# Preservation and storage management for Institutional Repositories

RSP Summer School 2008, Session 3

*Maureen Pennock*

*Steve Hitchcock*

REPOSITORIES SUPPORT PROJECT

# 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

RSP
www.rsp.ac.uk

# 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*

RSP
www.rsp.ac.uk

# 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

RSP
www.rsp.ac.uk

# Digital Preservation: Background

What are the risks?

- Loss of:
  - Content
  - Structure
  - Access
  - Investment
  - Ideas
  - Confidence

RSP
www.rsp.ac.uk

# Over to Steve ...

# Digital preservation for cab drivers

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

RSP
www.rsp.ac.uk

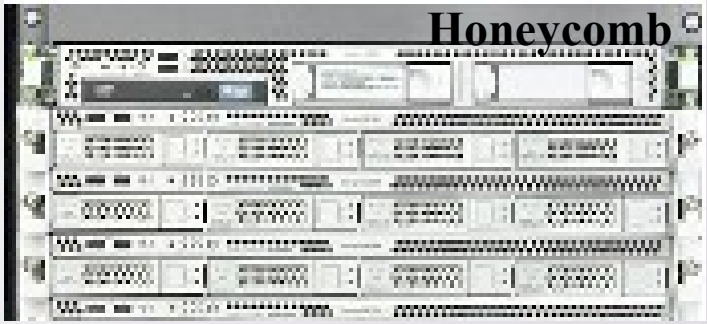
# Data proliferation

RSP
