Data management in a networked environment:

From Data Archives to Preservation Services

December 3rd, 2009

Gareth Knight

Centre for e-Research

Session Overview

1. Organisation background
2. Changing face of humanities research
3. New approaches to capture and curation
4. Interoperating with disparate systems
5. Curation management
6. Conclusions

Background

CeRch
Centre for e-Research

- Established in 2007.
- Incorporates AHDS Executive projects and staff
- Several objectives:
  - Research into e-infrastructures, e-research methods, and digital informatics, including the application of e-science to research;
  - Host national and international projects and services
  - Teaching and consultancy

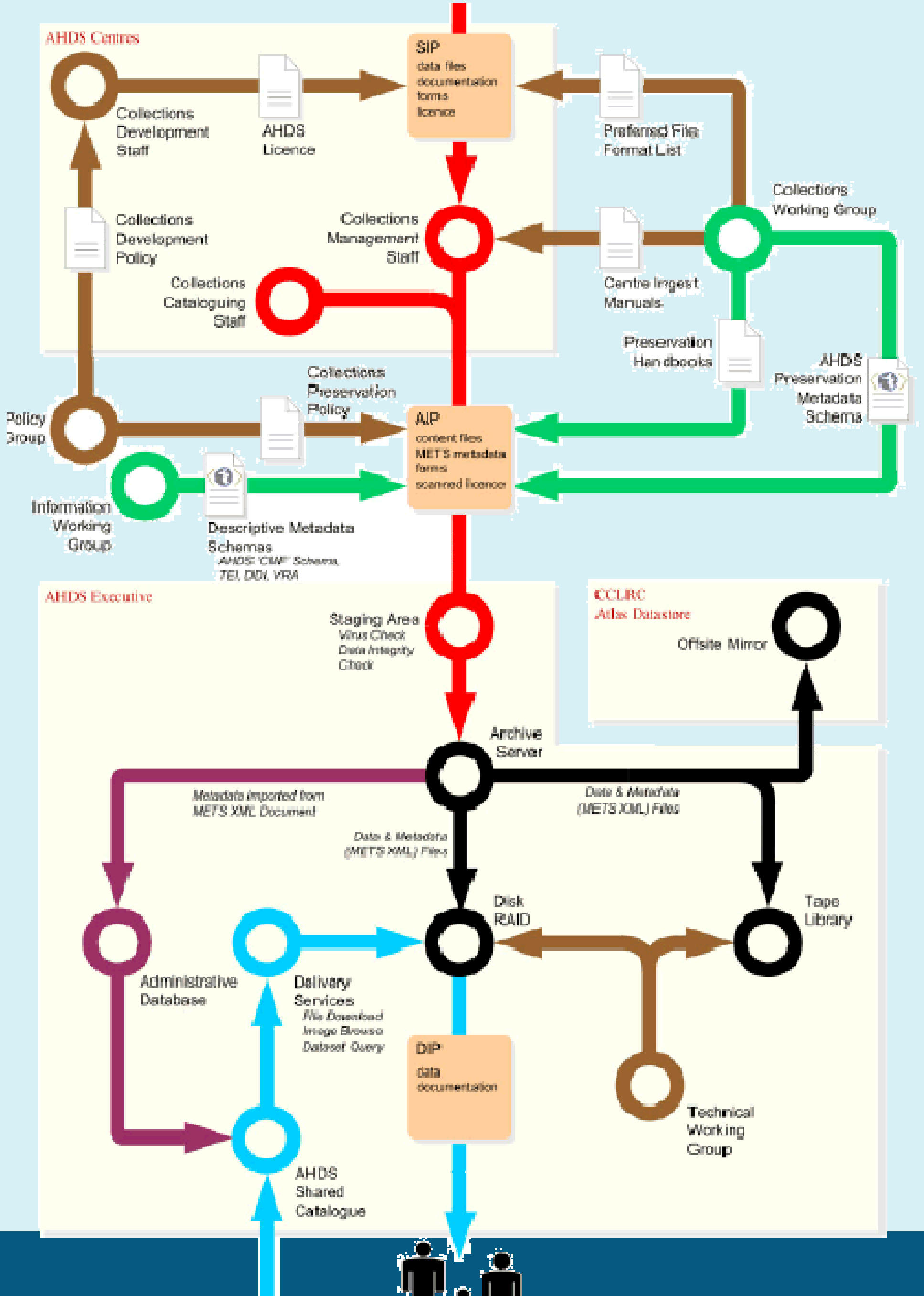
ahds arts and humanities data service

- Set-up in 1996, funded until 2008
- Research data repository for arts & humanities research
- Distributed structure:
  - Managing Executive
  - History
  - Visual Arts
  - Performing Arts
  - Archaeology
  - Literature, Languages, Linguistics
- 1000 digital collections

Characteristics of humanities research

LONDON
OPERA
HOUSE
KINGSWAY
ANNIVERSARY OF THE BATTLE OF JUTLAND
Great Thanksgiving Meeting
SAILORS' SOCIETY
SUNDAY AFTERNOON. JUNE 3rd
THE DEAN of DURHAM
Dr C. W. SALEEBY
ADMIRAL WILLIAM SIMS (U.S.A.
ADMIRAL SIR JOHN & LADY JELLICOE

- Qualitative human-centric data that requires novel selection methods
- Learning objectives vary between research communities
- Digital collections:
    - highly diverse in terms of type and size.
    - Complex internal structures
    - Require discipline-specific knowledge to process
    - Intrinsic, though poorly recorded semantics



