

Digital Curation

Getting digital objects into the archive

Ross Harvey

25 June 2013

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

Digital Curation: getting digital objects into the archive

Introduces a lifecycle approach to digital curation

Notes initial processes needed to provide high quality curation

Topics:

1. Challenges
2. What are we aiming to do?
3. Two models
4. The importance of planning
5. 'Preservation-friendly' digital objects
6. The role of metadata
7. Selecting digital objects
8. Ingest procedures

Topic 1: Challenges

- Obsolescence
- Quantity of digital objects
- Nature of digital objects
- Reproducing authentic digital objects
- Keeping digital objects over time

Challenges of digital curation: Obsolescence



Osborne portable computer 1981

CP/M Operating System, 64 KB memory,
two 5¼-inch floppy disk drives

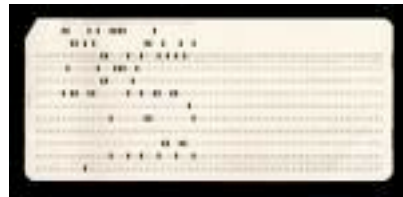


iPod Touch 2012

iOS operating system, 32 GB
memory, unlimited cloud storage

HARDWARE CHANGES FAST

Challenges of digital curation: Obsolescence



Lost your data?



STORAGE MEDIA DETERIORATES FAST

What to do with old media



Challenges of digital curation: Obsolescence

THE *SOFTWARE* CHANGES FAST

What is this?

How would you open it?

THE *FILE FORMATS* CHANGE FAST

What is this?

How would you open it?



Challenges of digital curation:

Quantity of digital objects

Quantities

We create and handle *lots* of digital materials in LIS work, e.g.

