

# Digital Preservation for Digital Repositories

David Tarrant

University of Southampton (UK)

[dct05r@ecs.soton.ac.uk](mailto:dct05r@ecs.soton.ac.uk)





Arts & Humanities  
Research Council

bbsrc  
Biotechnology and Biological  
Sciences Research Council



Science & Technology  
Facilities Council

E·S·R·C  
ECONOMIC  
& SOCIAL  
RESEARCH  
COUNCIL

EPSRC

Engineering and Physical Sciences  
Research Council



**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**?

# eprints: History of...

- Proposed as a 'build your own repository' solution
- Enable institutions and groups to participate in OAI metadata sharing initiative.
- First released April 2000 (to coincide 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 Access and Institutional Repositories with EPrints

The screenshot shows the EPrints website homepage. At the top, there are navigation links for HOME, SOFTWARE, COMMUNITY, SERVICES, and OPEN ACCESS. Below these are several sections: 'EPrints Services' (with a sub-section for 'Open Access to Research'), 'EPrints Repository Software' (with a sub-section for 'Open Access Archives'), and 'Training Latest' (with a link to the 'EPrints Training Library'). On the right side, there is a 'Harvard Law School' news item about adopting a Green OA self-archiving mandate, and a 'Updated Definition of OA' section. The footer contains a 'Weak' and 'Strong' OA definition.

UNIVERSITY OF  
**Southampton**

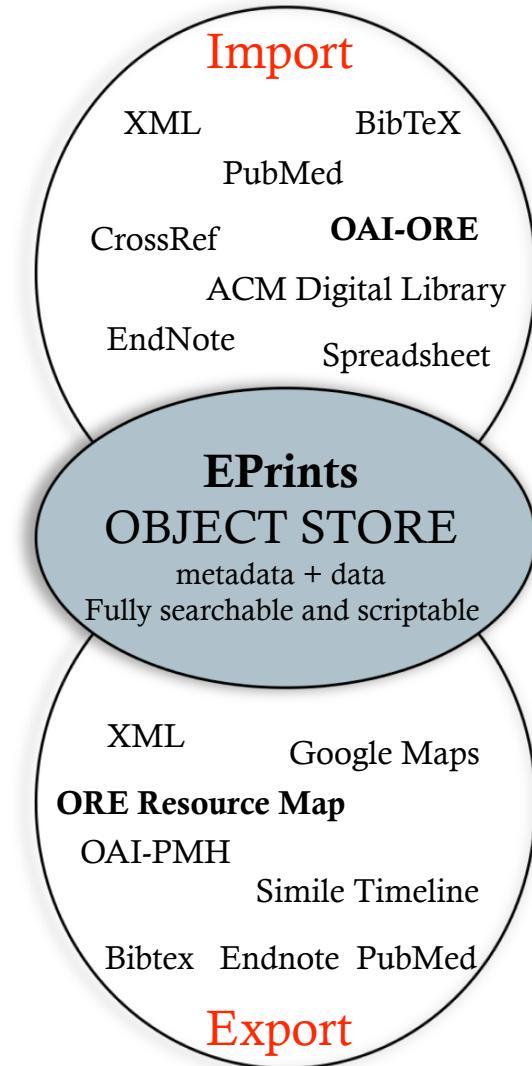
(32169 Records)

