Open Repositories 2007 – EPrints User Group

S.Antonio, Texas - USA *23/26* January 2007

A World-Wide Repository: The Technical Challenges of E-LIS

Zeno Tajoli – tajoli@cilea.it

The key points

- More beyond Latin 1
- What done for editors
- What done for submitters (authors)
- What more for end users
- SQL scripts and tuning
- Statistics done in batch way
- What we expect from EPrints 3

Where Scripts, fixes and patches are

http://eprints.rclis.org/softw.html

More beyond Latin 1

- **Modifications to use DOM module**
  - E-LIS uses DOM, not GDOME
  - Description in http://www.eprints.org/tech.php/1948.html
  - Patch in http://eprints.rclis.org/fixsoft/XML.pm.gz

More beyond Latin 1

- **The simplification on the search can't be used**
  - E-LIS has records in different scripts.

The standard simplification is not correct.

  - The explication: http://www.eprints.org/tech.php/2418.html
  - The patched file: http://eprints.rclis.org/fixsoft/Name.pm.gz

More beyond Latin 1

- **Too long file names in browsing for non-Latin scripts**
  - An hack on the subroutine that generates file names solves the problem.
  - You must have a file system that supports utf-8 in file names (like ext3)
  - The hacked routine (escape_filename in EPrints::Utils.pm): http://eprints.rclis.org/fixsoft/Utils.pm.gz
  - The explication: http://wiki.eprints.org/w/Files/FileNamesUTF8

More beyond Latin 1

- **Problems on indexing non-ASCII chars**
  - We are still working on the problem
  - No one knows every script in the word
  - A draft solution here:

http://wiki.eprints.org/w/Files/IndexNoLatin

What done for editors

- **Show all metadata without logging**
  - In the splash page there is a link “Show all fields”
  - The linked page shows all metadata
  - To check metadata more quickly
  - Instruction and configuration: http://wiki.eprints.org/w/Files/ShowAll
  - The code: http://eprints.rclis.org/fixsoft/showall.tar.gz

What done for editors

- **Submission buffer-page with languages**
  - Multi-language country
  - Editors don’t know all languages
  - To see immediately the situation of the paper
  - Technical discussion:

http://wiki.eprints.org/w/Files/SubBuffLang

  - The code:

http://eprints.rclis.org/fixsoft/buffer.gz

What done for editors

- **A Bcc when a paper is rejected**
  - When editors reject a paper they send a mail to the submitter
  - Editors want a copy of this mail
  - To do this we do an hack on the edit_buffer cgi
  - Technical discussion: http://wiki.eprints.org/w/Files/EditBufHacks
  - The code: http://eprints.rclis.org/fixsoft/edit_eprint.gz

What done for editors

- **A form to avoid spam**
  - We don’t insert e-mails of editors in the staff page
  - But we want to connect authors and editors
  - We use a PHP form
  - Credits: Rodríguez-Gairín, Josep-Manuel
  - Available on request
  - Technical info:

http://wiki.eprints.org/w/Files/EditorForm

What done for editors

- **More browsing views**
  - Some views are provided to help editors to check metadata
  - Conference
  - Book or Journal
  - Setup in the usual configuration

What done for editors

- **The special field "country"**
  - In the bibliographic metadata there is a field "country"
  - Optional, repeatable
  - It registers the countries of the authors
  - Every editor has a submission buffer that is filtered by one or more countries
  - Setup with the usual configuration

What done for submitters (authors)

- **An alert when the paper is online**
  - Some submitters want to know when their papers are gone on-line
  - The functionality is optional, as default it is not active.
  - When it is active, the submitter receives a mail
  - Technical discussion: http://wiki.eprints.org/w/Files/EditBufHacks
  - The code: http://eprints.rclis.org/fixsoft/edit_eprint.gz

What done for submitters (authors)

- **As few pages as possible**
  - In the submission process we compact the pages.
  - It seems that submitters want few pages
  - Done with standard configuration

What done for submitters (authors)

- **FAQ, Help and more**
  - The editorial staff do much work to help the submitters.
  - They write specific help, faq and tutorial on submission, copyright and other topics on static web pages
  - They answer to **many** specific requests

