Digital Curation II

Keeping digital objects over time

Ross Harvey

27 June 2013

Ready or Not? Enhancing Digital Resources Management
5th EABH Summer School

Digital Curation: Keeping digital objects over time

Topics:

1. Review: Getting digital objects into the archive
2. Keeping digital objects over time: processes
3. *Preservation Action*
4. Automating preservation actions
5. *Store*
6. Trusted Digital Repositories
7. *Access, Use & Reuse*
8. Putting it all together

Topic 1: Review: Getting digital objects into the archive

- What was covered in the 25 June session
  - Plan, Identify, Select, Ingest
- Coming up now: Store, Manage, Access

Review

Digital Curation: getting digital objects into the archive *(25 June session)*

- We looked at:
  - Challenges
  - Two models: OAIS, DCC Lifecycle
  - The importance of planning
  - 'Preservation-friendly' digital objects
  - Making digital objects preservation-friendly
  - The role of metadata
  - Ingest procedures
  - What are we aiming to do? Can we meet the aims?

Review

Plan

- Use preservation-friendly file formats
- Keep documentation about the data, formats, software, agreements about its use
- Scrupulously identify files
- Store files on appropriate media

Identify - What digital content do you have?

Select - What portion of your digital content will be preserved?

Ingest

- Get receipt or acknowledgement for transfer
- Calculate checksum
- Assign metadata
- Run antivirus checks

Review and looking forward

Store

- Store data on appropriate media
- Copy data to a reliable digital storage system

Manage

- Ensure data security, backups, error checking
- Refresh, check accuracy of results, document the process
- Maintain multiple copies of the bit stream
- Ensure you have the right to copy and apply preservation processes

Access

Comments from previous sessions

- Confusion between storage and preservation?
- For digital objects, storage is NOT the same as preservation
  - We are not talking only about servers – preservation involves a lot more

Topic 2: Keeping digital objects over time: processes

- The processes
- How long do we want to keep digital objects?

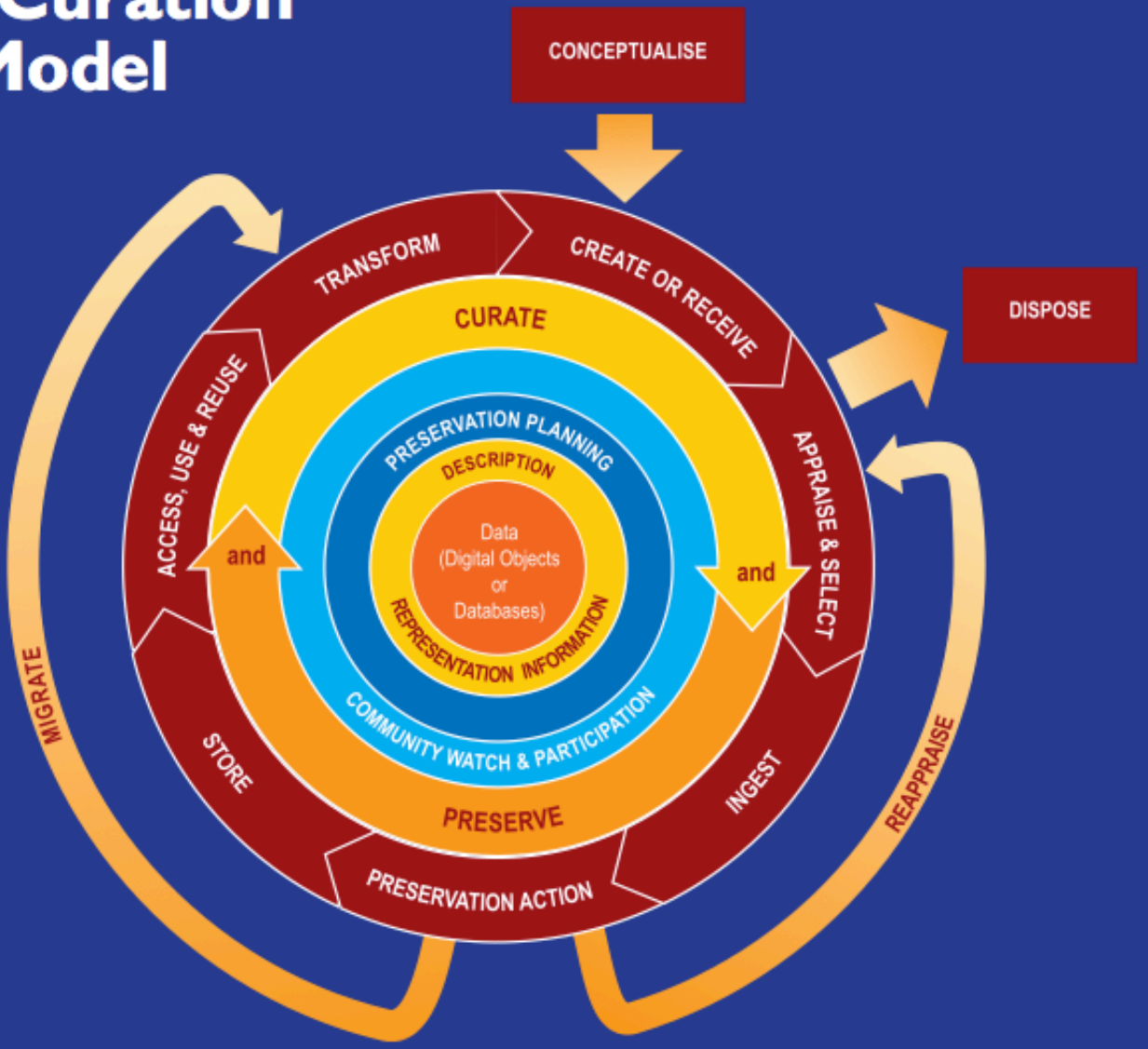
Keeping digital objects over time

What processes are needed to keep digital objects **over long periods of time?**

DCC Curation Lifecycle Model sequential actions:

- *Preservation Action*
  - Key strategies of normalization, migration, emulation
- *Store*
  - Requirements of archival storage, including cloud storage and security concerns
  - Trusted Digital Repository concept and ISO 16363:201
- *Access, Use & Reuse*
  - Strategies to ensure that archived digital objects can be located and used

The DCC Curation Lifecycle Model



http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf

Keeping digital objects over time

How long do we want to keep digital objects?

- Five years?
- Ten years?
- Twenty years?
- Longer?

Example of Roche: material has to be readable for 150 years

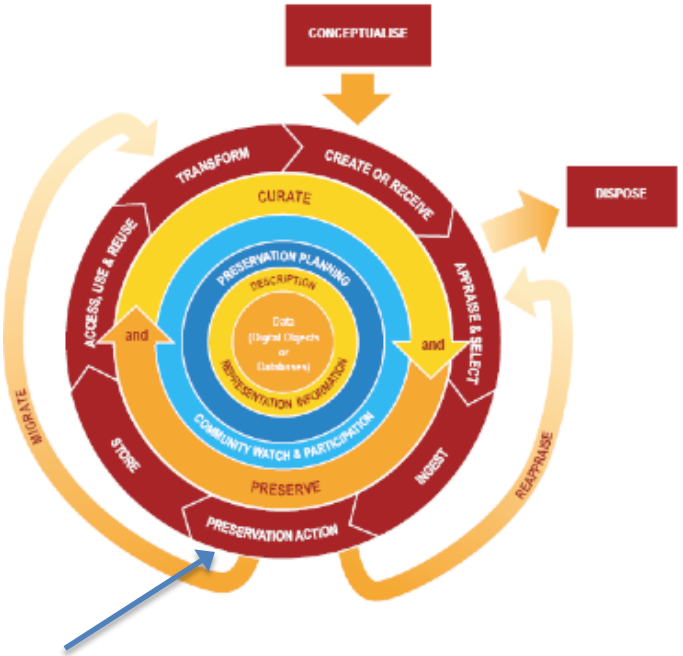
What is possible?