# eprints: Management

- 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

# eprints: 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



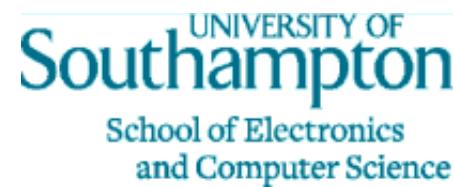
# Digital Preservation



**"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

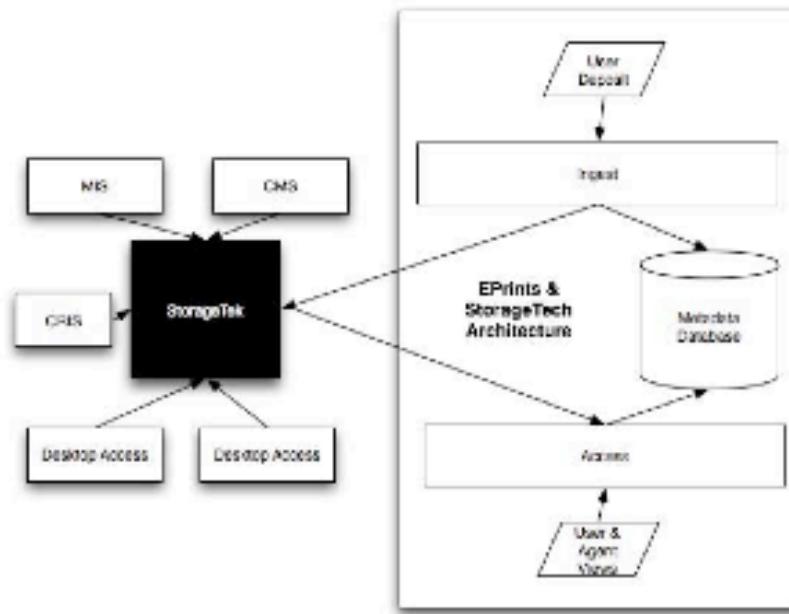
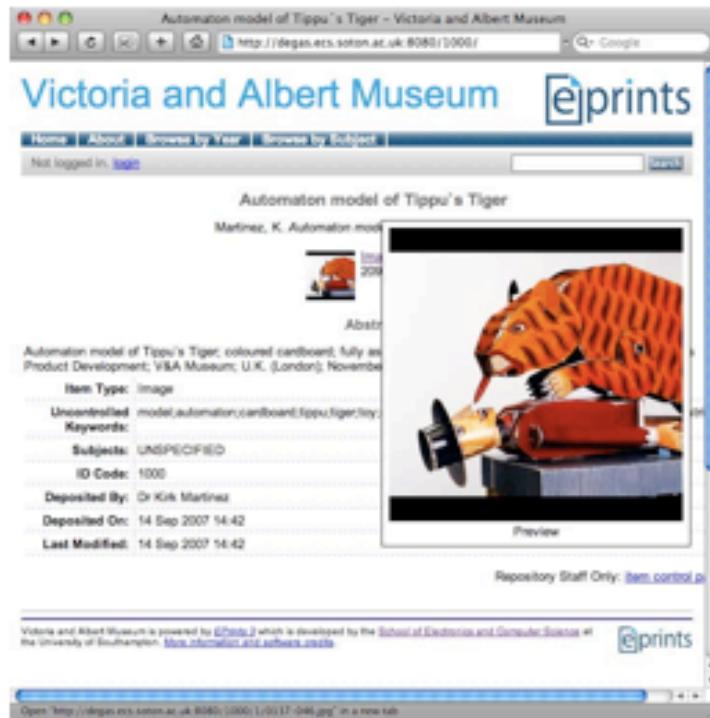


# Long Term Storage

- 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

*November 07*



- **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

# eprints: Core Objectives

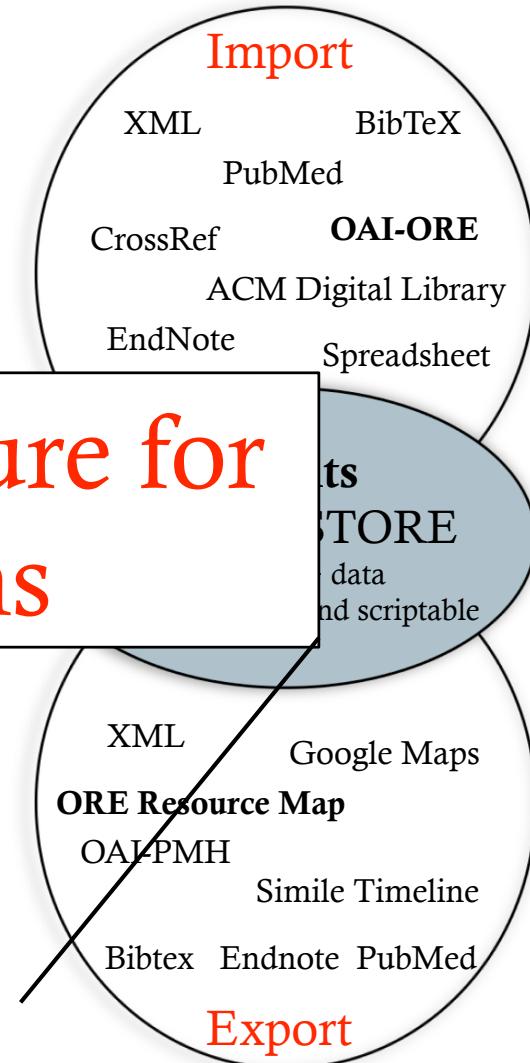
- Lower the barrier for depositors while improving metadata quality and ultimate collection value