AHDS operation

Specifications:

- OAIS RM compliant
- TRAC compliance (as expressed in TDR)

Issues:

- Manual process
- Disparate tools
- Time-consuming
- Small batch processing

The way we were:
Data transfer using the postal service

Extremely manual process:

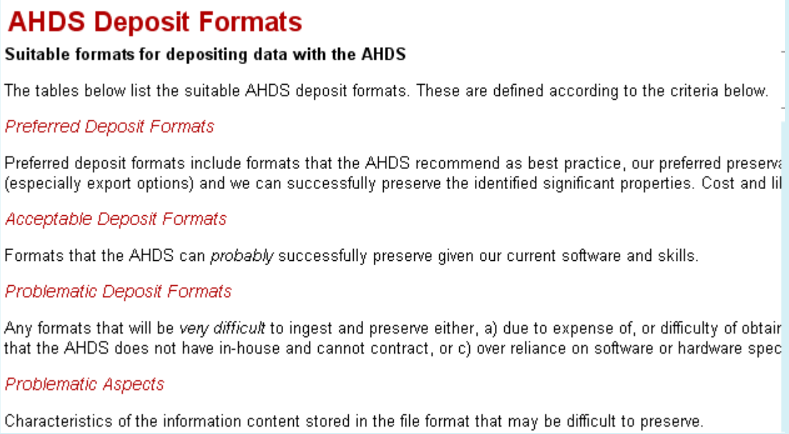
1. Review Deposit format list and prepare data for deposit
2. Complete a Data & Documentation Transfer that describes physical transfer
3. Complete a collection-level catalogue form
4. Complete and sign a licence form
5. Submit data via post, email, FTP

…Wait…

6. Receive receipt acknowledgement

…wait..

7. Confirmation of deposit and publication



AHDS Deposit Formats

Suitable formats for depositing data with the AHDS

The tables below list the suitable AHDS deposit formats. These are defined according to the criteria below.

Preferred Deposit Formats

Preferred deposit formats include formats that the AHDS recommend as best practice, our preferred preserva (especially export options) and we can successfully preserve the identified significant properties. Cost and lil

Acceptable Deposit Formats

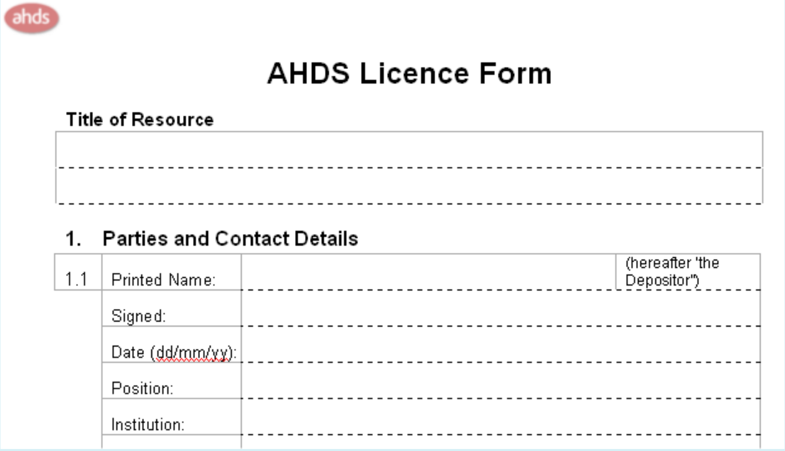
Formats that the AHDS can *probably* successfully preserve given our current software and skills.

Problematic Deposit Formats

Any formats that will be *very difficult* to ingest and preserve either, a) due to expense of, or difficulty of obtair that the AHDS does not have in-house and cannot contract, or c) over reliance on software or hardware spec

Problematic Aspects

Characteristics of the information content stored in the file format that may be difficult to preserve.