www.rsp.ac.uk

# More data – more storage

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


amazon web services™




RSP
www.rsp.ac.uk

# 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

RSP www.rsp.ac.uk

# 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.

RSP
www.rsp.ac.uk

# 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

RSP
www.rsp.ac.uk

# 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!

RSP
www.rsp.ac.uk

# After the exercise …

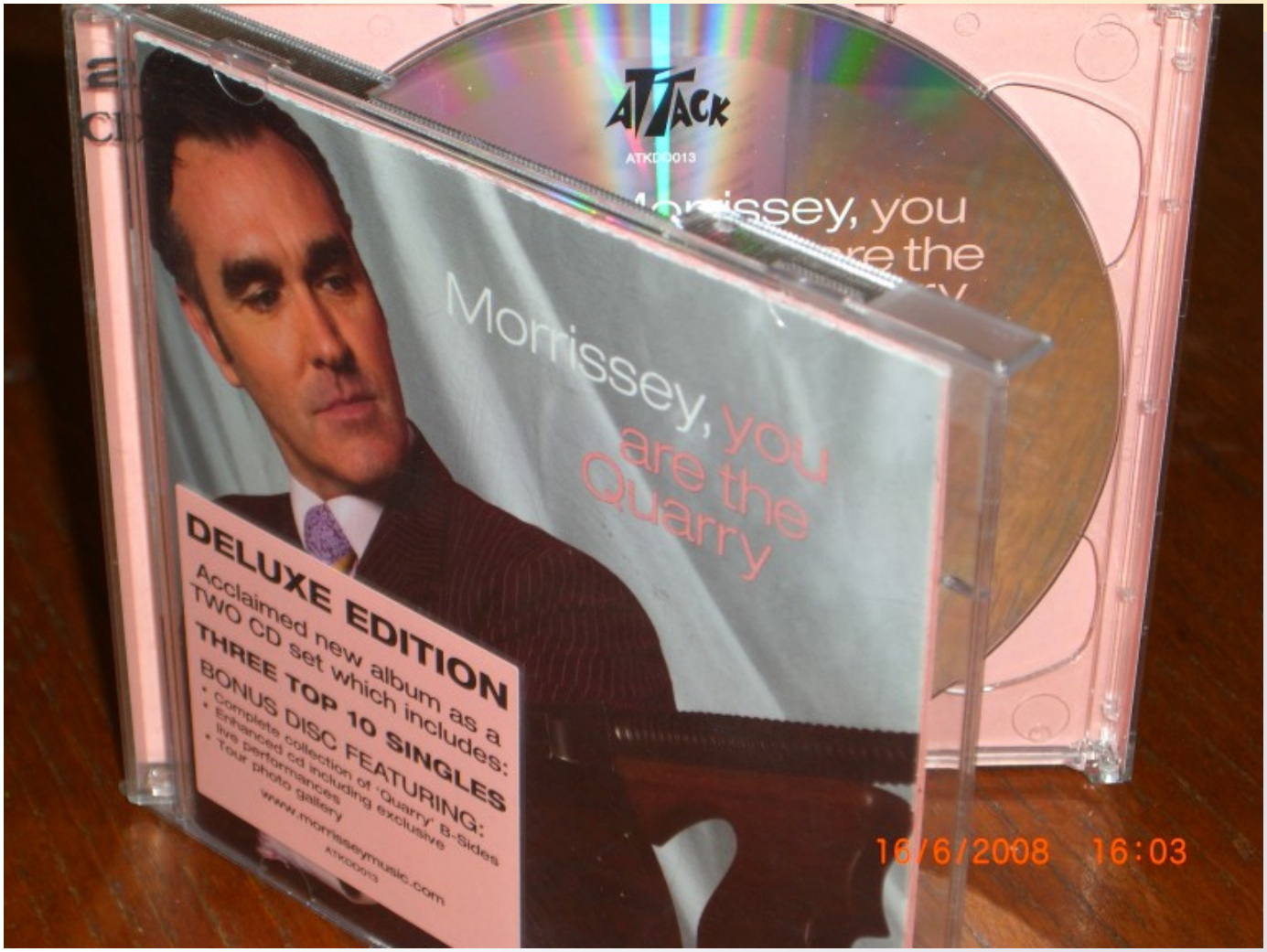
feedback and discussion…

# Digital preservation conundrum



www.rsp.ac.uk

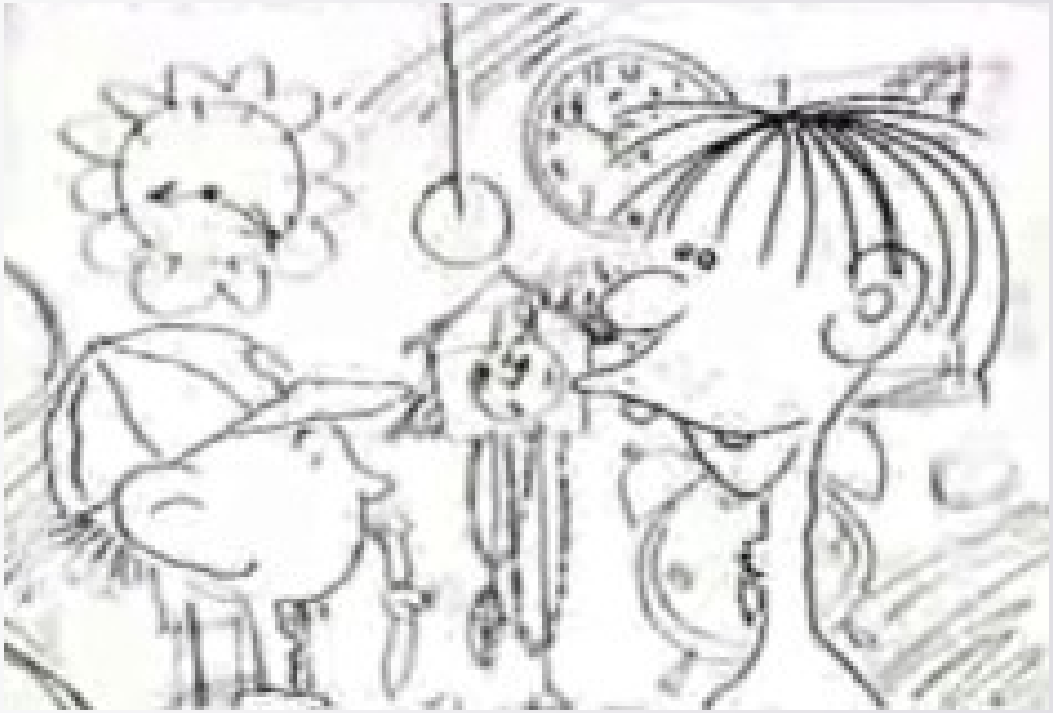
# Digital preservation conundrum




RSP
www.rsp.ac.uk

# Digital preservation conundrum

What time is it Eccles?



