Small steps and lasting impact: making a start with preservation

Sarah Jones
HATII, University of Glasgow
sarah.jones@glasgow.ac.uk

Outline

1. Principles, concepts and terminology
2. What goes on pre-preservation?
3. Practical steps to get started

1. Principles, concepts and terminology

What is digital preservation?

Digital preservation is the active management of digital information over time to ensure its accessibility.

Preservation of digital information is widely considered to require more constant and ongoing attention than preservation of other media.

Wikipedia, 23rd February 2011

How is digital different?

Digital objects break. They are bound to the specific application packages used to create them. They are prone to corruption. They are easily misidentified. They are generally poorly described.

Seamus Ross, *Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries,* ECDL, 2007

Digital objects… come in various formats

Formats may be:

Compressed - a shorthand way of writing out the bits to save storage space

With lossy or lossless compression

- Lossy accepts some loss of data (like rounding up numbers) e.g. JPEG
- Lossless is reversible so the original data can be reconstructed e.g. PNG, GIF

Open and/or proprietary

- Open means the format is an open, published standard e.g. ASCII, PDF, PNG
- Proprietary are commercial and typically closed, e.g. WMA, PSD, DOC etc
  (i.e. you need a licence and are reliant on the software provider continuing to support the format)

Different formats are good for different things

	Preservation Uncompressed, open, supported standards	Access In widespread use, open, small file-size for online hosting
Text	TXT, RTF, ODT, XML	DOC, PDF, ODT
Image	TIFF, PNG	JPEG, PNG
Audio	WAV, FLAC, AIFF	MP3, WMA, QuickTime
Video	MPEG-4, MJPEG 2000	MOV, AVI, WMV

Repositories may:

- prefer to take certain formats
- normalise data on ingest (i.e. convert to a standard format)
- keep data in multiple formats (e.g. a WAV preservation master & MP3 access copy)

Digital objects… are stored on different media

Media degrade – they need to be refreshed

Digital objects… can easily be copied

Lots Of Copies Keeps Stuff Safe

www.lockss.net/

Backup principle

Keep 2+ copies

on different types of media

in different locations (ideally one off-site)

If you use the same media twice, go for different manufacturers to avoid an error destroying both copies.

Digital objects… are not self-describing

Metadata is needed to understand digital objects

- Descriptive information (catalogue entry)
- Structural metadata (how digital objects fit together)
- Administrative context (technical details, preservation)

Metadata can be embedded (e.g. in TIFF header), or kept in a separate database (but need strong links!)

Standards can be used (e.g. Dublin Core and Thesauri like UKAT)