- Files created in digitizing projects
- Born-digital materials

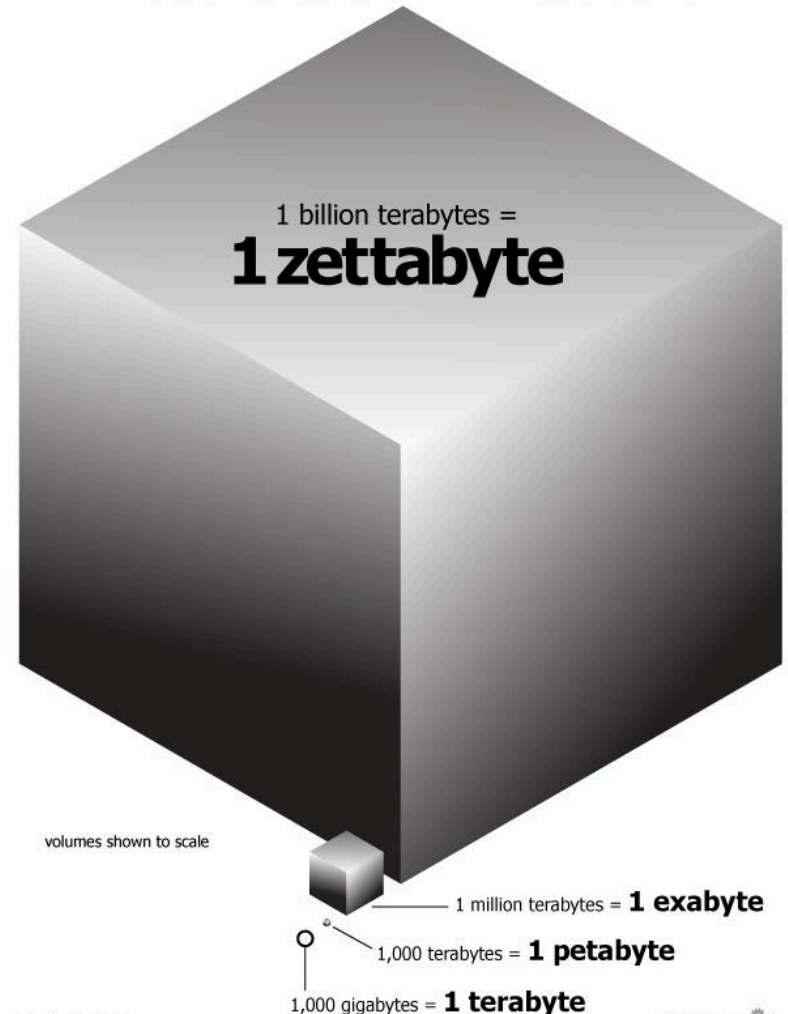
Internet-hosted materials

Quantities extremely large

BUT our procedures for archiving can currently handle only small quantities















Humanity Passes 1 Zettabyte Mark in 2010

A zettabyte is 1,000,000,000,000,000,000 bytes (that's 21 zeroes for those counting), or one trillion gigabytes. That's enough data to fill 75 billion 16-gigabyte-sized iPads.



Challenges of digital curation:

Nature of digital objects

Name	Date Modified
 APP1WP	Jan 29, 1991 10:59 PM
 CPRPERS.NDX	Jan 1, 1980 8:07 AM
 INANGHER	Dec 31, 1980 11:00 PM
 INTRO.RD	Jan 28, 1992 3:05 AM
 LCRIMAIL.RXD	Jan 1, 1980 9:52 AM
 LTFILES	Nov 27, 1992 5:22 PM
 NEWSP2.PRT	Jan 17, 1995 10:29 PM
 NZPNPERS.NDX	Dec 31, 1980 11:00 PM
 OTHEREXP	Jan 19, 1992 4:47 PM
 REGNOTES.DBF	Jan 1, 1980 9:56 AM
 SESSION8	Apr 15, 1992 4:29 AM
 TOWNNUMB	Jul 8, 1991 5:20 AM
 WANGEVHE.WP	Nov 15, 1991 2:30 AM
 Week11.ppt	Oct 18, 1997 1:01 AM

Some of my old files: how to open them?

Challenges of digital curation:

Nature of digital objects

Name	Date Modified
APP1WP	Jan 29, 1991 10:59 PM
CPRPERS.NDX	Jan 1, 1980 8:07 AM
INANGHER	Dec 31, 1980 11:00 PM
INTRO.RD	Jan 28, 1992 3:05 AM
LCRIMAIL.RXD	
LTFILES	
NEWSP2.PRT	
NZPNPERS.NDX	
OTHEREXP	
REGNOTES.DBF	
SESSION8	
TOWNNUMB	
WANGEVHE.WP	
Week11.ppt	

FILENAME	CREATING APPLICATION	OPENS WITH
APP1WP	WordPerfect	Open Office
INANGHER	WordPerfect	Open Office
INTRO.RD	WordPerfect	Open Office
LTFILES	WordPerfect	Open Office
NEWSP2.PRT	WordPerfect	Open Office
OTHEREXP	WordPerfect	Open Office
SESSION8	WordPerfect	Open Office
TOWNNUMB	WordPerfect	Open Office
WANGEVHE.WP	WordPerfect	Open Office
RENOTES.DBF	<u>dBase II</u>	??
CPRPERS.NDX	<u>dBase II</u>	??
NZPNPERS.NDX	<u>dBase II</u>	??
LCRIMAIL.RXD	Reflex	<u>HABit</u> Reflect viewer
Week11.ppt	Microsoft Office 95	Old version of Microsoft Office

Challenges of digital curation: Keeping digital objects over time



Source: *Atlas of Digital Damages*

Topic 2: What are we aiming to do?

- What are we aiming to do when we preserve digital objects?
- Meeting the aims

Aims? Can we meet them?

What are we aiming to do?

- **Authentic** digital records
 - To be what it purports to be, created or sent by the person purported to have created or sent it, created or sent at the time purported
- **Reliable** digital records
 - Contents can be trusted as a full and accurate representation of the transactions, activities or facts **Integrity**
 - Complete and unaltered
- **Usable** digital records
 - Can be located, retrieved, presented and interpreted

Aims? Can we meet them?

