GNU EPrints 2 Overview

Christopher Gutteridge

14th October 2002

Abstract

An overview of GNU EPrints 2. EPrints is free software which creates a web based archive and database of scholarly output and is intended for use by a university or university department. This article contains a description of the features and design philosophy of the software including OAI-PMH (Open Archive Initiative Protocol for Metadata Harvesting), internationalisation of the interface via XML files containing all text used in the interface, internationalisation of metadata fields which may be entered in more that one language and an API for adding your own features. This article also discusses the costs of setting up an eprints archive, including staff time and hardware.

1 What is GNU EPrints 2?

EPrints 1 was created to facilitate authors self archiving their work[10]. EPrints creates an on-line archive of "things", consisting of metadata and files, which may be contributed via the web. These "things" are usually scholarly material such as research papers and thesis, but could by anything from recipes to recordings of the Grateful Dead. EPrints also makes the data about records in the archives available for harvesting by the OAI-PMH interface.

Drawing from experiences of EPrints 1 [11, 12] we have learned that most users will want to use the software in similar, but distinctly different ways. This means that the software can never be "right". Instead we have designed the software with configuration and customisation in mind.

2 Technology

EPrints 2 is written in PERL, and runs as an apache module (using mod_perl). This means that the configuration does not have to be reloaded to serve each request. There are also a number of command line tools to build and maintain an archive.

More than one EPrints Archive can be served from a single installation of GNU EPrints but this increases the amount of RAM required.

EPrints uses MySQL to store the metadata about records and users. The actual files in the archive are stored in the UNIX file-system. A script allows the SQL database to be exported in a more meaningful XML structure.

The configuration files are a combination of XML and PERL.

The core PERL modules of EPrints 2 have been written in such a way as to make it possible to write new command-line and CGI scripts without having to directly deal with the SQL back-end.

3 Structure

3.1 What is a GNU EPrints "EPrint"?

An EPrint is a single record in the system. An EPrint consists of

- System Metadata Fields such as eprint-id and depositing-user-id. These are required by the software.
- Archive Metadata Fields such as title, author and year. These contain the information useful to people viewing and searching the archive. These may be customised when the archive is created. The default set of fields are more or less a super-set of the BibTEX fields.
- One or more Documents. This may be configured to be zero or more if the administrator wishes.

At any given moment an EPrint is in one of 4 "buffers". These are

- User Work-Area. For records which are not ready for submission yet
- Editor Buffer (a.k.a. Submission Buffer). For records which are awaiting approval by an editor
- Archive. For records which are approved. These appear on the public website, on normal searches and via the OAI interface.
- Deleted. For records which have previously been in the Archive buffer. This causes them to be listed as "deleted" on the public website and via OAI.

3.2 What is a GNU EPrints "Document"?

A Document represents a single format of the EPrint. A Document consists of

- System Metadata Fields such as doc-id, owning eprint-id, format. These are required by the software.
- One or more files in a directory in the UNIX file system.

Most documents are a single file, but some may contain many, such as an HTML document with diagrams or even a collection of images of the same item from many angles.

4 Public Site (The bits which don't require a password)

The public part of a website created by EPrints has four main sections. All pages are wrapped in the same template.

4.1 Static pages

These are just plain web pages, including the home page, and help page. They can be completely customised.

4.2 Browse Pages

These pages are created by a script periodically. The default fields which may be browsed are year and subject. Each distinct value has a static URL which contains a list of all items in that year or subject. Additional views may be added, for example it may be useful to add a "project" metadata field and a browse-by-project section. Then a project may link to a URL containing all the records associated with that project.

The XHTML of the views can also be generated without the normal wrappers. This allows it to be dynamically included in other pages, such as the main website for a project.

4.3 Abstract Pages

There is one abstract page for each record in the main archive. These pages summarise the metadata of the record and link to the documents. The way this page is generated may, and probably should, be customised. These pages have static URLs so they may be linked to and discovered by search engines.