Topic 3: Preservation Action

- DCC Lifecycle action: *Preservation Action*
- Principal activities
- Digital preservation methods
- Checklist

Preservation Action

- Fifth sequential action of the DCC curation lifecycle
- Activities: Undertaking actions to ensure long-term preservation and retention of the authoritative nature of data
- Aim: To ensure that data remains
  - Authentic, Reliable, Usable
  - While at the same time maintaining its integrity



Preservation Action

Specific preservation actions

- Commonly agreed-on specific preservation actions include:
  - Keep the original data bit stream as well as any "preservation version"
  - Use non-proprietary, well-documented data format standards
  - Apply good data management practices
  - Implement secure storage and institutional or organizational continuity
- These are prerequisites for implementing preservation methods such as migration and emulation

Digital preservation methods (2011)

- Analogue backups: output to permanent paper or microfilm
- Backwards compatibility
- Bit-stream copying
- Canonicalization: translating artifacts into standard or 'canonical' forms
- Data recovery
- Digital archaeology
- Mass storage systems
- Durable/persistent digital storage media
- Emulation
- Encapsulation
- Improving handling and storage of storage media
- Long-term formats
- Migration (format migration, normalization & migration, software migration, version migration)
- Normalization
- Persistent object preservation
- Policy development
- Refreshing (of data, of storage media)
- Replication (redundancy, keeping multiple copies)
- Reverse engineering of software
- Software repositories
- Standardizing data formats
- Standards
- Technology preservation
- Technology watch
- Universal Virtual Computer
- Viewers for obsolete formats
- Virtual machines
- XML

(Note: "method" is loosely defined).

RED = COMMON

BLACK = LESS COMMON

BLUE = RARELY ENCOUNTERED

Digital preservation methods

Most common

- Migration
- Emulation
- Standardizing data formats

Also used

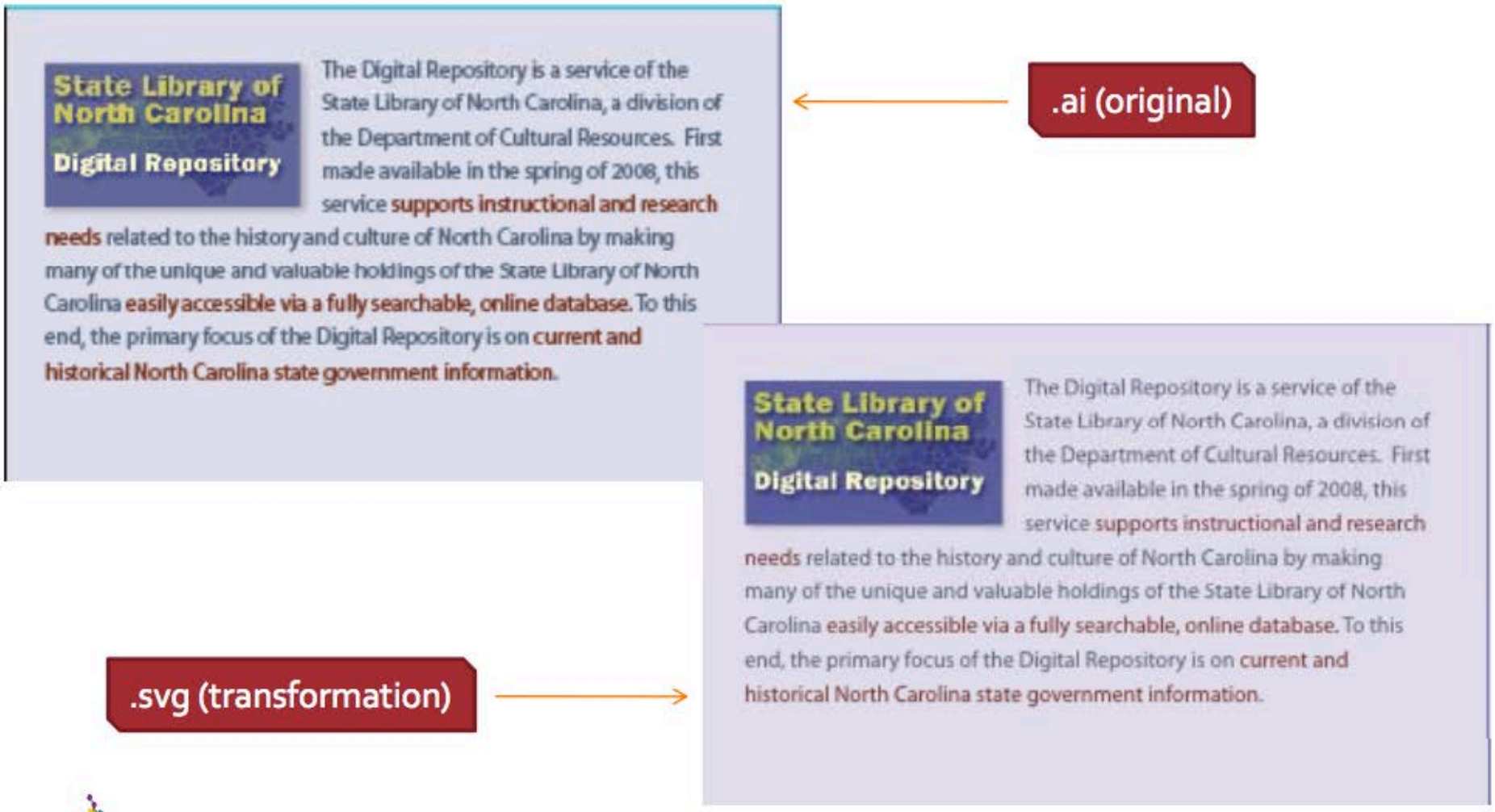
- Universal Virtual Computer
- Digital archaeology
- Technology preservation
- Encapsulation

