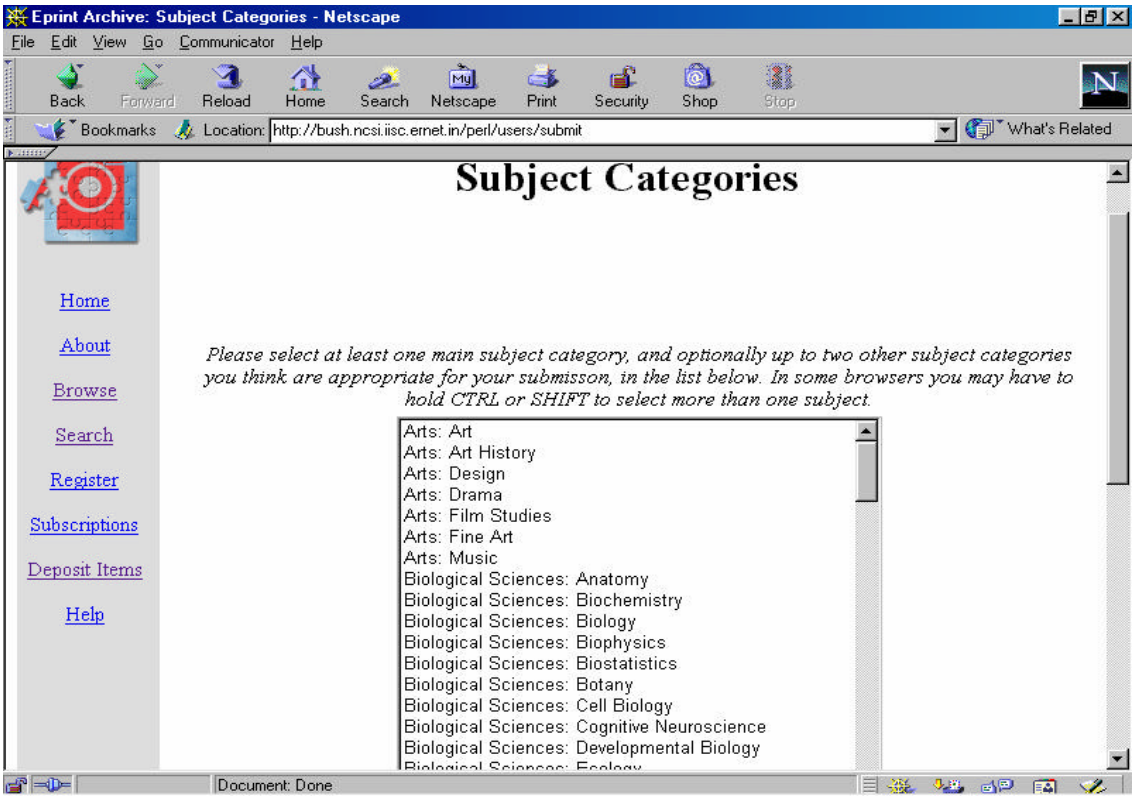
Eprint Archive: Subject Categories - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://bush.ncsi.iisc.ernet.in/perl/users/submit
What's Related
Subject Categories
Home
About
Browse
Search
Register
Subscriptions
Deposit Items
Help
Please select at least one main subject category, and optionally up to two other subject categories you think are appropriate for your submisson, in the list below. In some browsers you may have to hold CTRL or SHIFT to select more than one subject.
Arts: Art
Arts: Art History
Arts: Design
Arts: Drama
Arts: Film Studies
Arts: Fine Art
Arts: Music
Biological Sciences: Anatomy
Biological Sciences: Biochemistry
Biological Sciences: Biology
Biological Sciences: Biophysics
Biological Sciences: Biostatistics
Biological Sciences: Botany
Biological Sciences: Cell Biology
Biological Sciences: Cognitive Neuroscience
Biological Sciences: Developmental Biology
Document: Done

