Meeting Institutional Requirements

EPRINTS 3.2
Outline: Session 1 & 2

- Introductions

- What is the point
  - Mission, motivation and drive
  - Where have we got so far
  - Where do we want to be
  - Where is version 3.2 taking us

- Outline of version 3.2

- Discussions
  - Future technical directions
  - Future community possibilities
  - Relationships with other communities
EPrints Org Chart

Leslie Carr
(Technical Director)

EPrints Core
- Chris Gutteridge
  Chief Architect
- David Tarrant

EPrints Services
- Tim Brody
  Release Manager
- Tim Miles-Board

Funded Projects
- Sebastien Francois
- Patrick McSweeney
- Jessie Hey
- Debra Morris
- Steve Hitchcock
- et al.
EPrints Services

- For-pay commercial hosting, bespoke development, training, support
  - two fulltime repository developers/supporters

- Provides resources to sustain EPrints O/S development
- Provides customer contact
- Contributes code back to O/S
- Now starting to fund PhD studentships
What is the Point?

- Universities and researchers are knowledge producers and knowledge consumers
- The Web has radically altered the potential for knowledge dissemination in society in ways that we haven’t yet fathomed
- We want to understand and facilitate that change
  - Research and development
The Point 1: Open Access

- Scholarly comms have been outsourced
- Literally nothing to show as evidence of research activities
The Point 2: Knowledge Management

- Researchers have just enough IT support for daily activity
- Many problems intervene
  - Disk crashes
  - Stolen laptops
  - Software upgrades that go wrong
  - Backups that never quite get restored
- Life intervenes
  - Draws and folders full of old stuff that eventually fall off the radar
- “Lost in some research assistant’s computer, the data are often irretrievable or an undecipherable string of digits”

EPrints History

- Open Archiving Initiative - October 1999
  - Originally called UPS
- Among the Participants
  - Paul Ginsparg (Los Alamos, arXiv)
  - Carl Lagoze (Cornell, NCSTRL)
  - Stevan Harnad (Southampton, Cogprints)
- EPrints
  - proposed as a ‘build your own repository’ solution
  - enable institutions and groups to participate in OAI metadata sharing initiative
EPrints History

- First release written by Rob Tansley
  - PhD student at Southampton
  - Left for HP & wrote DSpace

- EPrints funded under NSF / JISC DLI2
  - With Cornell, who were writing Fedora
An EPrints repository is a set of services offered to all the members of an institution for open access, preservation, scholarly collections, teaching, e-learning, e-publishing, data sharing, collaboration, administrative reporting, research management, publicity, marketing, digital profile enhancement, research assessment 
and many, many other agendas.
EPrints Objectives

Lower the barrier for depositors while improving metadata quality and ultimate collection value
- Time saving deposits
- Import data from other repositories and services
- Autocomplete-as-you-type for fast data entry
- Name authorities

Enter once, reuse often
- Works with bibliography managers, desktop applications and new Web 2.0 mashups
- RSS feeds and email alerts keep you up to date
- Easily integrate reports, bibliographic listings, author CVs and RSS feeds into your corporate web presence
- Used for corporate reporting and national Research Assessment

Simple platform for open source contributions
- Tightly-managed, quality-controlled code framework
- Flexible plugin architecture for developing extensions
Goals for an EPrints Repository

Quality

Value

Control
Running a repository is becoming a challenge in keeping up

Satisfy the demands of an entire institution

- diversity
- quantity
- quality
- deadline
EPrints 3: What Have We Got

- Everything builds on the bottom layer
- Major part of v3.2 is strengthening the bottom layer
  - Improved data model
    - Enhanced data facilities
    - Enhanced metadata facilities
  - Improved programming & API
EPrints Core

- Datasets
- Data storage
- Utility methods
  - Data types
  - Configuration control
  - Query language
  - Control language etc
EPrints Mantle

- External interactions
  - web pages
  - web services
  - other protocols
EPrints 3.2 Outline

- Improved Storage
- Strengthened Data Model
- Inter-object Relations
- Extended Data Repertoire
- User Improvements
- Management Improvements
EPrints 3.2 Outline

Strengthened Data Model

- Improved Storage
  - Storage Controller
  - Virtual File System
  - Preservation Analysis

- Extended Data Repertoire
  - Extendable Datasets
  - CRIS reporting
  - Citation Framework
  - IRStats

- Inter-object Relations
  - Complex Thumbnails
  - Coverpage Capabilities
  - OpenXML support

- Extended Datasets

- CRIS reporting

- Citation Framework

- IRStats

- Complex Thumbnails

- Coverpage Capabilities

- OpenXML support
Files are now data objects

- Previously, eprints & documents were data objects. Files were just items in a file system.