- Time saving deposits
- Import data from other repositories and services
- Autocomplete-as-you-type for fast data entry
- Name authorities

- Enter data via web interface
- Flexible plug-in architecture for developing extensions

- 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**



# eprints: Architecture

- EPrints is expanding the number places in which plug-ins can be utilised.

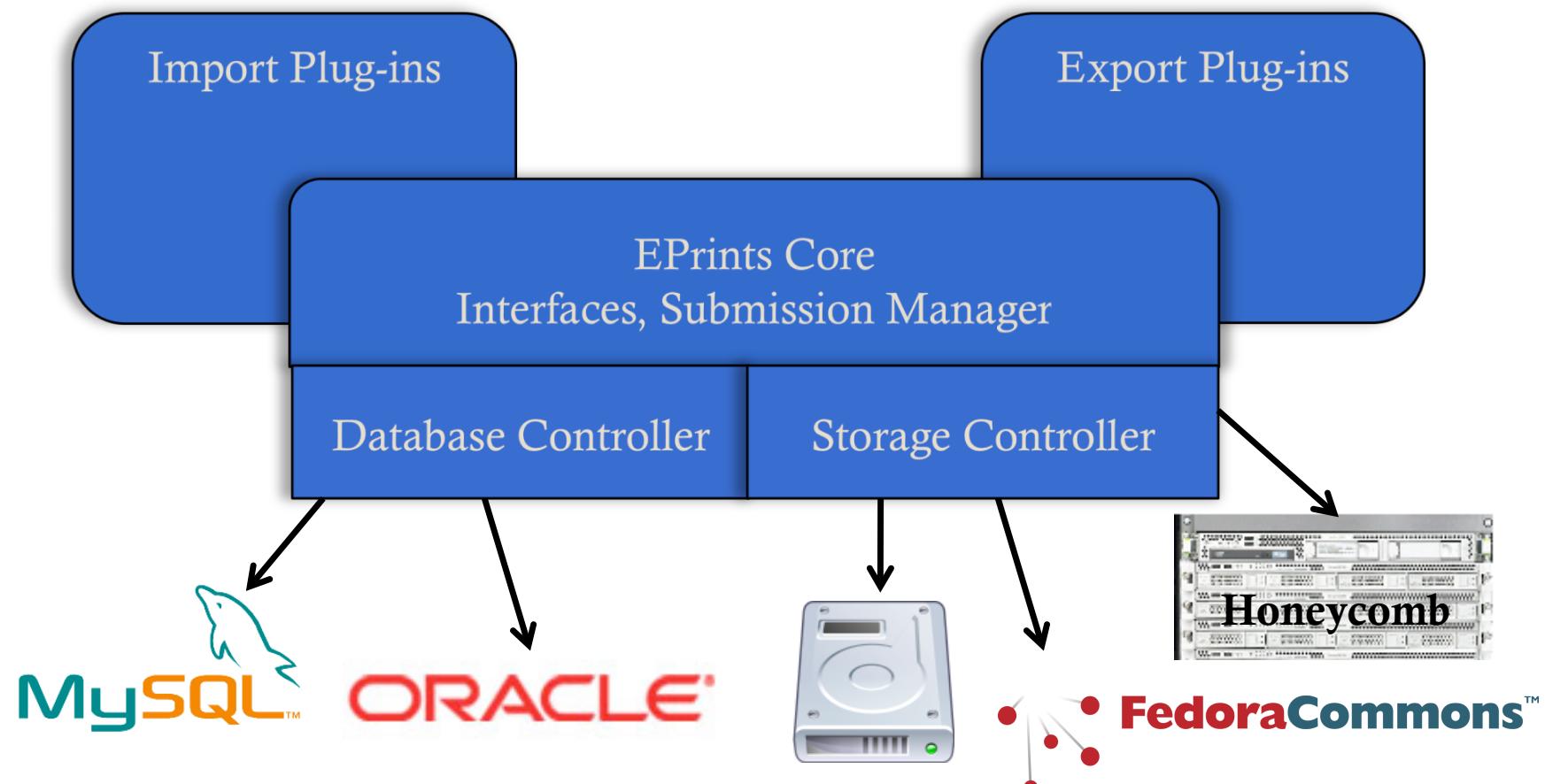