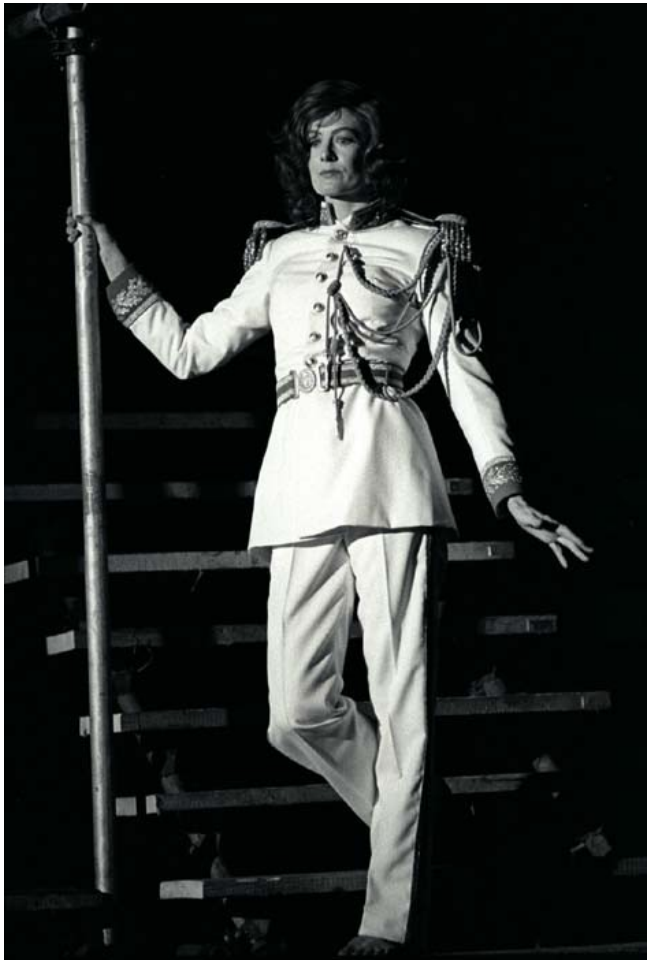
Dublin Core metadata example

Dublin Core elements

Creator: Donald Cooper
Role=Photographer
Subject: Shakespeare, William, 1564-1616, Antony and Cleopatra [LC]
Description: Vanessa Redgrave as Cleopatra
Date: 1973-08-09
Type: Image
Format: JPEG
Identifier: 4150 [catalogue no]
Source: negative no 235
Relation: Antony and Cleopatra: Thompson/73-8
IsPartOf
Coverage: Bankside Globe
Role=Spatial
Rights: Donald Cooper

Standardised input (thesauri, ISO)

