EPrints & Preservation

David Tarrant

University of Southampton (UK)

dct05r@ecs.soton.ac.uk

Preserv.org.uk
Repository Preservation and Interoperability

eprints

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”

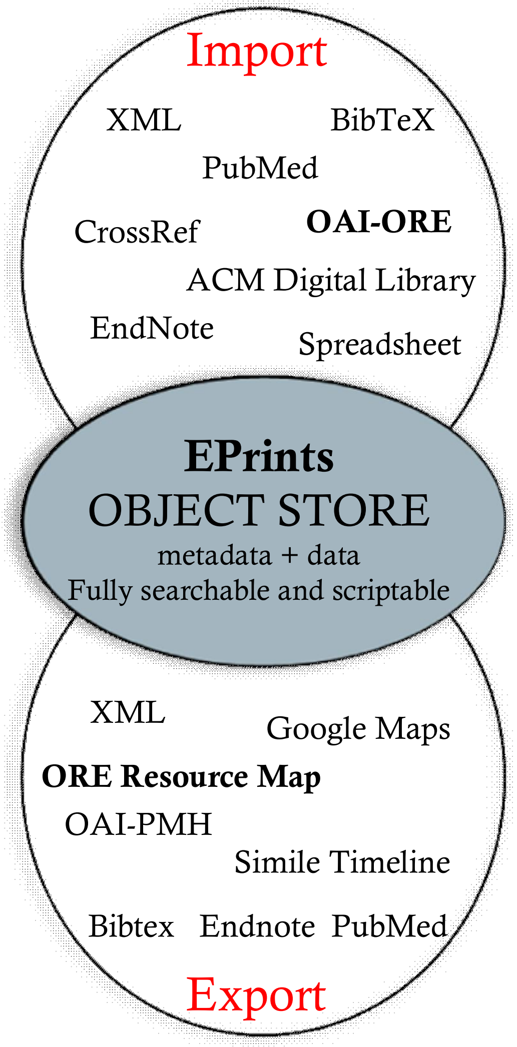
“Grassroots” preservation for Institutional and Small Business Outputs