Diagram Represents Proposed EPrints 3.2 Architecture

# 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.



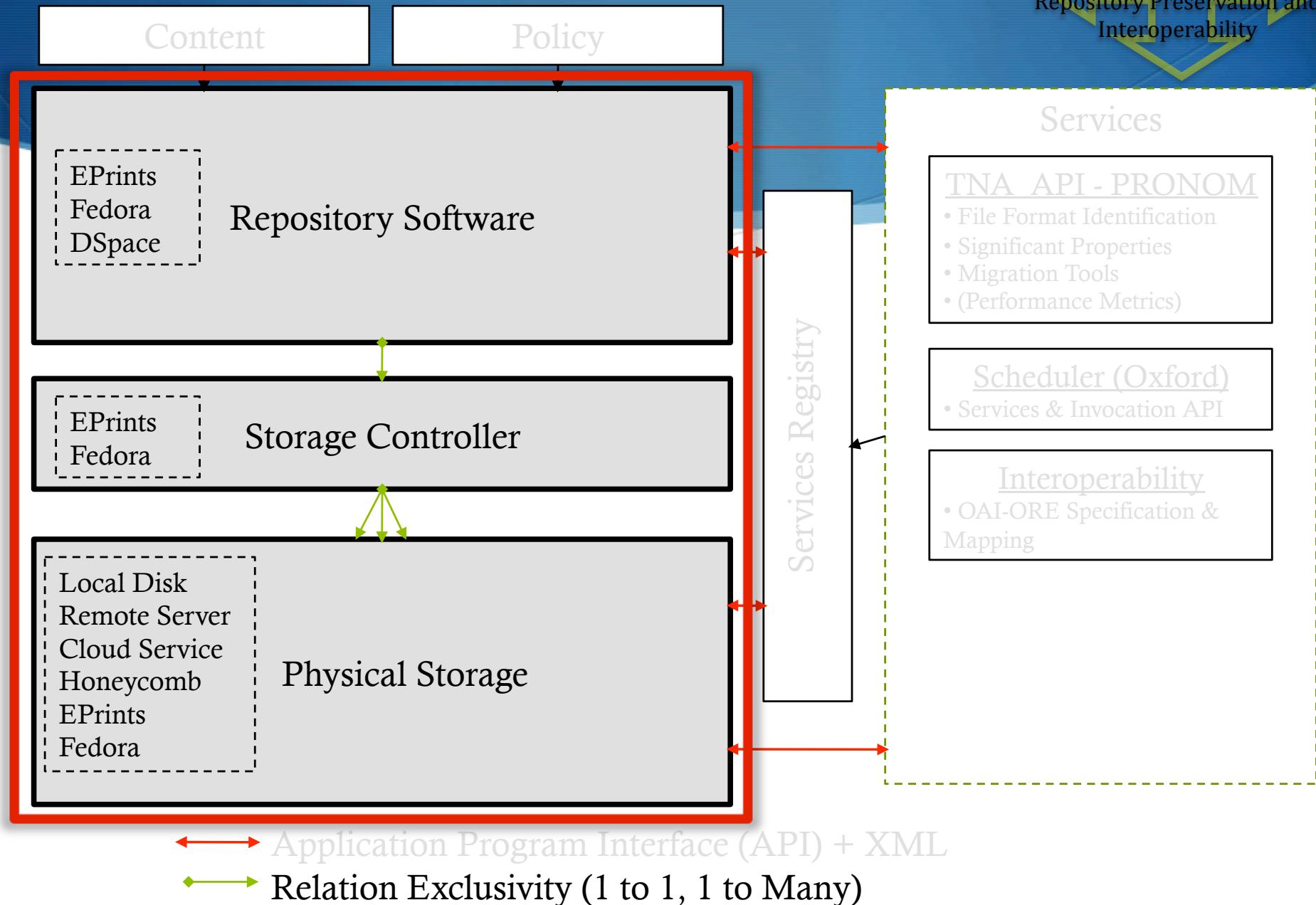
- Allows a repository to use a 3<sup>rd</sup> party storage platform
  - Direct deposition into a honeycomb etc
- Great enabler for preservation
  - Let the repository control the deposit process.
  - Ensures that the complete object is preserved and not just the “harvested” bits

