

Preservation and storage management for Institutional Repositories

RSP Summer School 2008, Session 3

Maureen Pennock

Steve Hitchcock



Digital Preservation

Aims of this session:

- ☐- Show why preservation matters
- ☐- Encourage you to engage with preservation planning
- ☐- Introduce you to some tools and services that may help

Digital Preservation: Background

What is Digital Preservation?

“ the series of actions and interventions required to ensure continued and reliable access to authentic digital objects for as long as they are deemed to be of value.”

JISC Briefing paper on Digital Preservation, 2006

Digital Preservation: Background

Why is it an issue?

- ❑ 'Fragility' of digital objects
- ❑ Evolution of technologies
- ❑ Underestimated challenge
- ❑ Organisational & cultural issues
- ❑ Supporting, rather than core, activity



Digital Preservation: Background

What are the risks?

☐ Loss of:

- Content
- Structure
- Access
- Investment
- Ideas
- Confidence



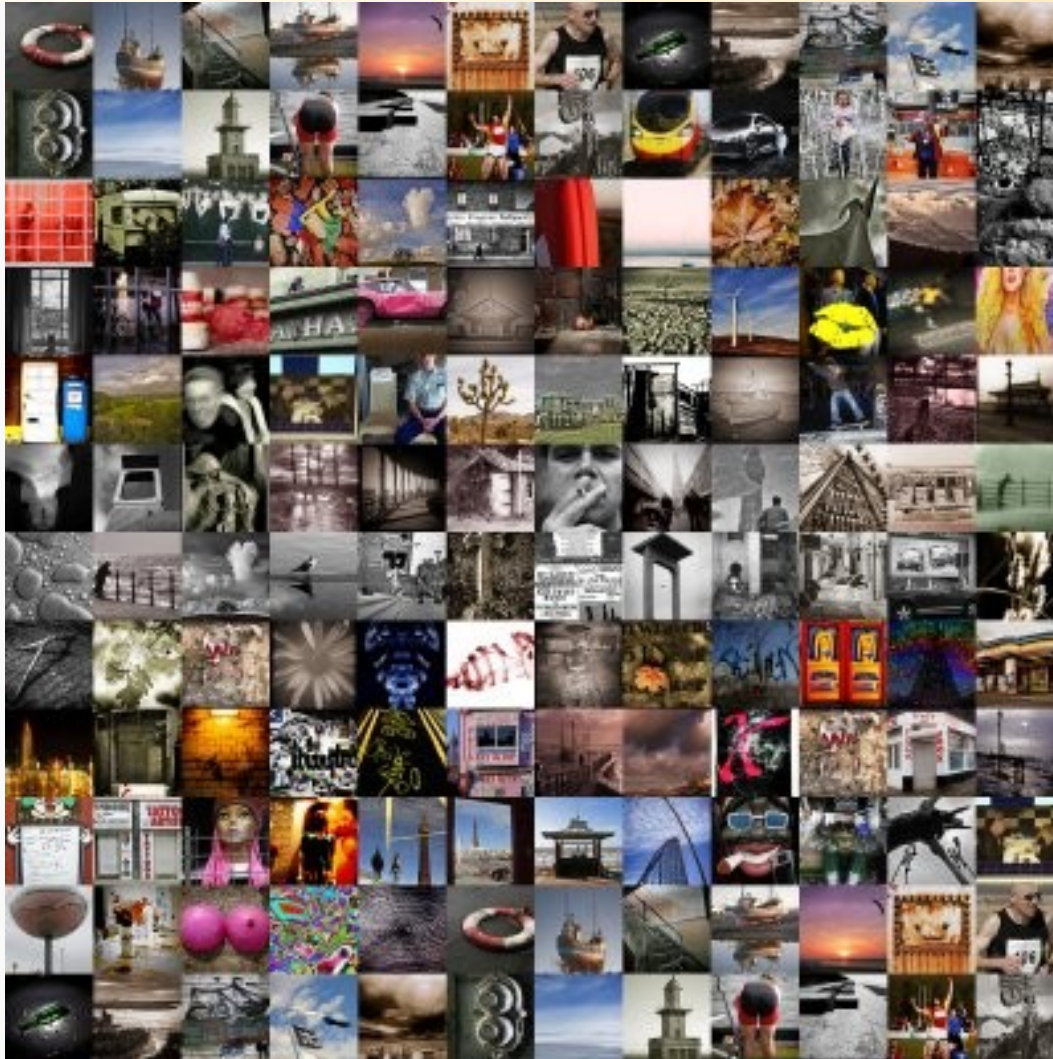
Over to Steve ...

