Preservation as a Process of a Repository

David Tarrant

University of Southampton (UK)

dct05r@ecs.soton.ac.uk

Preserv .org.uk

Repository Preservation and Interoperability

A Few Definitions

Repository: A repository is a place where data is stored and maintained.

Wikipedia

IR: A repository captures and preserves the intellectual output of an institution.

The Case for Institutional Repositories – Raym Crow (SPARC 2002)

IR: In my view, a university-based institutional repository is a set of **services** that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members. It is most essentially an organizational commitment to the stewardship of these digital materials, including long-term preservation where appropriate, as well as organization and access or distribution.

Institutional Repositories: Essential Infrastructure For Scholarship In The Digital Age - Clifford A. Lynch

Service: A service is something provided directly to a user or 3rd party agent.

David Tarrant, 2008

Process: A process is something which is invisible to the user or agent.

David Tarrant, 2008

The Library

The Library

- A building to store books in.
- A means by which new books/publications can be acquired.
- An indexing system to give order.
- Provides a mean by which books can be found.
- Provides a way to borrow & return books.
- A preservation process, e.g. rebind books when they get damaged/worn.
- …

The Digital Library

The Library	The Digital Repository
• A building to store books in.	• A server to store resources on.
• A means by which new books/publications can be acquired.	• A way to ingest new resources.
• A indexing system to give order.	• A database of resources and metadata.
• Provides a mean by which books can be found.	• A search engine and dissemination pages.
• Provides a way to borrow & return books.	• Open access and downloads.
• A preservation process, e.g. rebind books when they get damaged/worn.	• A preservation process, e.g. check that the file on the server can still be read/accessed.
• …	• …

> “In my view, a university-based institutional repository provides a set of services. The repository itself consists of a set of PROCESSES …”

Processes

Service: A service is something provided directly to a user or 3rd party agent.

Process: A process is something which is invisible to the user or agent.