Fig. 6

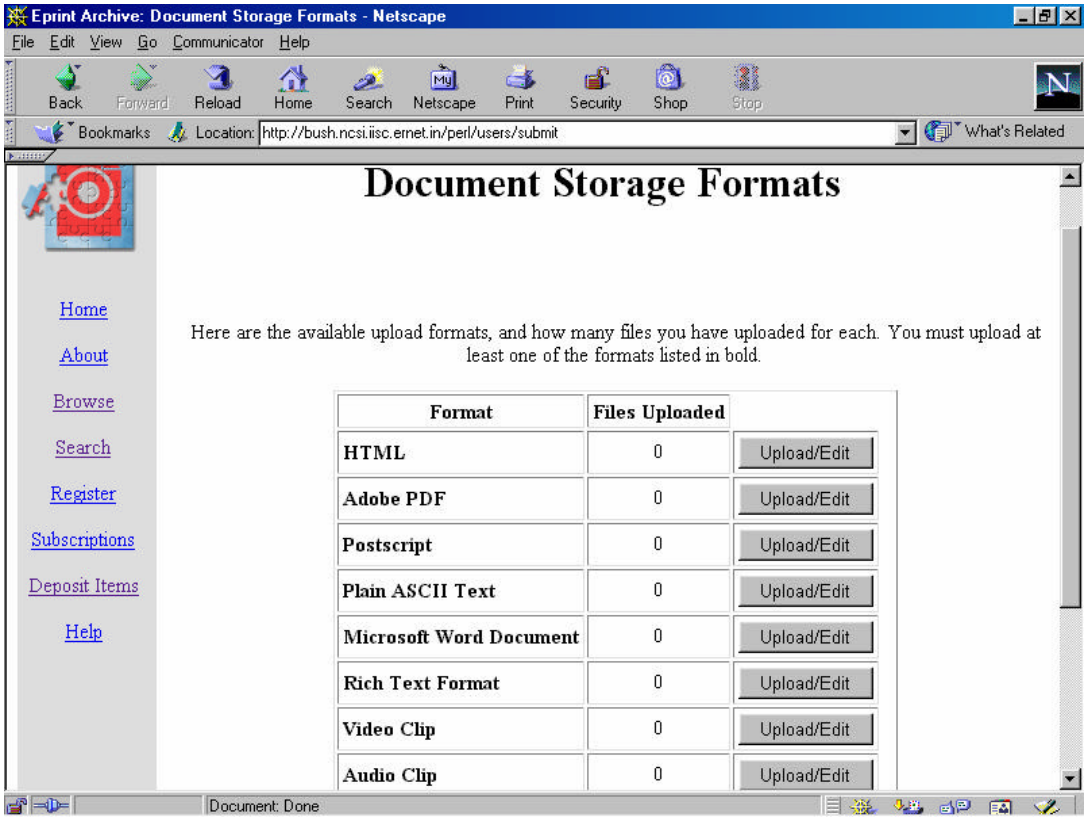
Eprint Archive: Document Storage Formats - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://bush.ncsi.iisc.ernet.in/perl/users/submit
What's Related
Document Storage Formats
Home
About
Browse
Search
Register
Subscriptions
Deposit Items
Help
Here are the available upload formats, and how many files you have uploaded for each. You must upload at least one of the formats listed in bold.
Format
Files Uploaded
HTML 0 Upload/Edit
Adobe PDF 0 Upload/Edit
Postscript 0 Upload/Edit
Plain ASCII Text 0 Upload/Edit
Microsoft Word Document 0 Upload/Edit
Rich Text Format 0 Upload/Edit
Video Clip 0 Upload/Edit
Audio Clip 0 Upload/Edit
Document: Done

Fig. 7

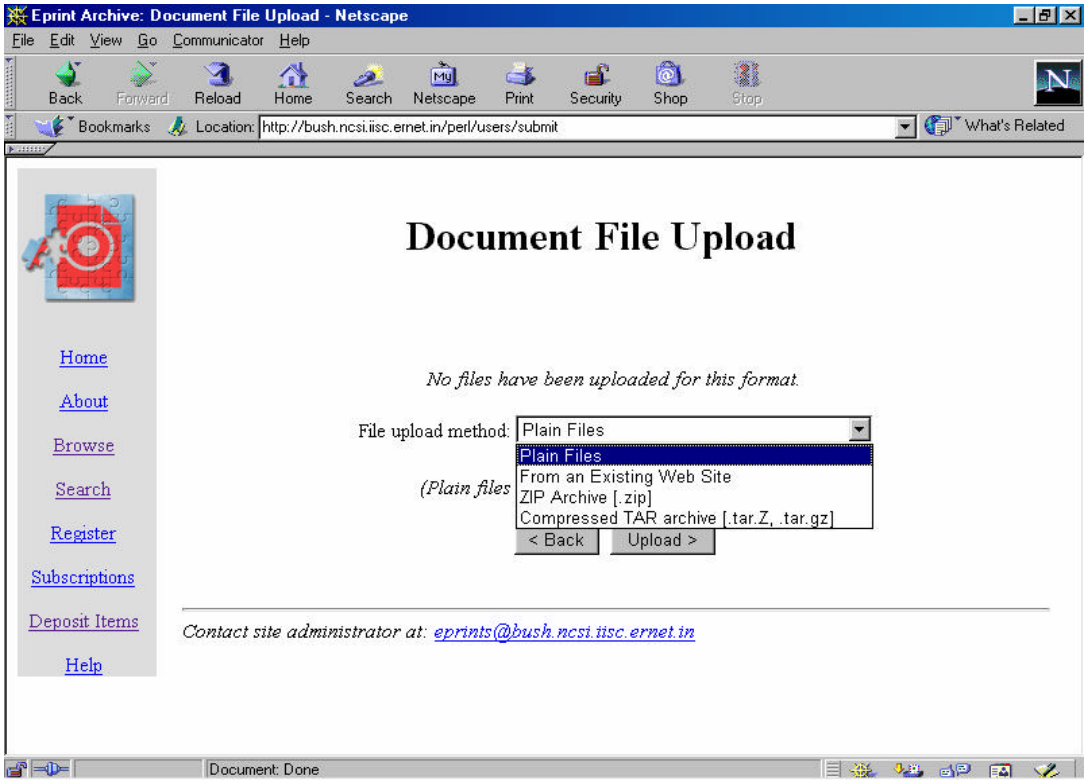
Eprint Archive: Document File Upload - Netscape
File Edit View Go Communicator Help
Back Forward Reload Home Search Netscape Print Security Shop Stop
Bookmarks Location: http://bush.ncsi.iisc.ernet.in/perl/users/submit
What's Related
Document File Upload
Home
About
Browse
Search
Register
Subscriptions
Deposit Items
Help
No files have been uploaded for this format.
File upload method: Plain Files
Plain Files
From an Existing Web Site
ZIP Archive [.zip]
Compressed TAR archive [.tar.Z, .tar.gz]
(Plain files
< Back
Upload >
Contact site administrator at: eprints@bush.ncsi.iisc.ernet.in
Document: Done

Fig. 8