- Now files are data objects too, and file objects are abstractions of data streams.
Storage Controller

- Files can be stored on multiple storage services
  - Local disk, SAN, NAS, Honeycomb, Cloud
- Various policies can be enacted through an EPControl XML files.
- Services can be monitored and files swapped between services.
Storage Architecture

EPrints Core

Storage Controller

Local Disk

Cloud Storage

Service Provider
<choose>
  <when test="datasetid = 'document'">
    <choose>
      <when test="$parent{relation_type} = 'isVolatileVersionOf'">
        <plugin name="Local"/>
      </when>
      <otherwise>
        <plugin name="AmazonS3"/>
      </otherwise>
    </choose>
  </when>
  <otherwise>
    <plugin name="Local"/>
  </otherwise>
</choose>
Storage Manager

Amazon S3 storage
There are 217 total files stored using this back-end, taking 3126 Kb.
Documents: 217

Copy to
Delete Copies

Local disk storage
There are 289 total files stored using this back-end, taking 1649 Kb.
History: 289

Copy to
Delete Copies
Compressed local disk storage

There are 85 total files stored using this back-end, taking 293 Kb.
History: 85

Copy to
Delete Copies
Virtual File System

- A VFS module invents a view of the repository’s files and structures as a hierarchical file system.

- WebDAV and FTP servers are implemented to allow client desktops to mount the repository as a pseudo-filesystem.
Preservation

- Creating abstractions of the files allows better preservation management.

- A risk analysis process
  - catalogues all the available files and their file types
  - weighs the evidence for their preservability
  - assigns a score to each file
This EPrints install is referencing a trial version of the risk analysis service. None of the risk scores are likely to be accurate and thus should not be used as the basis for a program of action.

High Risk Objects
- OLE2 Compound Document Format
  - Risk Score: 1

Medium Risk Objects
- Microsoft PowerPoint Presentation (Version 97-2002)
  - Risk Score: 3

Low Risk Objects
- Portable Document Format (Version 1.4)
  - Risk Score: 3
- Portable Document Format (Version 1.3)
  - Risk Score: 2
- ZIP Format
  - Risk Score: 2
Object relationships already appeared in v3.1
Now we start to build new facilities using them
Complex Thumbnails

- Thumbnails were hidden files
- Now they are documents

- Now thumbnails can be different formats
  - png, jpg, flv...
- Thumbnails can be uploaded manually

Interobject Relations

Main Document

Preview Document

- haspreviewThumbnailVersion
- hasVolatileVersion
- ispreviewThumbnailVersionOf
- isVolatileVersionOf
Thumbnail Addressing

- Main document
  - http://eprints.com/123/4

- Thumbnail

- Chaining Relationships
Object IDs & REST

- All data objects have their own data set and id
  - http://eprints.com/id/eprint/123
  - http://eprints.com/id/user/123

- Type these into a browser URL and EPrints will redirect to the appropriate display object
  - e.g. the eprint that contains document id 123
Object IDs & REST

- EPrints objects are exposed as data for reading and writing through a Web interface

- [http://devel.eprints.org/cgi/rest/eprint/4.xml](http://devel.eprints.org/cgi/rest/eprint/4.xml)
  - The XML output of an eprint

- [http://devel.eprints.org/cgi/rest/eprint/4/creators/2/name/family.txt](http://devel.eprints.org/cgi/rest/eprint/4/creators/2/name/family.txt)
  - The text of the surname of the second author of eprint 4
A new document upload tab allows PPTX/DOCX files to be burst open

- Media files stored as related documents
- Metadata added to the eprint

Also happens with SWORD plugin
Potential application
- auto generation of copyright auditing form for teaching materials
- … or theses
Extendable Datasets

- A repository needs to be aware of more than just publications
  - users
  - projects

- The Current Research Information Systems (CRIS) community have a mature standard for describing this information
  - CERIF
Extendable Datasets

Extendable Datasets

- New Datasets can be created: e.g. project

Manage Metadata Fields

This tool allows you to add metadata fields to your repository. Only fields added using this tool can be removed by this tool. To start configuring fields choose a dataset to add/remove fields from the following list.