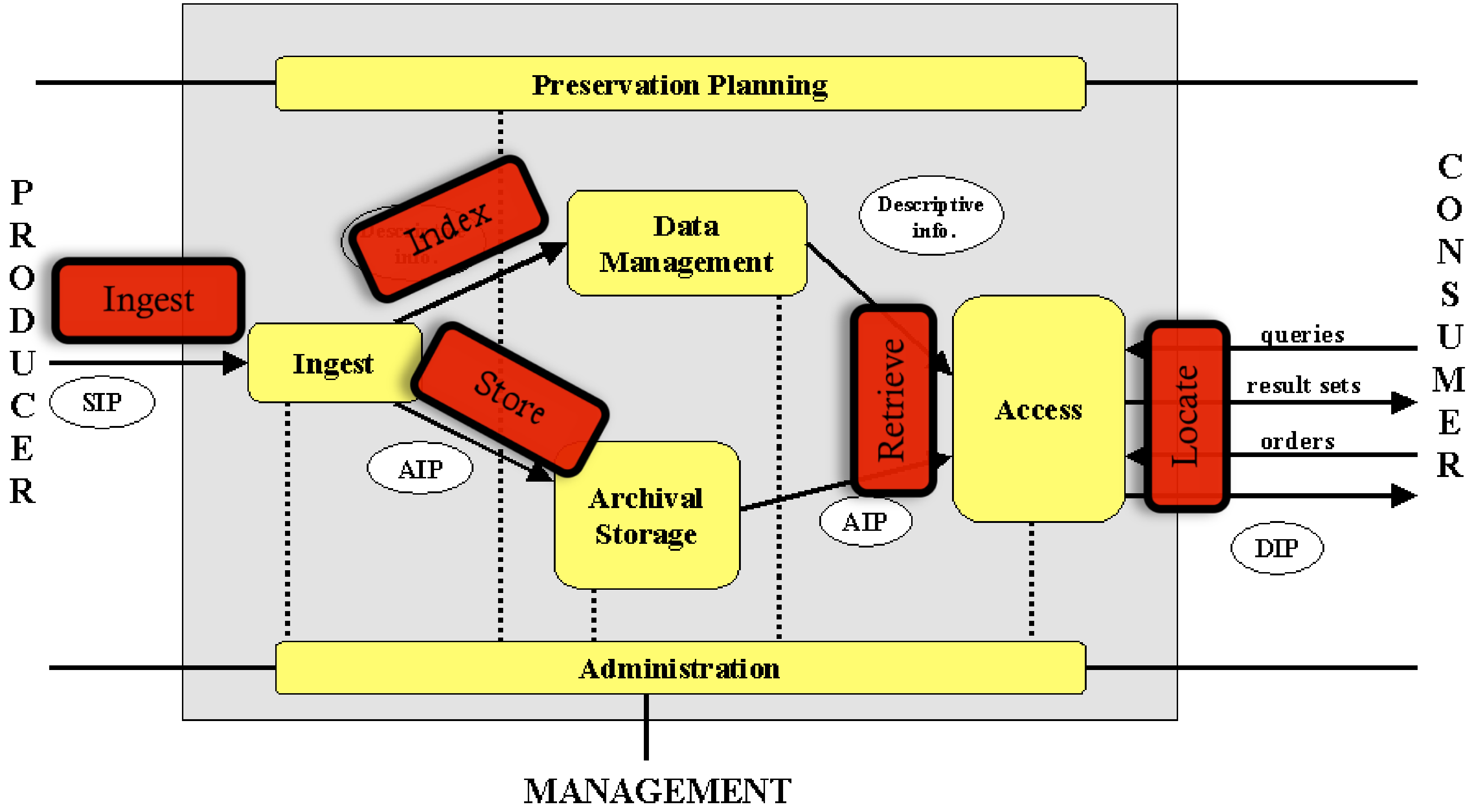
There are lots of Processes

Processes happen in parallel

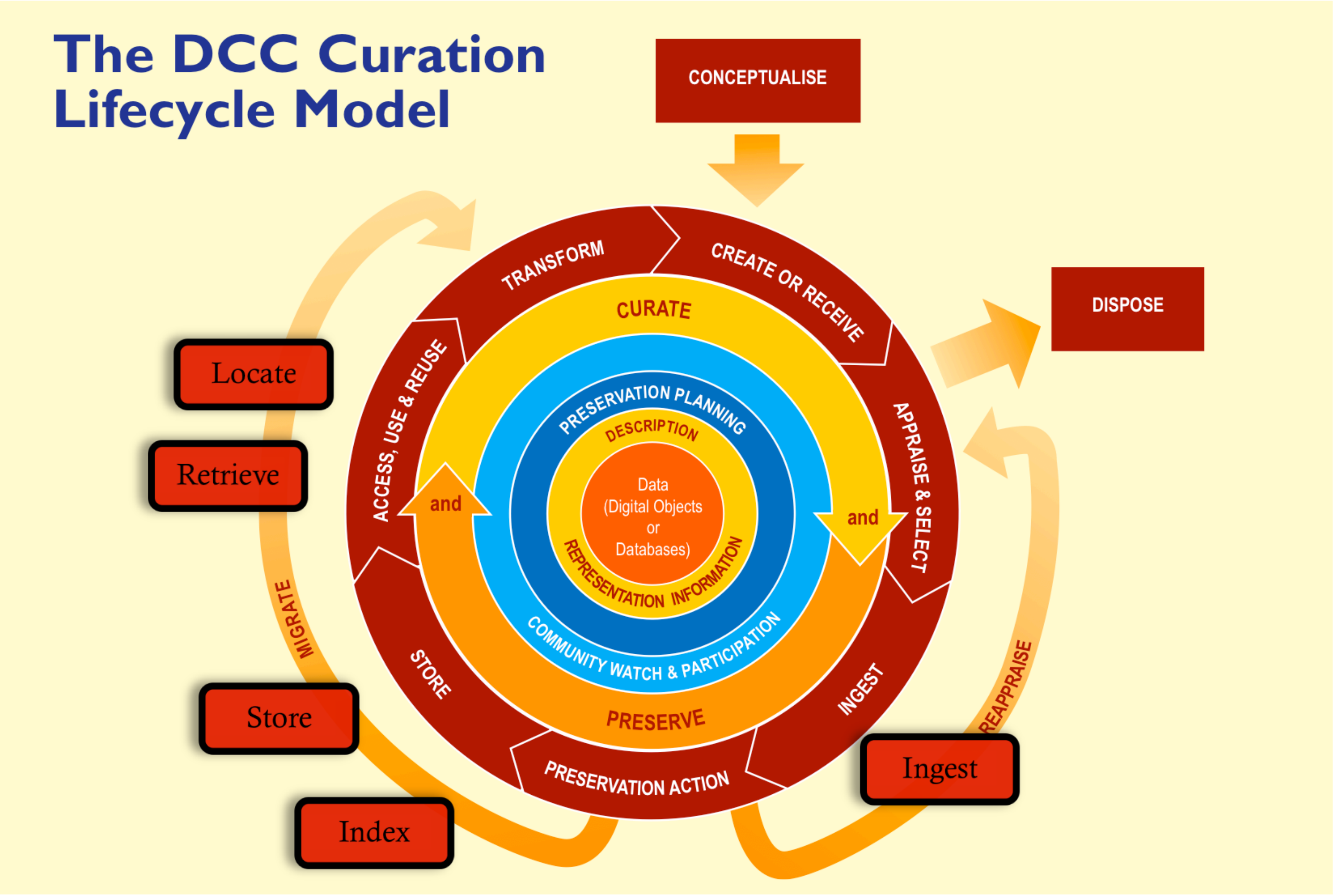
Processes happen in different orders

Processes and OAIS

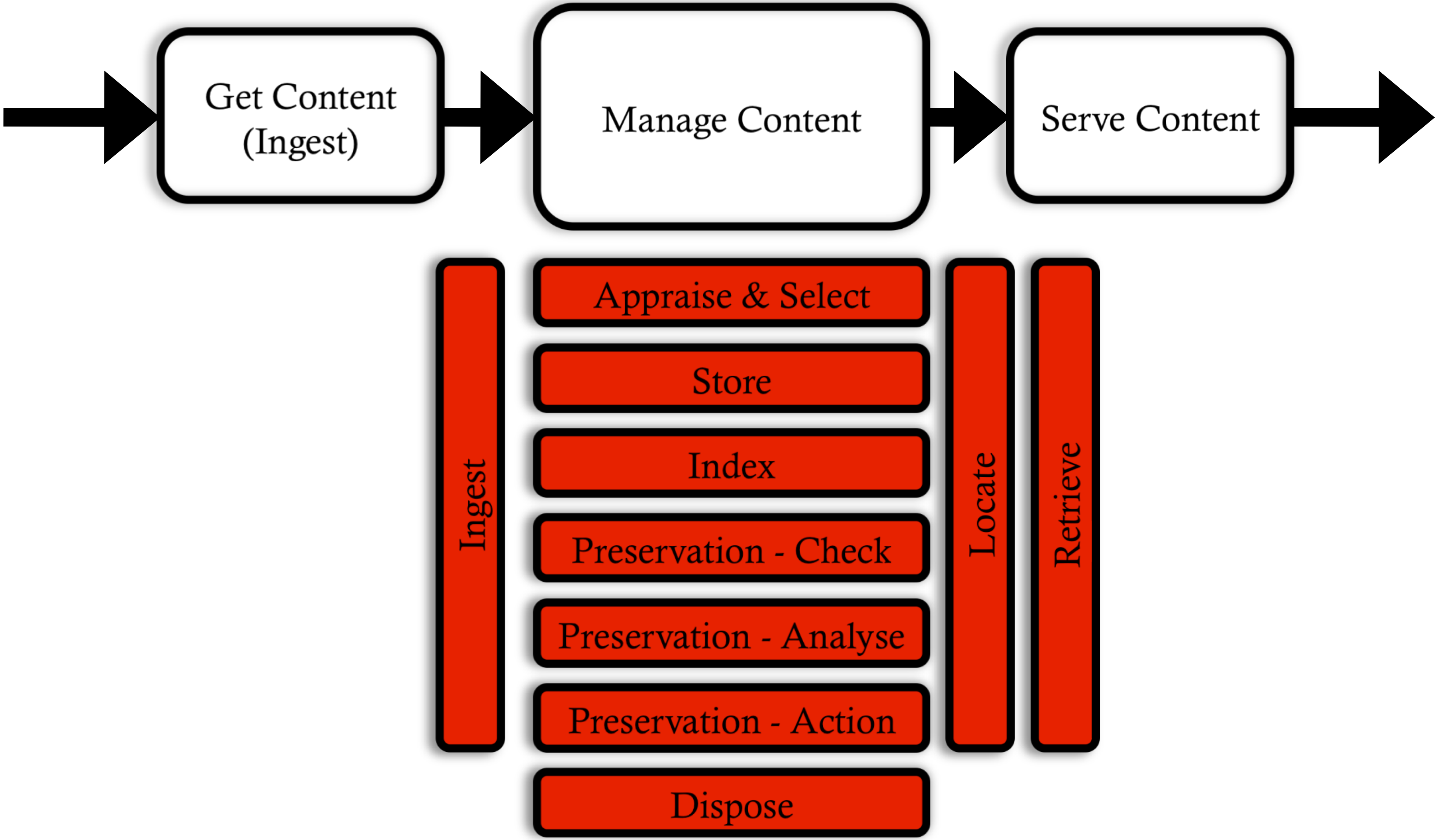
Many existing models contain this notion of processes and services, just not necessarily in a modern light. This doesn't however mean they are "wrong" or "right" they are just guiding principals.