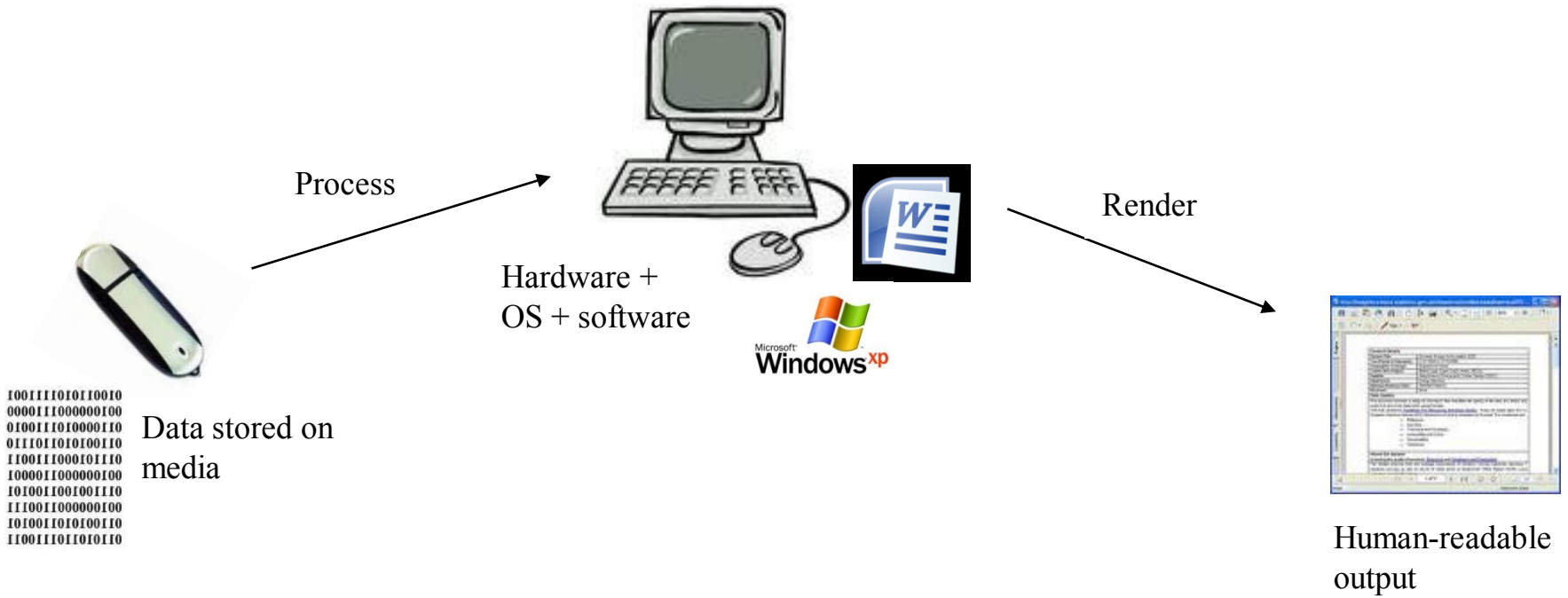
Optional extensions



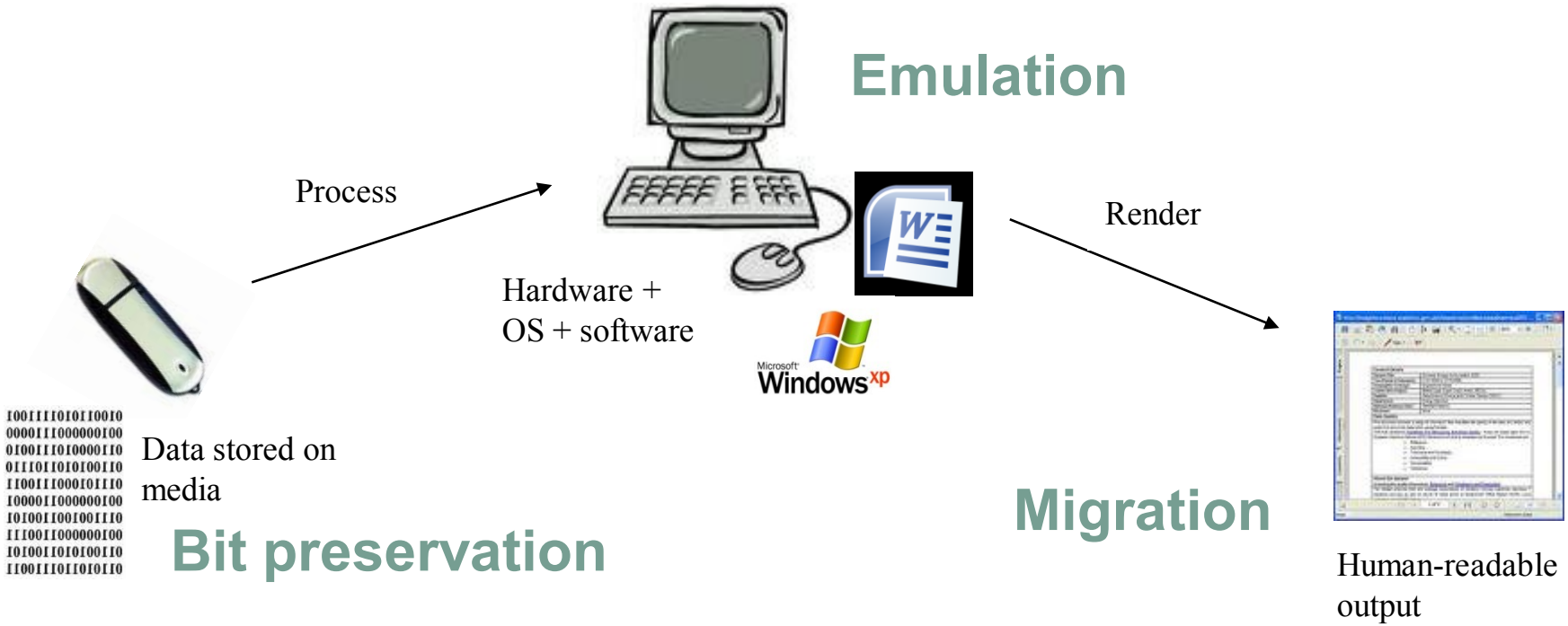
http://dublincore.org/

Digital objects… aren't tangible

We're not preserving the digital object, rather
<u>the ability to reproduce it</u>



What does this mean for preservation?



n.b. preservation approaches are not mutually exclusive. You may choose to migrate but also preserve the original bitstream so you can emulate later.

Bitstream preservation = basic level

- Capture information in its original form
- Follow basic archive processes
  – media refreshment, checksums to validate integrity etc