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

See also http://whattimeisiteccles.com/

RSP
www.rsp.ac.uk

# 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

RSP
www.rsp.ac.uk

# 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.

RSP
www.rsp.ac.uk

# 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

RSP
www.rsp.ac.uk

# 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.

RSP
www.rsp.ac.uk

# 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

RSP
www.rsp.ac.uk

# Repository preservation

What help is available?

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

RSP
www.rsp.ac.uk

# Ultimate interoperability: putting EPrints into Fedora, and back again

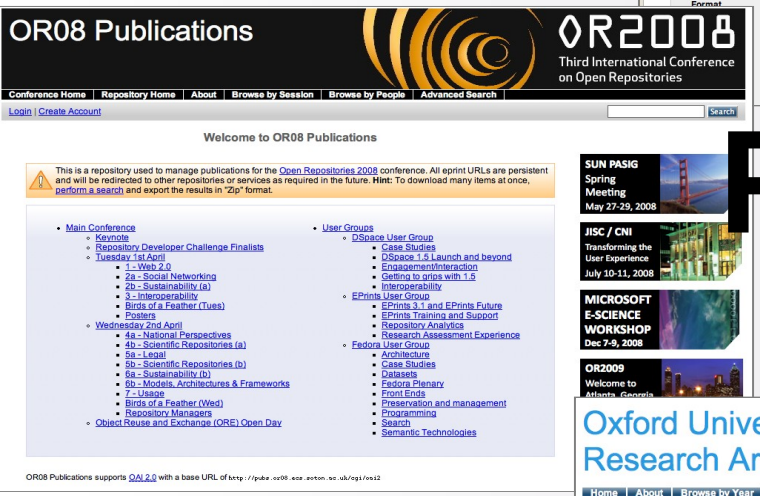

by Dave Tarrant, Ben O'Steen and Tim Brody, Preserv 2

From Blip TV http://blip.tv/file/866653

RSP www.rsp.ac.uk

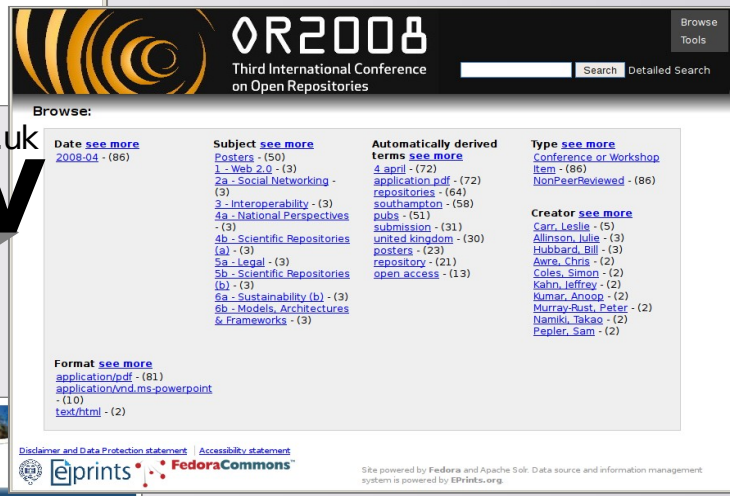
# Interoperability in Action

OAI-ORE
EPrints & Fedora

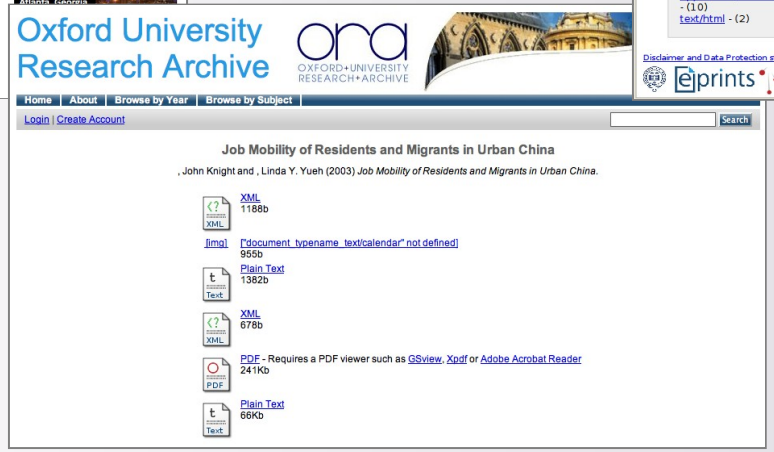







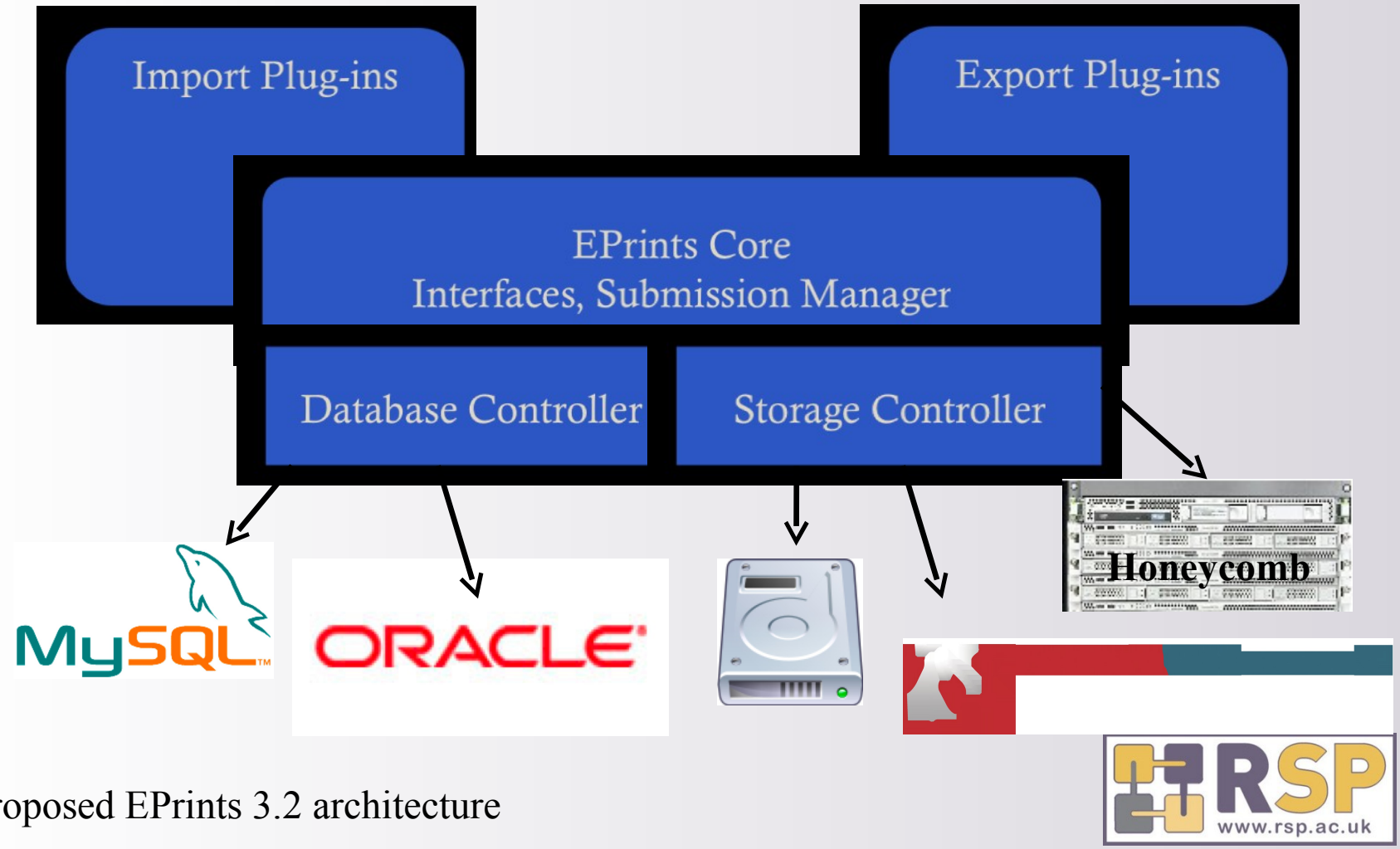








Which is which?