ahds

AHDS Licence Form

Title of Resource

1. Parties and Contact Details

1.1 Printed Name: (hereafter 'the Depositor')

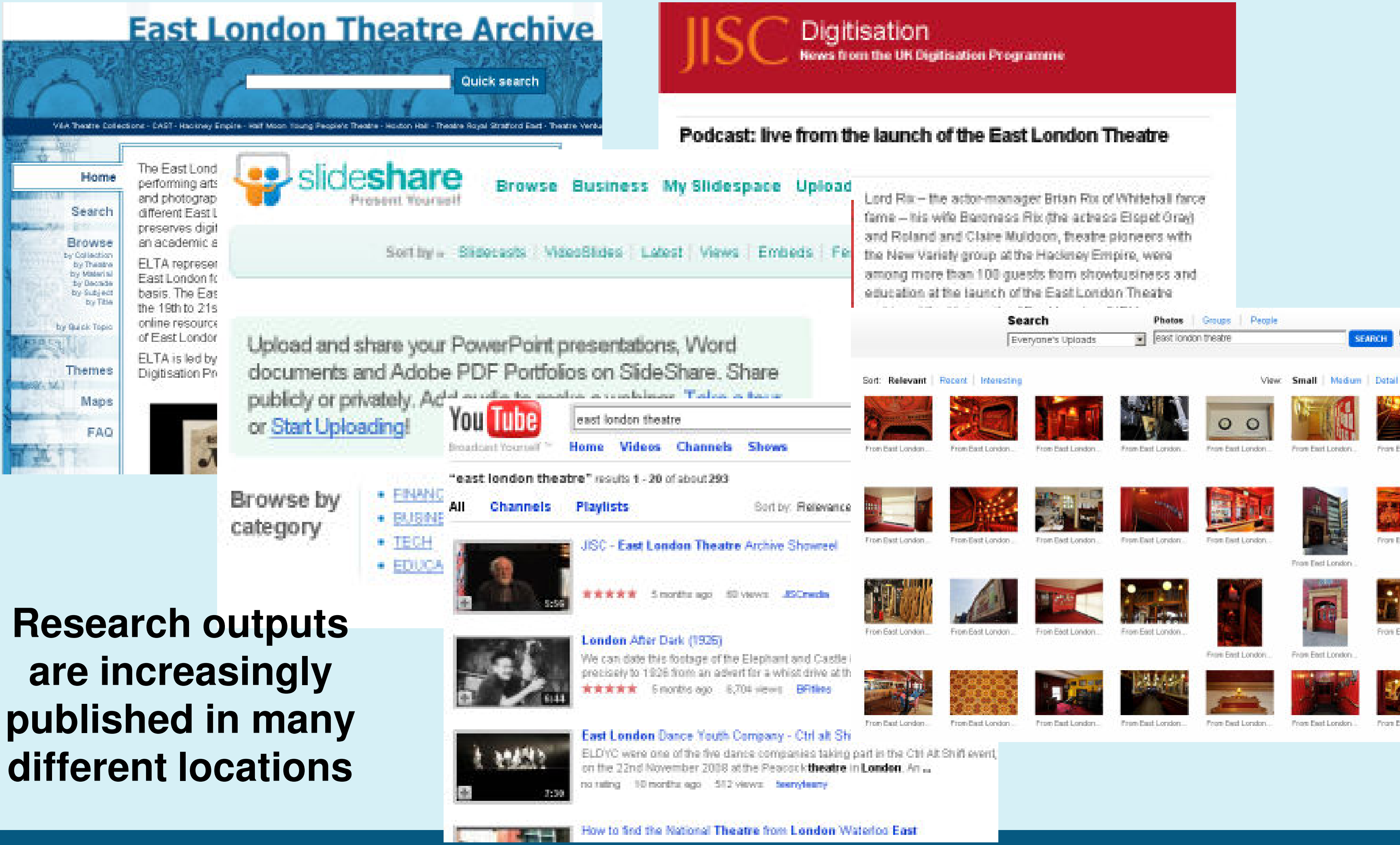
Signed:

Date (dd/mm/yy):

Position:

Institution:

Changing forms of humanities publication



East London Theatre Archive

Quick search

V&A Theatre Collections - CAST - Hackney Empire - Half Moon Young People's Theatre - Hoxton Hall - Theatre Royal Stratford East - Theatre Ventu

The East Lond performing arts and photograp different East L preserves digit an academic a

ELTA represen East London fo basis. The Eas the 19th to 21s online resource of East Londor

ELTA is led by Digitisation Pr

JISC Digitisation

News from the UK Digitisation Programme

Podcast: live from the launch of the East London Theatre

Lord Rix – the actor-manager Brian Rix of Whitehall farce fame – his wife Baroness Rix (the actress Elspet Gray) and Roland and Claire Muldoon, theatre pioneers with the New Variety group at the Hackney Empire, were among more than 100 guests from showbusiness and education at the launch of the East London Theatre

slideshare Present Yourself

Browse
Business
My Slidespace
Upload
Sort by » Slidecasts | VideoSlides | Latest | Views | Embeds | Fe

Upload and share your PowerPoint presentations, Word documents and Adobe PDF Portfolios on SlideShare. Share publicly or privately. Ad… or Start Uploading!

Browse by category

YouTube Broadcast Yourself

east london theatre

"east london theatre" results 1 - 20 of about 293

JISC - East London Theatre Archive Showreel

★★★★★ 5 months ago 50 views JISCmedia

London After Dark (1926)

We can date this footage of the Elephant and Castle precisely to 1926 from an advert for a whist drive at th

★★★★★ 5 months ago 6,704 views BFIfilms

East London Dance Youth Company - Ctrl alt Shi

ELDYC were one of the five dance companies taking part in the Ctrl Alt Shift event, on the 22nd November 2008 at the Peacock theatre in London. An ...

no rating 10 months ago 512 views teenyteeny

How to find the National Theatre from London Waterloo East

Search

Everyone's Uploads

east london theatre

Photos | Groups | People
SEARCH
Sort: Relevant | Recent | Interesting
View: Small | Medium | Detail

From East London...

Research outputs are increasingly published in many different locations

Do these resources require curation?

- (Most) third party services do not commit to storing data forever – may be deleted
- Data may be stored in form that causes significant properties to be lost
- Repository staff in an IR may be unable to perform preservation activities, due to lack of time or infrastructure
- Where are the boundaries for management of institutional data?