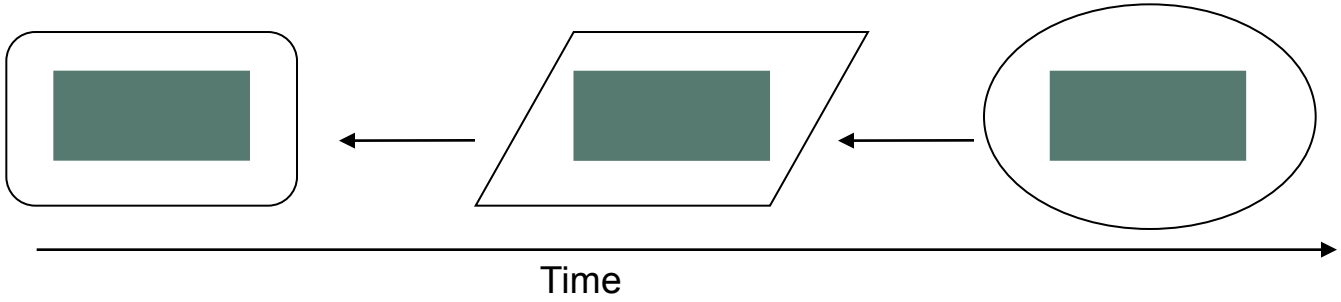
A checksum is a unique fingerprint which can be used to ensure that the file or program has not been changed during transfer or storage e.g. MD5

- ✓ scalable and practical
- ✓ works well so far

- × useful life of data unclear (format obsolescence)
- × not really future-proof given pace of change

Emulation = changing the environment

use emulators to mimic behaviour of obsolete systems

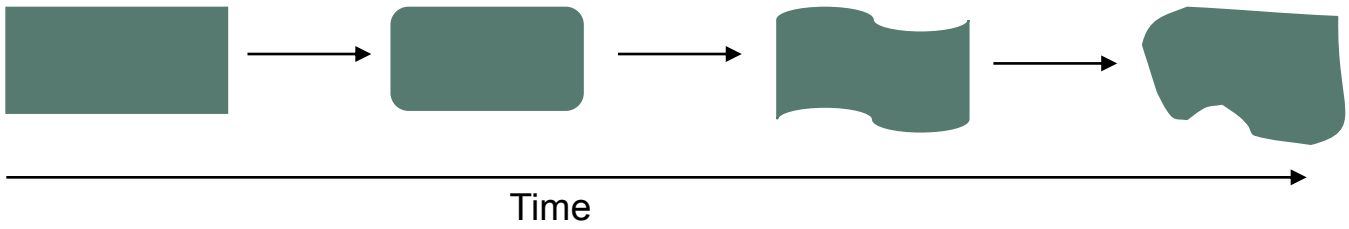


- ✓ No changes to the object are needed – more authentic?
- ✓ Keeps look & feel. Good if interactive e.g. computer games

- × Technically challenging
- × User has to know how to work in original environment
- × Quality Assurance is difficult

Migration = changing the object

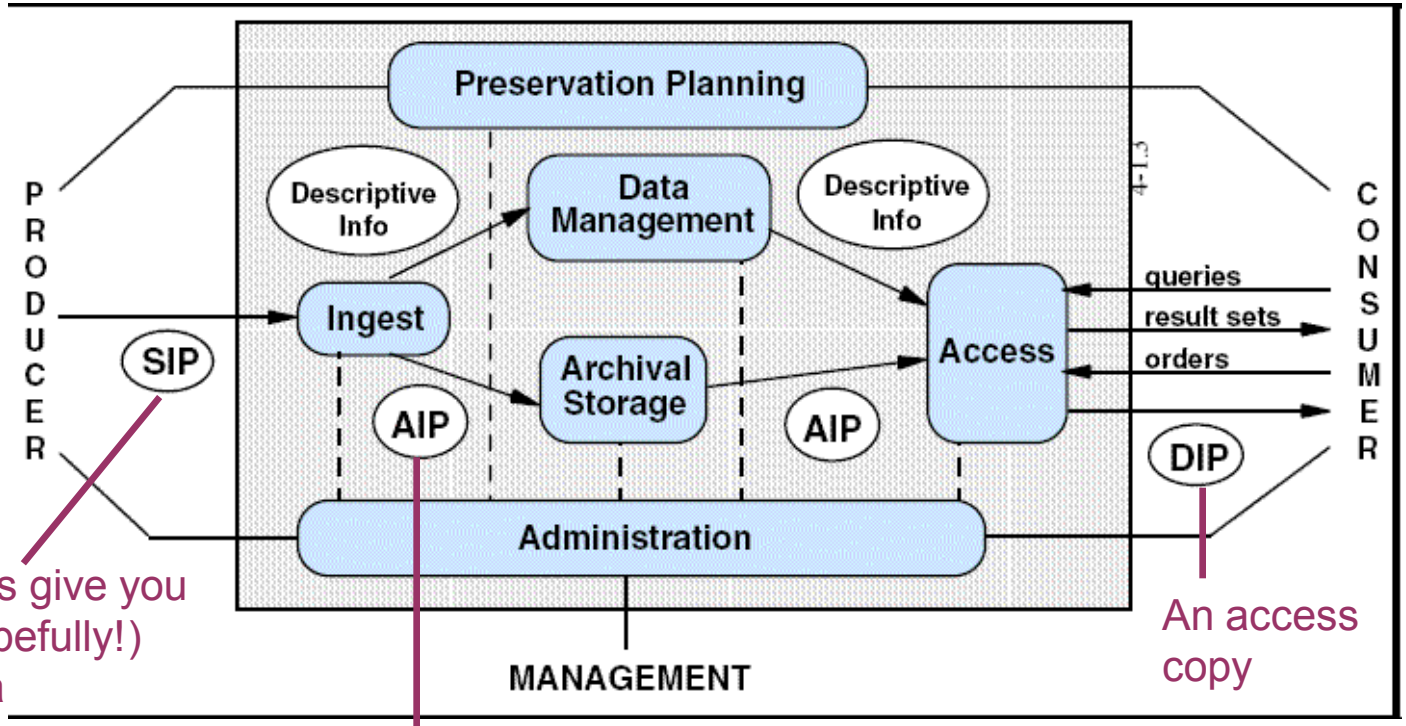
migrate object to new software/hardware environment