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**



Import

XML BibTeX

PubMed

CrossRef **OAI-ORE**

ACM Digital Library

EndNote Spreadsheet

EPrints

OBJECT STORE

metadata + data

Fully searchable and scriptable

XML Google Maps

ORE Resource Map

OAI-PMH

Simile Timeline

Bibtex Endnote PubMed

Export

eprints: Architecture

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

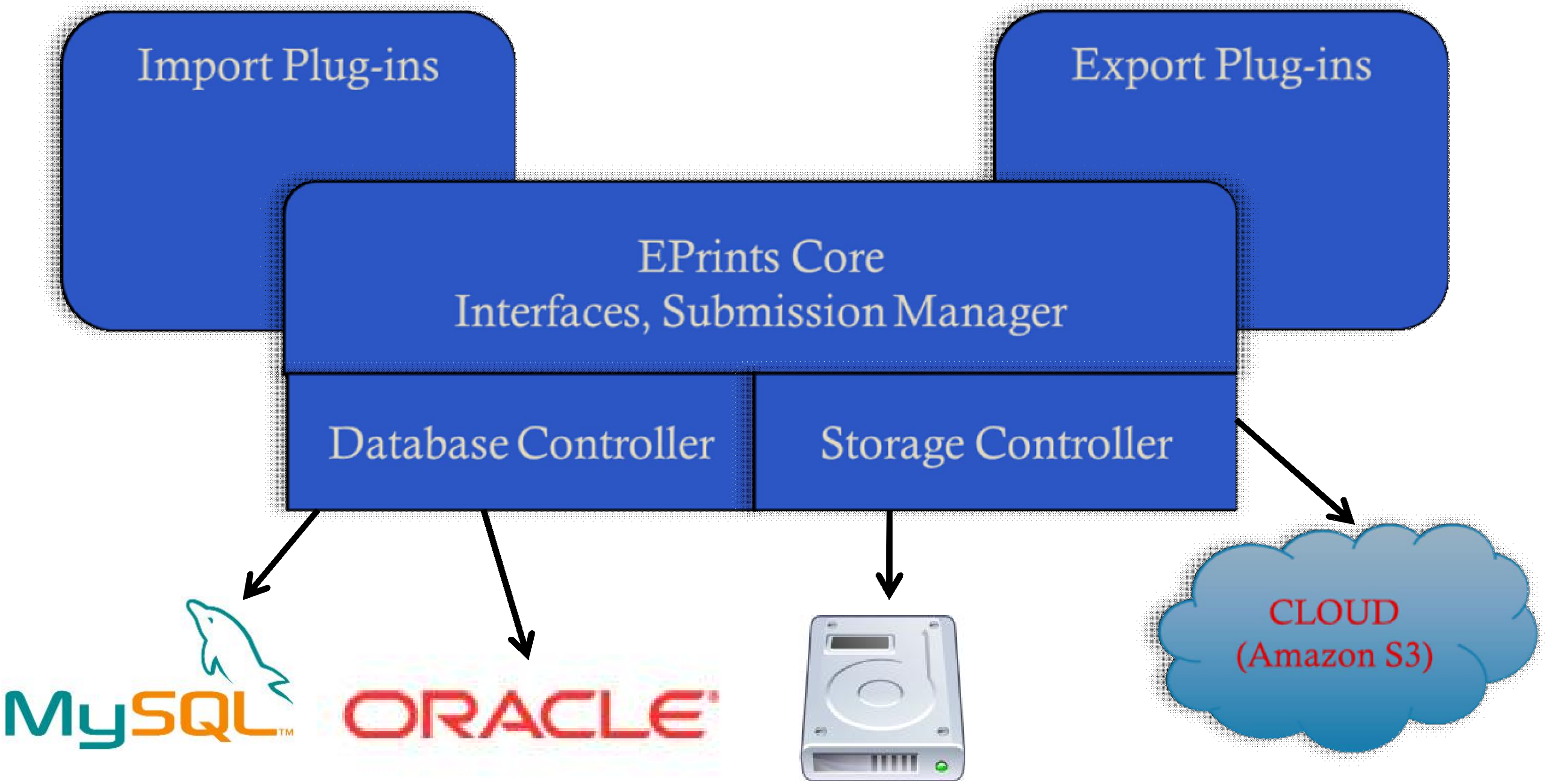


Diagram Represents Proposed EPrints 3.2 Architecture

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

amazon webservices™

- Simple to interface with Repository Software

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.

eCrystals

- 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

The Preservation Process

Preservation - Check

- Bit checking & checksum calculation

Preservation - Analyse

- What is the type of file, is the file valid?
- Is the file at risk of not having an editor/reader?
- Is there a better format available? Lossless or Lossy?

Preservation - Action

- File migration to avert risks found by analysis.
- Movement of file to new storage.

Preservation - Analysis

Preservation - Analyse

- What is the type of file, is the file valid?
  - Droid is a good classification tool for this.

The National Archives

- Is the file at risk of not having an editor/reader?
  - Functionality is being developed in PRONOM technical registry.