Curation projects

- **SHERPA Digital Preservation (1 & 2)**

  Investigated the curation and preservation requirements of research data that is encoded as varied content types and made available using many different technologies in disparate locations.

  http://www.kcl.ac.uk/iss/cerch/projects/completed/sherpadp2.html

- **SOAPI (Service-Oriented Architecture for Preservation and Ingest) of Digital Objects**

  Developed an architecture and toolkit for (partially) automating preservation and ingest workflows in digital repositories, based on a set of atomic web services, each encapsulating a unit of preservation functionality.

  http://www.kcl.ac.uk/iss/cerch/projects/completed/soapi.html

Curation of disparate resources

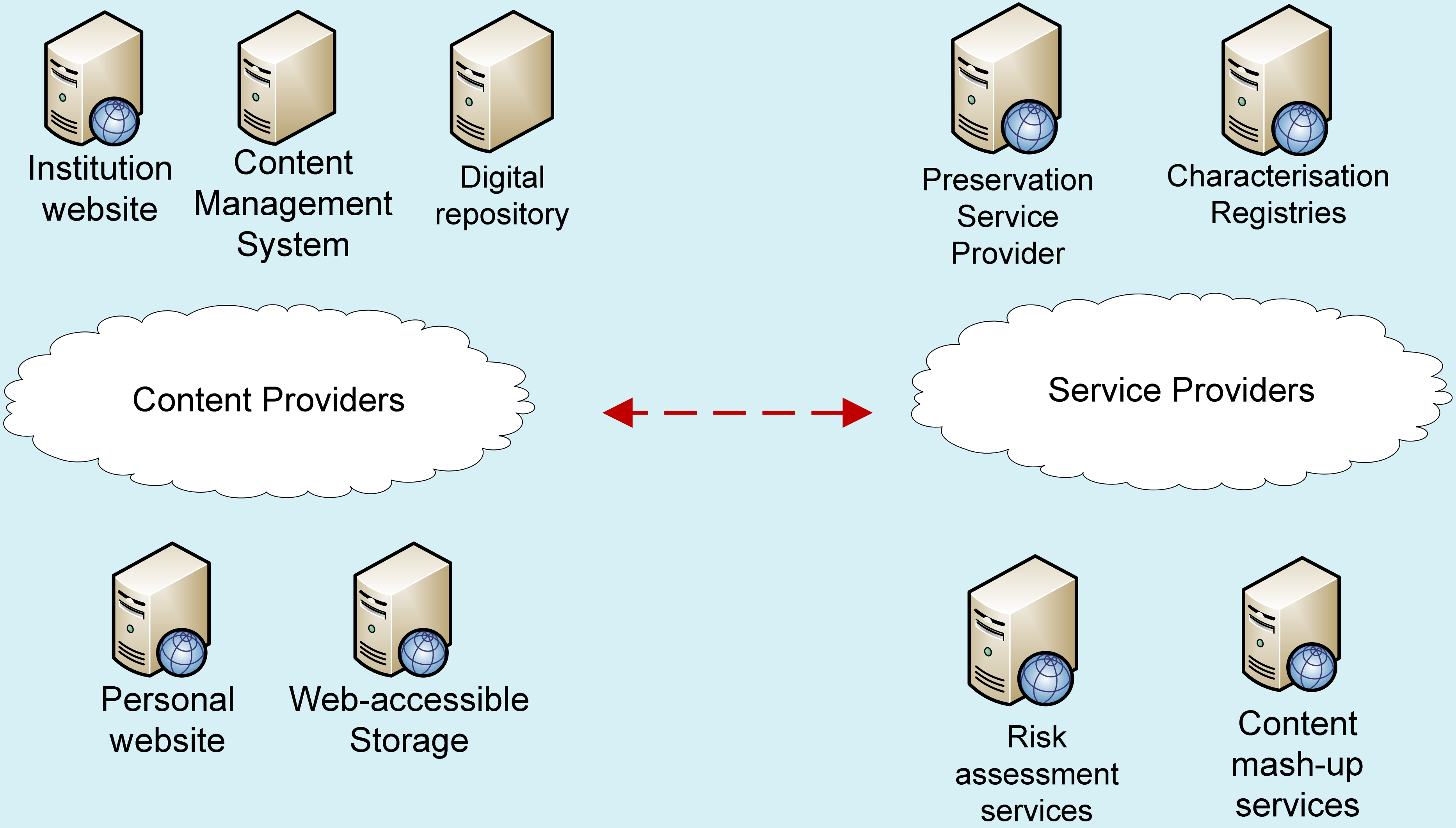
Basis:

- Institutional data management requirements extend beyond the confines of a digital repository.
- Preservation services must be able to interoperate with diverse types of technical systems and curate a wide variety of content types.

Benefits:

1. Maintain a record of research outputs of an institution/ dept that is not reliant upon a third-party that has no direct investment in maintaining the research data
2. Enables a uniform approach to curation and preservation of data that takes into account the significant properties of research data.
3. Provides an alternative method to populate a preservation repository with research data, while avoiding disruption to existing practices of research creation

A tale of two cities…



Institution website

Content Management System

Digital repository

Preservation Service Provider

Characterisation Registries

Content Providers

Service Providers

Personal website

Web-accessible Storage

Risk assessment services

Content mash-up services

Curation models

Scenarios considered:

- Storage failure, Data replacement, Data audit, System switch, Data enhancement, curation, preservation, migration

Services that a Preservation Service Provider may provide:

1. *Archiving service:* The PSP stores a complete/partial data backup in an offsite location.
2. *Migration service:* The PSP stores a complete/partial data backup offsite & creates enhanced DIPs for users.
3. *Preservation Service:* The PSP stores a complete/partial data backup offsite & creates normalised data objects, preservation metadata, or other content to support long-term preservation.