Digital preservation for cab drivers



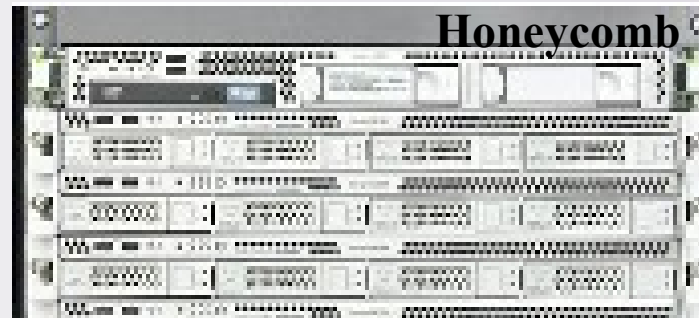
<http://www.flickr.com/photos/sweatyphotos/400920980/>

Data proliferation



More data – more storage

- Large
- Expandable
- Flexible
- Manageable
- Interoperable
- Open storage



Preservation @ RSP

Policy, Planning

Summer School 2007

<http://www.rsp.ac.uk/events/SummerSchool2007/programme>

Metadata

Professional briefings 2008 (BL, Bournemouth) <http://www.rsp.ac.uk/events/ProfBrief.php>

Formats, Services

Briefing paper (2pp)

<http://www.rsp.ac.uk/pubs/briefing-papers.php>

Today: Storage management

Storage management in the wider preservation picture

‘Passive’ preservation, storage-based

- ☐ bit-level storage, e.g. external storage, managed storage, backup

‘Active’ preservation, format-based

- ☐ Characterisation, e.g. which formats
- ☐ Planning, assesses implications of particular formats
- ☐ Action, e.g. transform, migrate at-risk objects

Active approaches are dynamic and continuous, probably requiring expert services, and involving dialogue with the content provider (repository) to assess and select the parameters to inform planning and trigger actions.

Storage management group exercise

Each group will consider a different data type:

Personal libraries

- ☐- Digital photographs
- ☐- Music (MP3s)

Storage management group exercise

Each group will consider a different data type:

Personal libraries

- ☐ Digital photographs
- ☐ Music (MP3s)

Digital libraries

- ☐ Digitised content
- ☐ Web sites
- ☐ Institutional repositories

Over to you...

Practical exercise to explore the issues

☐ Five groups

☐ Content types:

- Digital photos
- Digital music
- Digitised content
- Web sites
- Repositories

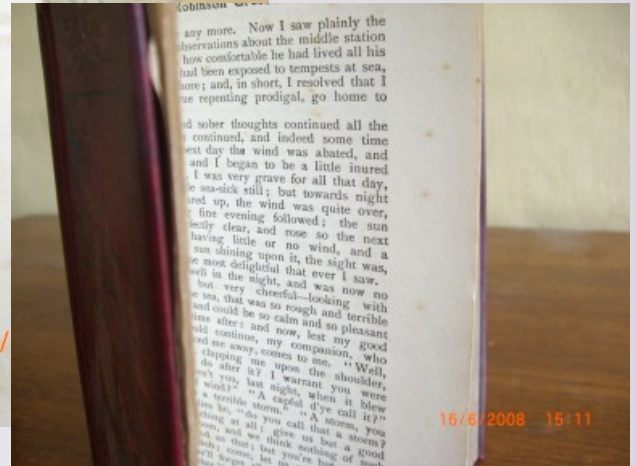
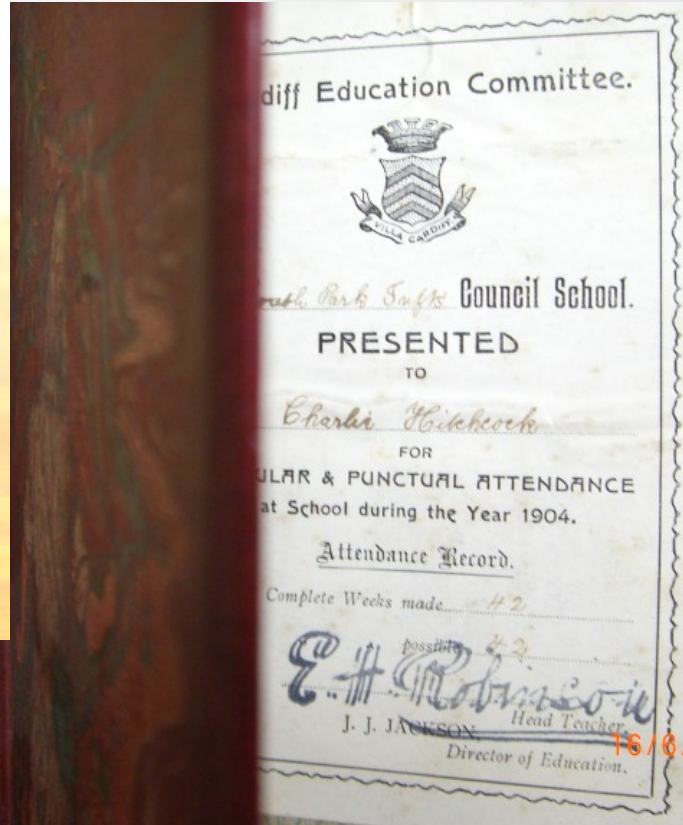