Digital preservation methods

Migration

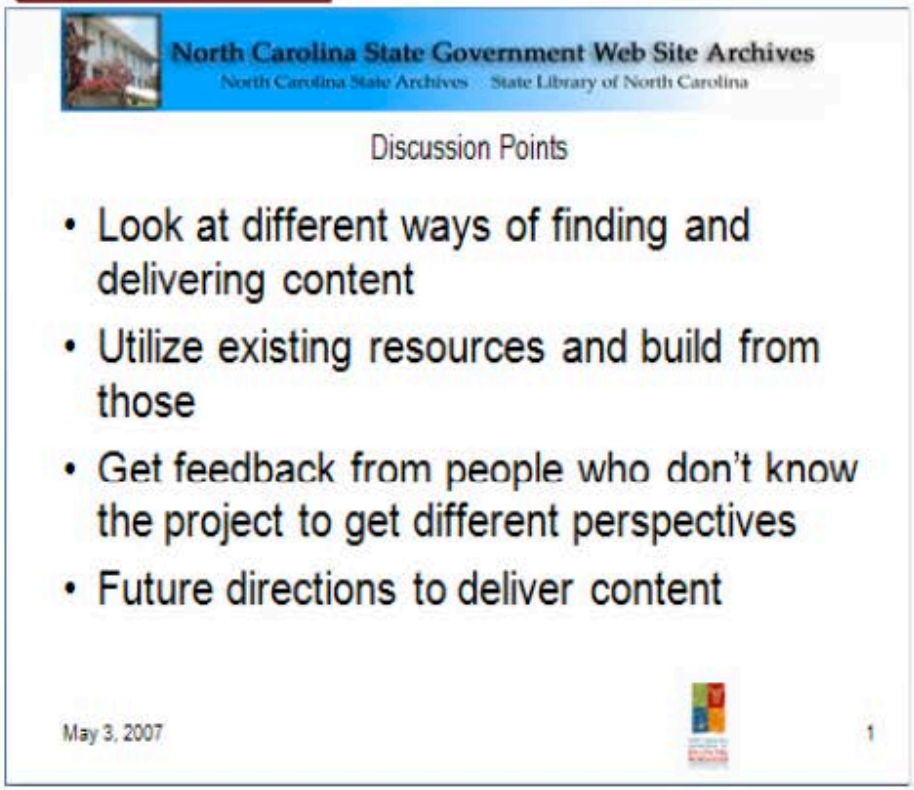
- *Migration* requires transforming data from one format to another successively as technologies change
- Well-understood process: eg occurs when systems are upgraded
- Likely to result in information loss and/or changes in systems' behaviors or computed results
- **All migration changes data**

Example: Font Change

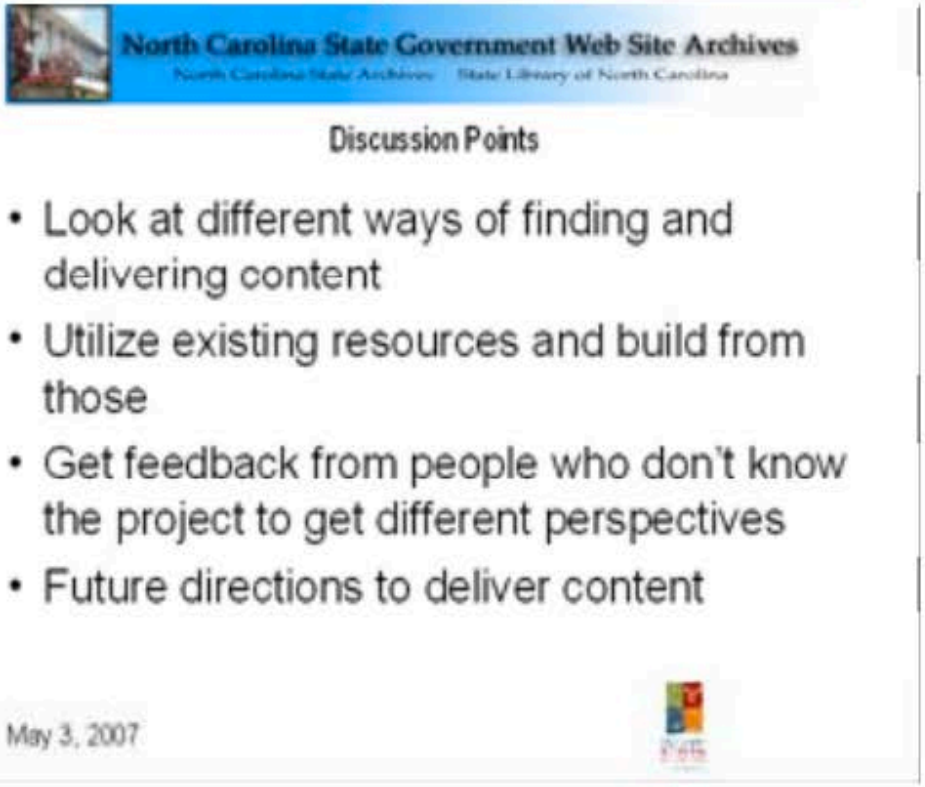


Example: Wonky Footer

.ppt (original)

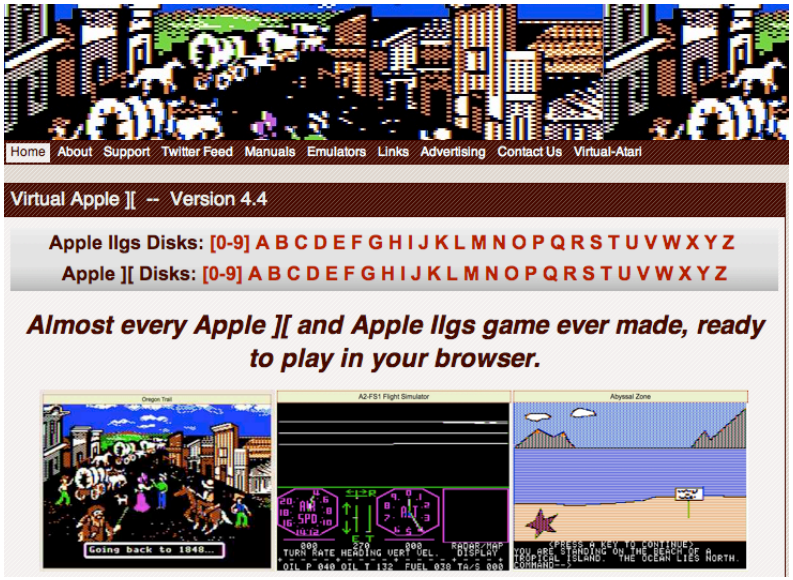


.odp (transformation)