How to meet the aims

- Copy digital objects to a reliable digital storage system
- Manage ongoing data protection in accordance with good IT practices for data security, backups, error checking
- Refresh (move files to a newer version of the same storage media, or to different storage media, with no changes to the bit stream), check accuracy of the results (for example, checksums), document the process
- Maintain multiple copies of the bit stream
- Ensure you have the right to copy and apply preservation processes, which may require negotiation with rights owners

Topic 3: Two models

- OAIS Reference Model
- Lifecycle models
- DCC Curation Lifecycle Model

Models

OAIS Reference Model (ISO 14721:2003)

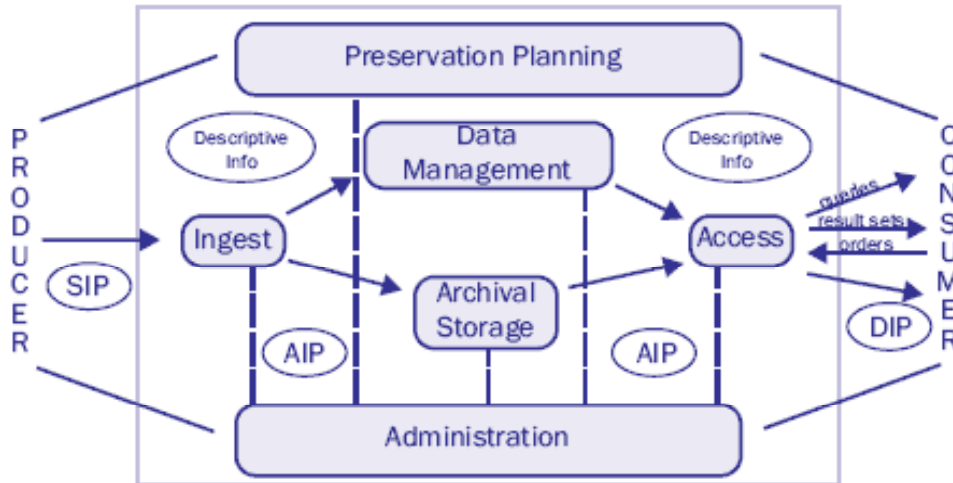


Figure 4: Seven OAIS Functions

Information Package:

1. The digital object to be preserved
2. The metadata required at that point in the system
3. Packaging information

OAIS information packages:

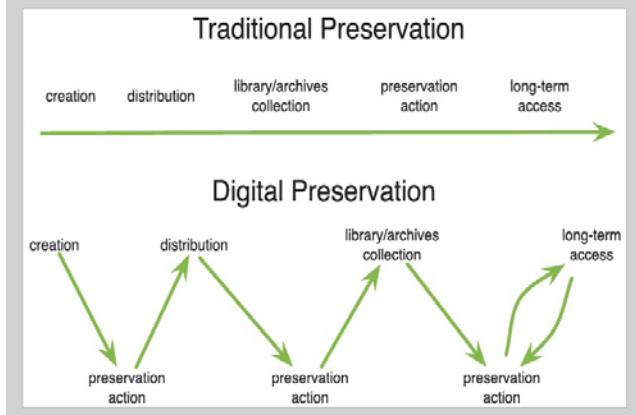
- *Submission Information Package (SIP)* – sent to the OAIS
- *Archival Information Package (AIP)* – what the OAIS produces for archival storage
- *Dissemination Information Package (DIP)* – what the OAIS delivers when there is a request for access

Models

Lifecycle models

Fig. 1. Traditional Preservation Versus Digital Preservation

Digital content requires active management throughout its entire period of use.