☐ Follow worksheet and discuss!

After the exercise ...

feedback and discussion...

Digital preservation conundrum



Digital preservation conundrum



Digital preservation conundrum

What time is it Eccles?



<http://www.youtube.com/watch?v=VSSGiA4f5cs>

See also <http://whattimeisiteccles.com/>

Digital preservation conundrum

What time is it Eccles?

Bluebottle (aka Peter Sellers): What time is it Eccles?

Eccles (aka Spike Milligan): Err, just a minute. I've got it written down on a piece of paper. A nice man wrote the time down for me this morning.

Bluebottle: Ooooh, then why do you carry it around with you Eccles?

Eccles: Well, um, if anybody asks me the time, I can show it to dem.

Bluebottle: Wait a minute Eccles, my good man.

Eccles: What is it fellow?

Bluebottle: It's writted on this bit of paper, what is eight o'clock, is writted.

Eccles: I know that my good fellow. That's right, um, when I asked the fella to write it down, it was eight o'clock.

Bluebottle: Well then. Supposing when somebody asks you the time, it isn't eight o'clock?

Eccles: Well den, I don't show it to 'em.

...

Bluebottle: Well how do you know when it's eight o'clock?

Eccles: I've got it written down on a piece of paper.

Transcript from *The Goons, The Mysterious Punch-Up-The-Conker*, first broadcast 7th February 1957

Digital preservation conundrum

This is the conundrum:

That we are used to preserving and presenting things that have some fixed, physical representation

Some data varies over time and requires some media to reproduce it (e.g. multimedia)

And some data simply becomes obsolete over time

We should beware trying to fit a physical representation to something when it isn't appropriate

We should be careful, especially with the Web, that we don't become Eccles, trying to write down the time.

Challenging digital preservation

"Unless the vexatious problem of digital preservation is solved, all texts "born digital" belong to an endangered species. The obsession with developing new media has inhibited efforts to preserve the old. We have lost 80 percent of all silent films and 50 percent of all films made before World War II. Nothing preserves texts better than ink imbedded in paper, especially paper manufactured before the nineteenth century, except texts written on parchment or engraved in stone. The best preservation system ever invented was the old-fashioned, pre-modern book."

Robert Darnton, *The Library in the New Age*, *New York Review of Books*, Vol 55, No 10, June 12, 2008

<http://www.nybooks.com/articles/21514>

Darnton is Director of the University Library, Harvard University

Challenging digital preservation: a response

How to read Darnton

- Books and print are a great preservation system.
- Not everything is in this form.
- There were no pre-digital halcyon days of preservation, apart from print.
- New media, new tools, new applications continue to emerge and are adopted.
- We can't stop the world and expect everyone to get off. In this sense preservation efforts will always be reactive.
- We will never 'solve' the vexatious problem of digital preservation. But we can attempt to manage it effectively with appropriate skills, services and resources.

Digital preservation conundrum

The key to all we do is *openness*:

- ☐- Open standards
- ☐- Open source
- ☐- Open Archives
- ☐- Open access
- ☐- Open storage
- ☐- Open repositories