# Open Storage for Repositories

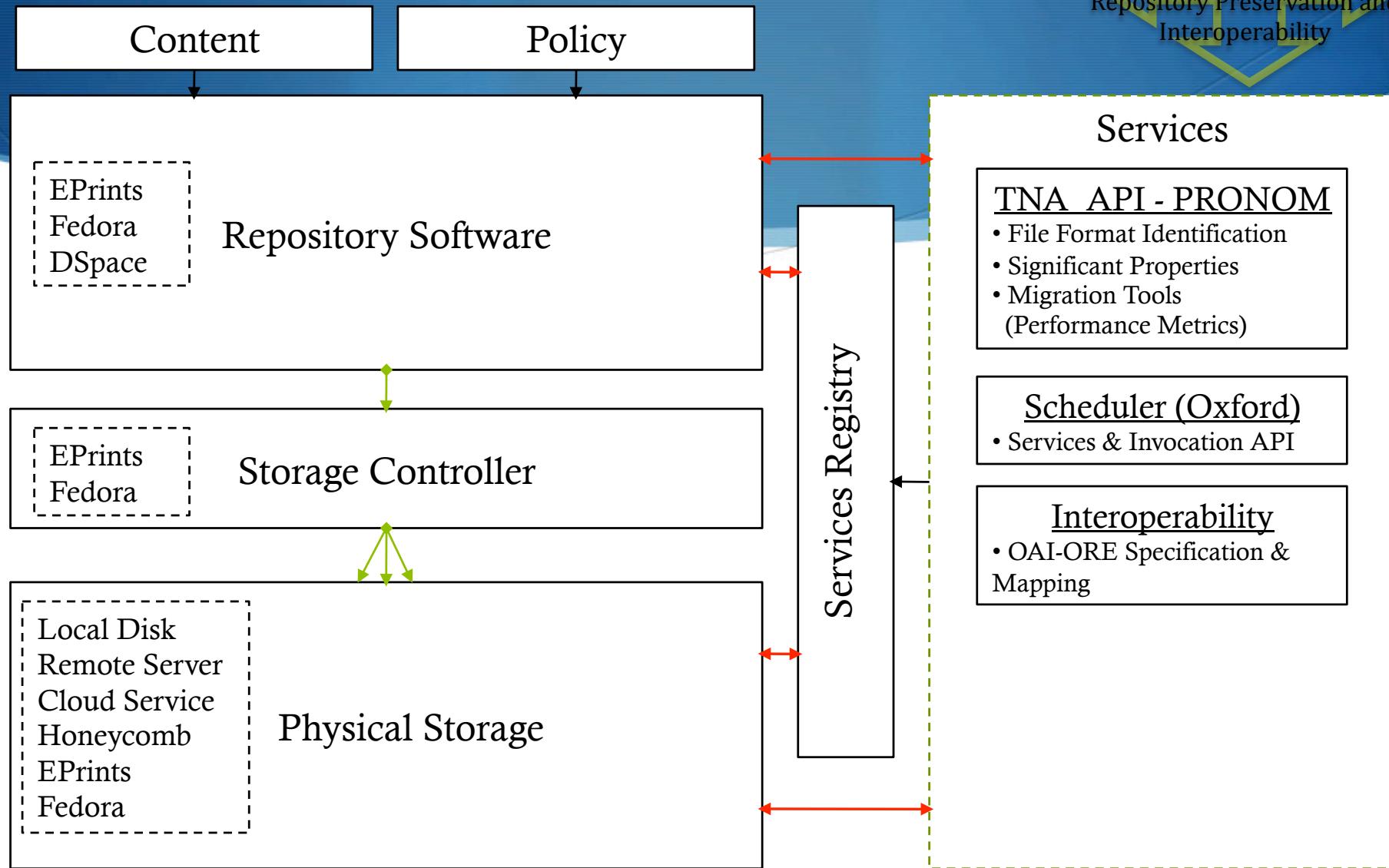
- Simple, open, managed storage.
- Advanced features built in:
  - ZFS
  - Error and Bit Shift Correction
  - Metadata Layer
- Simple API
  - Store
  - Retrieve
  - Delete
- Simple to interface with Repository Software



