eprints.org

GNU EPrints 2 Overview

Christopher Gutteridge

19th October 2002

CERN. Geneva, Switzerland.

Christopher Gutteridge

- Systems, Support and Research
- Department of Electronics and Computer Science, University of Southampton
- Inherited EPrints Version One

- support@eprints.org

What is GNU EPrints 2?

- Creates Online Archive
- Free Software
- OAI Compliant
- Targeted at Scholarly Material
- Adaptable

What is an EPrint?

- **System Metadata**
    - ☐ Eprint ID Number
    - ☐ Deposit date
- **Archive Metadata**
    - ☐ Title
    - ☐ Year
    - ☐ Authors
- **Zero or more Documents**

What is a Document?

- **System Metadata**
  - ☐ Document ID
  - ☐ Format

- **One or more files (in the filesystem)**

Workflow 1

- New user registers via web
- User is emailed a confirmation URL
- User visits URL to activate account

An alternative is to auto-create user accounts from your own institution database.

Workflow 2

- User configures personal metadata
  - Name
  - Address
  - Research interests

The Administrator may add remove and modify user-fields.

Workflow 3

- User creates a new EPrint

- Enters metadata

- Uploads documents

- Views preview

- User "deposits" EPrint

EPrint is now in "submission buffer"

Workflow 4

- **Editor views deposited EPrints**

- **Editor may:**
    - ☐ Approve EPrint into main archive
    - ☐ Edit EPrint
    - ☐ Return EPrint to user with email explaining problem
    - ☐ Delete EPrint

Workflow 5

- EPrint is now "Live"
- EPrint appears in searches
- EPrint appears in "views"
- EPrint has "abstract" page with metadata summary & links to documents
- Abstract page has static URL
- EPrint appears in OAI

How much will it cost?

- Free?
- Still need hardware & staff
- Not just $0.00
- Not free as in "lunch", but free as in "Willy"

Hardware

- Machine running UNIX (eg. Linux)
- 512M+ RAM
- 20Gb+ of disk space
- Remember to budget for backups!

A demo version could run on a much smaller machine than this.

Staff Costs

- **2 or 3 days of UNIX admin?**
  - A few hours of UNIX admin each configuration change
  - Months to be ready, need the admin on-and-off
- **Data Entry can be expensive and error-prone**
- **Low long-term costs are much lower**
  - Support to teach staff to enter their own records
  - 2 – 3 minutes to approve/reject each record

Design Philosophy

- It's never going to be perfect!
- Default is a good starting point
- Will need some changes
- EPrints is configurable rather than "perfect" out of the box
- Possible to build new tools & views

Technology

- **Any UNIX-like OS**
- **Developed under RedHat's GNU/Linux distro.**
- **Written in PERL**
    - □ Not very efficient in terms of RAM
    - □ Allows rapid development and modification
- **XML for import, export and parts of the configuration**
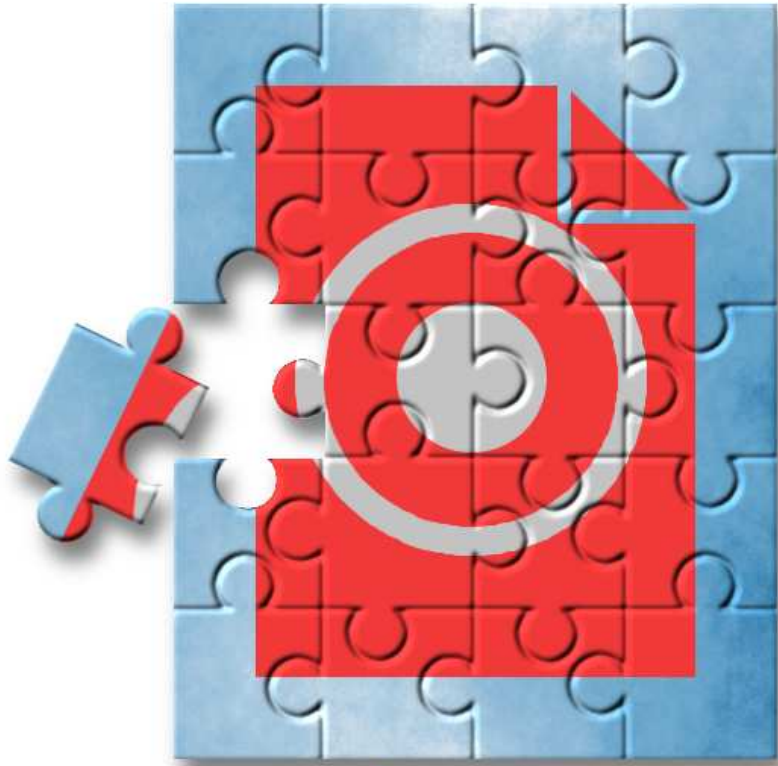- **Apache and mod_perl**
- **MySQL database**