RECORDS LIFE CYCLE



Data Life Cycle



Proposal Planning and Writing

Project Start-up and Data Management

Data Collection and File Creation

Data Analysis

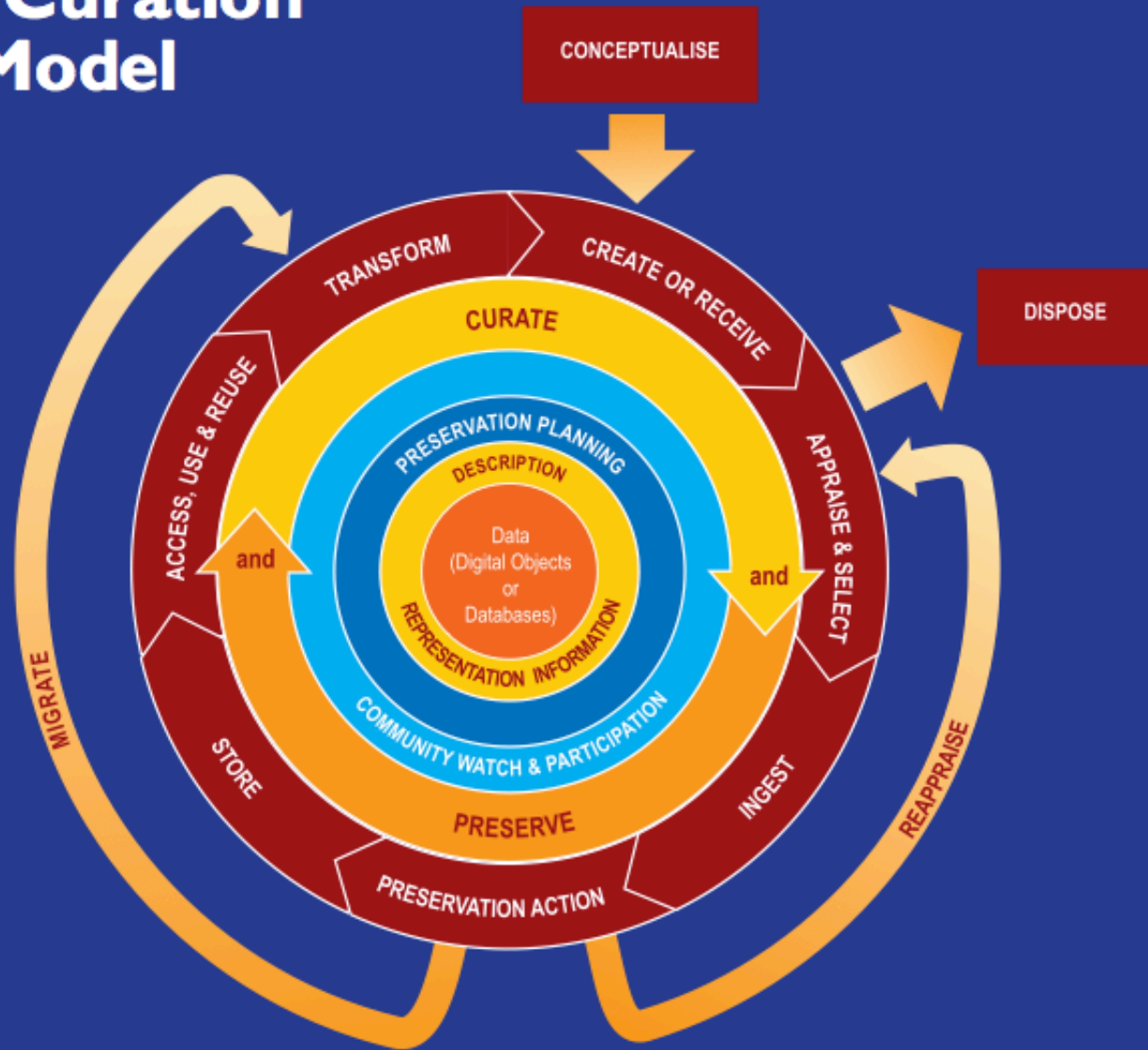
Preparing Data for Sharing

Depositing Data

After-Deposit Archival Activities

1. Anticipate archiving costs and challenges
2. Create a data management plan
3. Follow best practices for data and documentation
4. Manage master datasets and work files
5. Determine file formats to deposit
6. Comply with dissemination standards and formats

