Configuring EPrints

EPrints successfully runs on
- Linux
- Solaris
- Mac OS-X
- Grant from Microsoft for Windows version
- Installation process is standardised
- One site installation can run many separate repositories
  - mixing and matching EPrints and other Web-based services on the same host is possible

EPrints - the Administrator's View

- SQL database
- Web server
- Scripts to configure repository activities
- Configuration files

EPrints Home Directory

- Global configuration directory for the running EPrints server
- Holds the config subdirectories for each active archive (ie repository)
EPrints Home Directory

Directories for storing programs

EPrints documentation

The template for all new archives

Holds all the modules required by Perl scripting language

Temporary files
Contents of archives Directory

A subdirectory for every active repository

Plus a configuration file (XML) containing all the important information from the initial archive generation commands

Contents of individual archive directory

The configuration files for this archive

Contents of individual archive directory

PDFs etc.

Processed static webpages
Contents of individual archive directory

Temporary files

Contents of individual archive directory

Contents of individual cfg directory

...there are LOTS of configuration files!
XML, DTD, Perl modules, apache...
Rather than examine each individually, consider some common configuration tasks
Branding
Adding a new deposit type
Adding a new metadata field
One step at a time
examining cause and effect
not how you would normally do things!

Contents of individual config directory (2)

Task 1: Branding

The first thing most institutions do is brand their repository and fit it in with their existing look and feel

Branding: Which Configuration Files?

Site-wide HTML

template-en.xml

Site-wide HTML

template
Branding: Which Configuration Files?

- **static/en/*.xpage**
  - fixed content pages e.g. homepage, about page, help page
- **static/general/**
  - images and stylesheets

Aside: What is XML?

- Quick answer: it's a bit like HTML
  - HTML is for making pages for people to read
  - XML is for making data for computers to use
  - The syntax is very similar, just stricter
  - All tags must have a matching closing tag
  - All attributes must be quoted
  - It doesn't know anything about Web pages
  - or anything else, come to that!

Branding: *template-en.xml*

- EPrints lets you define an HTML template (outline) which is used to build every Web page
- Customise the look and feel of the whole site
  - header
    - title, logo, navigation menu
  - footer
- pins tell EPrints where on the page the title and page contents should be placed

Branding: *static directory*

- You can also define fixed content for:
  - homepage, “about” page, help pages, “error” page...
- These configuration files are stored in the **static** directory
  - one subdirectory per language
  - e.g. English files go in **static/en**
- Images and stylesheets (and other language independent files) are stored in **static/general**

Aside: *static/en/index.xpage*

```xml
<?xml version='1.0' encoding='iso-8859-1' standalone='no' ?>
<op:page
  xmlns:op='http://op.eprints.org'
  xmlns:xhtml='http://www.w3.org/1999/xhtml'
  title='Welcome to &archivename;'></op:page>
</op:body>
```

Welcome to Southampton Repository

Welcome to this generic eprints server, running the default configuration for GNU EPrints repository software.
Branding: Add a Logo

- Add the University of Southampton logo
  - first copy logo.gif into static/general/images
  - then add the logo to the header in template-en.xml

```
<div class="header noPrint">
  <table cellpadding="5" border="0" cellspacing="0" width="100%">
    <tr>
      <td>…

      <div class="archivename">Archivename;</div>
      -->
      <img src="$base_url/images/logo.gif" />
    </td>
  </tr>
</div>
```

Aside: entities-en.dtd

- Notice the entities in template-en.xml
  - &archivename;
  - &base_url;

- These are defined in entities-en.dtd
  - generated automatically by EPrints
  - definitions of character symbols e.g. copyright
  - contains useful symbolic names for various URLs and email addresses
  - lets you avoid hard-coding names and URLs

Aside: what is a DTD?

- A DTD is a definition file for XML
  - XML is just a naked standard for the syntax rules of a document or data file
  - A DTD provides it with the definitions needed for a particular vocabulary (e.g. HTML)
  - It also defines names for non-ASCII characters (e.g. copyright, euro, bullet)

Aside: generate_static Command

- We need to run the generate_static command
  - This takes the fixed content (.xpage) files in the static directory and wraps them in the template
  - The resulting HTML pages are written to the repository's html directory
  - myarchive/cfg/static/en/index.xpage
    - becomes
  - Images and stylesheets copied across as well

Branding: Check the Homepage

- Check the homepage... no logo!