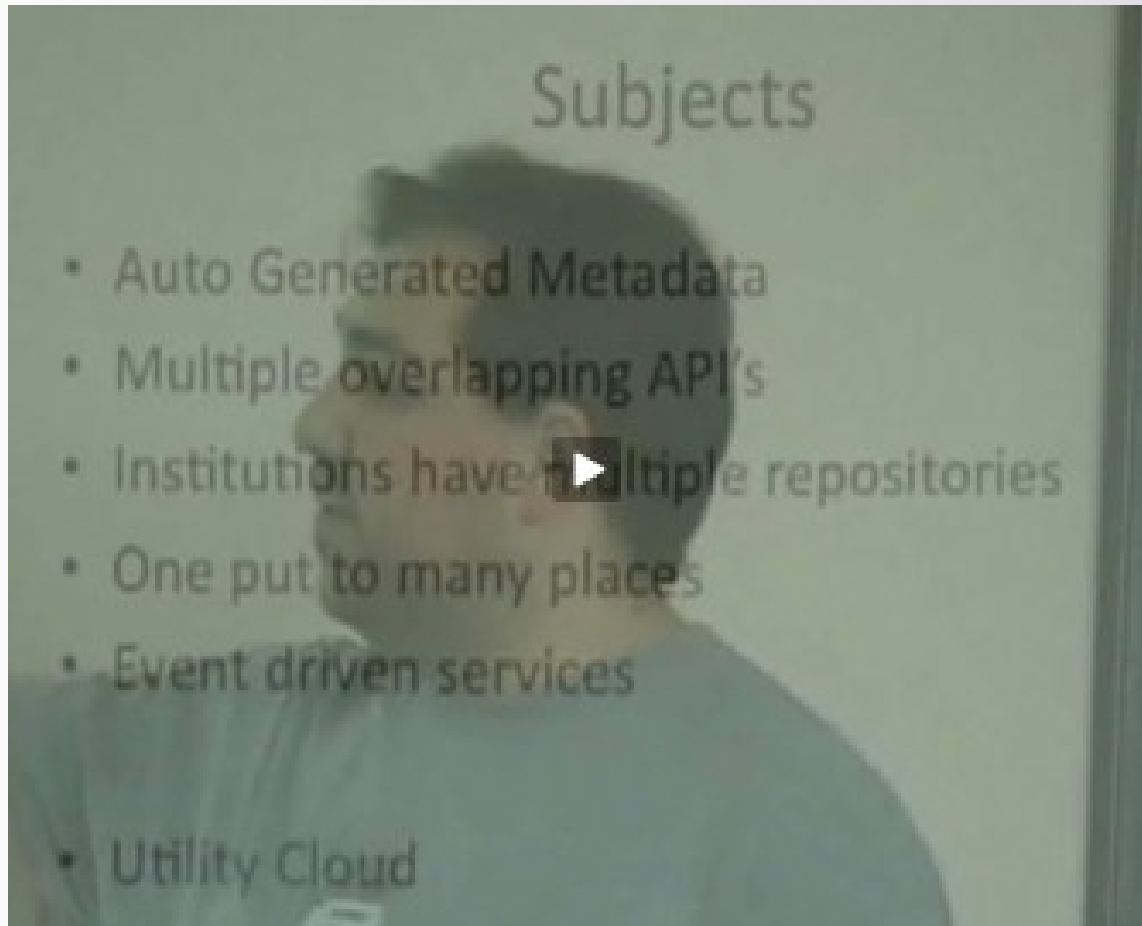
Don't lock into specific technologies

Repository preservation

What help is available?

- ☐ Storage
- ☐ Openness
- ☐ Interoperability
- ☐ Tools
- ☐ Services
- ☐ Service providers

Ultimate interoperability: putting EPrints into Fedora, and back again



by Dave Tarrant, Ben O'Steen and Tim Brody, Preserv 2
From Blip TV <http://blip.tv/file/866653>

Interoperability in Action



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://eprints.ox.ac.uk/3009/object/69c1744f-e435-4176-8969-8941003a8021>

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: O83 ; labour mobility ; labour turnover ; layoffs ; China ;
Format
Application/pdf ;
RePEc-Handle: RePEc:oxfwpaper:163
RePEc-number: 163
Original-URL: <http://www.economics.ox.ac.uk/Research/wp/pdf/paper163.pdf>
ora:1002
urn:uuid:69c1744f-e435-4176-8969-8941003a8021

Downloads
[Full Text as ASCII text of item](#)
[JOE HTML as item](#)

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

OAI-ORE EPrints & Fedora

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

OR2008
Third International Conference on Open Repositories

Search Detailed Search

Browse:

Preserv.org.uk
Repository Preservation and Interoperability

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. [Here](#): To download many items at once perform a search and export the results in "Zip" format.