- ✓ Object is available in current environment – good for users
- ✓ Homogeneous data easier to manage

- × Changes inevitably occur – may be hard to spot loss
- × Demands regular investment/activity – migrate on demand?
- × Unclear which migration paths are best

Open Archival Information System

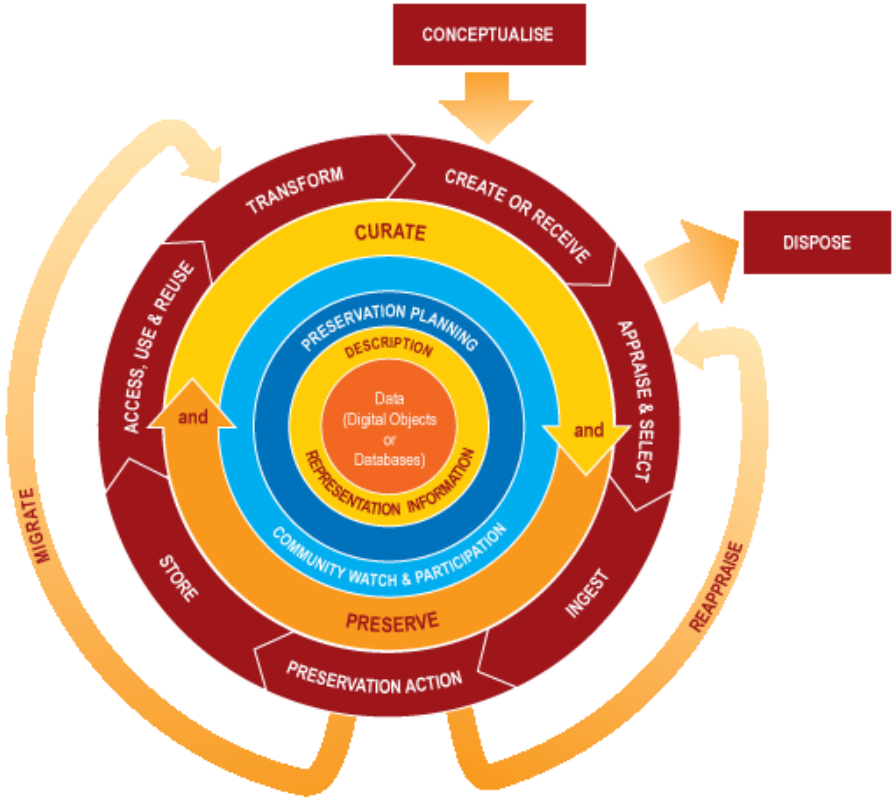


What depositors give you – objects + (hopefully!) some metadata

The object after checking, processing, cataloguing…

An access copy

2. Pre-preservation



How digital objects are created and looked after in the short-term affects how much work it is to ingest and preserve them