Datasets with Configurable Fields

- **Eprints:** Used to store eprints records
- **Documents:** Used to store documents metadata, for ALL of archive, inbox, etc.
- **Users:** Used to store info on eprints users
- **Saved Searches:** Used to store what searches users have and the frequency of alert emails
- **Import:** Stores the details of an import.
- **Files:** Used to store info about files
- **Projects:** Used to store CRIS-type information about projects

Projects

- **colis:** Co-Investigators
- **datestamp:** Datestamp
- **description:** Description
- **funder:** Funder
- **grantid:** Grant ID
- **pis:** Principal Investigators
- **projectid:** ID
- **title:** Title
- **value:** Grant Value
- **workers:** Project Team
New Datasets can be created: e.g. project

$eprint->{datasets}->{project} = {
    class => "EPrints::DataObj",
    sqlname => "project",
    datestamp => "datestamp",
};

$eprint->{fields}->{project} = [
    {name => "projectid", type => "int"},
    {name => "datestamp", type => "time"},
    {name => "title", type => "text"},
    ...
];

- Just like eprint_fields!
v3.2 will support CERIF

- Standard CERIF data types
  - Project
  - Person
  - Organisation
  - Funding
  - Citation
  - Metrics
  - ResultPublication

- Import, export and update via CERIF XML
Citation Framework

- EPrints has supported Google Scholar citation counts since 3.1
  - But we didn’t enable it due to licensing issues

- Now Thomson ISI are changing the licenses for Web of Science data
- Package to support Web of Science citation counts (works for 3.1)
  - http://files.eprints.org/446/
- (Similar package to be produced for GS)
Citation Framework

- Adds a WoS harvesting script, metadata fields and config changes
- In future will add citation data support for Scopus, ACM/IEEE etc
IRStats

- IRstats has been integrated into EPrints and is no longer a standalone install
  - Package available from files.eprints.org
  - Works on v3.1 repositories

- Is being extended to work with citation statistics and other evidence of use
  - ISI citation counts, h factors etc
Extended Data

IRStats

Top 10 papers by download

1. NUS: The Semantic Web Revision
2. 2742: An Object-Oriented Framework for Microcode
3. 1469: Conflict Overload Control
4. 1386: A Web of Community Trust - An Axiom of Online: A Case Study in Community-Focused Design for the Semantic Web
5. 1186: Conceptual analysis: An overview with applications to learning methods
6. 1149: Mobile Agent Architecture for Distributed Information Management
7. 942: Open Hypermedia and Streaming Audio
8. 707: A Framework for Web Services
9. 696: Structure in Text and Hypertext
10. 481: Ontology mapping: the state of the art

Top 10 papers by Google Scholar Citation counts

1. 942: Open Hypermedia and Streaming Audio
2. 707: A Framework for Web Services
3. 696: Structure in Text and Hypertext
4. 481: Ontology mapping: the state of the art
5. 469: Intelligent Agents: Theory and Practice
6. 554: Orthogonality of spectral methods and their application to non-linear system identification
7. 516: Toward a tool for the composition and enhancement of bioinformatics workflows
8. 328: A clustering technique for digital communications channel equalization using radial basis function networks
9. 215: GRASP: A Search Algorithm for Proportional Sensitivities
10. 515: Agents that Reason and Negotiate by Argument

Top 10 papers by Web of Science Citation counts

1. 859: Intelligent Agents: Theory and Practice
2. 554: Orthogonality of spectral methods and their application to non-linear system identification
3. 469: Intelligent Agents: Theory and Practice
4. 328: A clustering technique for digital communications channel equalization using radial basis function networks
5. 215: GRASP: A Search Algorithm for Proportional Sensitivities
6. 116: Agents that Reason and Negotiate by Argument
8. 103: Developing multiagent systems: the Gaia Methodology
9. 102: Identification of MEMO non-linear systems using a forward–agression orthogonal estimator
10. 78: A Methodology for Agent-Oriented Analysis and Design

The view determines how data is rendered and may provide additional data refinements (for example, showing a summary for authors).
Direct Page Editing

Training 12

Welcome to Training 12. Click here to start customising this repository.

- Latest Additions
  View items added to the repository in the past week.

- Search Repository
  Search the repository using a full range of fields. Use the search field at the top of the page for a quick search.

- Browse Repository
  Browse the items in the repository by subject.