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

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

Subject [see more](#)
Posters - (50)
1 - Web 2.0 - (3)
2a - Social Networking - (3)
2 - Interoperability - (3)
4a - National Perspectives - (3)
4b - Scientific Repositories (a) - (3)
5a - Legal - (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)
southernhampton - (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)
Hubbard, Bill - (3)
Awra, Ghis - (2)
Coles, Simon - (2)
Kahn, Jeffrey - (2)
Numar, Anoop - (2)
Murray-Put, Peter - (2)
Namiki, Takao - (2)
Pepler, Sam - (2)

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

OR08 Publications supports [OAI 2.0](#) with a base URL of <http://pubs.or08.ox.ac.uk/agent/ra52>

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 1188b
905b
Plain Text 1382b
XML 678b
PDF - Requires a PDF viewer such as [GSView](#), [Xpdf](#) or [Adobe Acrobat Reader](#) 241Kb
Plain Text 66Kb

Disclaimer and Data Protection statement | Accessibility statement

[Eprints](#) | [Fedora Commons](#)

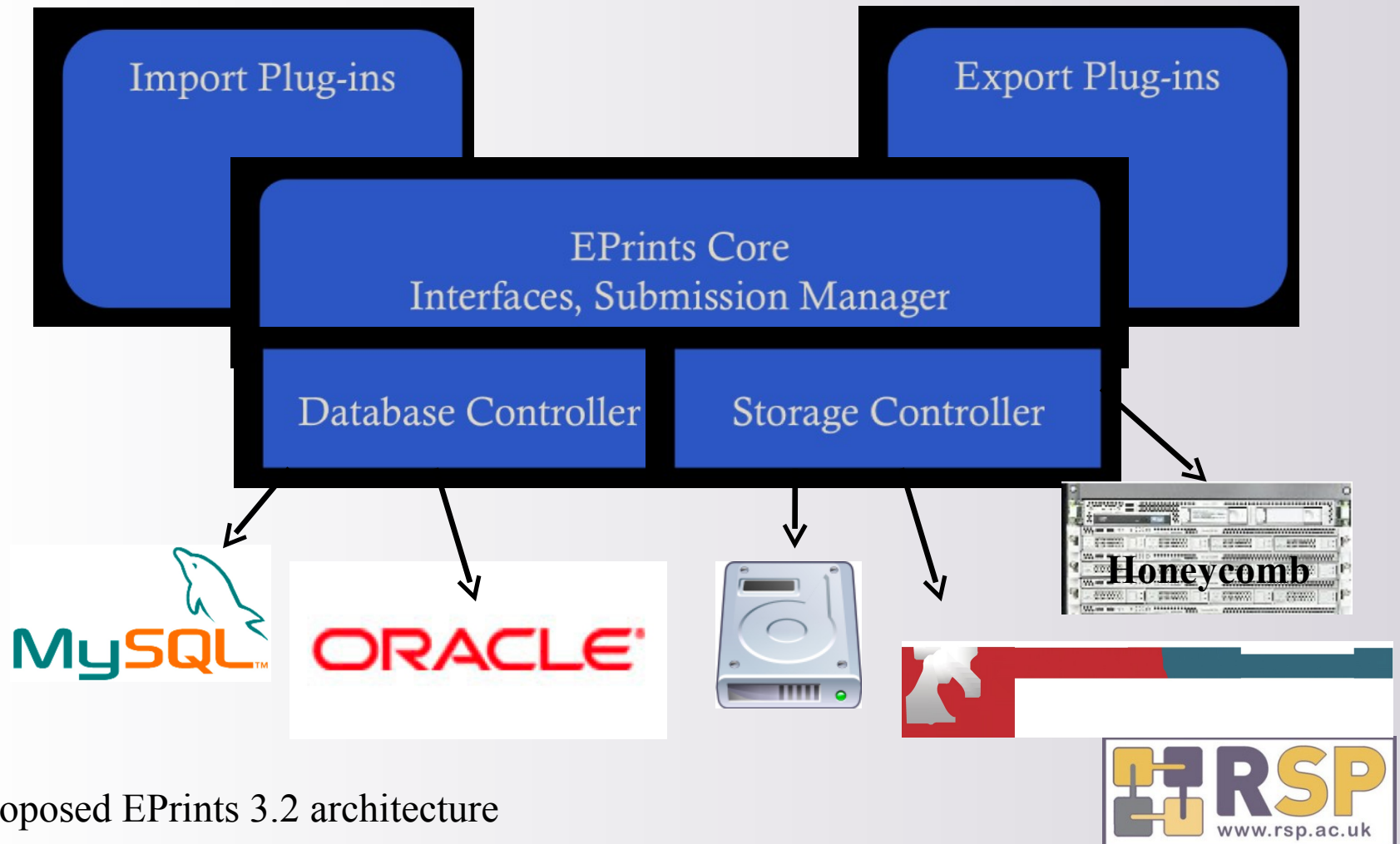
Site powered by Fedora and Apache Solr. Data source and information management system is powered by EPrints.org.