The technical registry PRONOM

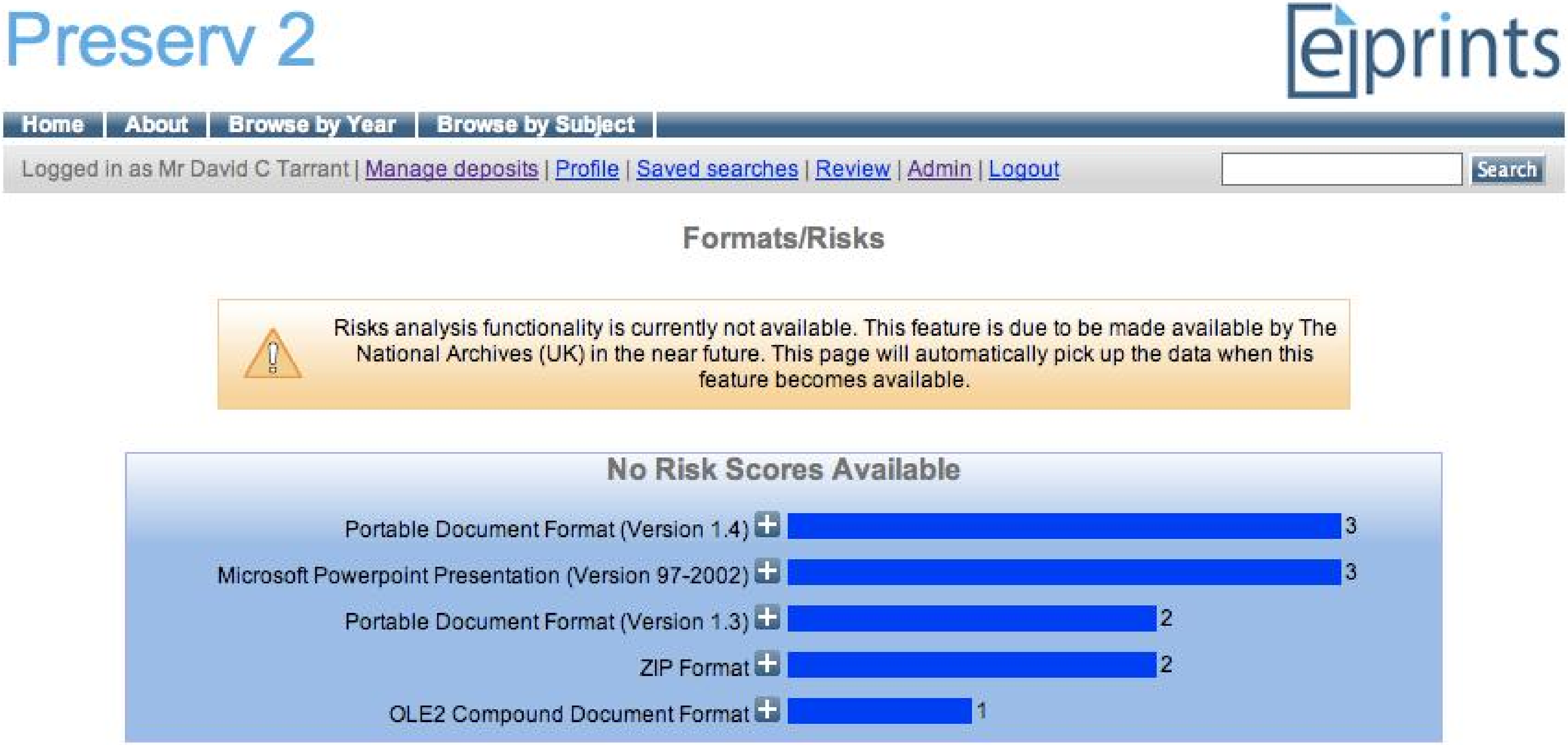
- Is there a better format available? Lossless or Lossy?
  - Planets registry of tools.

planets

Preservation - Analysis

Preservation - Analyse

EPrints File Classification


Preserv 2
eprints
Home
About
Browse by Year
Browse by Subject
Logged in as Mr David C Tarrant | Manage deposits | Profile | Saved searches | Review | Admin | Logout
Search

Formats/Risks

Risks analysis functionality is currently not available. This feature is due to be made available by The National Archives (UK) in the near future. This page will automatically pick up the data when this feature becomes available.

No Risk Scores Available

Portable Document Format (Version 1.4)	3
Microsoft Powerpoint Presentation (Version 97-2002)	3
Portable Document Format (Version 1.3)	2
ZIP Format	2
OLE2 Compound Document Format	1

Risk Analysis

Preservation - Analyse

The technical registry
PRONOM

- Is the file at risk of not having an editor/reader?
  - Functionality is being developed in PRONOM technical registry.