Branding: generate_static Command (2)

- Why?
  - these pages hardly ever change (hence "static")
  - for best server performance, serve static html pages
  - but, want to maintain "master" site template in single file
  - 3.0 has hybrid approach:
    - pin dynamic bits of content onto a static page
    - e.g. login status (logged off, logged on as ...)
Branding: Check the Homepage

After running `generate_static`:

Welcome to the Southampton Repository
Welcome to the Southampton Repository, running the default configuration for EPrints repository websites.

Branding: Check the View Pages

But the logo isn't showing up on the browse view pages!

The browse view pages change much more frequently than the homepage etc. but EPrints also serves these as static HTML pages for performance often visited by crawlers e.g. Google.

To regenerate the view pages, we need to run `generate_views` this is usually run nightly, or even hourly.

Branding: Check the Search Page

But the logo isn't showing up on the search pages!

The search pages, and also user home page, deposit pages etc. are dynamic created on-demand by EPrints.

For best performance, EPrints loads the template into memory at startup dynamic pages are wrapped in this in-memory copy so when we change the template, we need to get EPrints to refresh its copy run `force_config_reload` or restart the Web server.
Branding: Check the Search Page

After restarting the Web server:

1. Copy logo image to static/general/directory
2. Add logo to template-en.xml
3. Regenerate static pages: generate_static, generate_views, generate_abstracts
4. force_config_reload to show logo on dynamic pages

Branding: Summary

- Copy logo image to static/general/directory
- Add logo to template-en.xml
- Regenerate static pages: generate_static, generate_views, generate_abstracts
- force_config_reload to show logo on dynamic pages

Task 2: Adding a Deposit Type

- EPrints is pre-configured with several default deposit types
  - Article, Book, Book Section, Conference Item, Monograph, Other, Patent, Thesis
  - modeled on most common research outputs
- Each deposit type has a set of metadata associated with it
  - title, creators, editors, date of publication, abstract...
- And a set of document formats
  - PDF, PostScript, HTML, plain text

Task 2: Adding a Deposit Type

- Many institutions have other types of (research) output or collections/artefacts
  - data, teaching materials, multimedia
  - e.g. the University of Southampton has:
    - a School of Art
    - a Textile Conservation Centre
    - a Music division in the School of Humanities
- What kinds of deposit might be needed?
  - metadata fields? document formats?

Some Suggestions

- New Deposit Types:
  - Composition, Performance, Show/Exhibition, Artefact
- Metadata:
  - composers, conductor, medium (oil, pencil, ink, watercolour, gouache, marble, clay, scrap metal...), producer, sound engineer, commissioning body, creation dates, venues/dates, genre (opera, jazz...)
- Document formats:
  - image (JPG, TIFF...), audio (MP3, WAV, FLAC...), 3D model (?)

New Deposit Type: Which Config Files?

- metadata-types.xml deposit types and workflow, document formats
The configuration file which describes the deposit types is `metadata-types.xml`.
- different types of eprint (deposit type), user (users, editors, administrators) and document (PDF, PS, ...).
- defines the metadata fields that apply to each type.
- defines the order that the fields will appear in the deposit workflow.
- defines how the fields will be grouped into pages in the deposit workflow.
- 3.0 adds conditionals to workflow
  - e.g. different workflows for different departments.

Add a New Deposit Type
- Add a simple Composition deposit type
- use existing metadata fields for now
- Restart Web server to re-read configuration files

```xml
<type name="composition">
  <field name="title" required="yes" />
  <field name="creators" required="yes" />
  <field name="abstract" />
  <field name="abstract" />
  <field name="keywords" />
</type>
```

New Deposit Type: Check List of Types
- Begin a new deposit
  - the text for the Composition option looks strange!