Which is which?



Repository architecture: storage controller



Proposed EPrints 3.2 architecture

Combining active and passive storage: tools and service providers

- Accurately identify the formats of objects stored in the repository
- Adopt a trusted and current list of storage formats and their prospects for preservation
- Develop a plan of action based on the findings of 1 and 2

For 1 and 2 you can find tools and services on the Web:

☐- Format identification tools, e.g. DROID

<http://droid.sourceforge.net/wiki/index.php/Introduction>

☐- Repository registry services, e.g. ROAR has format profiles in development for over 200 repositories

<http://roar.eprints.org/index.php?action=profile&url=http://dspace.anu.edu.au/>

☐- Format reference sources, e.g. Library of Congress

<http://www.digitalpreservation.gov/formats/>

Prospective preservation service providers

Today's perspective

- ☐ **Preservation services**, National libraries, e.g. KB-DARE (Netherlands), German National Library (theses), BL (PubMed Central UK), Sherpa-DP
- ☐ **Institutional services**, e.g. Oxford
- ☐ **Repository software**
- ☐ **Repository services**
- ☐ **Library services**, e.g. OCLC
- ☐ **Cloud storage services**, e.g. Amazon, Google

Plato: Preservation Planning Tool

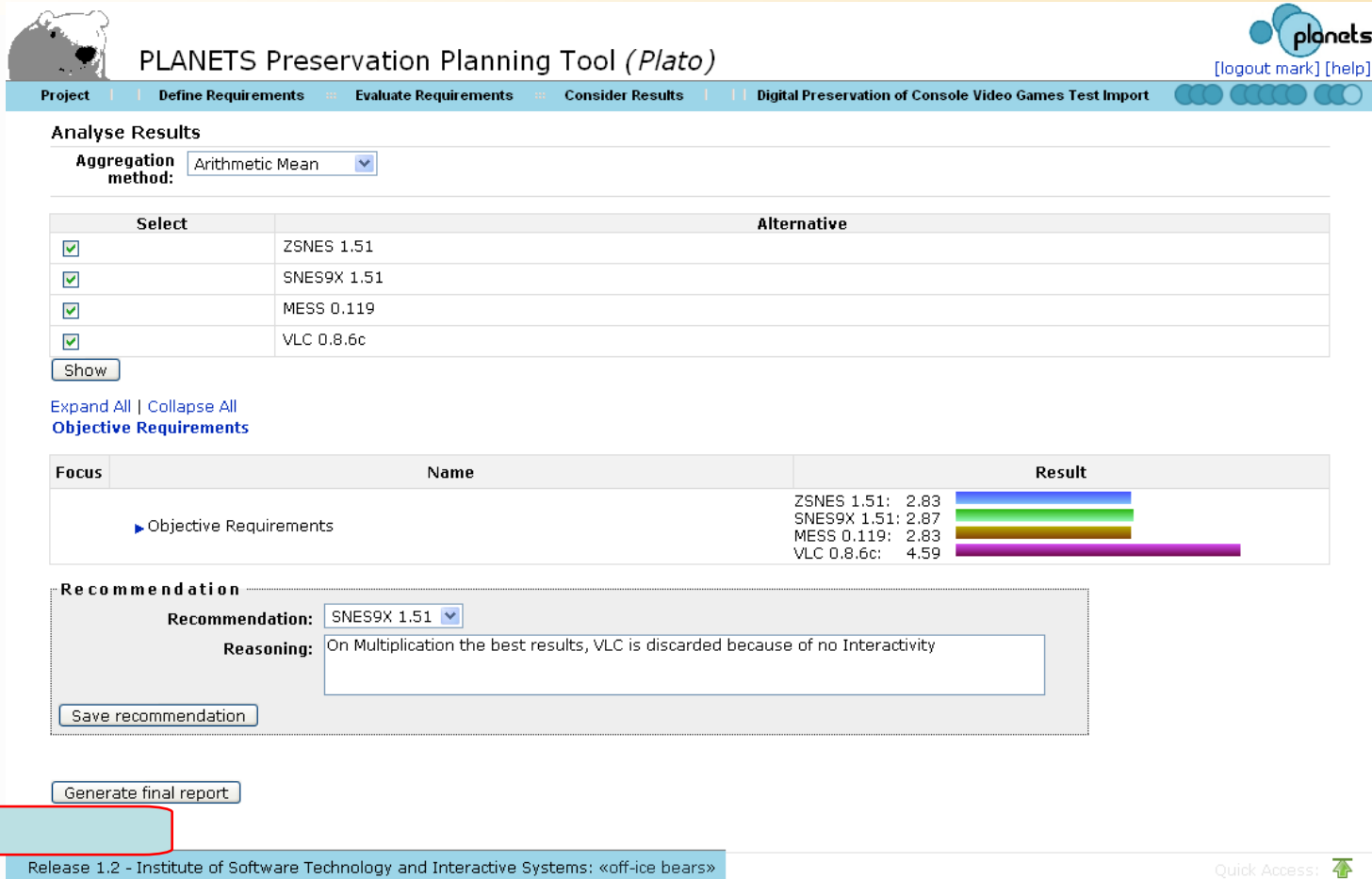


“Until now, preservation planning is largely a manual and tedious process where available solutions are evaluated against the specific requirements of a particular situation.”

- ❏ Implements a well-documented and validated preservation planning methodology
- ❏ Integrates registries and services for preservation action and characterisation
- ❏ Provides a Web-based interface to guide the planner through the process.



Plato: Analyze Results



The screenshot displays the PLANETS Preservation Planning Tool (Plato) interface. The top navigation bar includes links for Project, Define Requirements, Evaluate Requirements, Consider Results, and Digital Preservation of Console Video Games Test Import. The 'Analyze Results' section is active, showing an aggregation method of 'Arithmetic Mean'. A table lists four alternatives: ZSNES 1.51, SNES9X 1.51, MESS 0.119, and VLC 0.8.6c, all of which are selected. Below the table, a 'Show' button is visible. The 'Objective Requirements' section shows a bar chart comparing the alternatives based on their results. The 'Recommendation' section shows a recommendation for SNES9X 1.51, with reasoning that on multiplication, the best results, VLC is discarded because of no Interactivity. A 'Generate final report' button is located at the bottom left. A red box highlights a button in the bottom left corner of the interface.