# 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/

RSP www.rsp.ac.uk

# *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

RSP
www.rsp.ac.uk

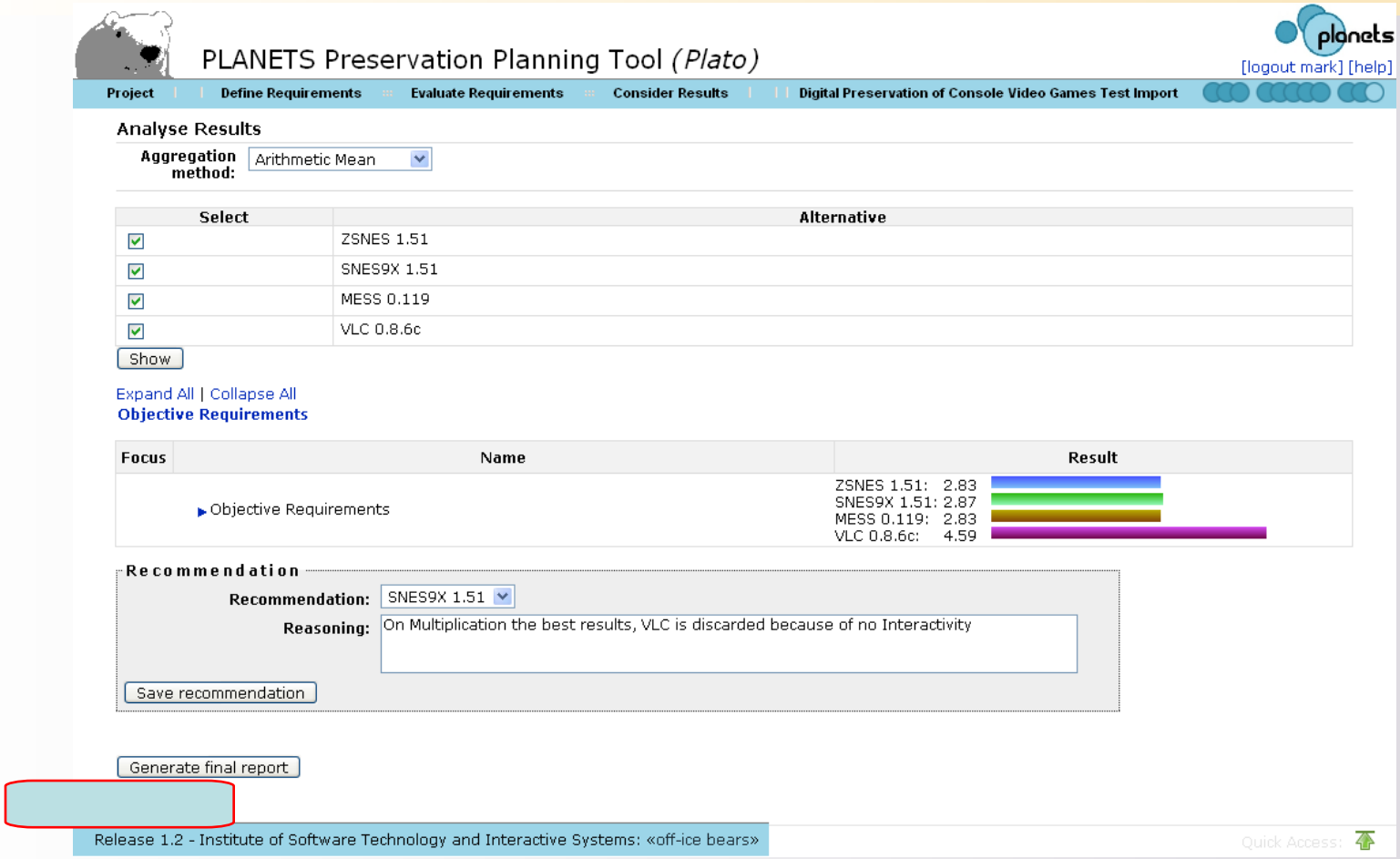
# Plato: Preservation Planning Tool

"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.