Processes in the DCC Model



The 3 Stage Model



Get Content (Ingest)

Manage Content

Serve Content

Ingest

Appraise & Select

Store

Index

Preservation - Check

Preservation - Analyse

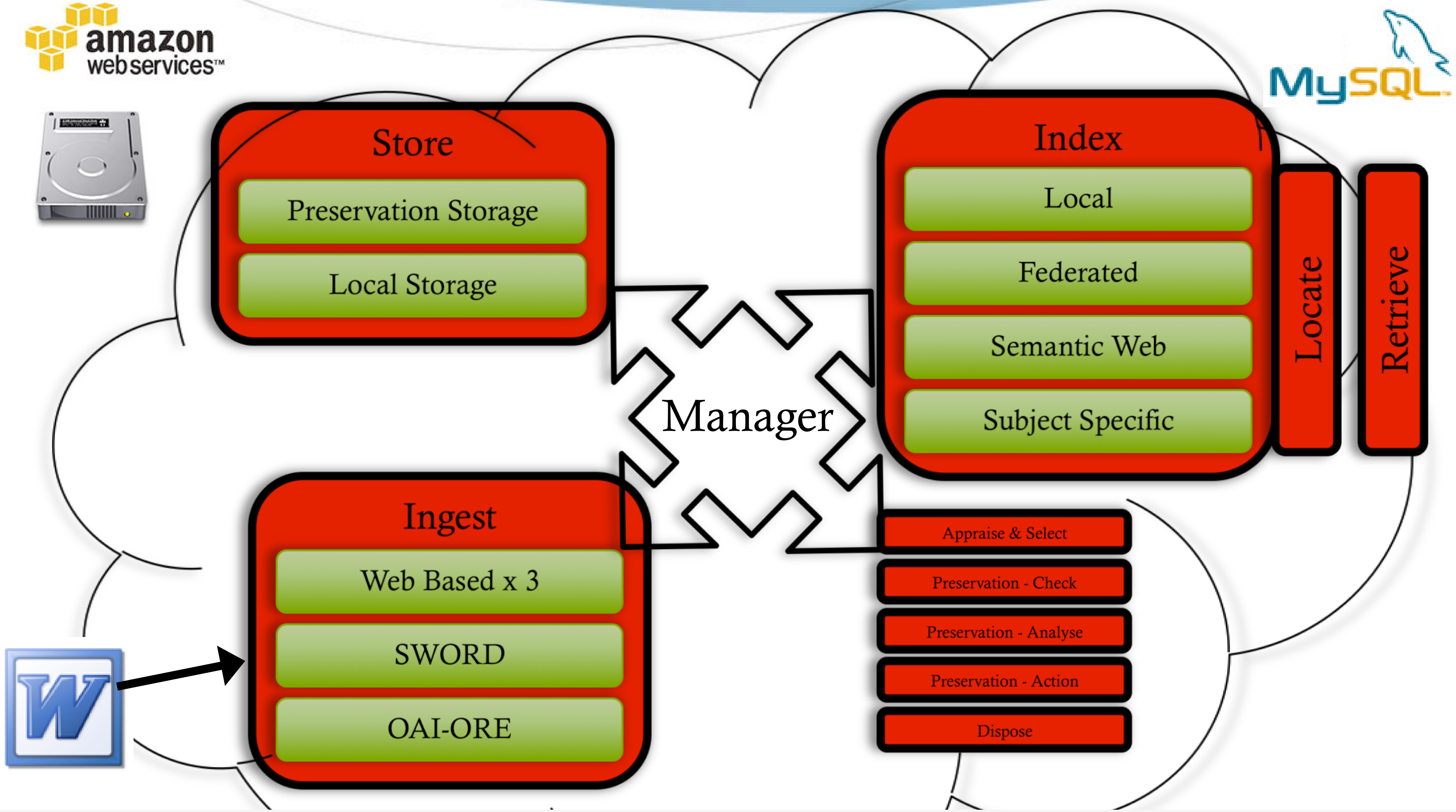
Preservation - Action

Dispose

Locate

Retrieve

Breaking up the Repository



The manager may provide capability to perform one or more of the processes. Typically the manager is all that is used.

Repository Management Software

A set of Pipes/Workflows* which know how to translate inputs into outputs.

Examples:

- OAI-ORE which contains Files and Metadata is split by the management software into File/Metadata storage and indexes.
- A request for a set of objects related to a single author is translated into a query to an index and a retrieve from the storage.

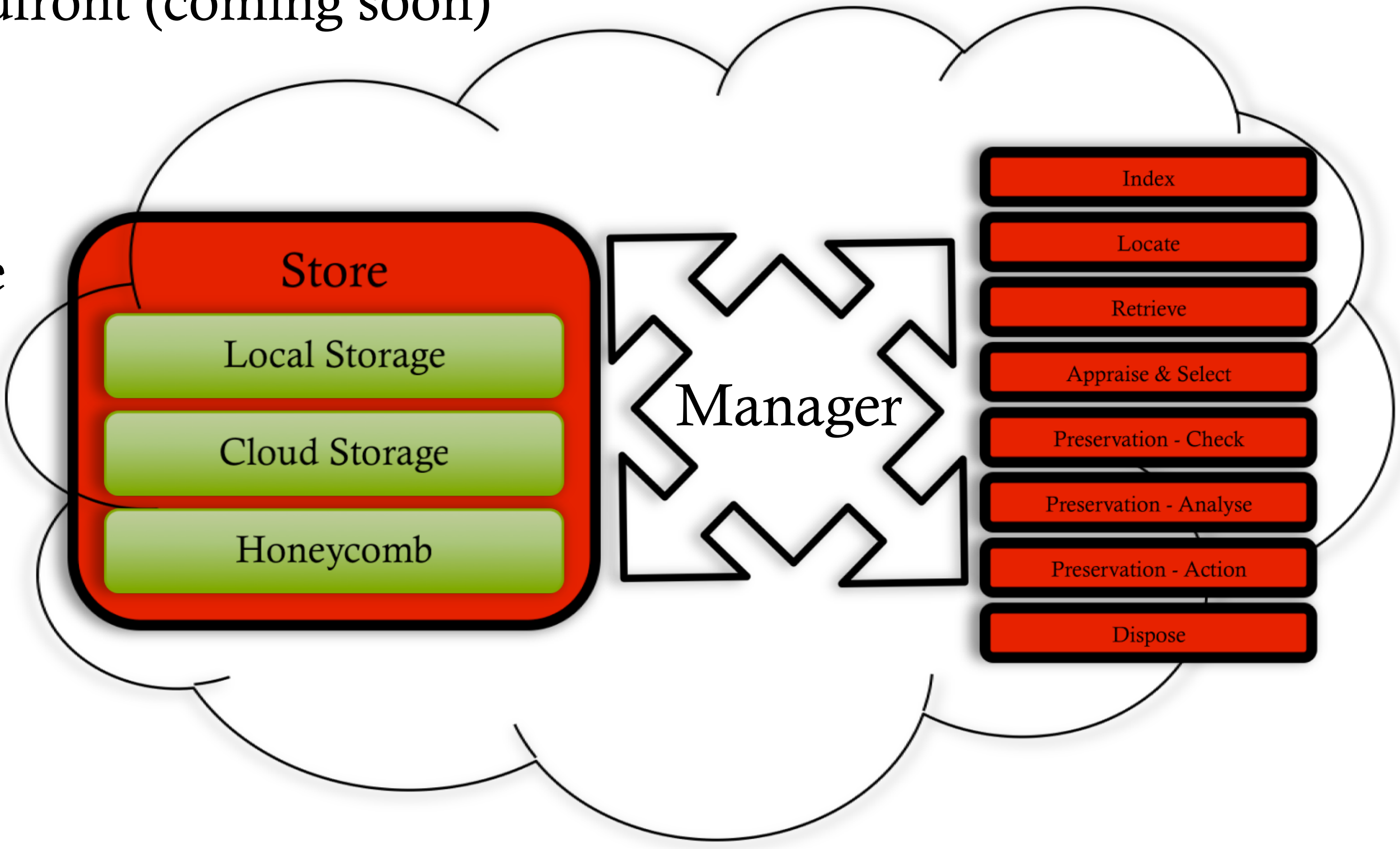
*Depending on your own definition you could also add "Middleware"

eprints :Storage Controller

- EPrints Storage Controller works!
  - Local Storage Plugin (legacy)
  - Honeycomb Storage Plugin
  - Amazon Cloudfront (coming soon)