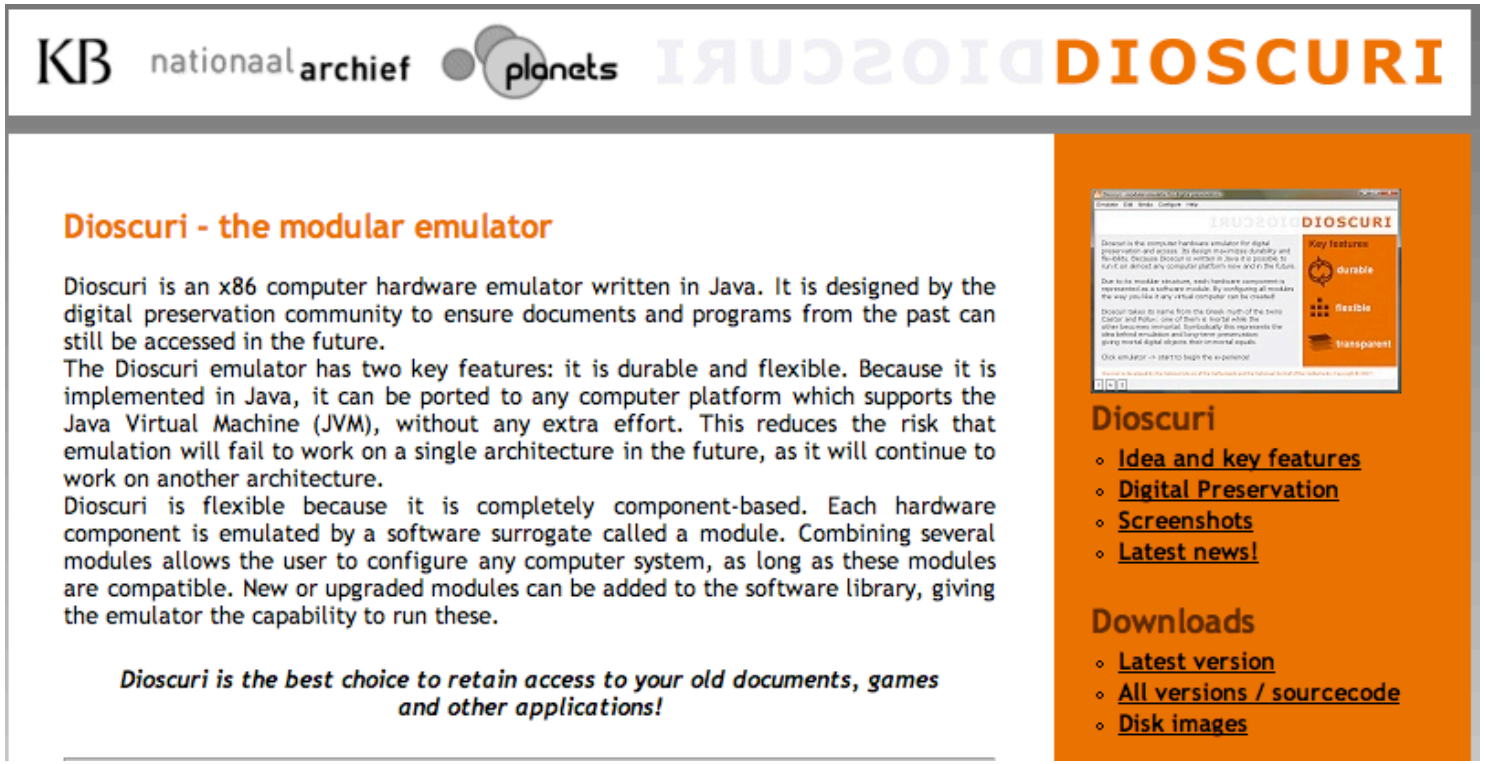


Digital preservation methods

- Emulation
  - Keep the original data and application software
  - Creat programs which emulate the behaviours of old computer systems
  - The original application and data can be processed – emulated – on contemporary architectures



Dioscuri (http://dioscuri.sourceforge.net/)



Dioscuri is an x86 computer hardware emulator written in Java. It is designed by the digital preservation community to ensure documents and programs from the past can still be accessed in the future.

More on migration and emulation

Two ends of a spectrum:

- Emulation *keeps data as is*, and builds software that enables a current computing platform to 'behave like' the original hardware, operating system, or software necessary to read the data
- Migration *transforms the data to modern formats*, so they can be read on modern hardware/software

Digital preservation methods

- Computer museums
  - Keep whole systems, including hardware and systems software, so that they can be used in the future
  - Issues: costs, lack of available expertise, physical decay of hardware

Digital preservation methods

- In practice
  - Migration and emulation are the primary preservation methods
  - No single method can be applied by itself (for example, they all depend on migration to some extent)
  - Selection of methods is determined by factors such as "the material to be preserved, the degree of technical success achieved and on economic and organisational factors" (Lord and Macdonald *Data Curation for e-Science in the UK* (2003) p.31).

DCC

Curation Checklists

Checklist for preservation action

☑	Know what you want people to be able to do with your data – this will impact many aspects (formats selected for long term storage, compression, etc…)
☑	Pin down the significant properties of your data and communicate them – make sure that the people carrying out preservation actions know what they are. This might be through metadata or other means.
☑	Don't be afraid to be critical when reviewing 'best practice' and recommended approaches. They might work for the specific scenario for which they were created but not for you. Do you know the criteria used to rate things like 'preferred' formats?
☑	Document preservation actions so that people know what has been done to the date over time.

Topic 4: Automating preservation actions

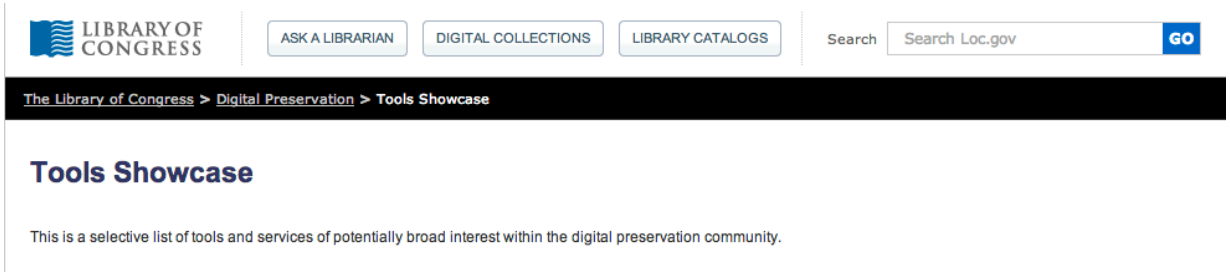
- Why automation is important
- Lists of tools
- Categories of tools
- Archivematica

Automating preservation actions

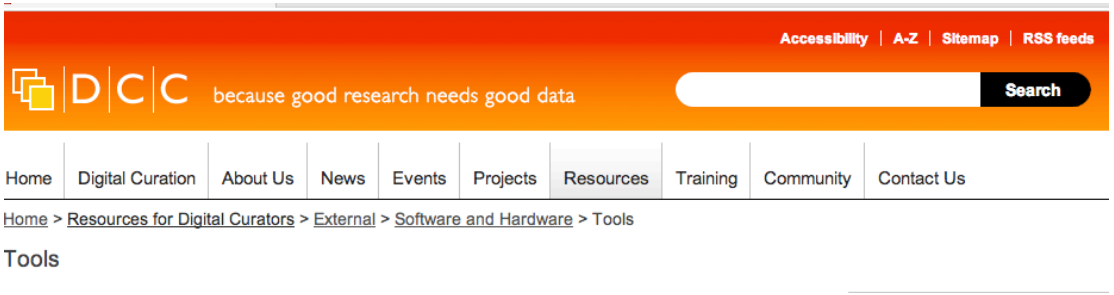
- Current digital preservation methods are labour-intensive – "artisan", "hand-crafted"
- We need ways of automating curation workflows & preservation actions
- **Automated curation procedures need tools**
  - Development by PLANETS, NDIIPP, KEEP