www.rsp.ac.uk

# Plato: Analyze Results




From Plato walkthrough slideshow
http://www.ifs.tuwien.ac.at/dp/plato/pres_plato-workshop.ppt

www.rsp.ac.uk

# 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

RSP
www.rsp.ac.uk

# 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

RSP
www.rsp.ac.uk

# DRAMBORA interactive

Digital Repository Audit Method Based on Risk Assessment

- 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

RSP www.rsp.ac.uk

# (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

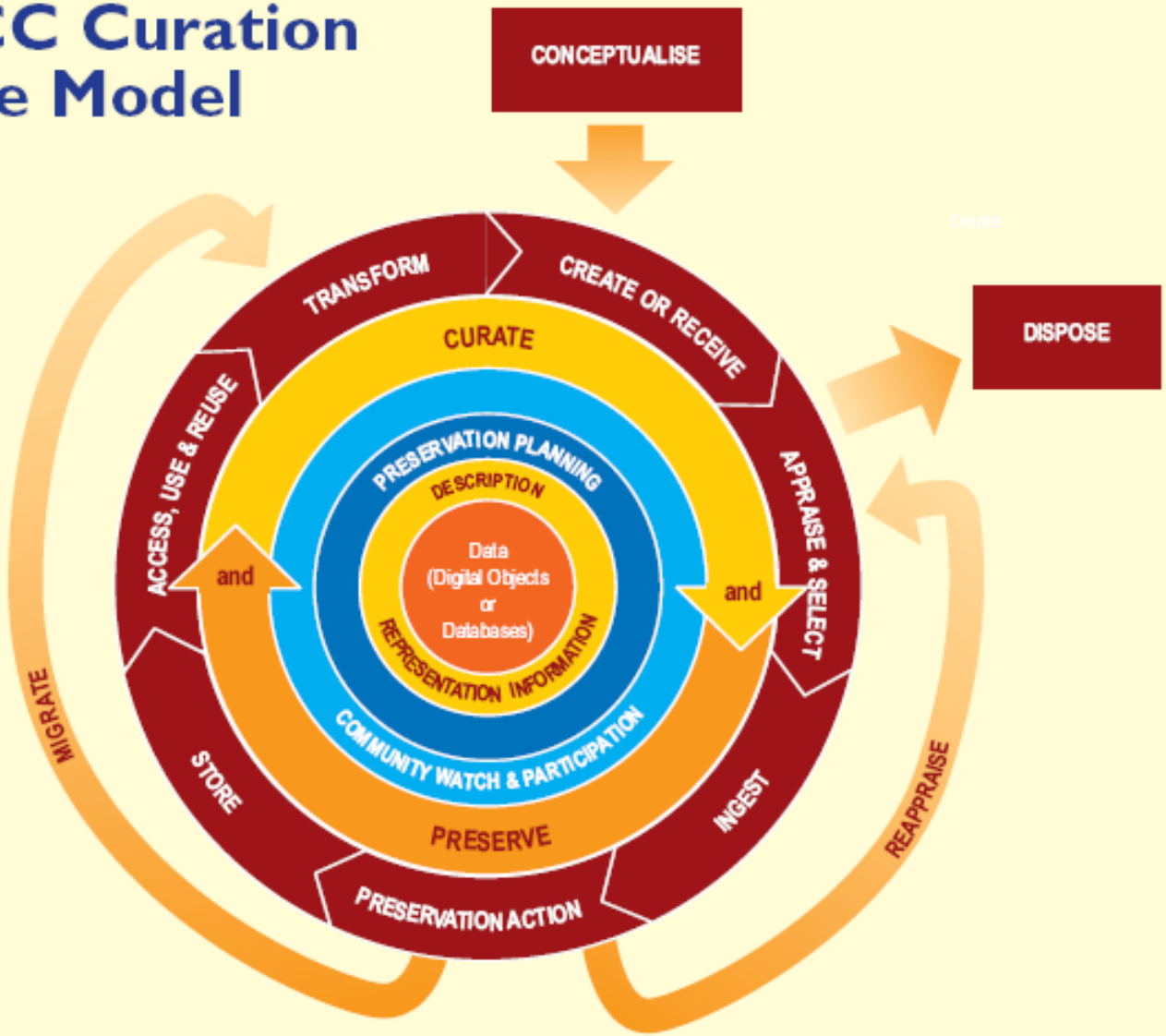
RSP
www.rsp.ac.uk

# 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

RSP
www.rsp.ac.uk

# The DCC Curation Lifecycle Model




RSP
www.rsp.ac.uk

This is where your hardware will end up

# Make sure your data doesn't!

Research outputs go in research repositories

eprints

Maureen Pennock
Steve Hitchcock

www.rsp.ac.uk

REPOSITORIES SUPPORT PROJECT

JISC RepositoryNet