- Simple SOAP web service

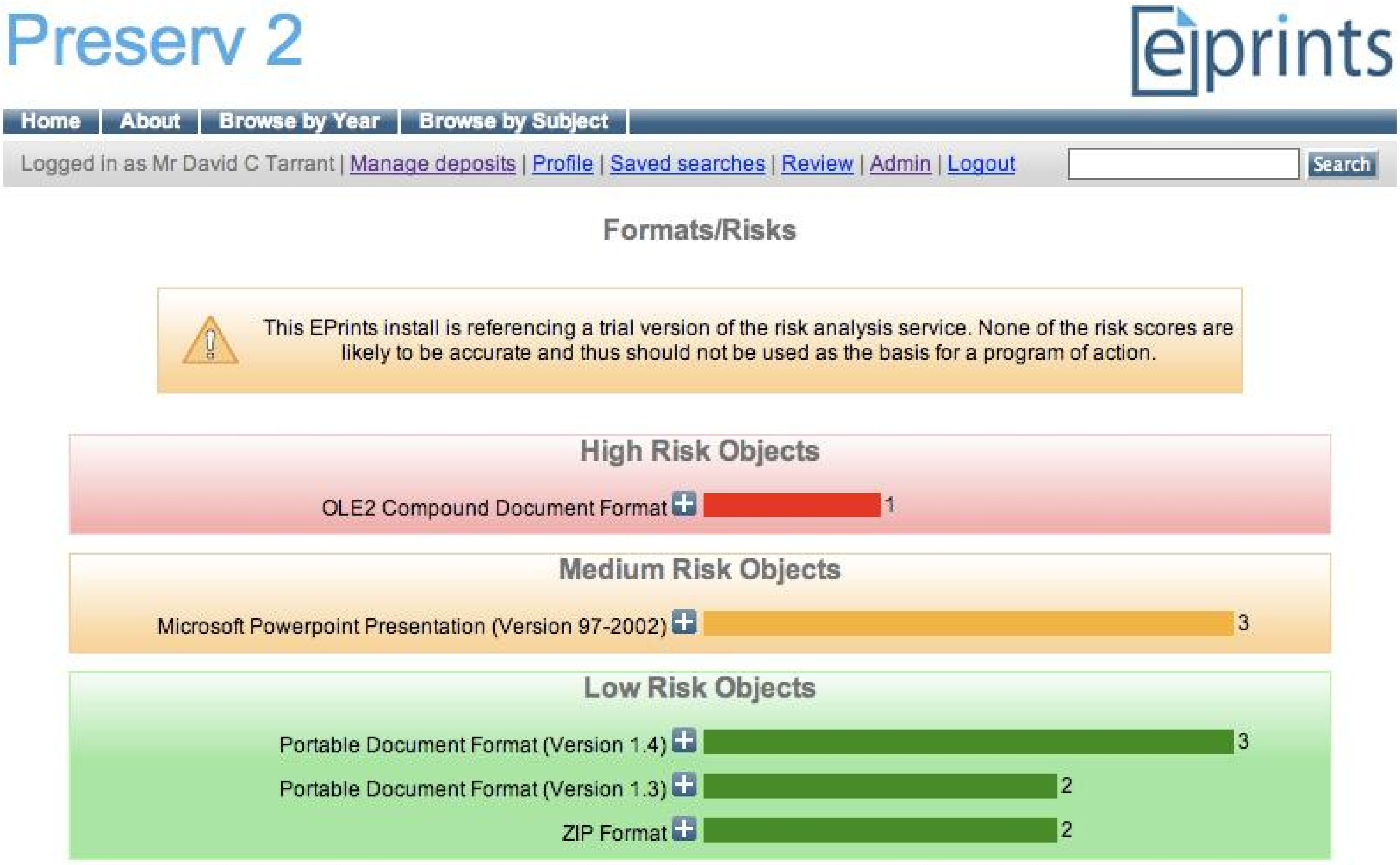
- Takes file format identification id's, hands back risk score.
- Breakdown of risk score may also be available in future releases.

- A stub you can download and run providing this functionality before the official release with mock up risk scores is available at http://preserv2.googlecode.com

Risk Analysis

Preservation - Analyse

EPrints File Classification + Risk Analysis


Preserv 2
eprints
Home
About
Browse by Year
Browse by Subject
Logged in as Mr David C Tarrant | Manage deposits | Profile | Saved searches | Review | Admin | Logout
Search