4.4 Dynamic Pages

The main dynamic page is the search (and advanced search) pages. There is also a "latest" page which lists all records added to the main archive over the past seven days.

5 Work-flow

This section describes the process of adding a record to the system.

5.1 Registration

Before contributing records to the archive, a user must be registered. The default configuration implements this via a form on a web page. The user enters a requested username, password and their email address. EPrints emails them a confirmation code to activate the password, thus ensuring that the given email address is valid and owned by the individual requesting the account.

The web-based registration may be disabled in favor of some other method. The University of Southampton Department of Electronics and Computer Science EPrints Archive[4] has a script which once an hour updates the users registered in EPrints from the department staff database.

5.2 User Information

After registering a user is asked to enter information about themselves such as name, institution etc. The data fields stored in the user record may be customised. The only field required, by default, is "name". If not registering users via the web this information could be imported rather than entered by hand.

5.3 User Area

The user area page gives a list of all functions available to the user. Certain users may be designated editors or administrators and will consequently see more options in the user area page.

The user area page, and the functions available from it, will require the user to log in using a username and password. By default the password is stored in the local database, but the system may be modified to use other authentication methods, such as LDAP.

5.4 Depositing a Record

From the user area a user may create a new EPrint record. They first specify a type, such as book, article, tech report. This affects what metadata fields are requested and which are required. They then fill in the metadata and may add one or more documents to the record.

Once they are finished, and the metadata appears valid, they may "deposit" the record. This moves it into the editorial buffer, awaiting approval.

5.5 Approving a Record

A user who has been given the powers of an editor may view the submission buffer. An editor may move records in the submission buffer to the main archive, modify them or return them to the depositing users work area, with a comment (sent by email) about what the problem is.

5.6 Subscriptions

A registered user may store searches to be run once a day, week or month. They will be emailed a list of any new records which match each search.

5.7 Administration Options

An administrator may search and modify all the information in the database, including user records. This is how normal users may be turned into editors.

6 Internationalisation

6.1 Unicode

EPrints uses UTF-8 (and encoding of Unicode) to store all metadata. All web pages are processed as XML and therefore internally Unicode. All configuration files, which contain information which will appear on the website or emails, are in XML. This means that GNU EPrints will have no trouble handling characters which are not in the commonly used ISO-LATIN-1 character set.

Of course, if you have Greek or Chinese characters in your data or site they will only be rendered correctly if the person viewing the site has the correct fonts installed.

6.2 Phrase Files

Internationalisation has been a key design focus in EPrints 2. All the phrases used by the system are stored in XML files which can be easily customised and translated. An single phrase is configured like this:

```
<ep:phrase ref="matchrange"><p>Searching for
matches in field <b><ep:pin ref="fieldname"
/></b> between <ep:pin ref="fromvalue" /> and
<ep:pin ref="tovalue" />.</p></ep:phrase>
```

The ep:pin tags will be substituted with appropriate values. Using these placeholders makes translation much easier than translating “Searching for matches in field”, “between”, “and” and “.”. It means that the entire sentence can be rearranged.

Note the use of XHTML elements. A useful side effect of the way the phrase file is implemented is the ability to include mark-up. In some phrases, such as the subject line of an email, XHTML mark-up is meaningless and will be ignored.

6.3 Multilingual Site

EPrints may be configured to run in more than one language at once. Configuration files which control the look of the site may be duplicated for each language. These include the website template, the static web pages, the phrase files, and the citation rendering configuration. A cookie determines which version of the site a person sees. At least one live site is using this feature[13].

6.4 Multilingual Metadata

It is possible to configure a field to be multilingual. This is most useful for titles and abstracts. For example we make the abstract field multilingual and require an

English version. There will be an option enter additional translations of the abstract in other languages. This could be implemented by just having two fields abstract_en and abstract_gr but this solution is more descriptive.

7 Open Archives Protocol

The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [5] is a system which allows a specialist search engine (a.k.a. Harvester) to query your archive and obtain a list of all items in the archive, and dublin core metadata about each record. Subsequent harvests will only need to query records which have been added or modified since the previous harvest.