# Preserv Project Structure (May 2008)



# Preserv Project Structure (May 2008)



↔ Application Program Interface (API) + XML  
→ Relation Exclusivity (1 to 1, 1 to Many)

# Characterisation & Migration Services

The **technical registry**  
**PRONOM**



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.

Under Investigation: Significant Properties Registry, Migration Tools Registry, Risk Analysis and Feedback.



# Interoperability in Action



**OR08 Publications**

**OR2008**  
Third International Conference on Open Repositories

Conference Home | Repository Home | About | Browse by Session | Browse by People | Advanced Search | Login | Create Account

Welcome to OR08 Publications

This is a repository used to manage publications for the [Open Repositories 2008](#) conference. All eprint URLs are persistent and will be redirected to other repositories or services as required in the future. Hint: To download many items at once, perform a search and export the results in "Zip" format.

Main Conference

- Keynote
- Tuesday 1st April
  - 1 - Web 2.0
  - 2a - Social Networking
  - 2b - Sustainability (a)
  - 3 - Interoperability
  - 4 - National Perspectives
  - Birds of a Feather (Tues)
  - Program Committee
- Wednesday 2nd April
  - 4a - National Perspectives
  - 4b - Scientific Repositories (a)
  - 5a - Models
  - 5b - Scientific Repositories (b)
  - 6a - Sustainability (b)
  - 6b - Models, Architectures & Frameworks
  - 7 - Use Cases
  - Birds of a Feather (Wed)
  - Repository Managers
- Object Reuse and Exchange (ORE) Open Day

SUN PASIG Spring Meeting May 27-29, 2008

JISC / CNI Transforming the User Experience July 10-11, 2008

MICROSOFT E-SCIENCE WORKSHOP Dec 7-9, 2008

OR2009 Welcome to Atlanta, Georgia

OR08 Publications supports [OAI 2.0](#) with a base URL of <http://pubs.or08.eess.neton.ac.uk/cgi/oai2>



**ora**  
OXFORD+UNIVERSITY  
RESEARCH+ARCHIVE

Search | Detailed Search

ORA Basic item: " Job Mobility of Residents and Migrants in Urban China"

Reference: John Knight ; Linda Y. Yueh , (2003) Job Mobility of Residents and Migrants in Urban China .

Link to this archived copy: <http://webarchive.eee.ox.ac.uk:5000/objects/uuid:0ec1744f-e435-4176-86e9-941003a8021>

Title: Job Mobility of Residents and Migrants in Urban China ;

Creator: John Knight ; Linda Y. Yueh ;

Date: 2003 ;

Subject: Classification-JEL: J21 ; Classification-JEL: J60 ; Classification-JEL: J63 ; Classification-JEL: Q53 ; labour mobility ; labour turnover ; layoffs ; China ;

Format: Application/pdf;

RePEc-Handle: RePEc:oxf:wpaper:163

RePEc-number: 163

Original-URL: <http://www.economics.ox.ac.uk/Research/wp/pdf/paper163.pdf>

ora:1202  
urn:uuid:0ec1744f-e435-4176-86e9-941003a8021

Browse | Contribute | Tools | Help

Downloads: FULLTEXT | ASCII Text of item | JOURNAL.pdf Item

Terms of Use: The copyright of this item rests with the author and/or other copyright holder(s). Click here for our Terms of Use