Formats/Risks

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	1

Medium Risk Objects

Microsoft Powerpoint Presentation (Version 97-2002)	3

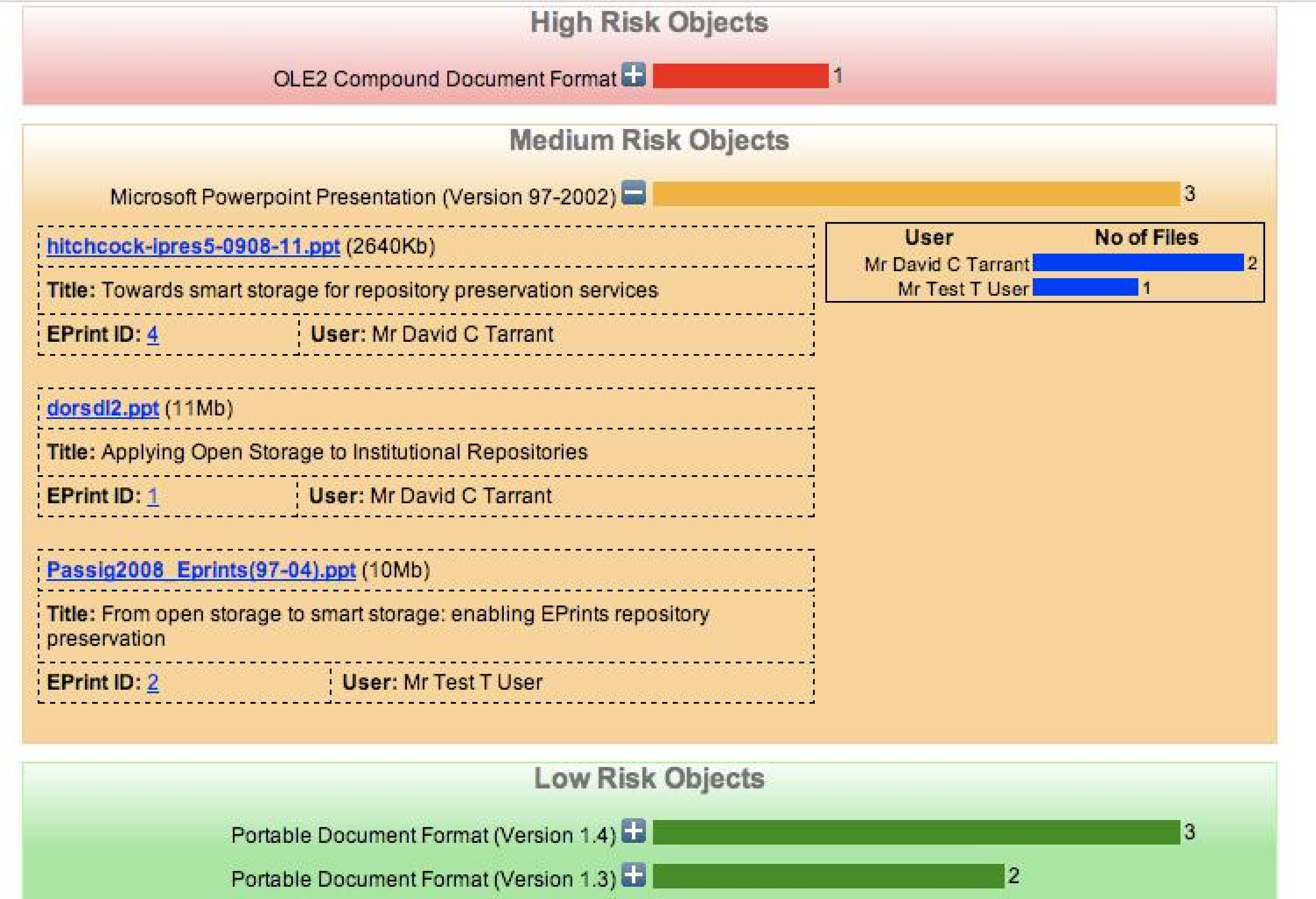
Low Risk Objects

Portable Document Format (Version 1.4)	3
Portable Document Format (Version 1.3)	2
ZIP Format	2

Risk Analysis

Preservation - Analyse

EPrints File Classification + Risk Analysis



High Risk Objects

OLE2 Compound Document Format ⊞ 1

Medium Risk Objects

Microsoft Powerpoint Presentation (Version 97-2002) ⊟ 3

hitchcock-ipres5-0908-11.ppt (2640Kb)

Title: Towards smart storage for repository preservation services

EPrint ID: 4 **User:** Mr David C Tarrant

dorsdl2.ppt (11Mb)

Title: Applying Open Storage to Institutional Repositories

EPrint ID: 1 **User:** Mr David C Tarrant

Passig2008_Eprints(97-04).ppt (10Mb)

Title: From open storage to smart storage: enabling EPrints repository preservation