Automating preservation actions

- Lists of data curation tools
  - NDIIPP



  - DCC list *Digital Curation Tools*

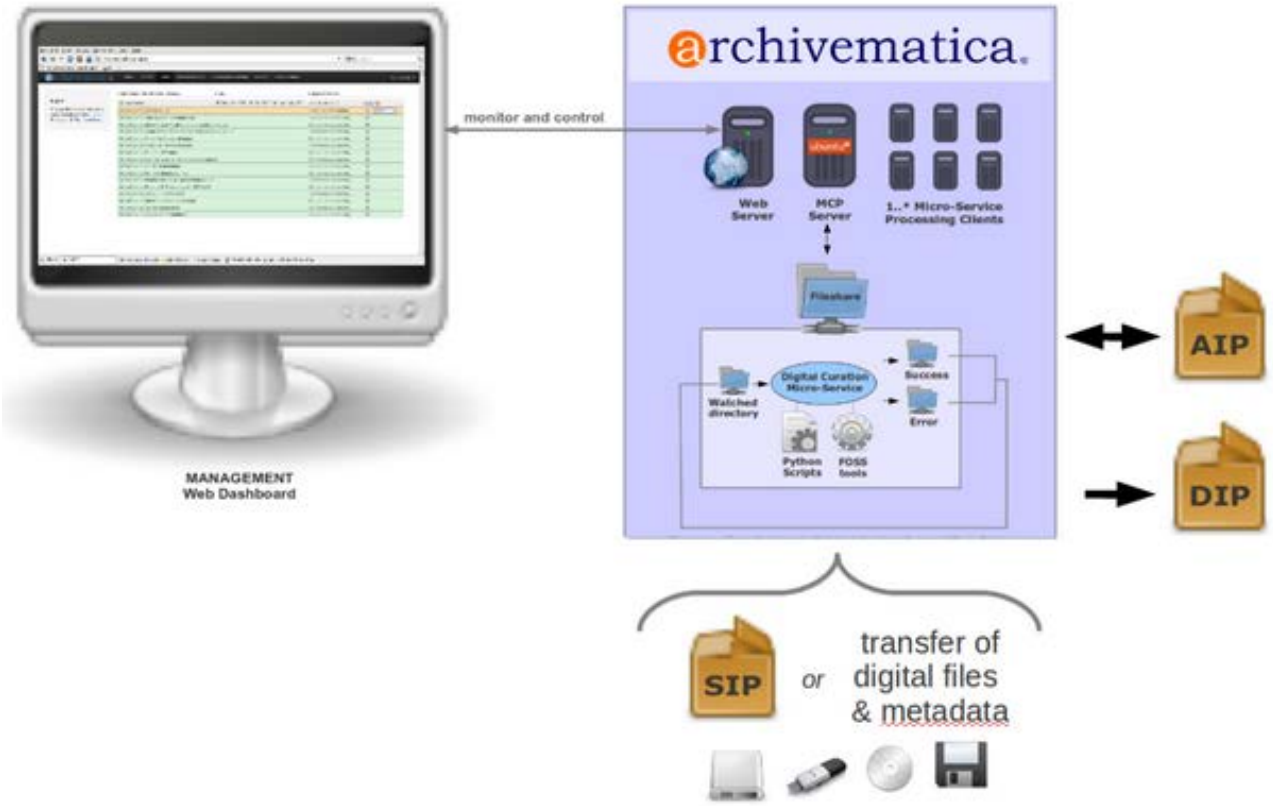


Categories of tools

- **Metadata tools** - extract technical metadata from digital objects, convert extracted metadata into XML
  - National Library of New Zealand Metadata Extraction Tool
  - Ecological Metadata Language (EML) editor
- **Format validation, format registry** and **obsolescence notification tools** - identify file formats, validate them
  - DROID , JHOVE
  - PRONOM Technical Registry
- **Web archiving tools**
  - Web Curator Tool
  - IIPC Web Archiving Toolset, NetarchiveSuite
- Also emulation tools, normalising/encapsulation tools, workflow tools

Archivematica

- Potentially very influential
- What is it?



What is Archivematica?

- digital preservation/curation system
- designed to maintain standards-based, long-term access to collections of digital objects
- uses a micro-services design pattern
- free and open-source (AGPLv3)
- supported by Artefactual Systems Inc.

https://www.archivematica.org/

What is Archivematica?

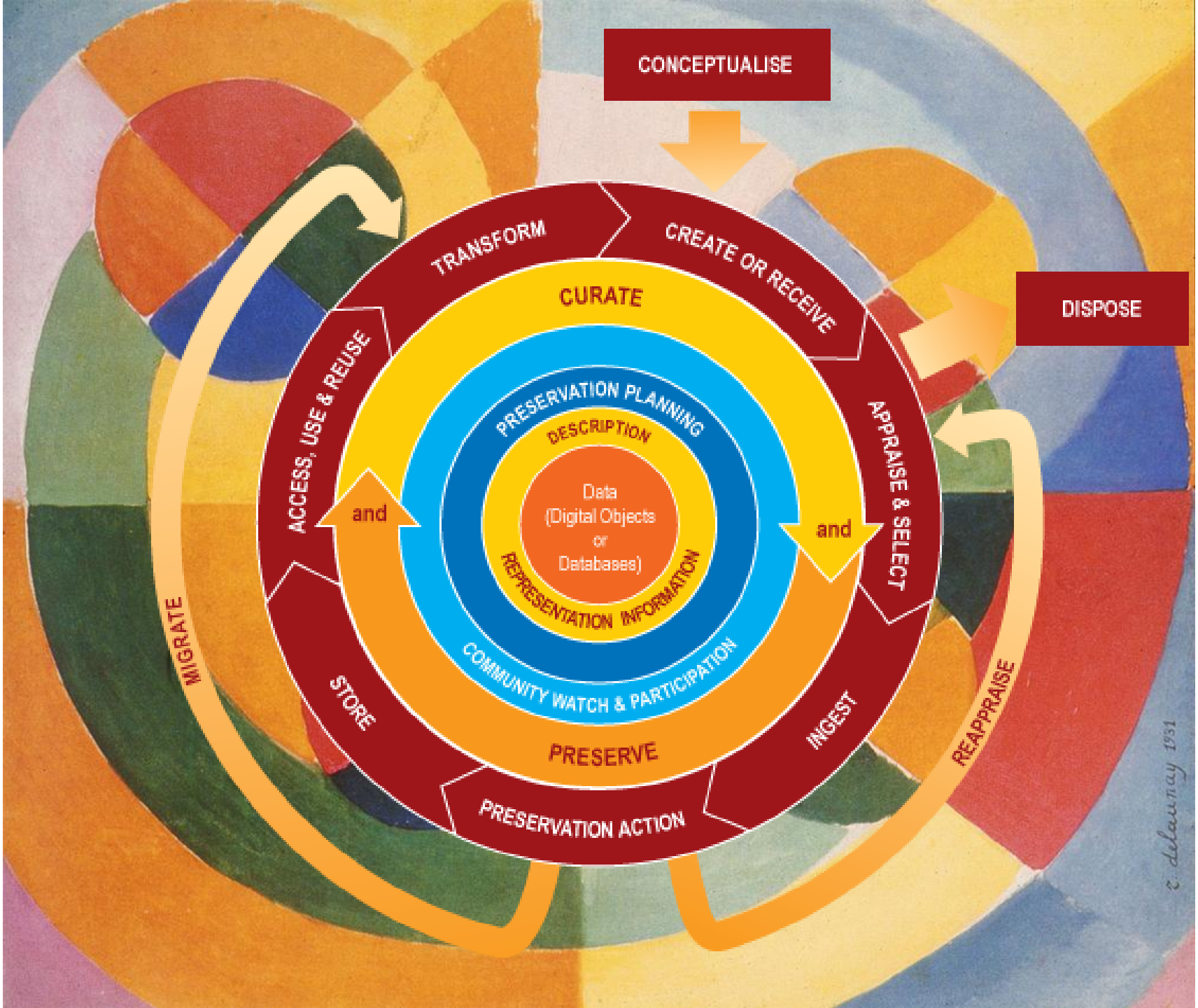
- allows users to process digital objects from ingest to access in conformance with the ISO-OAIS functional model
- creates high-quality, standards-compliant Archival Information Packages (AIP)
- provides
  - an architecture for implementing preservation strategies
  - a framework for evaluating and implementing format policies

Topic 5: Store

- DCC Lifecycle Action: *Store*
- Storage requirements
- Store: practices
- Cloud storage

Store

- Sixth sequential action of the data curation lifecycle
- Activity:
  - Storing the data in a secure manner adhering to relevant standards
- Storage facilities should ensure that:
  - Data are stored securely over time: their integrity is not compromised
  - The storage is reliable: data is not lost
  - They meet the requirements of relevant standards
  - Data in storage can be accessed: they can be got out of storage for use and reuse


