SETTING UP A MULTILINGUAL INFORMATION REPOSITORY : A CASE STUDY WITH EPRINTS.ORG SOFTWARE

Nagaraj N Vaidya Francis Jayakanth

Abstract

Today 80 % of the content on the Web is in English, which is spoken by only 8% of the World population and 5% of Indian population. There is wealth of useful content in the various languages of the world other than English, which can be made available on the Internet. But, to date, for various reasons most of it is not yet available on the Internet. India itself has 18 officially recognized languages and scores of dialects. Although the medium of instruction for most of the higher education and research in India is English, substantial amount of literature by way of novels, textbooks, scholarly information are being generated in the other languages in the country. Many of the e-governance initiatives are in the respective state languages. In the past, support for different languages by the operating systems and the software packages were not very encouraging. However, with the advent of Unicode technology, operating systems and software packages are supporting almost all the major languages of the world that have scripts. In the work reported in this paper, we have explained the configuration changes that are needed for Eprints.org software to store multilingual content and to create a multilingual user interface.

Keywords : Institutional Repository, Eprints, digital asset management system, Multilingual content, OAI-PMH, Unicode

1. Introduction

The world is in the midst of a technological revolution nucleated around Information and Communication Technologies (ICT). Advances in Human Language Technology will offer nearly universal access to information and services for more and more people in their native languages. The Internet and Unicode have provided a wonderful opportunity to the governments, non-governmental organizations and individuals to publish information in native languages.

With the availability of user-friendly open source and multilingual repository creation and maintenance software like Eprints.org and Dspace, creation of multilingual repository is becoming a reality. These software are basically meant for creation and maintenance of institutional or discipline-based repositories to provide open access to the research papers produced by the institutions and organizations. Several hundreds of Institutional Repositories (IR) are already in existence with new ones emerging very other day. A partial list of registered repositories is available at http://www.eprints.or. Providing open access to the research work will help the researchers, especially in the developing world, to have access to the research literature from across the world. Also, studies have shown that the citations for the open access scholarly literature is higher than the subscription based scholarly literature. [1]

Eprints. Dspace and other software meant for providing open access to scholarly information can also be used as general-purpose digital asset management systems. As these software are Unicode compliant, the digital assets can be in any of the native languages. With the wide spread usage of computers in every sphere of life, substantial amount of literature are being born digital. And, in a country like India, which has 18 official languages based on 10 different scripts, a multilingual digital asset

management system is very essential to manage and provide access to digital assets that are in the native languages. In this article, we have explained the changes we have made to the EPrints.org software configuration files, so that it can be used for managing and provide access to digital assets in two of the Indian languages, namely Hindi and Kannada. The same procedure can be followed for other languages as well.

2. What is an Institutional Repository ?

The definition of an IR as given in wikipedia, (http://en.wikipedia.org/wiki/Institutional_repository) is:

"The Institutional Repository, as a concept, is to capture and make available as much of the research output of an institution (i.e. a university) as possible. In the first instance this might include material such as research papers and electronic versions of documents such as theses, but may also include many of the digital assets generated by normal campus life, such as administrative documents, course notes, or learning objects" An IR software typically aims to:

- Capture and describe digital material using a workflow
  - Provide interface for online submission of research material (Intranet)
- Provide access to this material over the web (metadata and/or full publication)
- Preserve digital material for posterity
- Share metadata with other Institutional Repositories

3. Need for Multilingual Information Repository in India

India is a multi-religious, multi-cultural and multi-lingual country. It produces significant amount of literature in its native languages. Many organizations are showing keen interest in digitizing and archiving these valuable literatures.. It is essential that information processing and digital repositories of knowledge resources should be available for wider proliferation of research and progress to benefit the people at large. This can become a reality only if the knowledge resources available in the native languages are captured and made available through multilingual repositories. With the availability of high quality, reliable, easy to install and use open source software for repository creation and maintenance, it is cost effective for an institution or an organization to use one such open source software for maintaining its multilingual repository.

3.1 About GNU Eprints 2

GNU EPrints is generic repository software under development by the University of Southampton http://www.soton.ac.uk. It is intended to create a highly configurable web-based archive. GNU EPrints' primary goal is to facilitate an open archive for research papers, and the default configuration reflects this. But, it could be easily used for other things such as images, research data, audio archives - anything that can be stored digitally, by making changes to the configuration files.

