Digital Preservation for Digital Repositories

David Tarrant
University of Southampton (UK)
dct05r@ecs.soton.ac.uk
Congratulations on your new research project!
This is where your hardware will end up

Make sure your data doesn’t!

Research outputs go in research repositories
Grassroots Preservation

Small Science > Big Science
“The sum of the smaller parts adds up to a greater number than that of the bigger parts combined”

• “Grassroots” preservation for Institutional and Small Business Outputs

• Until now EPrints has mainly been focused on encouraging acquisition of Data.

• How do we create our Global Collection?
• Proposed as a ‘build your own repository’ solution

• Enable institutions and groups to participate in OAI metadata sharing initiative.

• First released April 2000 (to co-inside with OAI-PMH)
• Version 3.1 release at recent Open Repositories Conference 2008

• Used by over 240 registered repositories

Number of Records captured from the Registry of Open Access Repositories (ROAR)
http://roar.EPrints.org
• Open source (GNU license)

• EPrints development model is more centralised than DSpace / Fedora
  • Faster turnaround on development cycles
  • More focused
  • Easier quality management
  • Better for Support Model

• EPrints Services:
  • Repository hosting, bespoke development & training
  • Sustain the development team
Core 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 plug-in architecture for developing extensions
"It is important to build the concept of preservation from the outset. In the digital era, the 'outset' for most new research and educational materials will be the institutional repository."
Digital Preservation

• Long term reliable storage
  • Open Storage

• Maintaining readability
  • Migration / Emulation

• Interoperability for multiple usage scenarios
Long Term Storage

• Reliable
  • Self Checking and Self Healing file System

• Resilient
  • Must have the capability to be robust in the case of part failure

• Simple & Expandable
  • Must be made of parts which are easy to expand / upgrade way into the future.

• Open
  • Any software developed to enable all of the above must be open, same with any hardware specifications.
**EPrints + Honeycomb**

- **Jam today** - large self-managing storage extends repository bang for library buck
  - New chemistry & artistic objects to be collected
- **Jam tomorrow** - potentially take over part of repository responsibility

November 07
• 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

• Easier integration 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 plug-in architecture for developing extensions**
- EPrints is expanding the number places in which plug-ins can be utilised.
The **eprints** Storage Controller

- Each item can be stored using a different storage plug-in (hence in a different place) dependant on file or metadata properties and values.
  - e.g. Large binary files of scientific data (raw machine result data) can be stored in a large disk (slower access) system and sent to a tape company for long term storage.
  - Processed results can be stored locally and on a honeycomb server where they are preserved.

- Allows a repository to use a 3rd party storage platform
  - Direct deposition into a honeycomb etc

- Great enabler for preservation
  - Let the repository control the deposit process.
  - Ensures that the complete object is preserved and not just the “harvested” bits
Open Storage for Repositories

- Simple, open, managed storage.

- Advanced features built in:
  - ZFS
  - Error and Bit Shift Correction
  - Metadata Layer

- Simple API
  - Store
  - Retrieve
  - Delete

- Simple to interface with Repository Software
Preserv Project Structure (May 2008)

Repository Software
- EPrints
- Fedora
- DSpace

Storage Controller
- EPrints
- Fedora

Physical Storage
- Local Disk
- Remote Server
- Cloud Service
- Honeycomb
- EPrints
- Fedora

Services Registry

Services
- TNA_API - PRONOM
  - File Format Identification
  - Significant Properties
  - Migration Tools
  - (Performance Metrics)
- Scheduler (Oxford)
  - Services & Invocation API
- Interoperability
  - OAI-ORE Specification & Mapping

Application Program Interface (API) + XML

Relation Exclusivity (1 to 1, 1 to Many)
Preserv Project Structure (May 2008)

Application Program Interface (API) + XML
Relation Exclusivity (1 to 1, 1 to Many)
The online registry of technical information. PRONOM is a resource for anyone requiring impartial and definitive information about the file formats, software products and other technical components required to support long-term access to electronic records and other digital objects of cultural, historical or business value.

Free PRONOM tools and services to support digital preservation, including DROID, the automatic file format identification tool, together with links to relevant external tools and services.

EPrints will provide one of the first platforms for the development of preservation services where direct interaction takes place between the Repository Software and Preservation Services.
One More Thing…

Smart Storage

Storage which has the capability to perform actions directly upon the objects stored within it.

Autonomous classification and migration of objects

No reliance on repository software for processor time, yet same results.

Services

- TNA API - PRONOM
  - File Format Identification
  - Significant Properties
  - Migration Tools
    (Performance Metrics)

- Scheduler (Oxford)
  - Services & Invocation API

- Interoperability
  - OAI-ORE Specification & Mapping
Many Thanks!

Christopher Gutteridge
Tim Brody

Steve Hitchcock

Neil Jeffries
Ben O’Steen

Adrian Brown
Questions...?

David Tarrant
University of Southampton (UK)
dct05r@ecs.soton.ac.uk

http://www.preserv.org.uk