CONCEPTUALISE
DISPOSE
TRANSFORM
CREATE OR RECEIVE
APPRAISE & SELECT
INGEST
PRESERVATION ACTION
STORE
ACCESS, USE & REUSE
MIGRATE
REAPPRAISE
CURATE
PRESERVE
and
and
PRESERVATION PLANNING
DESCRIPTION
REPRESENTATION INFORMATION
COMMUNITY WATCH & PARTICIPATION
Data
(Digital Objects
or
Databases)
c. delaunay 1931

Storage requirements

- Ten basic characteristics of digital preservation repositories identified by participants at a 2007 meeting

http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re

1. The repository commits to continuing maintenance of digital objects for identified community/communities.
2. Demonstrates organizational fitness (including financial, staffing structure, and processes) to fulfill its commitment.
3. Acquires and maintains requisite contractual and legal rights and fulfills responsibilities.
4. Has an effective and efficient policy framework.
5. Acquires and ingests digital objects based upon stated criteria that correspond to its commitments and capabilities.
6. Maintains/ensures the integrity, authenticity and usability of digital objects it holds over time.
7. Creates and maintains requisite metadata about actions taken on digital objects during preservation as well as about the relevant production, access support, and usage process contexts before preservation.
8. Fulfills requisite dissemination requirements.
9. Has a strategic program for preservation planning and action.
10. Has technical infrastructure adequate to continuing maintenance and security of its digital objects.

Store: practices

Appropriate practices

- Store the data in formats that:
  - do not apply any form of manipulation which causes data loss or loss of authenticity
  - are widely implemented and supported
  - preferably are open or non-proprietary
  - have a potentially long life
  - are most likely to have migration pathways to the next format available

Store: practices

- Store enough metadata and representation information to support identification, access and preservation processes
- Use a reliable storage format on at least two types of carrier
- Make multiple copies, which are checked and verified regularly
- Replace carriers and software as the market demands, with plans to migrate the content to the next type of reliable carrier

Store: cloud storage

- Where do we "physically" locate the data?
- Under our immediate control, e.g. local servers
  - Pros?
  - Issues?
- Commercial storage, e.g. Cloud
  - Pros?
  - Issues?

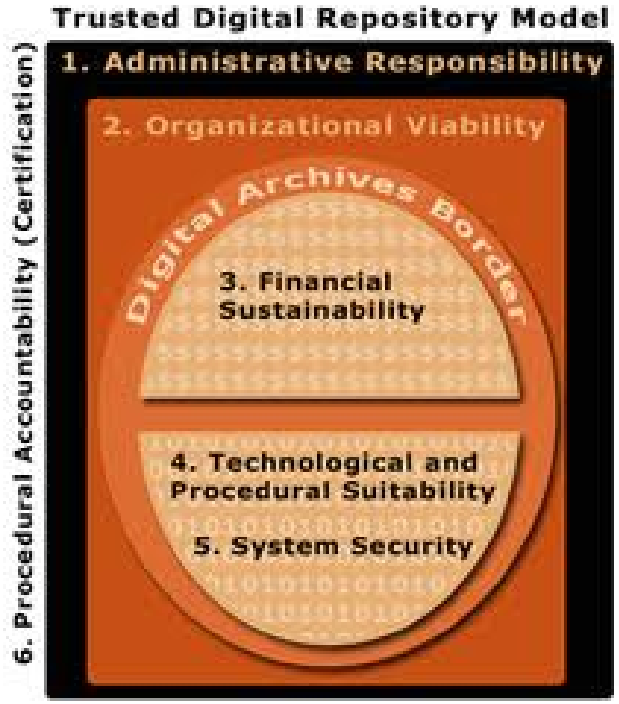
DURASPACE.ORG

Topic 6: Trusted Digital Repositories

- Definitions
- Requirements
- Audit and certification
- HathiTrust

TDRs (Trusted Digital Repositories)

– One whose mission is to provide reliable, long-term access to managed digital resources to its designated community, now and in the future (RLG, *Trusted Digital Repositories: Attributes and Responsibilities* (2002))



TDRs (Trusted Digital Repositories)

- A TDR must meet certain requirements
  - Compliance with OAIS Reference Model
  - Accept responsibility for long-term maintenance of digital resources
  - Have a sustainable organizational system
  - Demonstrate fiscal responsibility and sustainability
  - Design its system(s) in line with commonly accepted conventions and standards
  - Establish methodologies for system evaluation
  - Be trusted to carry out its long-term responsibilities
  - Have policies, practices, and performance that can be audited and measured

TDRs (Trusted Digital Repositories)

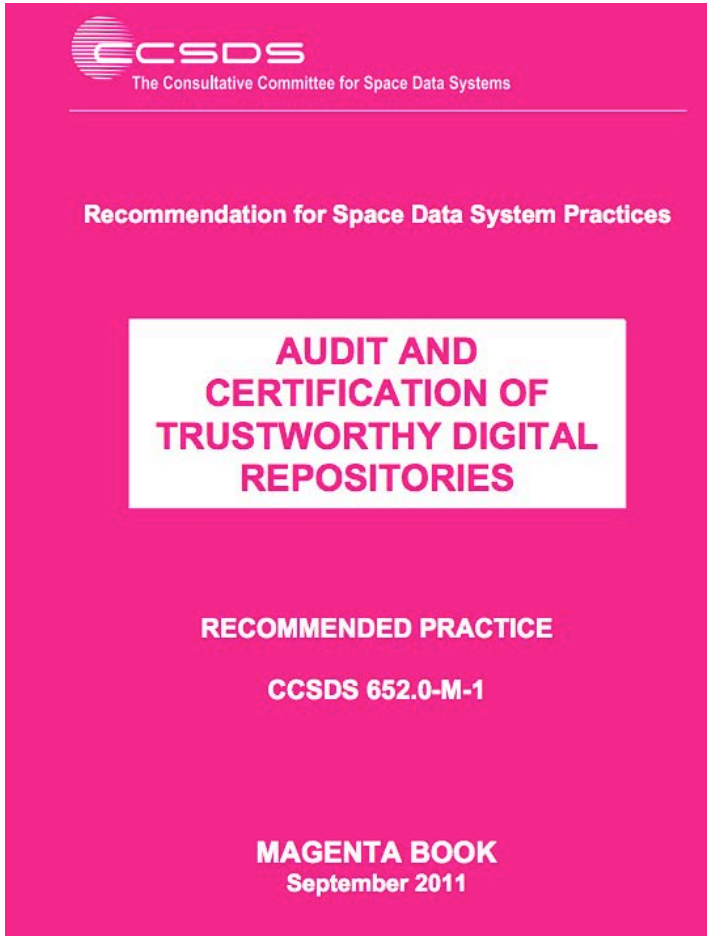
How do you know if the digital repository can be trusted to maintain its contents over time?