PLANETS Preservation Planning Tool (*Plato*)

[logout mark] [help]

Project | Define Requirements | Evaluate Requirements | Consider Results | Digital Preservation of Console Video Games Test Import

Analyze Results

Aggregation method: Arithmetic Mean

Select	Alternative
<input checked="" type="checkbox"/>	ZSNES 1.51
<input checked="" type="checkbox"/>	SNES9X 1.51
<input checked="" type="checkbox"/>	MESS 0.119
<input checked="" type="checkbox"/>	VLC 0.8.6c

Show

Expand All | Collapse All

Objective Requirements

Focus	Name	Result
► Objective Requirements	ZSNES 1.51:	2.83
	SNES9X 1.51:	2.87
	MESS 0.119:	2.83
	VLC 0.8.6c:	4.59

Recommendation

Recommendation: SNES9X 1.51

Reasoning: On Multiplication the best results, VLC is discarded because of no Interactivity

Save recommendation

Generate final report

Release 1.2 - Institute of Software Technology and Interactive Systems: «off-ice bears»

Quick Access:

From Plato walkthrough slideshow

http://www.ifs.tuwien.ac.at/dp/plato/pres_plato-workshop.ppt

OpenDOAR policy tool

- ❏ Policies are an important element supporting sustainability
- ❏ OpenDOAR policy tool allows repository managers to easily implement publicly accessible and machine readable policies on:
 - Metadata Data
 - Content & Submission
 - Preservation
- ❏ www.opendoar.org




JHOVE

- ❏ JSTOR/Harvard Object Validation Environment
- ❏ Extensible software framework for performing digital object:
 - Format identification
 - Format validation
 - Format characterisation
- ❏ Outputs in XML; can be used as desired (eg in conjunction with further technology watch)
- ❏ hul.harvard.edu/jhove

- ❏ Auditing your repository is a highly effective means to ensure your activities will satisfy your goals, particularly when they include preservation.
- ❏ DRAMBORA methodology:
 - Provides internal auditors with completed risk register
 - Helps prepare for external audit (and certification?)
 - Facilitates retrospective reflection & proactive planning
- ❏ www.repositoryaudit.eu

(Don't) PANIC!