+ additional advisory capacity

http://ie-repository.jisc.ac.uk/395/1/sherpadp2_finalreport_v1.pdf

Workflow management requirements

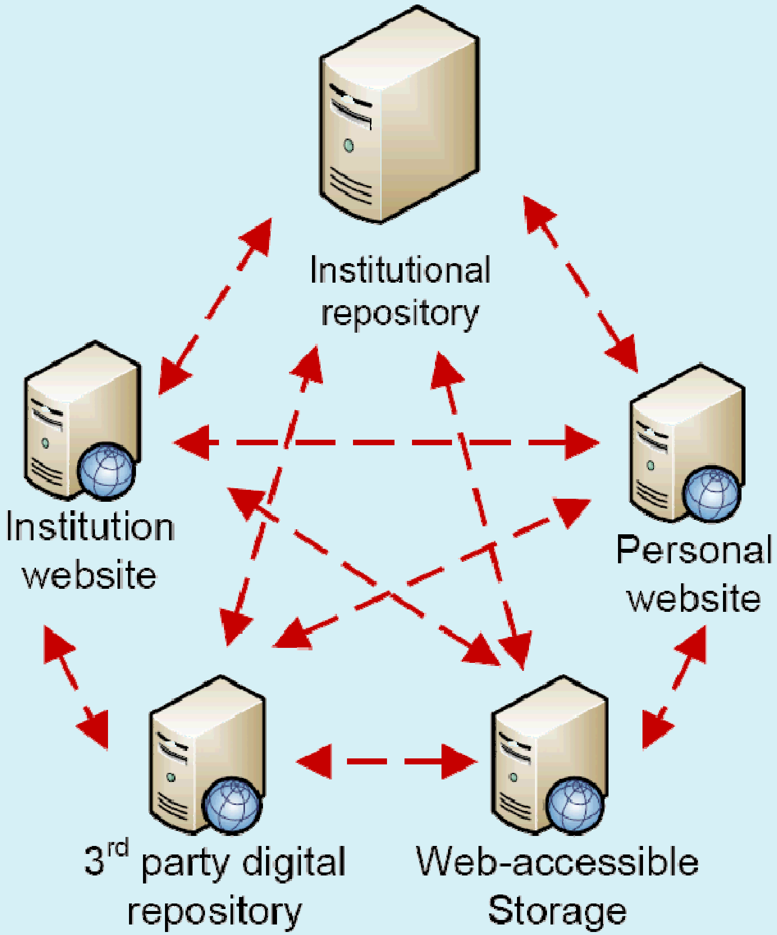
Stages:

1. Monitor resource for updates or other changes
2. Capture
3. Validate
4. Curate
5. Preserve
6. Re-submit (if required)

Requirements:

- Automate large sections of workflow
- Scalable approach
- Integration of multiple-third-party tools
- Policies and procedures for handling

Characteristics of Content Providers (1)



- Set of Content Providers providing value-added services for access, e.g. cloud storage, high powered computing
- Each provides services for interacting with resources.
- Many digital resources are dynamic, providing no fixed form.

Characteristics of Content Providers (2)

CC, Attribution 2.0, generic

http://www.flickr.com/photos/s_y_s/2305290082/

- Can curation action be performed on remote system?
- Does data need to be captured?
- Where is the data for capture located?
- How is it distinguished from data that should not be captured?

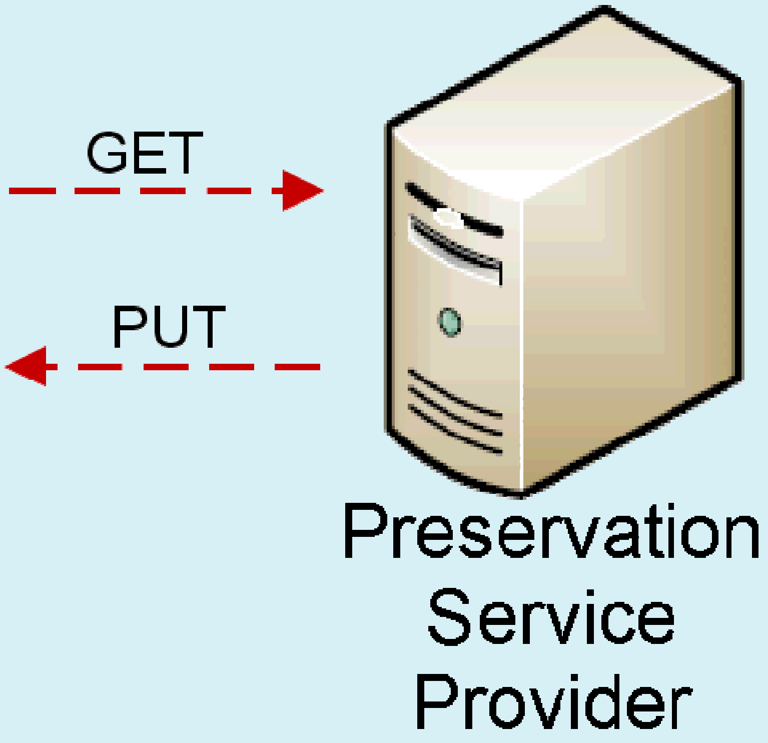
Case Study: Monitoring/capture/deposit

Testbed systems:

- Repositories: Fedora, EPrints, DSpace, CERN Document Server
- CMS: DigiTool
- Other: Subversion, Web sites