What more for end users

- **URLs are the best links in the reference**
  - Many references have URLs inside.
  - This version of Paracite and Paratools uses URL as first search.
  - Code and configuration:

http://files.eprints.org/48/

  - Credits: Alessandro Tugnoli for CILEA

What more for end users

- **Adding abstract field in alerts**
  - More info in alerts
  - With the abstract field is easier to understand the topic of the paper
  - No need for a huge citation
  - You need to modified Eprints:Subscription.pm
  - The configuration: http://wiki.eprints.org/w/Files/AbsIntoAlerts

What more for end users

- **Count the papers**
  - Many users want to know how many papers are into archive
  - A dynamic solution with a SSI
  - Inserted into the home page
  - Code and configuration:

http://files.eprints.org/47/

What more for end users

- **List the last 8 papers in the home page**
  - The latest update is important for users
  - With the standard tools there are the latest 20 papers with RSS and latest week with a cgi
  - We wrote a special SSI – starting from code of Aneesh Joy
  - Technical discussion:

http://eprints.rclis.org/fixsoft/whatsnew.pl.gz

  - The code:

http://eprints.rclis.org/fixsoft/whatsnew.pl.gz

SQL scripts and tuning

- **Check subjects**
  - To detect the bad subjects in our Eprints
  - At the end you have a list of all eprintsid with bad subjects
  - The code:

http://files.eprints.org/35/

SQL scripts and tuning

- **Metadata with full-text**
  - To check if metadata are connected with at least one full-text
  - To ask full-texts to old submitters
  - Now the archive is set with full-text mandatory
  - The code:

http://eprints.rclis.org/fixsoft/check-vuoti.pl.gz

SQL scripts and tuning

- **Delete the false users**
  - Many robots on the web create “dummy” users
  - The registration could then be “false”
  - The script deletes incomplete users after one week
  - The code: http://eprints.rclis.org/fixsoft/erase_users_unfinished.pl.gz

SQL scripts and tuning

- **To delete "passive" users**
  - A relevant number of people register themselves but they don't do anything
  - No alerts
  - No upload
  - They are deleted once per year
  - The code: http://eprints.rclis.org/fixsoft/elimina-utenti-passivi.pl.gz

SQL scripts and tuning

- **List users e-mail addresses**
  - To create a list of e-mail addresses
  - To send a message to every user
  - It is possible to extract more data for statistic purposes
  - The code:

http://eprints.rclis.org/fixsoft/estrai-email.pl.gz

SQL scripts and tuning

- **To delete a specific eprint**
  - To purge buffers from errors
  - It works on command-line level
  - As input it requires an eprint id
  - The code:

http://eprints.rclis.org/fixsoft/elimina-doc-morti.pl.gz

SQL scripts and tuning

- Use MySql 4.x for the cache
- Attention with indexer and generate_views
- Monitoring CPU load

Statistics done in batch way

- **Tasmania software doesn't fit E-LIS**
  - It uses dynamic pages with PHP
  - And it generates a too huge load
  - We generate static pages one time every night
  - Done with Perl

Statistics done in batch way

- **To purge logs from robots**
  - We use the ‘user-agent’ value of apache log
  - We built a list reading who calls the page ‘robots.txt’
  - Many person call robots.txt with a browser
  - We need to check the list by hand
  - Done every 3 months

Statistics done in batch way

- **Data warehouse**
  - We insert data about downloads and abstract views only
  - The downloads of the same paper need to have a span of 180 seconds
  - The same for abstracts views
  - Technical discussion:

http://wiki.eprints.org/w/Files/BatchStats

  - The code:

http://eprints.rclis.org/fixsoft/stats.tar.gz

What we hope from Eprints 3

- More documentation on API
- To use AJAX to control metadata during submission
- Support for Creative Commons licenses
- More support for multi script pages (Arabs chars with Latin numbers, unusual Asian languages like Nepali)
- More flexible indexing

We have finish !!

Questions ?

Thank for your attention

Code written by Zeno Tajoli. Some code written by Chris Gutteridge, Aneesh Joy, Rodríguez-Gairín Josep-Manuel, Alessandro Tugnoli.