Audit and certification:

- TRAC: the Trustworthy Repositories Audit and Certification (TRAC) Criteria and Checklist:http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf
- DRAMBORA: the Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) toolkit: http://www.repositoryaudit.eu/

ISO 16363

- *NEW! ISO standard for auditing and certification of TDRs based on TRAC, ISO/DIS 16363*



Example of a TDR: HathiTrust

HATHI TRUST
a shared digital repository

http://www.hathitrust.org/

Currently Digitized

10,701,922 total volumes
5,613,722 book titles
278,108 serial titles
3,745,672,700 pages
480 terabytes
127 miles
8,695 tons
3,357,616 volumes(~31% of total) in the public domain
View visualizations of HathiTrust
call numbers, languages, and dates

statistics information ››

Preservation Formats

HathiTrust currently ingests only documented acceptable preservation formats, including TIFF ITU G4 files stored at 600dpi, JPEG or JPEG2000 files stored at several resolutions ranging from 200dpi to 400dpi, and XML files with an accompanying DTD (typically METS). HathiTrust supports these formats because of their broad acceptance as preservation formats and because the formats are documented, open and standards-based, giving HathiTrust an effective means to migrate its contents to successive preservation formats over time, as necessary. The Repository Administrators have undertaken such transformations in the past; moreover, HathiTrust offers end-user services that routinely transform digital objects stored in HathiTrust to "presentation" formats using many of the widely available software tools associated with HathiTrust's preservation formats. HathiTrust gives attention to data integrity (e.g., through checksum validation) as part of format choice and migration.

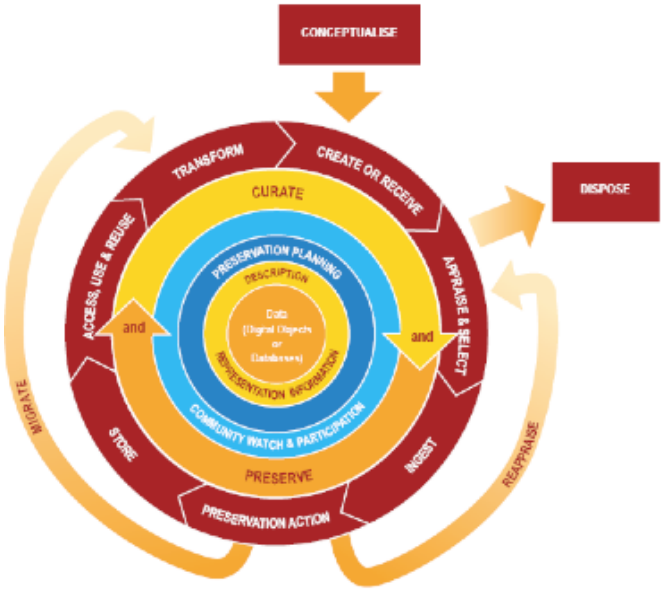
Topic 7: Access, Use & Reuse

- DCC Lifecycle Model: *Access, Use & Reuse*
- Key activities
- Standards
- Legal permission
- Access controls and authentication
- Checklist

Access, Use & Reuse

Strategies to ensure that archived digital objects can be located and used

- Standards
- Legal permission
- Collaboration
- Access controls



Guiding questions:
Who are the users?
How do they use/reuse digital objects?
Why is re-use so important?
Are our digital archives usable?
How can we improve their usability?

Access, Use & Reuse

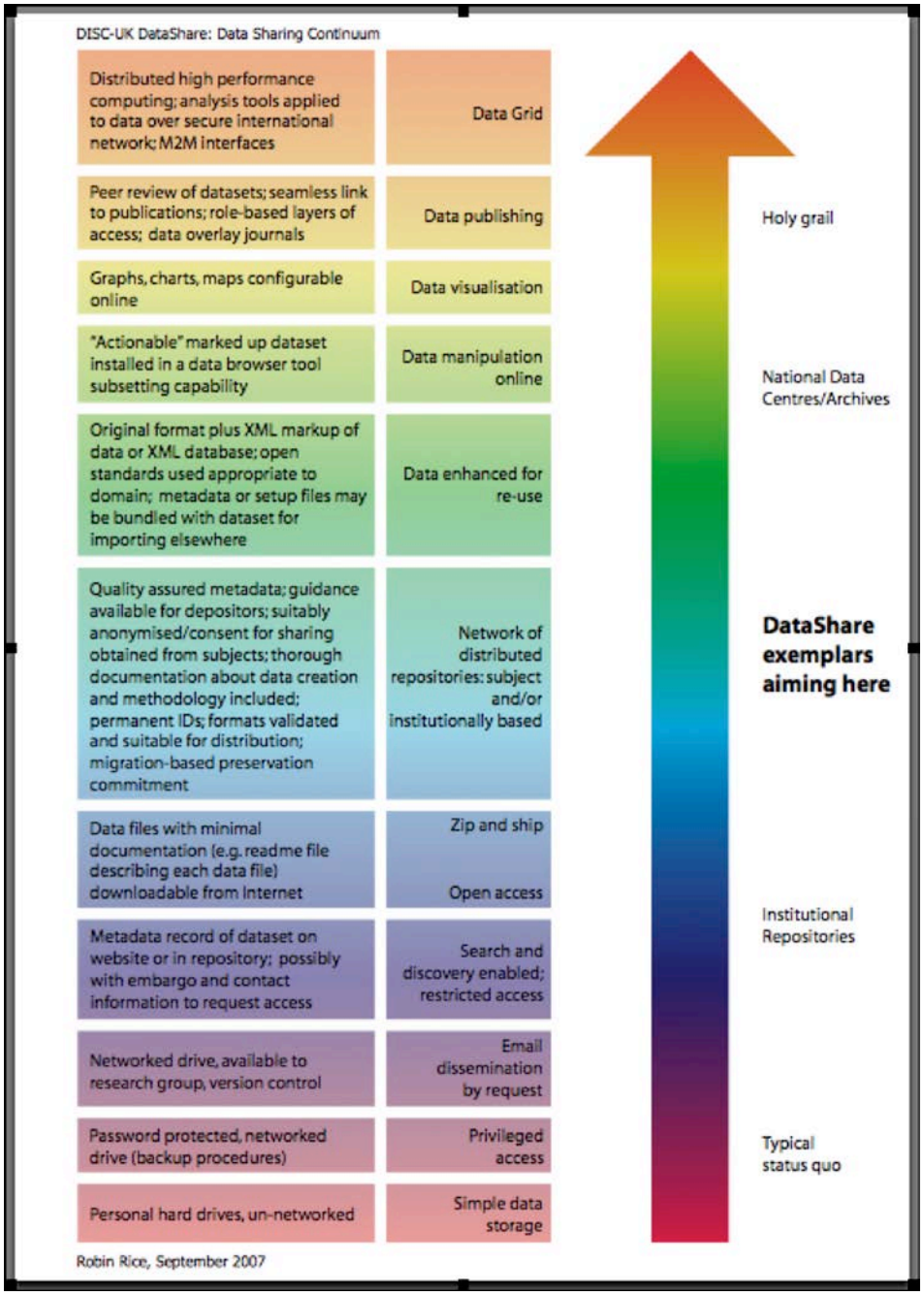
- The seventh sequential action of the data curation lifecycle
- Activities:
  - Ensuring that data is accessible to both designated users and reusers, on a day-to-day basis, usually (but not necessarily) in the form of publicly available published information
  - Applying robust access controls and authentication procedures where applicable

Access, Use & Reuse

- Aim of these actions - to ensure that:
  - Digital objects can be located
  - Digital objects can be used and reused by legitimate users

Access, Use & Reuse

- Key activities:
  - Ensuring digital objects can be **discovered** (located) by applying standards that ensure appropriate metadata are present and allow interoperability
  - Ensuring that the required **legal permissions** are available for digital objects to be used and reused, and that legal restrictions on their use and reuse are adhered to
  - Providing tools that allow **collaboration** in the use and reuse of digital objects
  - Ensuring digital objects are **accessible** only by authorized users, by applying access controls and authentication procedures



Data sharing

DISC-UK DataShare: Data Sharing Continuum

http://www.disc-uk.org/docs/data_sharing_continuum.pdf

Legal permissions

- Legal restrictions that mean digital objects can't be shared:
  - Owner may not allow re-use IP rights may be so restrictive that access to that material is not possible
- Legal requirements fall into three categories:
  - Requirements of funding bodies
  - Legislation about confidentiality & privacy
  - IP rights, digital rights management

Access controls and authentication

- Rights issues
  - Who has the right to access that data?
  - Who does not?
- Authorizing users :
  - Users can be authenticated - standard good practice for computing facilities

The Shibboleth System is a standards based, open source software package for web single sign-on across or within organizational boundaries. It allows sites to make informed authorization decisions for individual access of protected online resources in a privacy-preserving manner.