Technologies:

- OAI-PMH
- Web Feeds (RSS, Atom)
- Database backup
- Versioning system check-out/check-in
  - SVNKit
- OAI-ORE (partially)
- SWORD



GET

PUT

Preservation Service Provider

Data transfer issues

- Inconsistent metadata output across repositories
  - Simple DC – yes, but what else?
- Difficulty in obtaining all metadata associated with an Object
- Changes to the content models within a collection
- Unable to validate transfer, in most cases
  - Lack of checksums

Transfer package requirements

Content

- Manifest/inventory of the page contents
- Relationship metadata
- Structural metadata describing composition of the object

Description

- Descriptive metadata
- Information about agents (people, organizations, software) that have a relation to the object

Preservation

- General/format-specific technical metadata
- Significant properties of the object
- Event metadata describing actions performed

Legal/contractual

- Rights metadata indicating access & use
- Business information regarding the producer's desired or contracted-for treatment of the object

http://www.dlib.org/dlib/november08/caplan/11caplan.html

Transfer package Issues

- **Commonality:**
  - Packaging format (e.g.METS, MPEG21)
  - Metadata formats (e.g. Dublin Core, MODS, PREMIS, MIX)
- **Consistency:**
  - MD format in packaging (e.g. PREMIS in METS)
  - http://www.loc.gov/standards/premis/guidelines-premismets.pdf
- **Handling redundancy:**
  - Handling duplicate elements, but potentially contradictory information

Transfer Package examples

- Repository eXchange Package (RXP)

http://wiki.fcla.edu:8000/TIPR/21

- BagIt File Packaging format

http://www.cdlib.org/inside/diglib/bagit/bagitspec.html

- Kopal Universal Archive Format

http://kopal.langzeitarchivierung.de/downloads/kopal_Universal_Object_Format.pdf

- ECHO METS profile

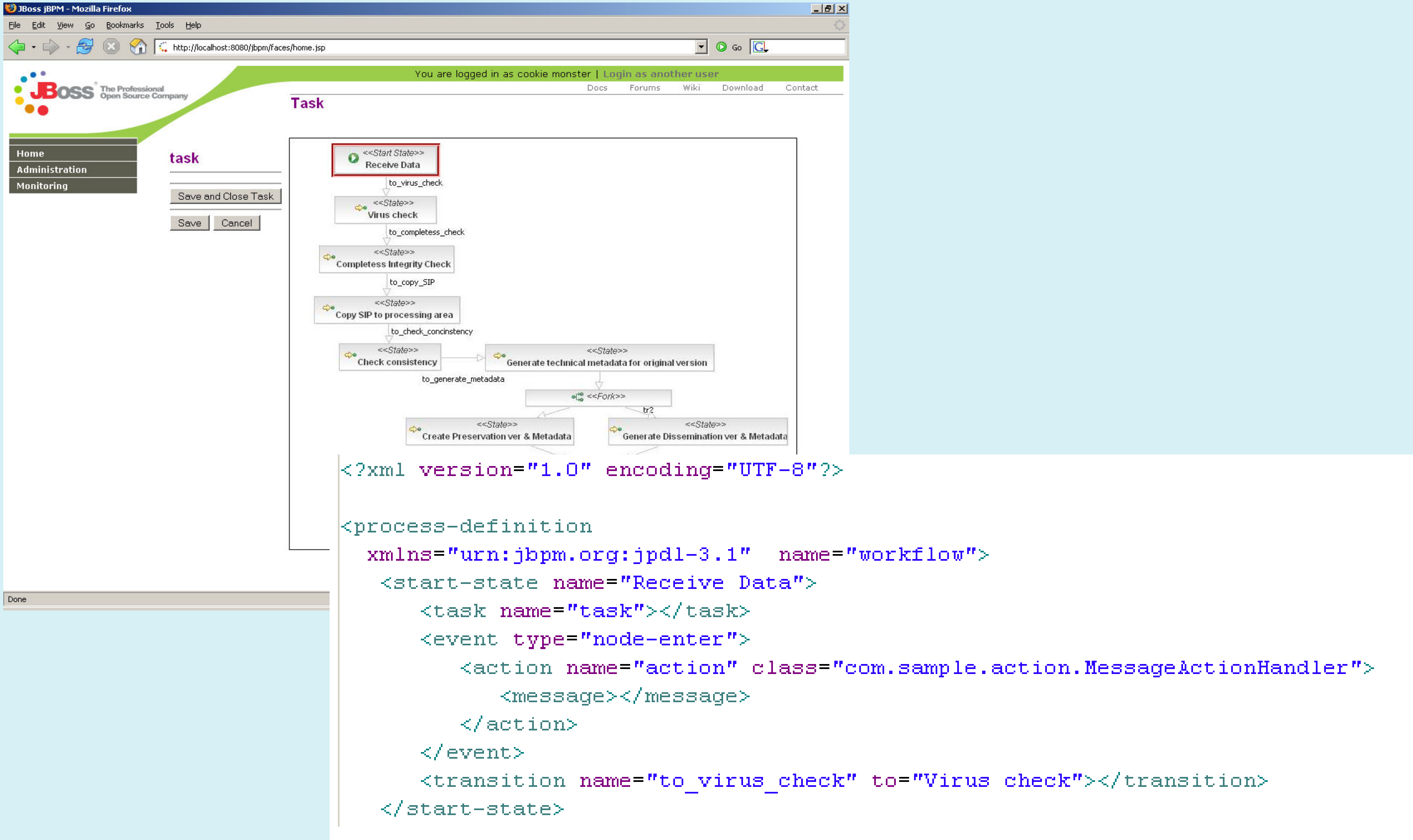
http://www.ndiipp.illinois.edu/

… And others

Digital Curation management

- Workflow management engine
- Evaluated several workflow engines: Taverna, BPEL (Active BPEL), jBPM, others. Settled on jBPM
- Chain together automated actions and user tasks to form a workflow or “Business Process”
- Generic interfaces to encapsulate functional units
- Generic interfaces to wrap third-party tools.
- Web service (SOAP & REST) and local implementations