Configuring and registering the OAI interface will increase the visibility of your records as people will be able to search across many, or all, OAI compliant archives rather than having to search each archive in turn.

EPrints 2 supports OAI-PMH version 1.1 and 2.0.

At the time of writing there are about 12 OAI harvesters[7, 6].

8 Cost

EPrints is distributed freely under the GNU Public License. This does not just mean that it costs $0.00, but also that it may be freely modified and distributed. Southampton University owns the entire copyright and at a later date may release the software under a different license *in addition* to the GPL. All software on which EPrints 2 depends is also available at no cost.

EPrints is part of the GNU project[8]. This means that it conforms to most of the GNU guidelines for free software and GNU standard for software, such as installation and command line options.

The software is free, but this does not mean it costs nothing to implement a working archive. Both hardware and staff time cost money.

8.1 Hardware Costs

For testing purposes any spare GNU/Linux (or other UNIX) machine will do. For a production system I would recommend:

- 1 GHz or more PC running GNU/Linux, or equivalent UNIX server
- 512Mb or more of RAM
- 20Gb or more of disk space with possibility of adding more if your archive gets big.

In the United Kingdom, at the time of writing, such a server would cost about £1000.

Also remember to budget for a backup strategy.

8.2 Staff Costs

In theory a UNIX administrator with experience in installing software, and some understanding of Apache and PERL could get an archive up and running in a day. In practice almost every site running EPrints has wanted to make customisations. Some sites have made only "cosmetic" changes, such as text and colour, other sites have made extensive customisations including building additional scripts. This is not due to a failing in the software, just the opposite. EPrints is designed so the local extensions may be added. For example interfacing with a sites local databases.

You may also wish to enter existing records. Data entry takes a significant amount of time. 5 minutes per record does not sound that long, until you have 500 records.

Configuring and customising the system is usually an iterative process much like designing a website. Once this initial stage is complete the only costs are the time taken by users to deposit and editors to check and approve (or reject) records. At this stage little UNIX administration will required, accept to perform occasional patches and upgrades.

9 Further Information.

For more information on the software see the EPrints website[1] and try the demonstration server[3]. For more information on the project, related projects, and the motivations behind the software see the main eprints.org site[2].

For more information and advice on the process of setting up an archive, please see my article on the subject[9].

References

[1] EPrints Website, including documentation
http://software.eprints.org/

[2] University of Southampton eprints.org project, and related projects
http://www.eprints.org/

[3] EPrints demonstration server
http://demoprints.eprints.org/

[4] University of Southampton Department of Electronics and Computer Science EPrints Archive
http://eprints.ecs.soton.ac.uk/

[5] Open Archives Initiative (OAI)
http://www.openarchives.org/

[6] List of OAI Service Providers
http://www.openarchives.org/service/listproviders.html

[7] Citebase. A search across multiple OAI sources.
http://citebase.eprints.org/

[8] GNU Project
http://www.gnu.org/

[9] Gutteridge, Christopher and Harnad, Stevan. Applications, Potential Problems and a Suggested Policy for Institutional E-Print Archives (preprint).
http://eprints.ecs.soton.ac.uk/archive/00006768/

[10] Self-Archiving FAQ for the Budapest Open Access Initiative (BOAI)
http://www.eprints.org/self-faq/

[11] Stephen Pinfield, Mike Gardner and John MacColl "Setting up an institutional e-print archive" 11-April-2002, Ariadne Issue 31
http://www.ariadne.ac.uk/issue31/eprint-archives/intro.html

[12] William Nixon "The evolution of an institutional e-prints archive at the University of Glasgow", 8-July-2002, Ariadne Issue 32
http://www.ariadne.ac.uk/issue32/eprint-archives/intro.html

[13] Papyrus - Institutional Eprints Archive of Universitéde Montréal
http://papyrus.bib.umontreal.ca/