The DCC Curation Lifecycle Model



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

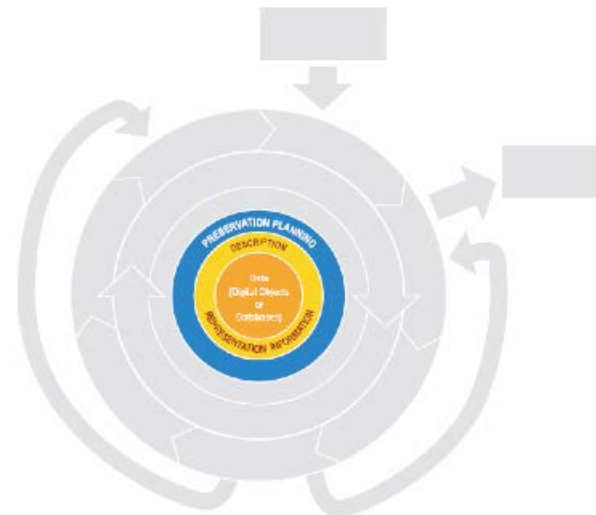
Topic 4: The importance of planning

- Planning in the DCC Lifecycle Model
- Planning tools

The importance of planning

Planning in the DCC Lifecycle Model:

- Specified in *Preservation Planning* action
- Embedded in all lifecycle actions
 - Planning for preservation throughout the curation lifecycle of digital material
 - Developing and applying plans for management and administration of all curation lifecycle actions



Planning tools

- DMP Online: <http://dmponline.dcc.ac.uk/>



- DMPTool: <https://dmp.cdlib.org/>



Topic 5: 'Preservation-friendly' digital objects

- 'Preservation friendly': what is it?
- *Conceptualise*
- Three examples
- Checklists
- Making digital objects preservation-friendly

'Preservation-friendly' digital objects

- 'Preservation friendly': what is it?
- *Preservation-friendly file formats*: open, well-supported standard formats for which access tools are more likely to remain available in the future

DOC *or* RTF *or* ODT?

PDF *or* PDF/A?

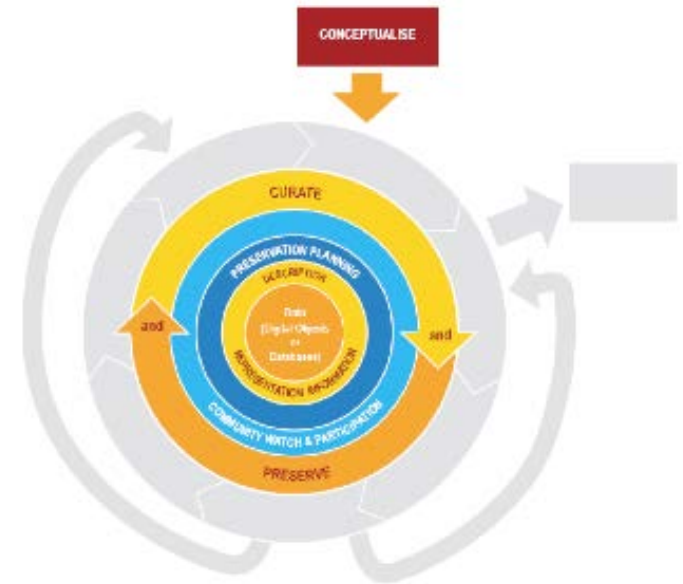
**Sustainability of Digital Formats
Planning for Library of Congress Collections**