**Item Type**
- Please select the most appropriate type for your deposit.
  - [ ] eprint_tlmnue_composition (not defined)
  - [ ] eprint_tlmnue_composition (not defined)
  - [ ] Article
    - An article in a journal, magazine, newspaper. Not necessarily peer-reviewed as an inline journal or news website.
New Deposit Type: **phrases-en.xml**

- EPrints needs to know how to display the type.
- The **phrases-en.xml** configuration file is where all the phrases which appear in the EPrints Web interface are defined.
- Each **ep:phrase** element has a **ref (id)** often structured: `eprint_fieldname_abstract`

Why?
- phrases are not embedded in EPrints code
- single file for editing phrases
- referring to phrases by id enables multi-language support

---

New Deposit Type: **Add Phrases**

- Add phrases for the Composition deposit type

```xml
<ep:phrase ref="eprint_typename_composition">Composition</ep:phrase>
<ep:phrase ref="eprint_metadata_type_composition">A musical composition.</ep:phrase>
```

Item Type:
- Please select the most appropriate type for your deposit.
  - **Composition**: A musical composition.
  - **Article**: An article in a journal, magazine, newspaper. Not necessarily peer-reviewed online journal or new website.

---

New Deposit Type: **Check Citation**

- As you work through the deposit process, EPrints displays the “citation” at the top of the screen.
- this shows you how the citation will appear on other pages.
- For our new deposit type, we get an error.

Core Bibliographic Information

Current item: Error: Unknown Citation Style "eprint_composition"

Please enter the information about this item. Fields marked with a * are fields that must be accepted.

---

New Deposit Type: **citations-en.xml**

- The citation style for each deposit type is defined in the **citations-en.xml** configuration file.
- Very powerful and flexible but a bit hard to read.
- Add citation style for Composition and restart.

```xml
<ep:citation type="eprint_composition">
  <span class="citation">
    @creators#<ep:linkhere><em>title:magictop</em></ep:linkhere>  
    <ep:ifset name="keywords">(0keywords)@ep:ifset>  
  </span> <!--citation-->
</ep:citation>
```

---

New Deposit Type: **Check Citation**

- Citation OK

But default deposit formats not helpful!
New Deposit Type: Document Types

- Adding extra document types is a similar process to adding a new deposit type
  - add extra formats to `metadata-types.xml`
  - add phrases to `phrases-en.xml`
    - `document_typename_mp3`
    - `document_typename_wav`
  - also need citations
  - Can now deposit MP3/WAV
    - but also need to configure required document formats for Compositions

You must upload at least one of the following formats: HTML, PDF, Postscript, ASCII.

New Deposit Type: Add Formats

- Add new document types to the list of required formats
  - `$c->{(required_formats)} = [``html``,
    ```pdf``,
    ```ps``,
    ```ascii``,
    ```mp3``,
    ```wav``,``];`

You must upload at least one of the following formats: HTML, PDF, Postscript, ASCII, MP3, WAV.

New Deposit Type: ArchiveConfig.pm

- Required document upload formats is just one of the many settings in the `ArchiveConfig.pm` configuration file
  - Perl syntax, but easy to change simple things
    - skip submission buffer
    - web signup for depositing users
    - metadata input defaults
    - submission form customisation
    - definition of browse views, search forms and user privileges

New Deposit Type: Test Deposit


New Deposit Type: Summary

1. Add type definition and any extra document formats to `metadata-types.xml`
2. Add phrases for type and formats to `phrases-en.xml`
3. Add citations for type and formats to `citations-en.xml`
4. Specify which formats are required for each type in `ArchiveConfig.pm`
5. `force_config_reload`

Task 3: Add a New Metadata Field

- Continuing our theme, add an extra field to the Composition type called `composition_genre`
New Metadata Field: Which Config Files?

- metadata-types.xml
  - which fields apply to which types

ArchiveMetadataFieldsConfig.pm
- defines type and properties of all fields

New Metadata Field: Which Config Files?

- phrases-en.xml
  - display names and help text for fields, display names for field options

New Metadata Field: Check Workflow

- But when we restart the Web server...

Task 3: Add a New Metadata Field

- Add the new field to metadata-types.xml

```xml
<type name="composition">
  <page name="core"/>
  <field name="title" required="yes"/>
  <field name="composition_genre" required="yes"/>

  <page name="abstract"/>
  <field name="abstract"/>
  <field name="keywords"/>
</type>
```

New Metadata Field: ArchiveMetadataFieldsConfig.pm

- We've used a field in metadata-types.xml that EPrints doesn't know about
- All metadata fields must be defined in the ArchiveMetadataFieldsConfig.pm configuration file
New Metadata Field: ArchiveMetadataFieldsConfig.pm

- **ArchiveMetadataFieldsConfig.pm** defines:
  - types and properties of all metadata fields for eprints, users and documents
  - e.g. creators, title, abstract
  - default field values
  - automatic metadata fields e.g. calculating the number of authors
  - Perl intensive

New Metadata Field: Add Definition

- Add a definition for the `composition_genre` field to ArchiveMetadataFieldsConfig.pm

```perl
{
  name => "composition_genre",
  type => "set",
  options => [ "classical", "rock", "jazz", "blues", "folk", "world", "pop", "easy" ],
},
```

New Metadata Field: Check Workflow

- Web server restarts OK, but

**Internal Server Error**

The server encountered an internal error or misconfiguration and was unable to complete your request.