Ingest is biggest cost in preservation

KRDS studies

How researchers manage their data

- Naming & filing varied wildly – issues retrieving content
- Lots of duplication across different folders
- Metadata creation big burden so not always done
- Not enough storage so data put anywhere to hand

www.data-audit.eu

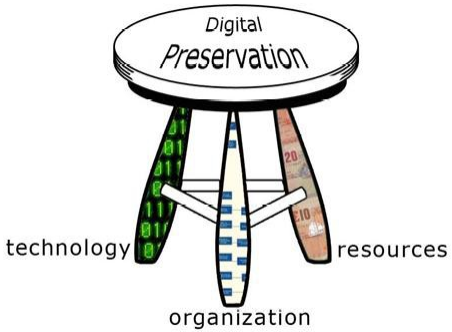
Incremental

www.lib.cam.ac.uk/preservation/incremental/

Digital objects need attention quickly – can't leave on shelf for 20 years
If they're disorganised, input from creators will be key

→ **Can't afford a huge digital accessions / cataloguing backlog**

3. Practical steps to get started



- Don’t be phased by technology
  - It’s only one aspect
  - Work with IT professionals
  - Add your library / information skills to the mix
- Keep things in proportion
  - Do you need a full bells and whistles set-up?
  - Remember that digital preservation is in infancy
- Have a go!

MLA archive case study, Alex Eveleigh

Accessioning digital material from MLA Yorkshire when closing

Tight timeframes – steep learning curve

Use of free tools to run checksums, identify duplicate files etc

http://www.dpconline.org/training/roadshows-0910

Gloucestershire Archives project

Project using existing digital collections to develop approaches for modern digital records likely to be deposited

Developed the SCAT tool for curation and trust

SCAT provides an interface to various curation tools for archivists to try out

http://futurearchives.blogspot.com/2010/03/scat-gloucestershire-archives.html

Gloucestershire COUNTY COUNCIL

How to set up & run a data service, UKDA

Slides are online covering all processes, including:

- acquisition
- ingest
- data management / archival storage
- preservation
- access / promoting reuse
- administration

www.data-archive.ac.uk/news-events/events.aspx?id=2576

Blog reports and event notes at:

- http://pekin.cerch.kcl.ac.uk/?p=97
- www.dcc.ac.uk/news/how-run-data-service

UK • DATA ARCHIVE

Summary

Try things out and develop clear policies and procedures

Key questions to ask

- Will you only accept certain formats?
- Do you plan to normalise data at ingest?
- What metadata will you create and how?
- Where will you store the data – on what media?
- How will the archive be managed? (checksums, refreshment, backup)
- What approach to preservation is best for you / your users?
- How will access be provided? (online, authenticated…)

Ask for help

Training

- DPTP, 16th-18th May 2011, Glasgow

www.dptp.org/2011/02/16/next-dptp-course-confirmed-for-may-2011/

- DCC Roadshow, June, Glasgow

www.dcc.ac.uk/events/data-management-roadshows

Community

- Join listservs and discuss your ideas
  - digital-preservation@jiscmail.ac.uk
  - JISC-repositories@jiscmail.ac.uk
  - research-dataman@jiscmail.ac.uk

Thanks – any questions?

sarah.jones@glasgow.ac.uk

Image credits:

George Service House © HATII http://www.gla.ac.uk/departments/hatii/

Media refreshing image © Patricia Sleeman http://www.ulcc.ac.uk/digital-preservation/current-activities/digital-preservation-training-programme-dptp.html

Vanessa Redgrave as Cleopatra © Donald Cooper http://www.ahds.ac.uk/performingarts/collections/designing-shakespeare-info.htm

Migration / emulation diagram concept © Sara Van Bussell, Planets project http://www.planets-project.eu/training-materials/3-van-bussel-how_to_preserve/

OAIS model © NASA, http://public.ccsds.org/publications/archive/650x0b1.PDF

DCC lifecycle © DCC, http://www.dcc.ac.uk/resources/curation-lifecycle-model

Three-leg stool © Nancy McGovern & Ann Kenny, Cornell University http://www.library.cornell.edu/