## OAI-ORE EPrints & Fedora

**OR2008**  
Third International Conference on Open Repositories

Conference Home | Repository Home | About | Browse by Session | Browse by People | Advanced Search | Login | Create Account

Welcome to OR2008

Repository Preservation and Interoperability

**Preserv.org.uk**

**Browse:**

Date [see more](#)  
2008-04 - (86)

Subject [see more](#)  
Posters - (50)  
1 - Web 2.0 - (3)  
2a - Social Networking -  
2 - Interoperability - (3)  
4a - National Perspectives - (3)  
4b - Scientific Repositories (a) - (3)  
5a - Social - (3)  
5b - Scientific Repositories (b) - (3)  
6a - Sustainability (b) - (3)  
6b - Models, Architectures & Frameworks - (3)

Automatically derived terms [see more](#)  
4 april - (72)  
application/pdf - (72)  
repositories - (64)  
southampton - (58)  
pubs - (51)  
submission - (31)  
united kingdom - (30)  
posters - (23)  
repository - (21)  
open access - (13)

Type [see more](#)  
Conference or Workshop Item - (86)  
NonPeerReviewed - (86)

Creator [see more](#)  
Carr, Leslie - (5)  
Allinson, Julie - (3)  
Lubbard, Bill - (3)  
Alderson, Mark - (2)  
Jones, Simon - (2)  
Kahn, Jeffrey - (2)  
Kumar, Anoop - (2)  
Murray-Rust, Peter - (2)  
Namiki, Takao - (2)  
Pepler, Sam - (2)

Format [see more](#)  
application/pdf - (81)  
application/vnd.ms-powerpoint - (10)  
text/html - (2)

Disclaimer and Data Protection statement | Accessibility statement

**eprints** • **FedoraCommons**

Site powered by **Fedora** and Apache Solr. Data source and information management system is powered by **EPrints.org**.

**Oxford University Research Archive**

**ora**  
OXFORD+UNIVERSITY  
RESEARCH+ARCHIVE

Home | About | Browse by Year | Browse by Subject | Login | Create Account

Job Mobility of Residents and Migrants in Urban China

John Knight and Linda Y. Yueh (2003) Job Mobility of Residents and Migrants in Urban China.

XML 1186b

Image [document\_type name\_text/calendar not defined] 955b

Plain Text 1382b

XML 678b

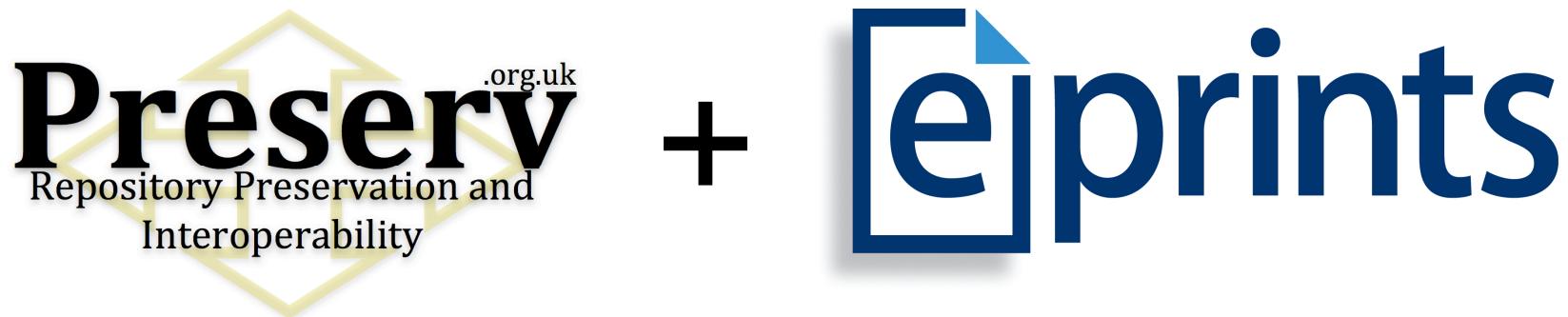
PDF - Requires a PDF viewer such as [GSview](#), [Xpdf](#) or [Adobe Acrobat Reader](#) 241Kb

Plain Text 66Kb

Text

Which is which?