- Honeycomb Stats
  - 4MB/s ingest*
  - 200MB/s retrieve

*USB2.0 max speed



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

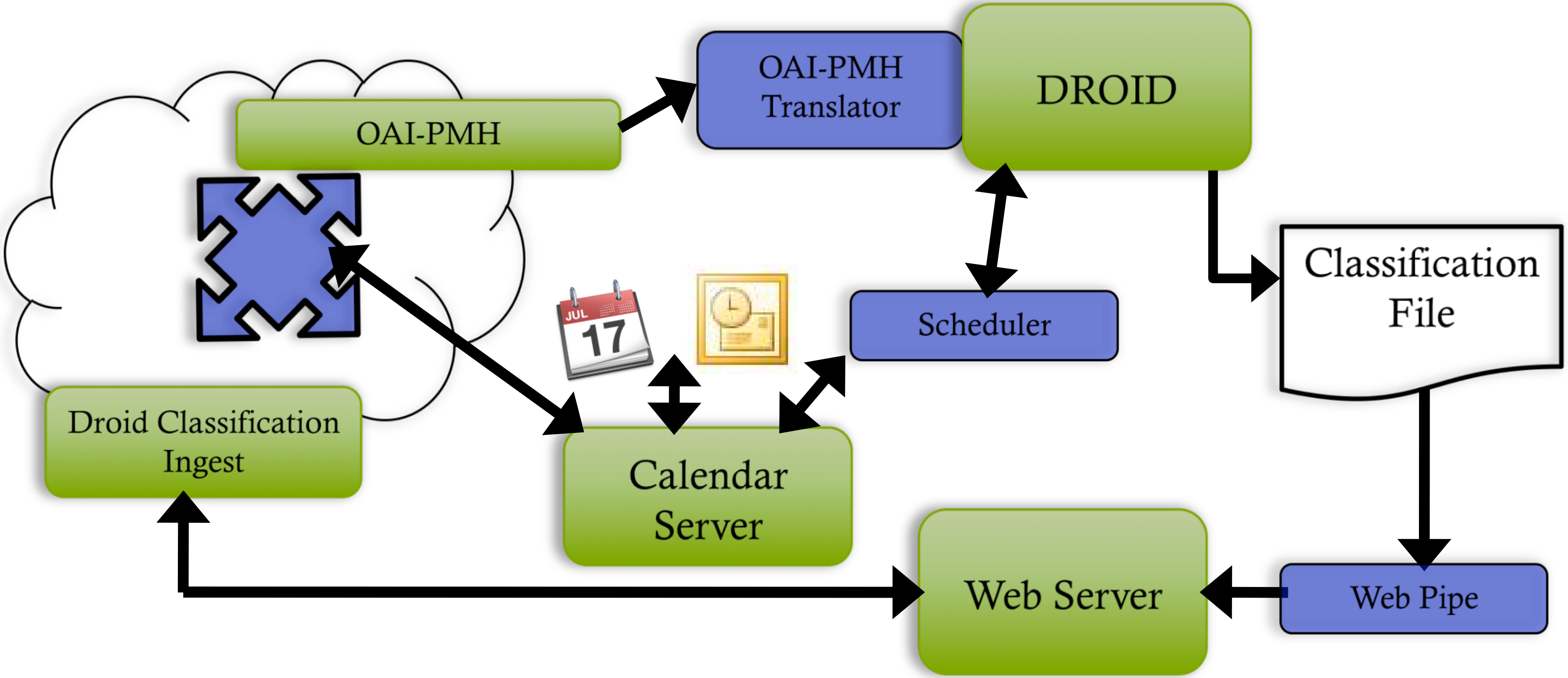
Preservation - Analyse

Key

Pipe / Workflow

Service

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



OAI-PMH

OAI-PMH Translator

DROID

Classification File

Scheduler

Droid Classification Ingest

Calendar Server

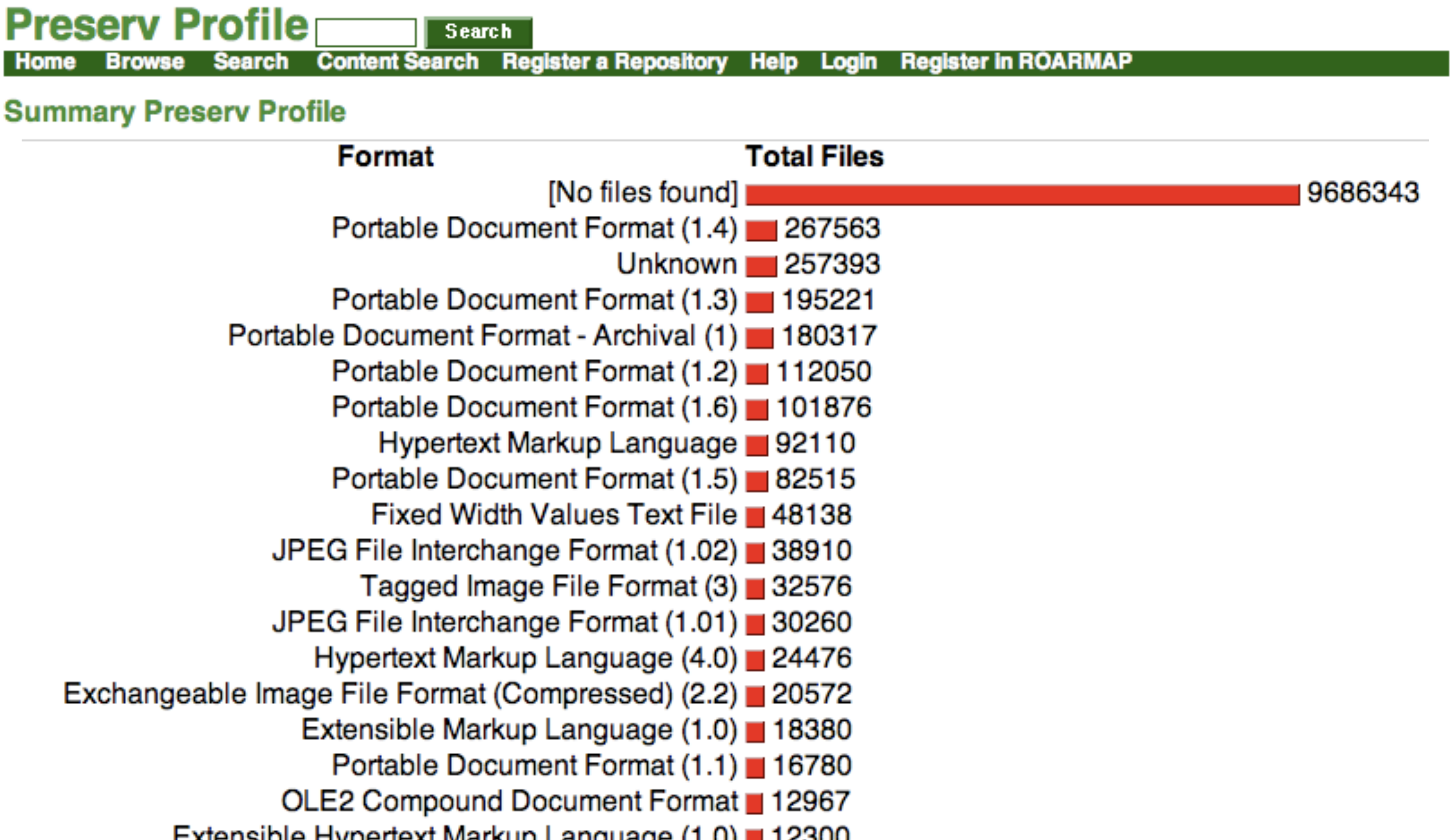
Web Server

Web Pipe

Preservation - Analysis

Preservation - Analyse

PRONOM-ROAR (Preserv 1)
http://roar.eprints.org



Preserv Profile

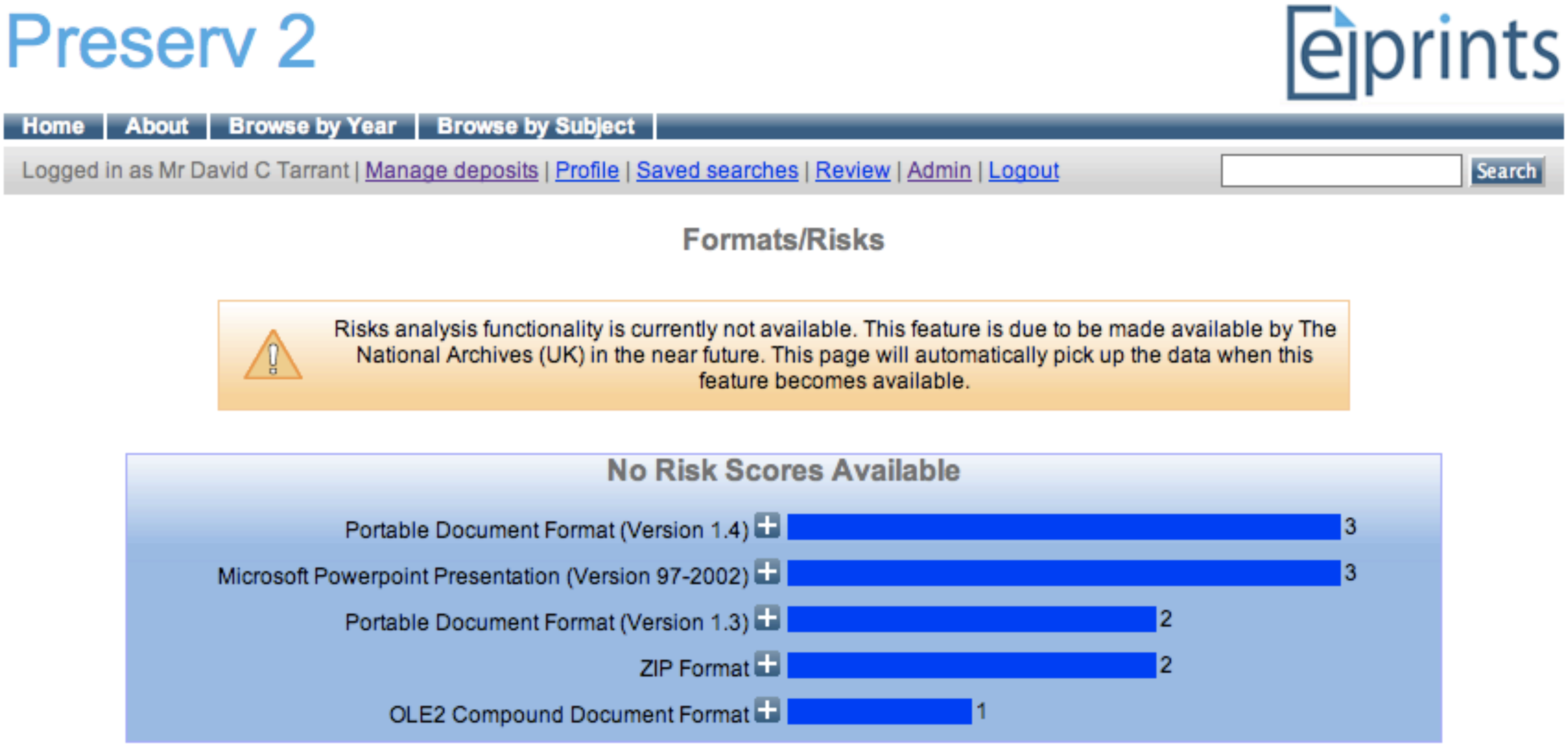
Summary Preserv Profile

Format	Total Files
[No files found]	9686343
Portable Document Format (1.4)	267563
Unknown	257393
Portable Document Format (1.3)	195221
Portable Document Format - Archival (1)	180317
Portable Document Format (1.2)	112050
Portable Document Format (1.6)	101876
Hypertext Markup Language	92110
Portable Document Format (1.5)	82515
Fixed Width Values Text File	48138
JPEG File Interchange Format (1.02)	38910
Tagged Image File Format (3)	32576
JPEG File Interchange Format (1.01)	30260
Hypertext Markup Language (4.0)	24476
Exchangeable Image File Format (Compressed) (2.2)	20572
Extensible Markup Language (1.0)	18380
Portable Document Format (1.1)	16780
OLE2 Compound Document Format	12967
Extensible Hypertext Markup Language (1.0)	12300

Preservation - Analysis

Preservation - Analyse

EPrints File Classification


Preserv 2
eprints
Home
About
Browse by Year
Browse by Subject
Search

Logged in as Mr David C Tarrant | Manage deposits | Profile | Saved searches | Review | Admin | Logout