EPrints 2 is written in PERL, and runs as an apache module (using mod_perl) EPrints uses MySql to store the metadata about records and users. The actual files in the archive are stored in the OS file-systems. Being Unicode compliant, it can handle multilingual content and the user interface can also be multilingual. It is an OAI-complaint (http://www.openarchives.org) software facilitating harvesting of metadata automatically by OAI-based service provider (http://arc.cs.odu.edu).

4. About Unicode Standard

Fundamentally, computers just deal with numbers. They store letters and other characters by assigning a number for each one. Before Unicode was invented, there were hundreds of different encoding systems for assigning these numbers. No single encoding scheme could contain all the letters and other characters that make up different languages of the world. Also, when more than one encoding scheme is in use, there is a possibility of conflict among the encoding schemes. That is, two encoding schemes can use same numbers for two different characters, or use different numbers for the same character. The advent of Unicode is changing all these issues by providing a unique number for every character; no matter what the platform, what the program, or what the language is (http://www.unicode.org/standard/WhatIsUnicode.html).

4.1 Unicode Standard for Indian Scripts

Presently Unicode supports 10 Indic scripts, which are based on ISCII (Indian Standard Code for Information Interchange, 1988). Devanagari is a script for writing classical Sanskrit and its modern historical derivative, Hindi. Devanagari script is also used to write many other languages like Marathi, Grahwali, Gondi etc.

Devanagari block of Unicode is based on ISCII 1988. Unicode hexadecimal numbers for Devanagari starts from U+0900-U+97F.

Kannada is derived from ancient Dravidian language and it is one of the south-Indian scripts used for writing Kannada. Unicode hexadecimal number for Kannada starts from U+0C80-U+0CFF.

5. Configuring E-Prints Software to Handle Native Languages

The development platform of Eprints software is Red Hat Linux, but it can be installed on the other Linux distributions like Debian or Suse. It also works on Solaris and other UNIX-based operating systems. For our work, we have installed Eprints version 2.3.12 on a PIII system, with 256 MB of RAM and running GNU Red Hat Fedora Core 3 operating system. EPrints software is dependent on number of other software like Apache web server (with mod_perl), MySql, Perl and number of Perl modules. It is also dependent on number of standard Linux commands like ‘wget’, ‘tar’ and ‘unzip’.

The latest version of EPrints.org software can be downloaded from the site

http://software.eprints.org. It is bundled with the documentation.

The installation of EPrints is quite simple. The documentation provides the required details. By default EPrints gets installed in the directory named /opt/eprints2. The directory structure of the EPrints system after the basic installation is as below.

/opt/eprints2/bin This directory contains various command line scripts used for creation and maintenance of repositories.

/opt/eprints2/cfg Contains the plain text-based configuration files that constitute part of the site-specific portion of the Eprints archive. They contain information about what metadata to hold about each EPrints and user, the initial subject hierarchy.

/opt/eprints2/cgi	Contains the Perl scripts that are invoked by the Apache WWW server. These scripts create pages with dynamic content, for example, search results and the eprint depositing interface.
/opt/eprints2/perl_lib	Contains the core Eprints Perl library files
/opt/eprints2/defaultcfg	This file describes configuration files for a default archive**.**

5.1 Language issues :

EPrints uses utf-8 encoding internally. There is no problem to store documents or entering metadata in any native languages. EPrints can interpret characters using utf-8 encoding directly.

Fedora Core 3 has built in support for 9 Indian languages, namely – Bengali, Gujarati, Hindi, Marathi, Kannada, Tamil, Telugu, Oriya and Malayalam. It has GNOME 2.8 and KDE 3.3 as desktop managers. Keyboards can be set up on GNOME, to type Hindi and Kannada characters. To enable Hindi keyboard,

- Right Click on the Panel > Add to Panel > Keyboard Indicator> Add

To add Hindi keyboard layout using Keyboard Indicator Applet

- Right Click on Keyboard indicator > Open keyboard preferences > Layout > Hindi > Add

Note that this operation is OS specific and has nothing to do with EPrints Software.

EPrints software may be configured to have user interfaces in different languages. To accomplish this functionality, the static and configuration files, which control the look of the site and rendering of dynamic content, needs to be duplicated for each of the languages. These files include the website template, the static web pages, the phrase files, and the citation rendering configurations.