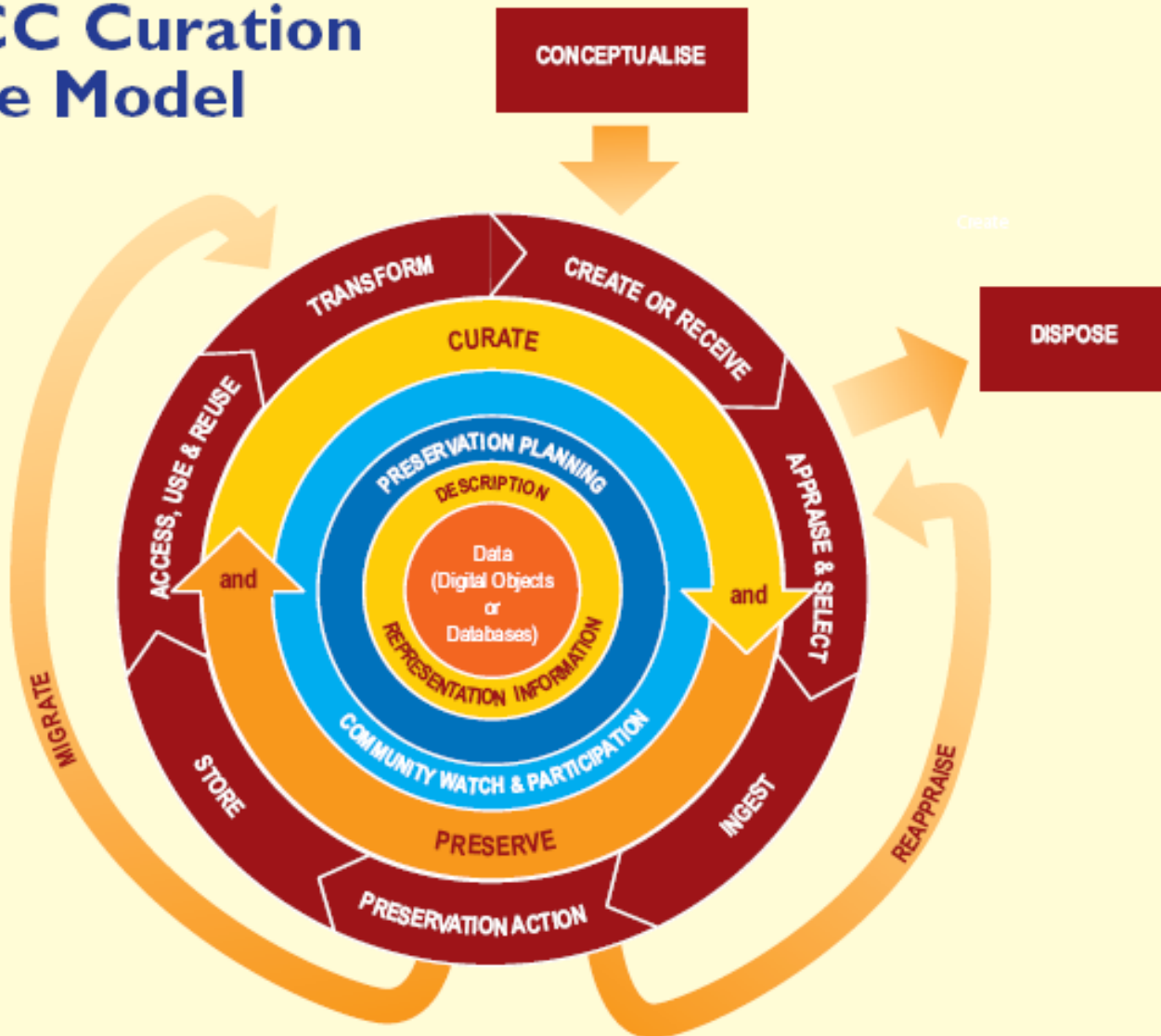
PANIC PREMINT: Preservation Metadata Input Tool

-  Designed to collect information regarding a digital object so that it can be archived and preserved.
-  Takes into account the current state of the digital object, intention behind the creation of the object & attitude of the creator regarding preservation of the digital object.
-  See www.itee.uq.edu.au

NLNZ Metadata Extractor Tool

- ❏ Programmatically extracts preservation metadata from a range of file formats like PDF documents, image files, sound files Microsoft office documents, and many others.
- ❏ Outputs metadata in a standard format (XML) for use in preservation activities.
- ❏ Can also be used in other activities, including resource discovery
- ❏ Available from Sourceforge

The DCC Curation Lifecycle Model





**This is
where your
hardware
will end up**

Make sure your data doesn't!

Research outputs go in research repositories



Maureen Pennock
Steve Hitchcock

www.rsp.ac.uk

 **REPOSITORIES**
SUPPORT PROJECT