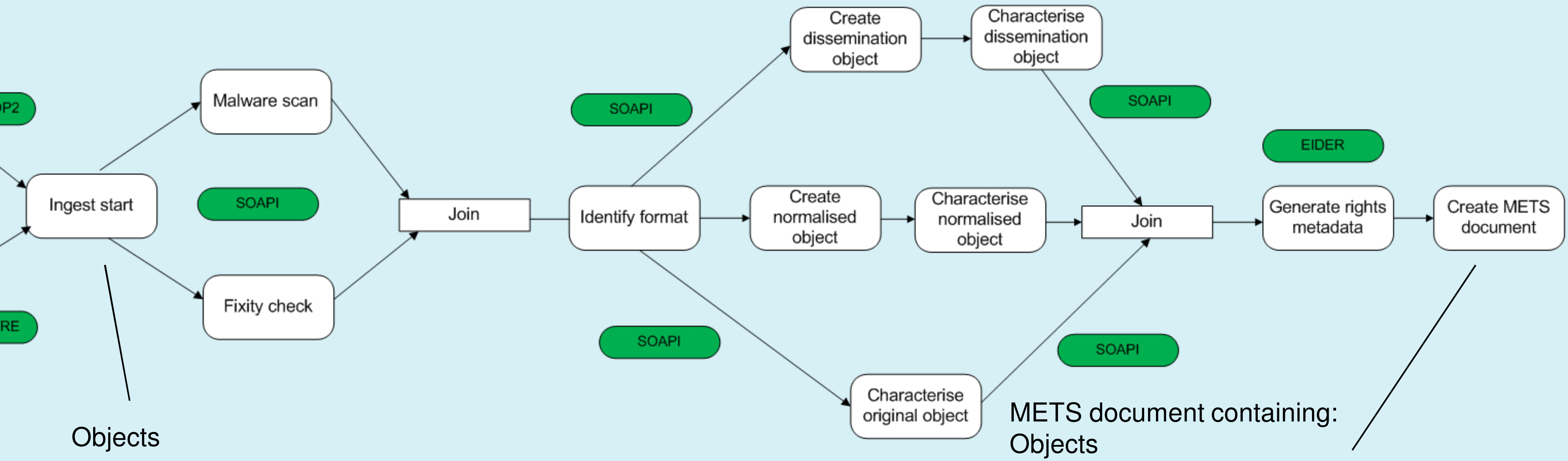
Workflow in jBPM


JBoss jBPM - Mozilla Firefox
File Edit View Go Bookmarks Tools Help
http://localhost:8080/jbpm/faces/home.jsp
Go
JBoss The Professional Open Source Company
You are logged in as cookie monster | Login as another user
Docs Forums Wiki Download Contact
Task
Home
Administration
Monitoring
task
Save and Close Task
Save
Cancel
<<Start State>> Receive Data
to_virus_check
<<State>> Virus check
to_completess_check
<<State>> Completess Integrity Check
to_copy_SIP
<<State>> Copy SIP to processing area
to_check_concinstency
<<State>> Check consistency
<<State>> Generate technical metadata for original version
to_generate_metadata
<<Fork>>
tr2
<<State>> Create Preservation ver & Metadata
<<State>> Generate Dissemination ver & Metadata
Done

```xml
<?xml version="1.0" encoding="UTF-8"?>

<process-definition
  xmlns="urn:jbpm.org:jpdl-3.1"  name="workflow">
   <start-state name="Receive Data">
      <task name="task"></task>
      <event type="node-enter">
         <action name="action" class="com.sample.action.MessageActionHandler">
            <message></message>
         </action>
      </event>
      <transition name="to_virus_check" to="Virus check"></transition>
   </start-state>
```

… Or to put it another way…



Objects

METS document containing:
Objects

- Descriptive metadata
- PREMIS object metadata (technical)
- PREMIS event metadata
- PREMIS relationship metadata
- Format-specific technical metadata (e.g. MIX)

Workflow tools and standards

Activities:

- *Object identification* – what is it?
- *Characterisation* – What does it contain?
- *Validation* – Does it conform to standard?
- *Format conversion* – convert to normalised and migrated derivatives
- *Verify conversion* – Does it contain everything that was in original?
- *Validate conversion* –Does it conform to standard?