To create a repository, ‘configure_archive’ script located in ‘bin’ directory needs to be executed. This script expects various inputs like ‘archivID’, ‘archive name’ etc. Once this script is successfully executed, the configuration files related to this repository gets created in a sub directory. The name of the sub directory is same as the archiveID. The path for the repository thus created will be, /opt/eprints2/archives/archiveID

The default user interface of EPrints software is English. However, it is quite simple to create user interfaces in other languages as well. We explain below steps involved to configure EPrints to handle content and creation of user interface in Hindi.

5.2 Customizing Static Content :

All the static web pages for a repository gets created in a sub-directory under the directory path /opt/eprints2/archives/archive_id/cfg/static. The name of this sub-directory is a two-lettered code for the languages as per the ISO 639-1 standard. For example, for the English interface the name of the directory is ‘en’. There are five static files in this directory, namely

- **index.xpage** : the welcome page of EPrints archive
- **information.xpage** : the “about” page

- **error401.xpage** : returned in case of authentication failure on the pages with restricted access
- **vlit.xpage** : the page where is exposed the transcopyright permission.
- **help/index.xpage** : the help page for the users of the archive.

The above pages are XML files and the actual homepage, about page, and help pages gets generated from the above files.

5.3 Customizing Language :

To create static pages in Hindi, copy the contents from the 'en' directory to 'hi' directory. 'Hi' is two-lettered ISO standard for Hindi language. The contents of all the static web pages in the 'hi' directory are to be translated into Hindi language.

The rest of the user interface is generated dynamically as per the users' requests and they are dependent on number of XML files located in /opt/eprints2/archives/archive_id/cfg and other directories. For the Hindi interface, the files citations-en.xml, template-en.xml, phrases-en.xml are to be copied to citations-hi.xml, template-hi.xml, phrases-hi.xml respectively and equivalent translation in Hindi needs to be done.

All the Hindi files should start with the following two lines:

```
<?xml version="1.0" encoding="utf-8" standalone="no" ?>

<!DOCTYPE phrases SYSTEM "entities-hi.dtd">
```

The DTD file, '*entities-hi.dtd',* is generated automatically, and contains all standard definitions from the file 'xhtml-entities.dtd', and other EPrints specific definitions, such as '&adminemail;' '&archivename;'..

Also, in the /opt/eprints2/cfg/ directory, the following two files are to be modified:

- **system-phrases-en.xml:** it contains the common pieces of phrases used in the archive (independent from the nature of the archive). It has to be modified by copying the content to system-phrases-hi.xml
- **languages.xml:** this file doesn't need any translation, but the language specific line has to be uncommented. For example, to enable Hindi and Kannada, the following lines are to be uncommented.

<lang id="hi" supported="yes"> Hindi </lang>

<lang id="kn" supported="yes"> Kannada </lang>

Since the installation script does not prompt for a language option, the first archive has its language set to English. Easiest way to make provision for other languages is by modifying archiveID.xml file. This file gets created automatically when the script, 'bin/configure_archive' is executed.

For example, the file, 'archiveID.xml', has the following lines for English. It needs to be modified to accommodate other languages.

<language >en</language>·

<archivename language=”en”>eprints@ncsi</archivename>

<defaultlanguage >en</defaultlanguage>

To include Hindi, it should be modified to:

<language >en</language>

<language >hi</language>

<archivename language=”en”>eprints@ncsi</archivename>

<archivenamelanguage=”hi”>eprints@ncsi</archivename>

<defaultlanguage >en</defaultlanguage>

Next step is to enable multi languages by modifying the following line in the ‘ArchiveConfig.pm’ file located at /opt/eprints2/archives/archiveID/cfg/ArchiveConfig.pm

By default, the multi-language option is set to 0:

$c -> {multi_language_options) = 0;

This has to be set to “1” for archive to be multilingual.

Changes made to the configuration files require the web server to be stopped and started again.

5.4 Duplicating Database Tables :

There are eight elements distributed in different tables of the repository database that are language sensitive: these tables are named “xxx__ordervalues_en”. These tables are to be duplicated and renamed for specific languges “xxx__ordervalues_lgid”. (eg: for hindi, “xxx__ordervalues_hi”). Then execute bin/configure_archive and bin/generate_views scripts.