# Digital Preservation



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*.

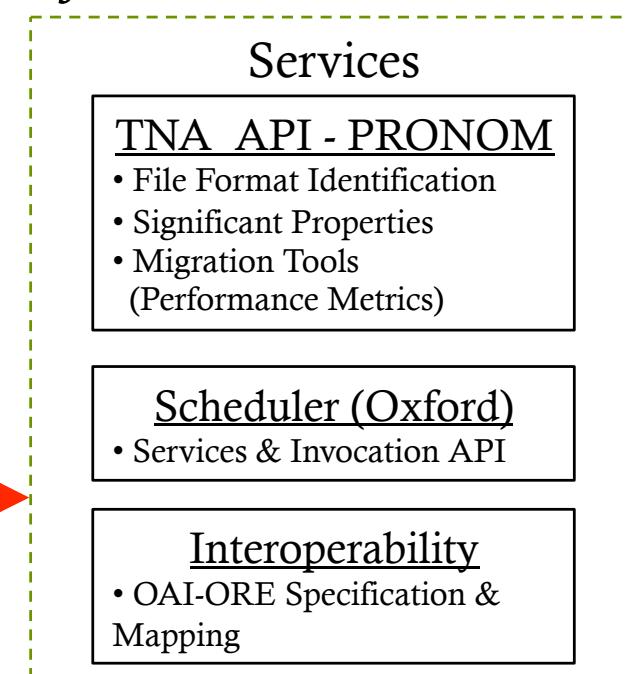
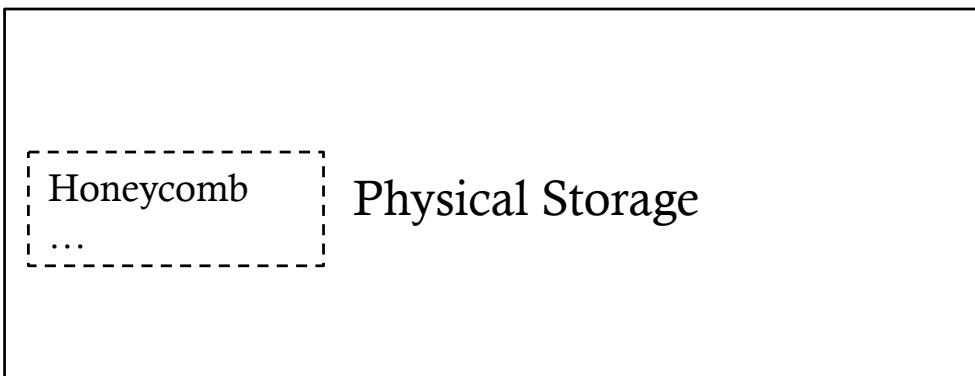
# One More Thing...

## *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.



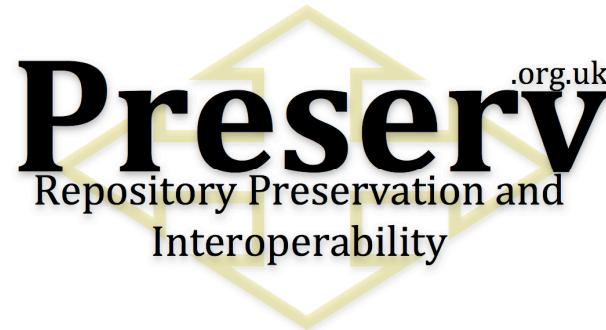
# Many Thanks!

Christopher Gutteridge  
Tim Brody

Steve Hitchcock

Neil Jeffries  
Ben O'Steen

Adrian Brown



# Questions...?

?



<http://www.preserv.org.uk>

David Tarrant

University of Southampton (UK)

[dct05r@ecs.soton.ac.uk](mailto:dct05r@ecs.soton.ac.uk)