DCC

Curation Checklists

Checklist for access and reuse

☑	Know what you want users to be able to do with your data and for how long.
☑	Pin down and communicate the significant properties of your data.
☑	Ensure that any restrictions on access and use are communicated and respected.
☑	Ensure that you provide enough context to ensure that your data can be located and used – either by the originally designated user community or new users over time.

Topic 9: Putting it all together

- Library of Congress DPOE (Digital Preservation Outreach and Education) framework
- List of starting points

Putting it all together

Library of Congress DPOE (Digital Preservation Outreach and Education) framework

1. **Identify** - What digital content do you have?
2. **Select** - What portion of your digital content will be preserved?
3. **Store** - What issues are there for long-term storage?
4. **Protect** - What steps are needed to protect your digital content?
5. **Manage** - What provisions are needed for long-term management?
6. **Provide** - What considerations are there for long-term access?

Putting it all together

I have added two more: **Plan,** and **Ingest:**

1. **Plan** - What do you need to plan before you start?
2. **Identify** - What digital content do you have?
3. **Select** - What portion of your digital content will be preserved?
4. **Ingest** - How do you get digital content into the archive?
5. **Store** - What issues are there for long-term storage?
6. **Protect** - What steps are needed to protect your digital content?
7. **Manage** - What provisions are needed for long-term management?
8. **Provide** - What considerations are there for long-term access?

1. Plan - What do you need to plan before you start?

- Know who you are creating your digital content for and what you want them to be able to do with it
- Use preservation-friendly file formats
- Keep documentation about the digital content, formats, software, agreements about its use
- Develop file-naming policy
- Be realistic - strike a balance between what is sufficient and what is ideal based on your practical realities

2. Identify - What digital content do you have?

- Survey your digital holdings

3. Select - What portion of your digital content will be preserved?

- Start on selection and appraisal as early as possible
- Know who you are creating your data for and what you want them to be able to do with it

4. Ingest - How do you get digital content into the archive?

- Get a receipt or acknowledgement of transfer
- Capture digital content in preservation-friendly file formats
- Calculate checksums
- Assign metadata
- Run antivirus checks

5. Store - What issues are there for long-term storage?

- Make multiple copies, which are checked and verified regularly
- Store files on appropriate media
- Store enough metadata and representation information to support identification, access and preservation processes
- Store digital content in preservation-friendly file formats

6. Protect - What steps are needed to protect your digital content?

- Use good IT practices for data security, backups, error checking
- Maintain multiple copies of the bit stream
- Refresh, check accuracy of results, document the process

7. Manage - What provisions are needed for long-term management?

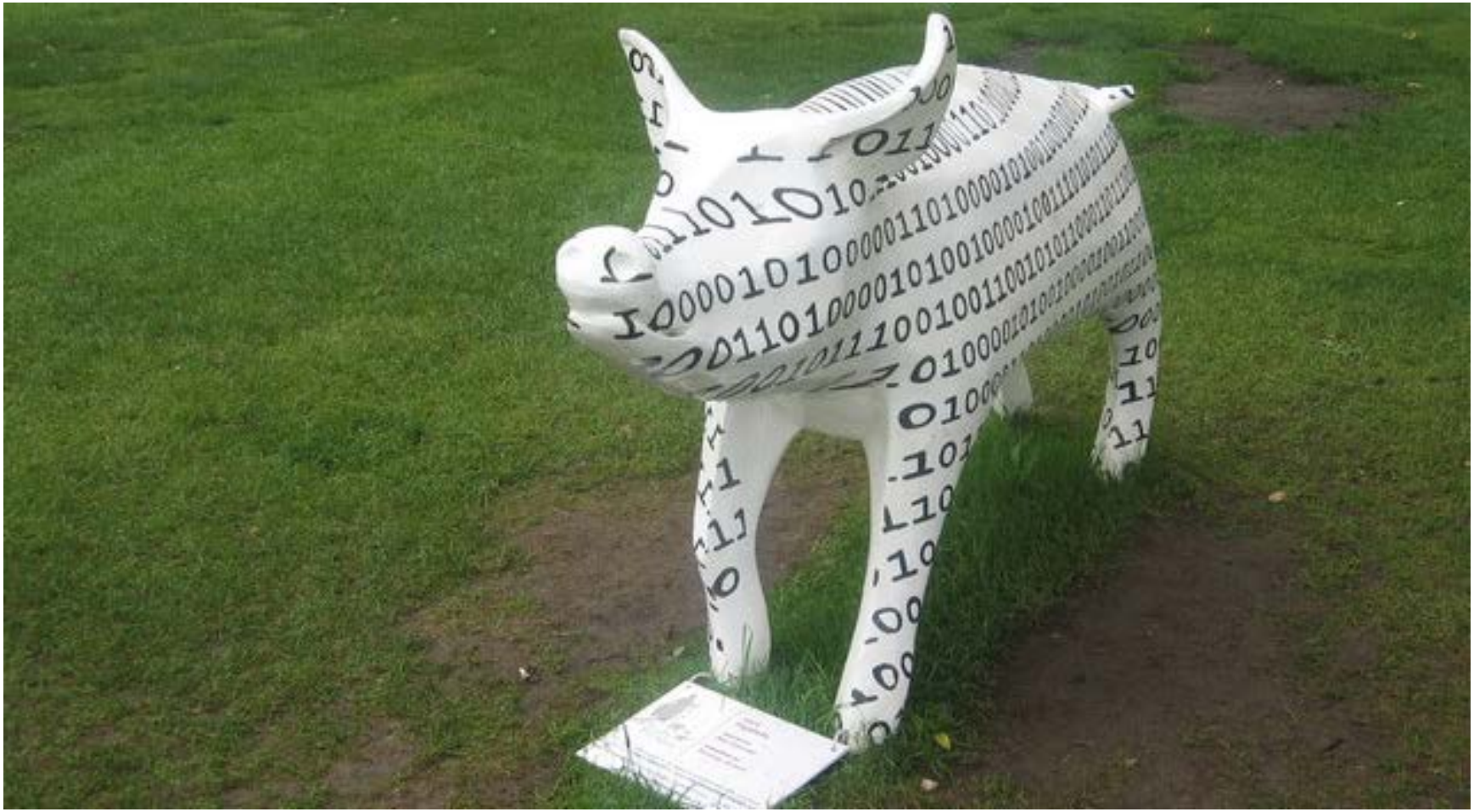
- Ensure you have the right to copy and apply preservation processes
- Document preservation actions so that people know what has been done to the digital content over time

8. Provide - What considerations are there for long-term access?

- Ensure that you know about any restrictions on access and use of your digital content
- Ensure that users know about any restrictions on access and use of your digital content

Just for fun: Pigabyte

http://www.kingbladudspigs.org/



Contact details

New email address from 1 July:

rharvey@gmx.com

Website: elibank.net