Next step is to translate subject tree, for that we need to use an XML-encoded subject file. To generate this file, use the bin/export_xml script, it will generate the XML file corresponding to the existing subject tree, and then we have to add the translation for each subject record. A sample subject tree record is shown below:

<record>

<field name=”subjectid”>HF</field>

<field name=”name”>

<lang id=”en”>HF Commerce</lang>

<lang id=”hi”>HF □□□□□□□ </lang>

</field>

<field name=”parents”>C</field>

<field name=”depositable”>TRUE</field>

</record>

The modified file has to be imported into the database using the 'bin/import_subjects script'. A link for language section should be added in the navigation bar of the repository homepage by including the following line in the 'index.xpage' file:

Language

Then, run the 'bin/generate_static' and 'bin/generate_views' scripts. Stop and restart web server.

All the configuration changes mentioned in the above paragraphs are based on the document available at http://sophia.univ-lyon2.fr/doc/eprints.org/2-customization-gb.htm

5.5 Virtual Keyboard Interface :

To search for Hindi documents in the archive, the end user has to provide search terms in Hindi. For some reason if a multilingual keyboard is not setup on the client systems, such clients can use a Java-based virtual keyboard interface to type Hindi characters. It is available at http://144.16.72.145/hindikeyboard

The virtual keyboard contains all the alphabets of Hindi language, so the end user can generate any combination of characters for his/her search requirements. One has to just click on the alphabet buttons on the virtual keyboard to form search terms. Provision for selecting the keyboard could be made from the homepage of Hindi language interface.

Some screenshots for Hindi and Kannada interfaces are given in the appendix.

6. Conclusion

Unicode complaint software facilitates searching and retrieval of content in native languages. User interfaces can be created easily in any of the languages that have corresponding Unicode character sets. User interfaces in native languages will be of great help to countries like India where there are eighteen officially recognized languages. In the days ahead, more and more of content that are in native languages will make their presence on the Internet. This will draw non-english speaking people to computers and Internet.

7. References

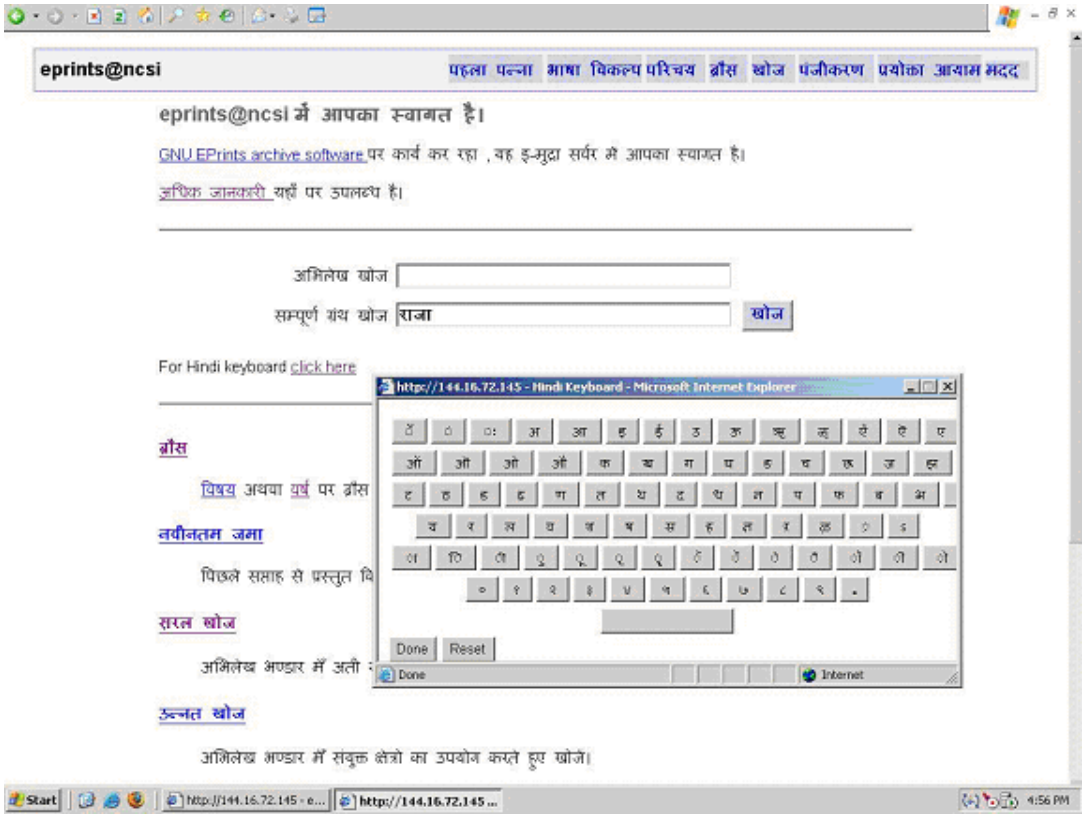
1. Steve Lawrence: "Online or invisible?" Nature 2001, 411:521