Internationalisation

- Core Design Issue

- XML Phrase File for easy translation

- Multilanguage support

- Uses UTF-8 encoding throughout (Unicode)

- Option of Multilingual Metadata

Other Features

- OAI 1.1 and 2.0 support

- Metadata search

- Subscriptions

- Export metadata as XML

Setting up an EPrints Archive

Some general advice, not limited to the GNU EPrints software

eprints.org

Decide what it's for

- Don't just assume that everyone assumes the same assumptions as you

Applications & Implications

- Long-term Preservation
- Online presence to make the documents available to people now
  - OAI
- Metadata applications
  - CV Generation
  - Reports
  - Statistics
- Research applications

Long-Term Archiving

- Future-Proof document format

- Backups are essential

- Most projects only have 1-5 years of funding

- Need a long-term plan

Making it available Online

- Activate the OAI interface
- Easy to view format:
    - Plain Text
    - HTML
    - PostScript
    - PDF
- Visible to Search Engines

Metadata Applications

- CV Generation
  - ☐ Page with all an authors papers listed
  - ☐ Good "carrot"
  - ☐ Extra records required (beyond the scope of your archive?)
- Data must be accurate and complete
  - ☐ More accurate = more staff costs
  - ☐ Requires all records, even if no full text

Interactions between Goals

- More staff time required to enter and check records

- More complex metadata

- May require multiple formats for each document (screen format and archive format)

More Information

- GNU EPrints Website
    - □ http://software.eprints.org/
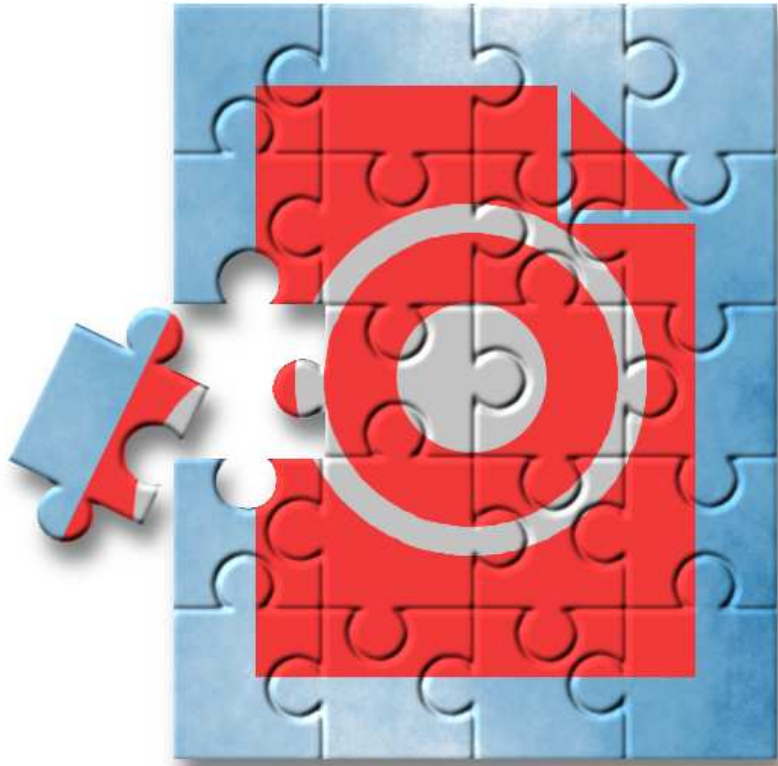
- EPrints Project Website
    - □ http://eprints.org/

Further Reading

- **GNU EPrints 2 Overview**
    - □ http://eprints.ecs.soton.ac.uk/archive/00006840/

- **Applications, Potential Problems and a Suggested Policy for Institutional E-Print Archives**
    - □ http://eprints.ecs.soton.ac.uk/archive/00006768/

Thank-you.

Christopher Gutteridge

19th October 2002

CERN. Geneva, Switzerland.