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

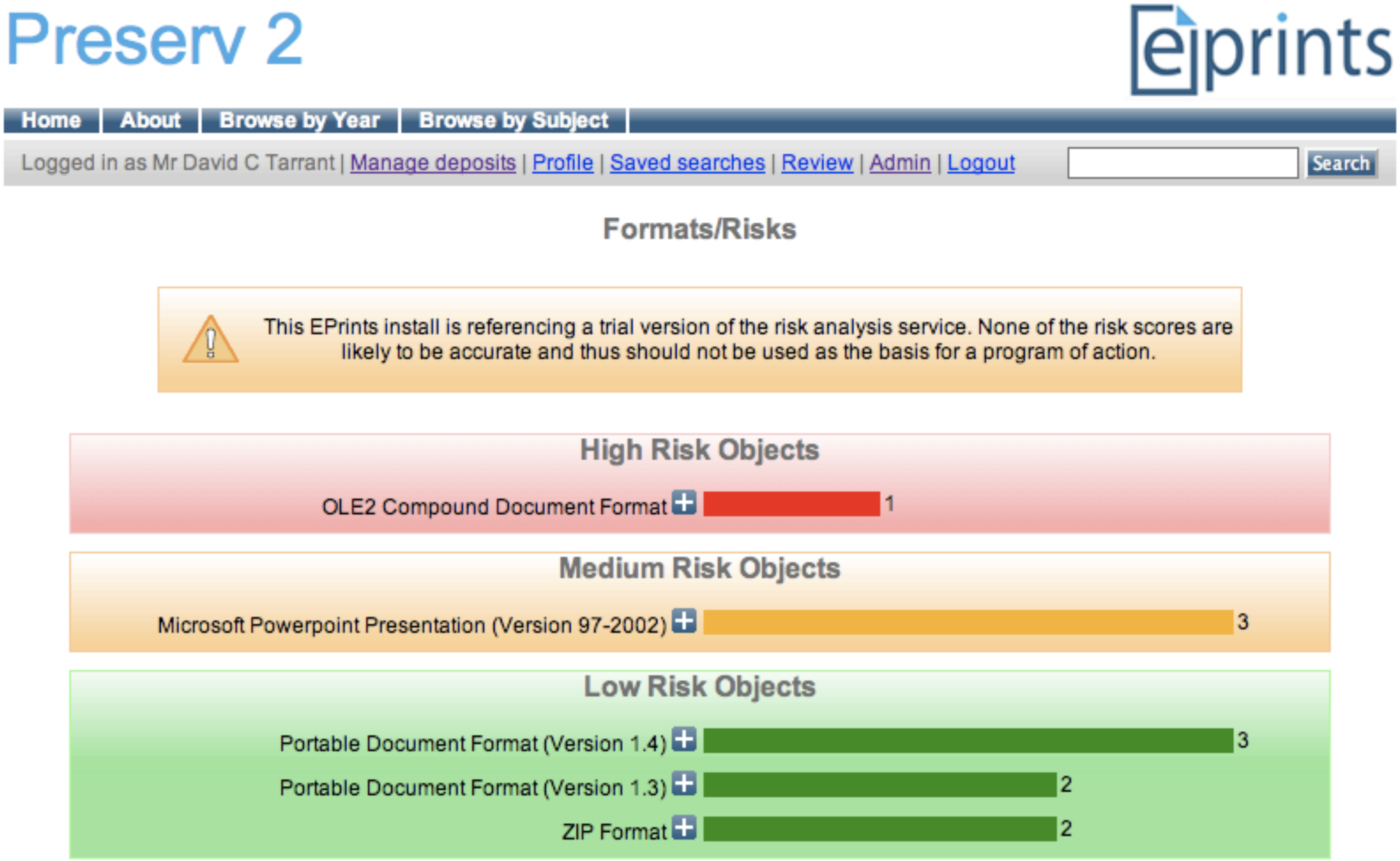
- Takes file format identification id's, hands back risk score.
- Breakdown of risk score may also be available in future releases.

- 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



Logged in as Mr David C Tarrant | Manage deposits | Profile | Saved searches | Review | Admin | Logout

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

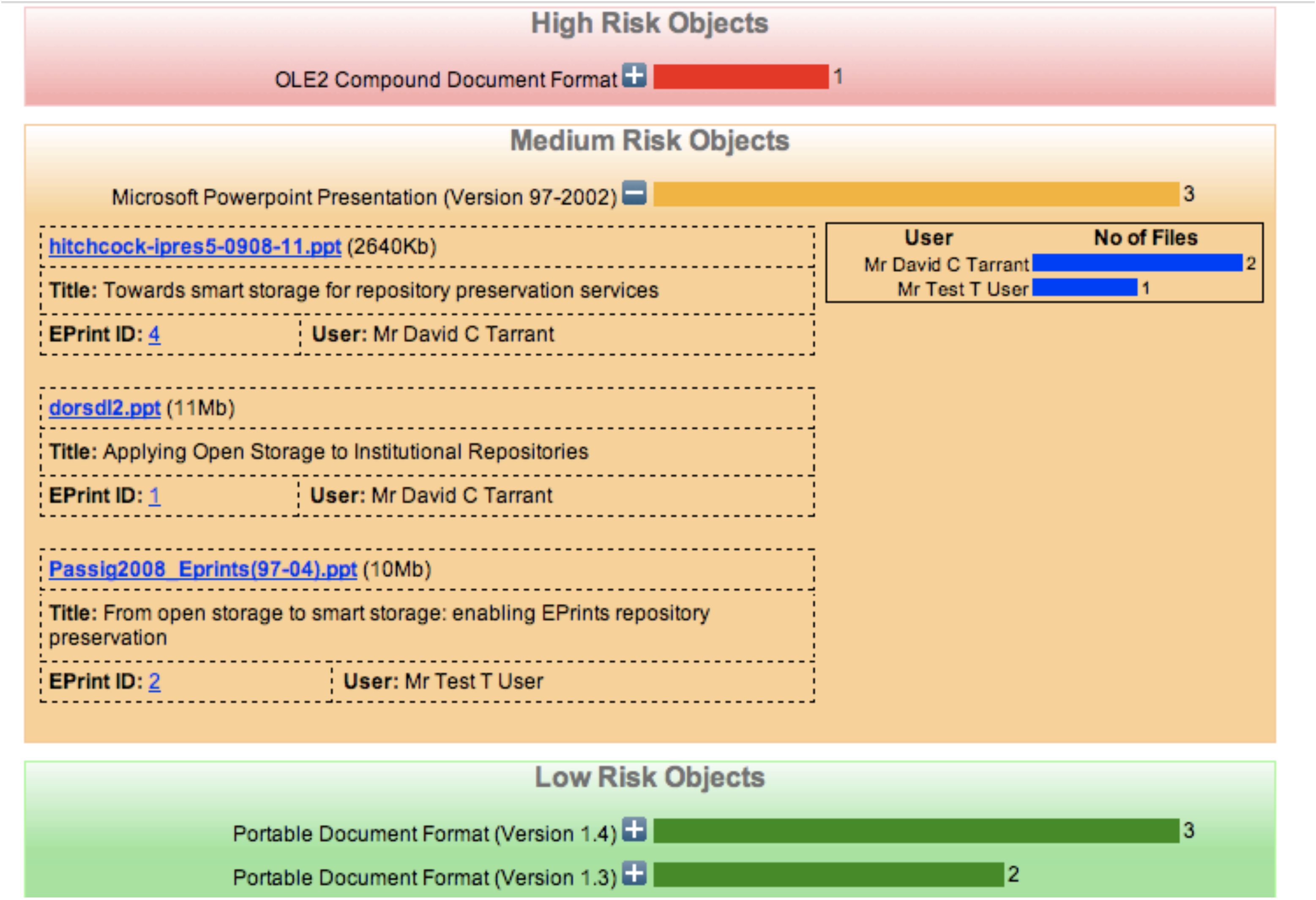
Portable Document Format (Version 1.3) 2