About Author

Sh. Nagaraj N. Vaidya, working as an Information and Knowledge Management Trainee at the National Centre for Science Information, Indian Institute of Science, Bangalore, India. He has Completed his masters degree in library and information science from Karnatak University, Dharwad, Karnataka, India
E-Mail : nagaraj@ncsi.iisc.ernet.in

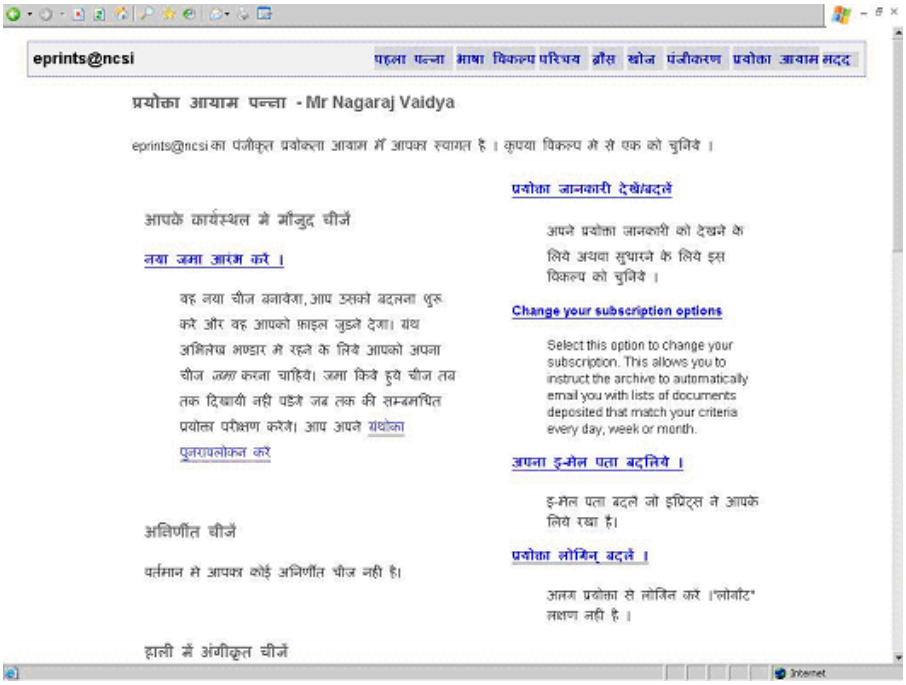
Sh. Francis Jayakanth, works as a Scientific Staff at the National Centre for Science Information, Indian Institute of Science, Bangalore, India. His research interests include digital libraries, OAI-compliant information systems, web enabling and OAI-compliance of legacy databases.
E-Mail : franc@ncsi.iisc.ernet.in