EPrint ID: 2 **User:** Mr Test T User

User	No of Files
Mr David C Tarrant	2
Mr Test T User	1

Low Risk Objects

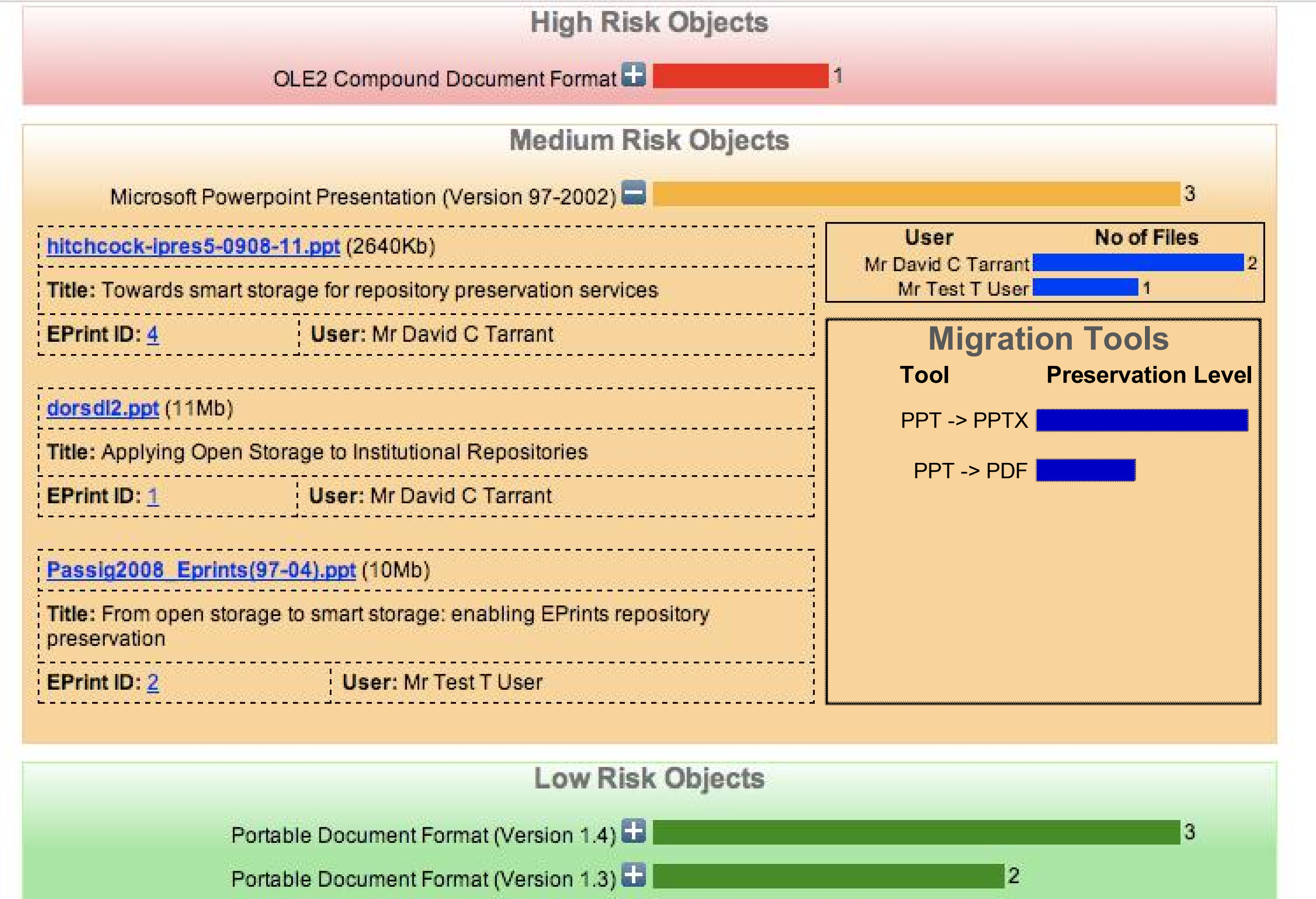
Portable Document Format (Version 1.4) ⊞ 3

Portable Document Format (Version 1.3) ⊞ 2

Transformation?

Preservation - Action

Mock up Transformation Interface



High Risk Objects

OLE2 Compound Document Format ⊞ 1

Medium Risk Objects

Microsoft Powerpoint Presentation (Version 97-2002) ⊟ 3

hitchcock-ipres5-0908-11.ppt (2640Kb)

Title: Towards smart storage for repository preservation services

EPrint ID: 4 **User:** Mr David C Tarrant

dorsdl2.ppt (11Mb)

Title: Applying Open Storage to Institutional Repositories

EPrint ID: 1 **User:** Mr David C Tarrant

Passig2008_Eprints(97-04).ppt (10Mb)

Title: From open storage to smart storage: enabling EPrints repository preservation

EPrint ID: 2 **User:** Mr Test T User

User	No of Files
Mr David C Tarrant	2
Mr Test T User	1

Migration Tools

Tool	Preservation Level
PPT -> PPTX	
PPT -> PDF	

Low Risk Objects

Portable Document Format (Version 1.4) ⊞ 3

Portable Document Format (Version 1.3) ⊞ 2

Many Thanks!

eprints

David Tarrant
Les Carr
Steve Hitchcock

UNIVERSITY OF Southampton

Steve Hitchcock
Tim Brody

Adrian Brown

The National Archives

Preserv .org.uk
Repository Preservation and Interoperability

Neil Jeffries
Ben O’Steen
Sally Rumsey

UNIVERSITY · OF · OXFORD