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

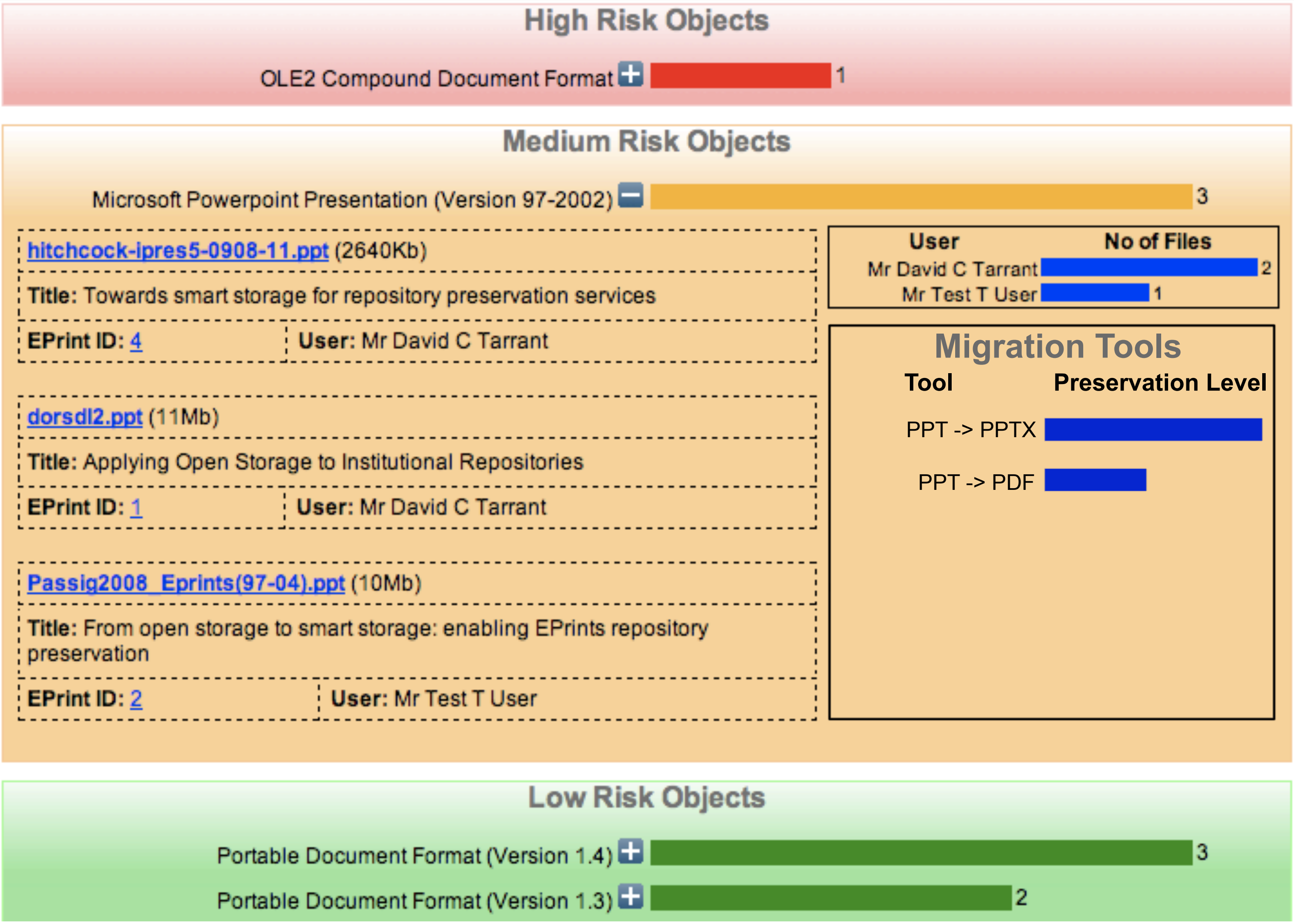
Portable Document Format (Version 1.4) ⊞ 3

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

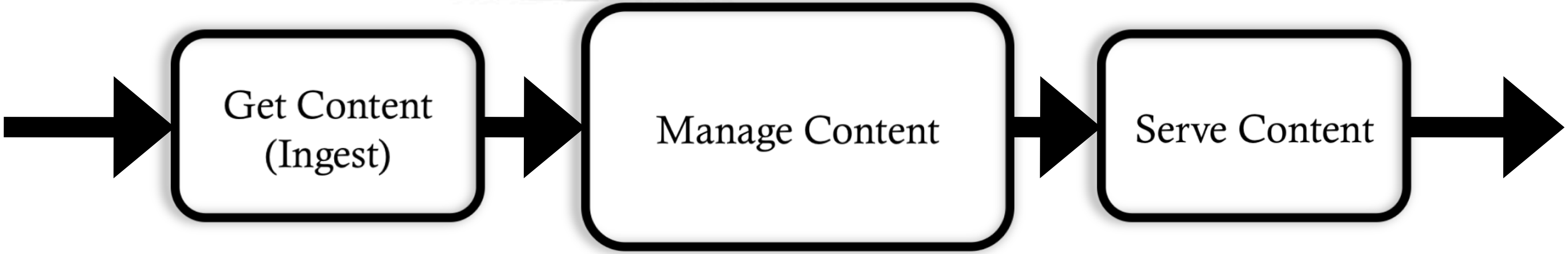
Tool	Preservation Level
PPT -> PPTX	🟦
PPT -> PDF	🟦

Low Risk Objects

Portable Document Format (Version 1.4) ⊞ 🟩 3

Portable Document Format (Version 1.3) ⊞ 🟩 2

Summary – 1/2



- Processes, Services and Glue
- Storage Controller provides an API you can glue to.
- Enabling preservation for any repository model by writing small bits of glue.
- Portable services are more powerful, faster and cheaper.
- Make use of existing and supported software where possible.

Summary – 2/2

Preserv .org.uk

Repository Preservation and Interoperability

+ eprints

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*.

Many Thanks!