Appendix -I

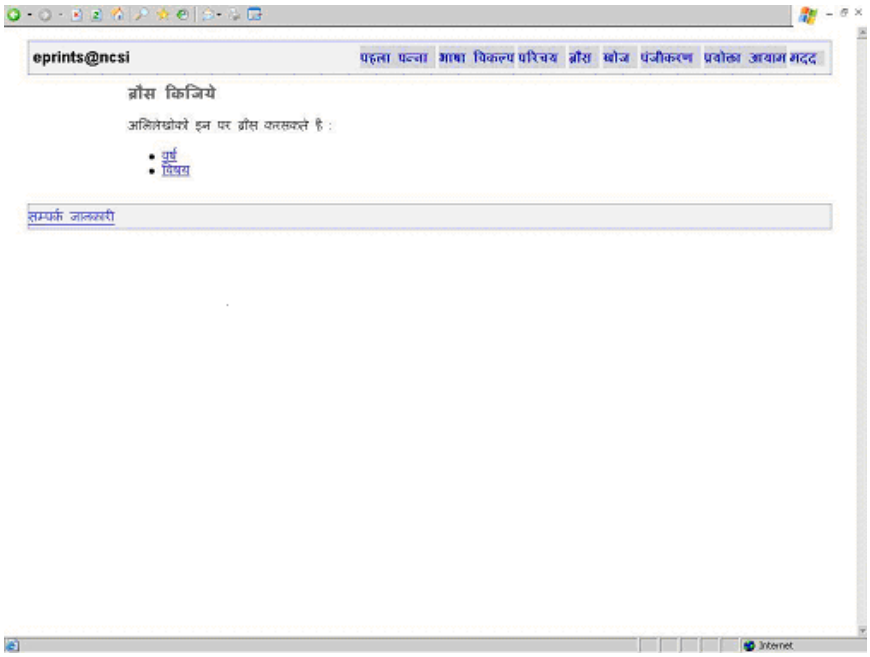


Hindi Homepage with Virtual Keyboard

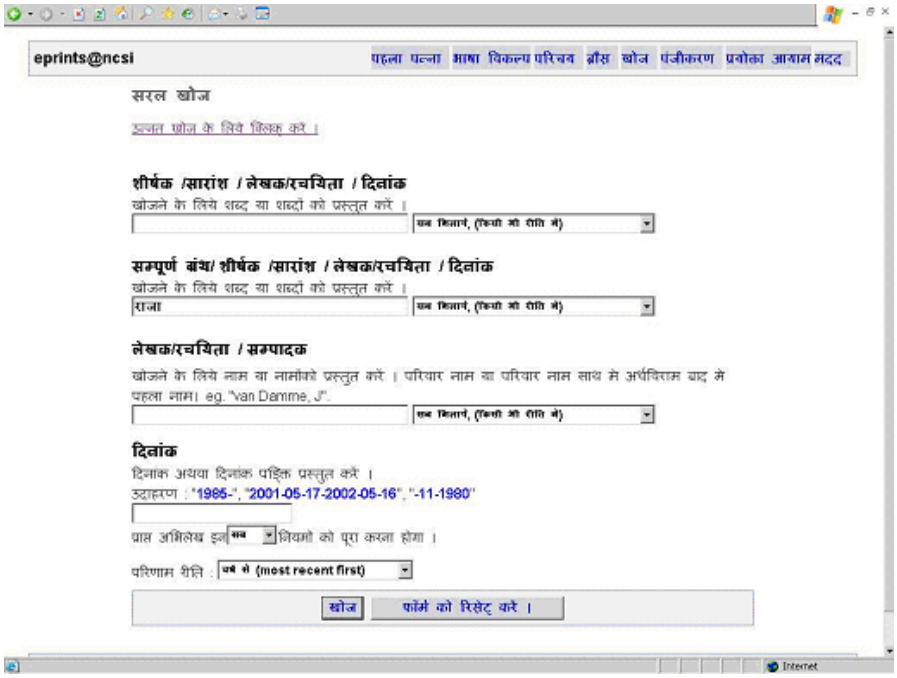
Hindi User Registration



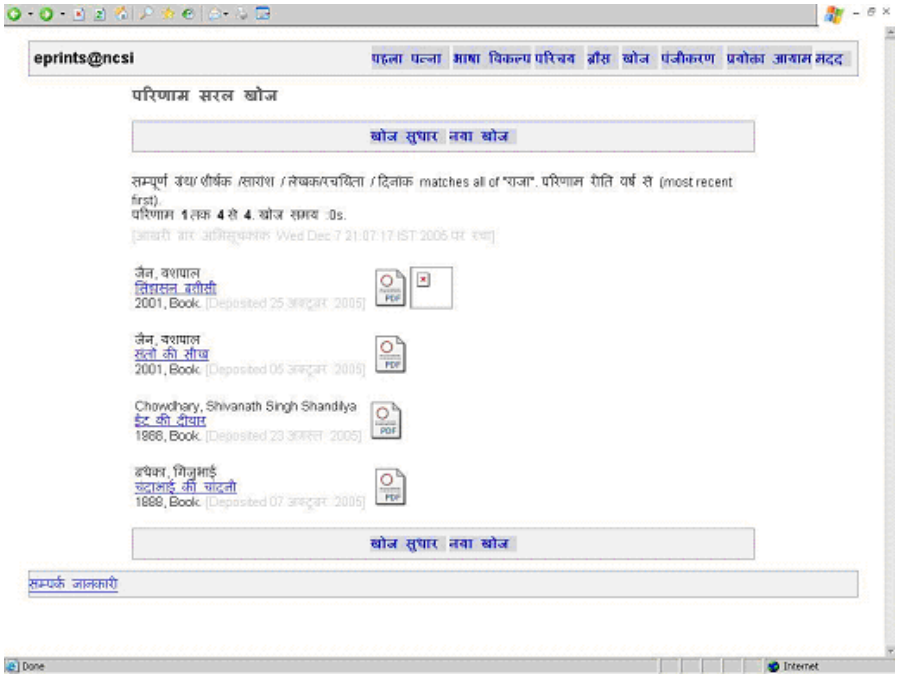