Tools:

- DROID, File, JHOVE, JHOVE2, NLNZ Metadata Extractor, XCL, others
- XENA, Open Office, SOX, ImageMagick, SIARD

Standards:

- PREMIS 1.0/2.0 Object, Event, MIX for images, AudioMD, DocumentMD, others

Integration with third-party services

Preservation services

- PRONOM, UDFR, Preserv2 Semantic preservation tool, PLATO, others
- Characterisation
- Risk assessment
- Preservation planning

Storage

- Grid technologies - originally SRB. Now iROD
- Extensive use of complex metadata formats stored within Fedora.
- Integrated, but changeable system rules
- Fedora repository discovery belonging to different administrative domains.
- Data resource discovery across Fedora repositories

Data management issues

- Lack of suitable tools in some areas – expensive, outputs unreliable
- Preserving content – what do we actually want to preserve?
- Significant properties – soft concept, hard to quantify (InSPECT, PLANETS)
- Problems with jBPM

Conclusions

- System interoperability extends beyond the repository domain
- Automation requires definition of rules. Sig props MD and other metadata requires further work
- Further work necessary to package data of various types and transport between systems
- Further integration is necessary between repository services and national approaches, such as PLANETS toolkit.

Some references

http://www.driver-support.eu/documents/DRIVER_Guidelines_v2_Final_2008-11-13.pdf

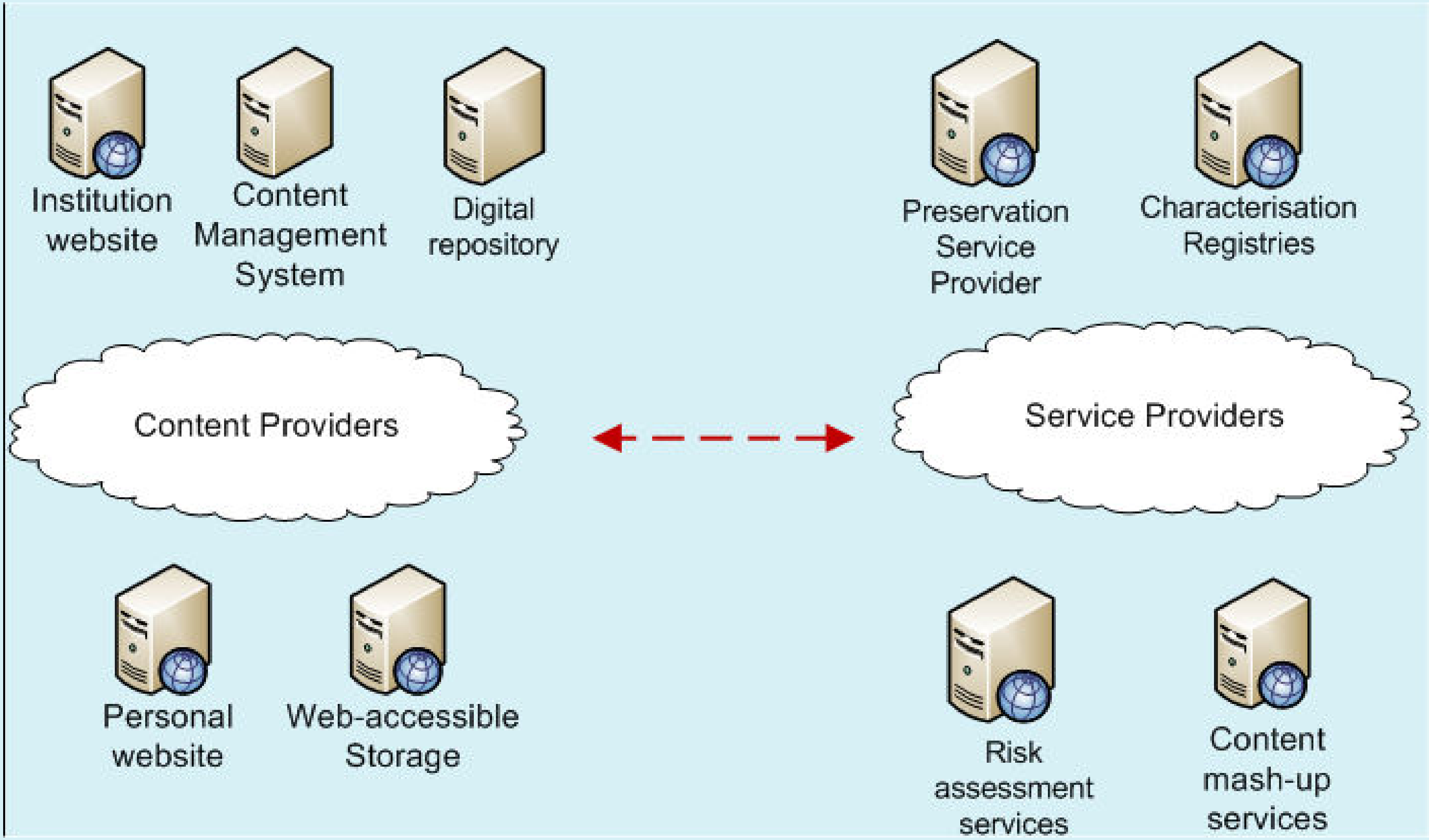
http://www.ukoln.ac.uk/repositories/digirep/images/a/a5/Introductoryecology.pdf

http://ie-repository.jisc.ac.uk/395/1/sherpadp2_finalreport_v1.pdf

http://wiki.fcla.edu:8000/TIPR

http://www.dlib.org/dlib/november08/caplan/11caplan.html

A tale of two cities…


Institution website

Content Management System

Digital repository

Preservation Service Provider

Characterisation Registries

Content Providers

Service Providers

Personal website

Web-accessible Storage

Risk assessment services

Content mash-up services