Preserv .org.uk
Repository Preservation and Interoperability

David Tarrant
Steve Hitchcock

UNIVERSITY OF Southampton

Neil Jefferies
Ben O'Steen
Sally Rumsey

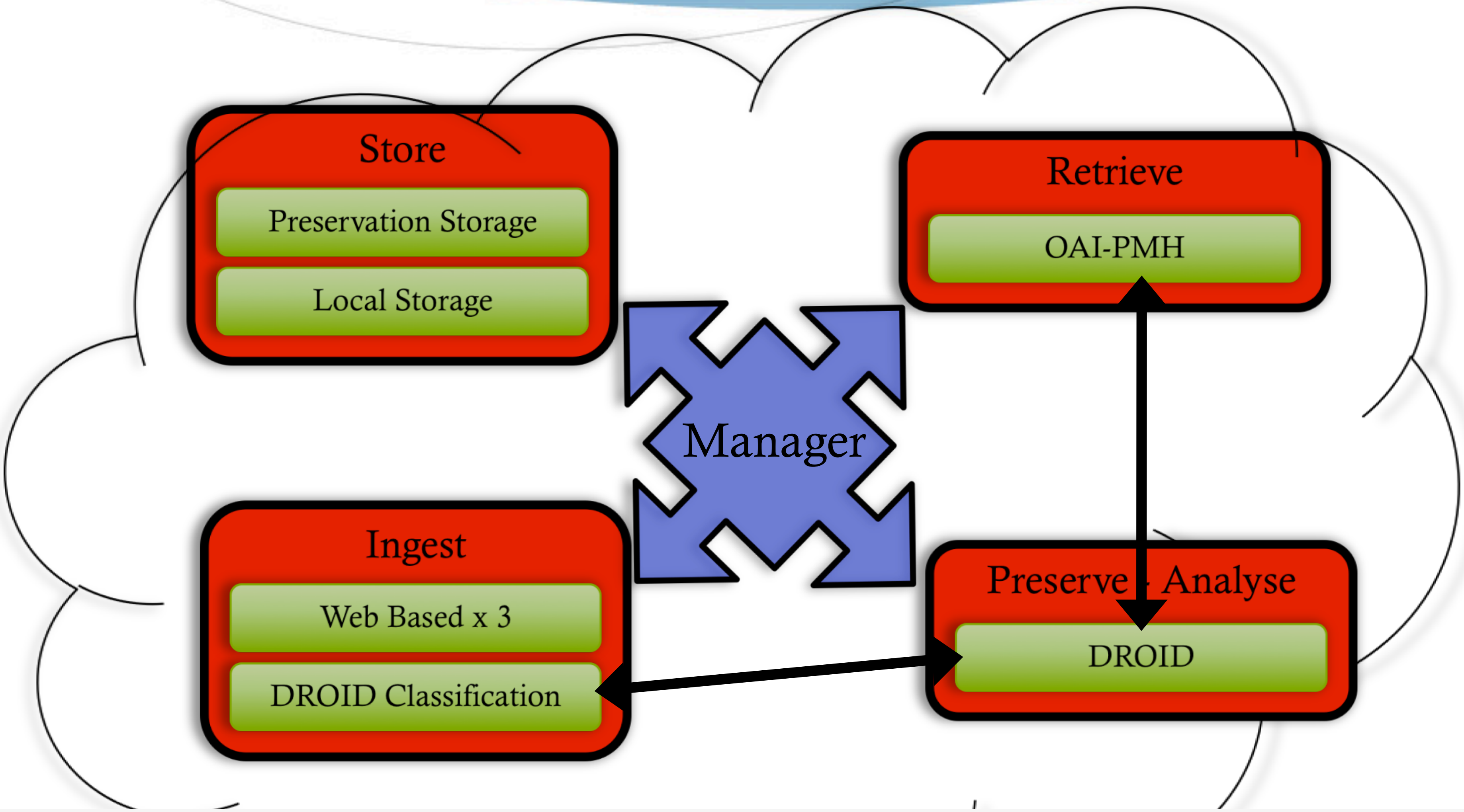
UNIVERSITY · OF · OXFORD

Adrian Brown

The National Archives

Appendix Slides

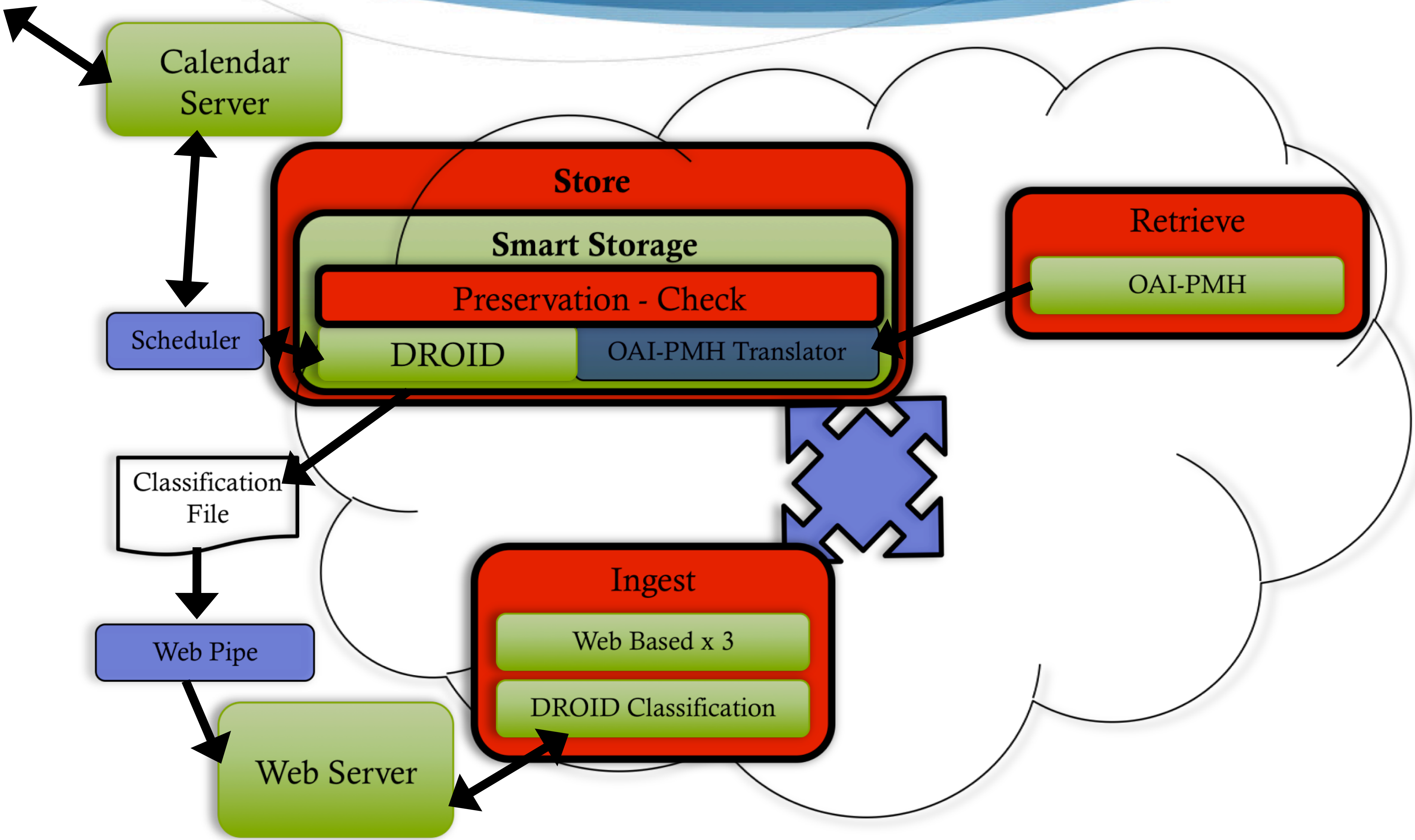
Other options for DROID Positioning



This is not the recommended solution as DROID is a 3rd party service for your repository.

All other services are provided by your repository.

DROID Alongside Your Resources


Calendar Server
Store
Smart Storage
Preservation - Check
Scheduler
DROID
OAI-PMH Translator
Retrieve
OAI-PMH
Classification File
Ingest
Web Based x 3
DROID Classification
Web Pipe
Web Server