User Page



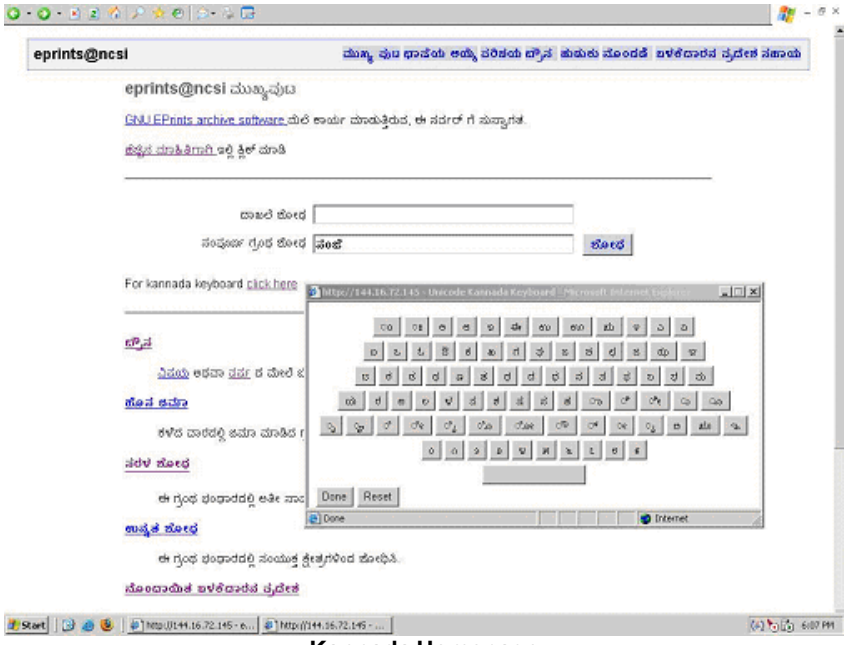
Browse page



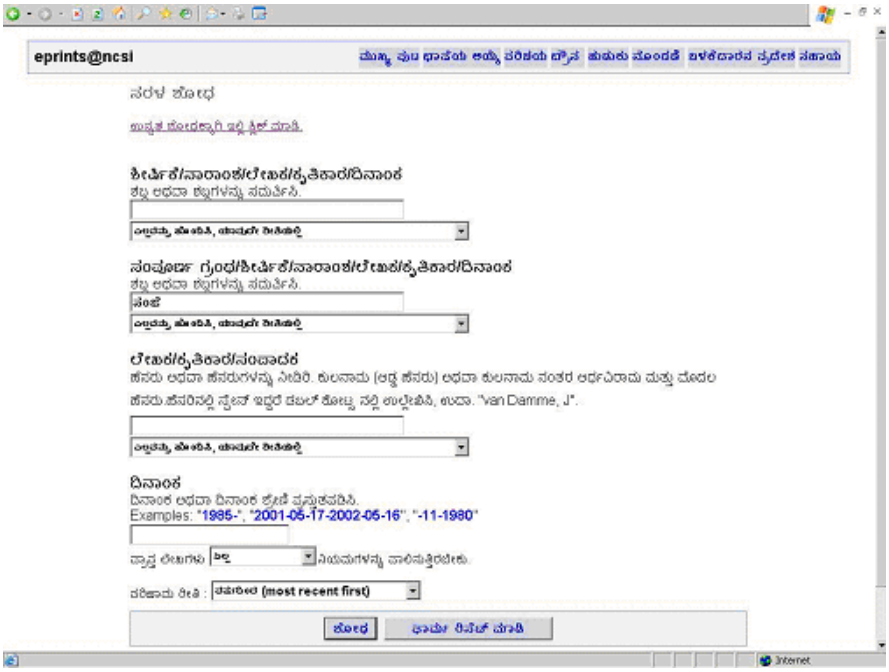
Search page



Search Result



Kannada Homepage



Kannada Search