- About this Repository

- Edit the contents of static pages directly
  - Home page, help pages etc
Either export to favourite HTML editor, or edit in place in a Web form. Images can be substituted.
Per-page Phrase Editing

- Instead of editing ALL phrases in system, can edit only those used on current (dynamic) page
Per-page Phrase Editing

- Instead of editing ALL phrases in system, can edit only those used on current (dynamic) page.
Import / Updates

- The commandline import script now accepts a new argument
  - `/opt/eprints3/bin/import –enable-updates`

- Any imported item that specifies an eprintid that already exists will update the fields in that item that are given in the imported data.
### Updates via Spreadsheet

- **New format ‘Multiline Excel’ makes updates easier**

Each eprint assigned several rows to accommodate multiple field values.

Sort by row_id to return to proper order.

Can delete cols for any field, but all compound fields must be either completely present or completely absent.

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
<th>J</th>
<th>K</th>
<th>L</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>9170</td>
<td>9170.0</td>
<td>conference</td>
<td>Argonne National Laboratory</td>
<td>DAMLS: A Semantic Markup Language For Web Services</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>9170</td>
<td>9170.1</td>
<td>Burstein, Mark</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>9170</td>
<td>9170.2</td>
<td>Hobbs, Jerry</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>9170</td>
<td>9170.3</td>
<td>Leopold, Dorothea</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>9170</td>
<td>9170.4</td>
<td>Martin, David</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>9170</td>
<td>9170.5</td>
<td>McIntosh, Sheila</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>9170</td>
<td>9170.6</td>
<td>Narayanan, Srinivasan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>9170</td>
<td>9170.7</td>
<td>Porcino, Massimo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>9170</td>
<td>9170.8</td>
<td>Payne, Terry R.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>9170</td>
<td>9170.9</td>
<td>Byera, Katta</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>9170</td>
<td>9170.10</td>
<td>Zang, Hongli</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>5741</td>
<td>5741.0</td>
<td>conference</td>
<td>Anthony, P.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>5741</td>
<td>5741.1</td>
<td>Hill, W.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>5741</td>
<td>5741.2</td>
<td>Delg, V.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>5741</td>
<td>5741.3</td>
<td>Jennings, N. R.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>6152</td>
<td>6152.0</td>
<td>conference</td>
<td>Bourne, R. A.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>6152</td>
<td>6152.1</td>
<td>Shrop, K.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>6152</td>
<td>6152.2</td>
<td>Jennings, N. R.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>19</td>
<td>4234</td>
<td>4234.0</td>
<td>conference</td>
<td>Bussmann, S.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>4234</td>
<td>4234.1</td>
<td>Jennings, N. R.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>21</td>
<td>4234</td>
<td>4234.2</td>
<td>Woodrige, M. J.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>22</td>
<td>5834</td>
<td>5834.0</td>
<td>conference</td>
<td>Carr, Leida</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>23</td>
<td>5834</td>
<td>5834.1</td>
<td>Hill, Wendy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>24</td>
<td>5834</td>
<td>5834.2</td>
<td>Bechhefer, Susan</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>25</td>
<td>5834</td>
<td>5834.3</td>
<td>Noble, Corina</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>26</td>
<td>1997</td>
<td>1997.0</td>
<td>conference</td>
<td>Davis, H. C.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>27</td>
<td>1997</td>
<td>1997.1</td>
<td>White, S. A.</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>28</td>
<td>1997</td>
<td>1997.2</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>29</td>
<td>1997</td>
<td>1997.3</td>
<td>60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>1997</td>
<td>1997.4</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>31</td>
<td>1997</td>
<td>1997.5</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>32</td>
<td>1997</td>
<td>1997.6</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>33</td>
<td>1997</td>
<td>1997.7</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>34</td>
<td>1997</td>
<td>1997.8</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>35</td>
<td>1997</td>
<td>1997.9</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>1997</td>
<td>1997.10</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>37</td>
<td>1997</td>
<td>1997.11</td>
<td>80</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>38</td>
<td>1997</td>
<td>1997.12</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Coming Soon

- Scheduler / Calendar interface to event planning
- Multistage Editorial Control
- Manual raising of Quality Issues
### Miscellaneous

- Geolocation-based autocompleters
  - cities / countries
  - research institutions
- Lightbox-based previews
- Progress bar
- Document ordering

- Coming soon
  - User profiles
  - User-defined collections
  - Funders / projects / grant id autocompleters
Conclusions

- EPrints as a competent and mature part of the researcher’s information environment
- Integrating with their desktop knowledge creation activities (making and editing documents)
- Syndicating rich summaries to Web portals, Web 2 and social networking environments
- Participating in linked data / semantic web
Conclusions

- EPrints as a competent and mature part of the institution’s information environment
- Providing open access, sustainable storage and management services to research community
- Leveraging information about research outputs for senior management to guarantee service sustainability