Please contact the server administrator, webmaster@eprints.org and show them the error message and all details from the server logs.

More information about this error may be available in the server error log.

Apache/2.2.11 (Fedora) Server at fig.eprints.org Port 80

New Metadata Field: Why it Failed

- EPrints uses the metadata configuration in ArchiveMetadataFieldsConfig.pm to:
  - construct its database tables
  - generate queries for selecting data from the database
  - EPrints expects to find a `composition_genre` column in the database

New Metadata Field: Update Database

- We need to either
  - rebuild the EPrints database tables for the new metadata configuration
  - will lose all data and uploaded files
  - use erase_archive and then create_tables
  - don't do this on a live repository!
  - useful development technique
  - add the field to the database by hand
  - won't lose any data
  - instructions for doing this on the EPrints wiki [http://wiki.eprints.org/w/Adding_a_Field_to_a_Live_Repository](http://wiki.eprints.org/w/Adding_a_Field_to_a_Live_Repository)

New Metadata Field: Check Workflow

- Field now appears in deposit workflow
  - Now just need to add some phrases!
    - field title and help text
    - name of each option
      ```
      ['eprint_fieldtitle_composition_genre' not defined] *
      ['eprint_fieldhelp_composition_genre' not defined] *
      ['eprint_fieldoption_composition_genre_pop' not defined] *
      ['eprint_fieldoption_composition_genre_classical' not defined] *
      ['eprint_fieldoption_composition_genre_classical' not defined] *
      ['eprint_fieldoption_composition_genre_jazz' not defined] *
      ['eprint_fieldoption_composition_genre_blues' not defined] *
      ```
1. Define type and properties of new field in ArchiveMetadataFieldsConfig.pm
2. Add field to deposit workflow in metadata-types.xml
3. Add display name and help text, and display names for each field option, to phrases-en.xml
4. force_config_reload
5. Erase and rebuild database
or manually add new field

New Metadata Field: Summary

Other Config Files: subjects
- Plain text file that defines the subject tree for the classification system
- By default contains the top 2 levels of the US Library of Congress classification
- "subjects" is actually a misnomer
  - other hierarchical classifications can be defined
  - our composition_type field could have taken its values from a hierarchy of musical genres

Other Config Files: ArchiveOAIConfig.pm
- Methods for handling Open Archive Initiative metadata harvesting protocol (OAI PMH)
  - main method eprint_to_unqualified_dc converts an EPrints data structure to an OAI structure
  - other informational definitions give policies etc
- This file should be extended for data archives, to allow non-DC information to be shared
- Perl intensive
- Hardly ever used (except for exotic data types)
  - we could expose composition_genre as dc:subject

Other Config Files: ArchiveRenderConfig.pm
- Methods for generating the abstract pages for each item
- Perl intensive
- eCrystals data repository heavily modified this configuration file
- We could embed an music player applet on each Composition page

Other Config Files: ArchiveValidateConfig.pm
- Methods for checking the metadata fields that a depositor is submitting
  - individual fields
  - a whole page (i.e. combination of fields)
  - a document (e.g. has the user submitted a format safe for preservation purposes?)
  - a complete eprint record
  - a user
- Perl intensive

Other Config Files: ArchiveTextIndexing.pm
- Methods for supporting free text indexing
  - definitions of lexical token separators
  - list of stop words
  - filter that translates a text into a bag of words
- Unlikely to be changed
### Web Server Config Files

- **auto-apache.conf** is the main workhorse
- defines where the archive files are, how to handle script requests and errors etc.
- usually not changed
  - some tweaks may be necessary if you are hosting other Web-based services on the same server

### Reflection: What do you need to do?

- Look back at the issues you raised for configuring EPrints
  - can you see where you would need to start working in the EPrints setup?
  - can you find some repositories which do things in the same way?
    - i.e. can you find someone to give you advice?