<http://www.digitalpreservation.gov/formats/>

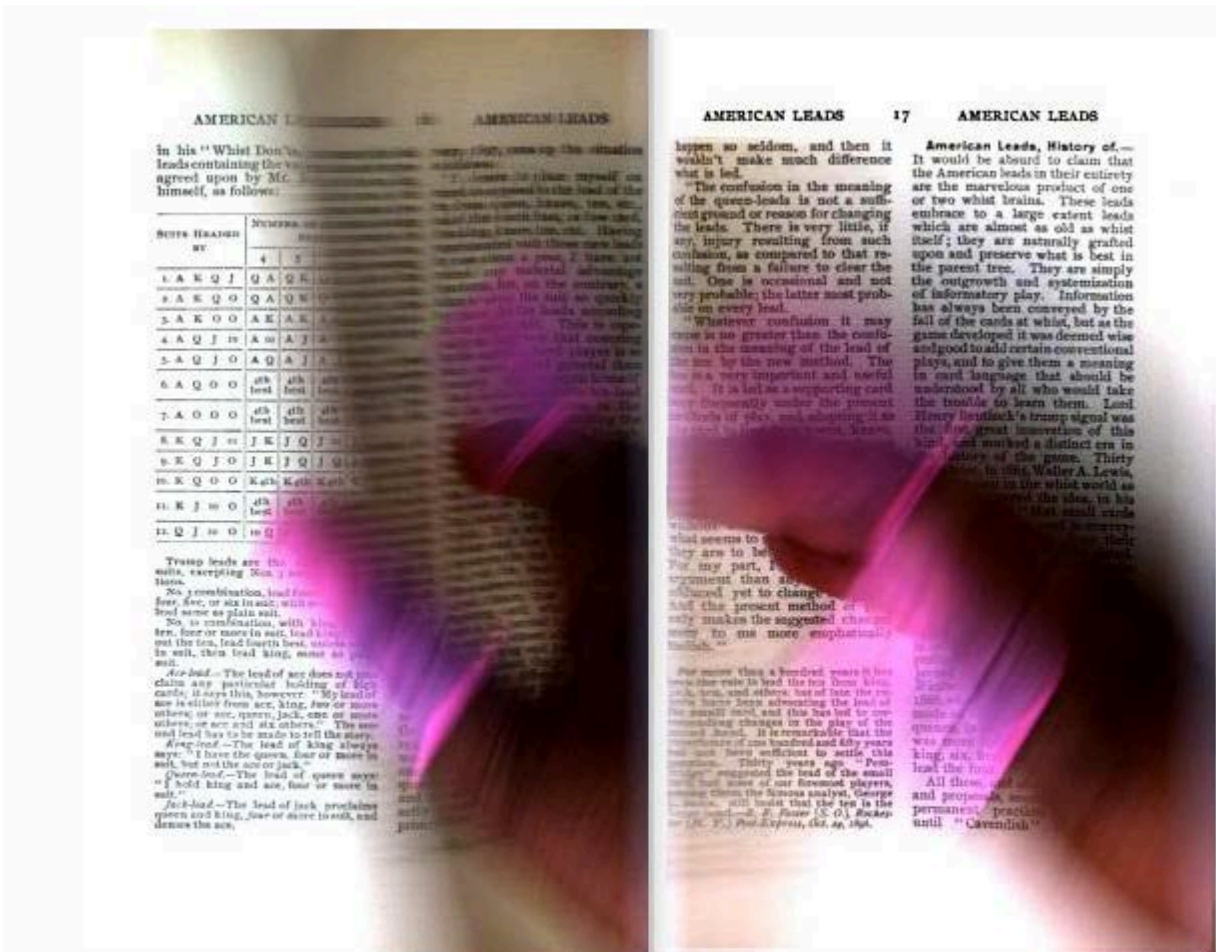
Planning again: Conceptualise

Conceptualise: the first sequential stage of the curation lifecycle

- Conceive and plan the creation of data
- Plan with digital curation processes, outcomes in mind



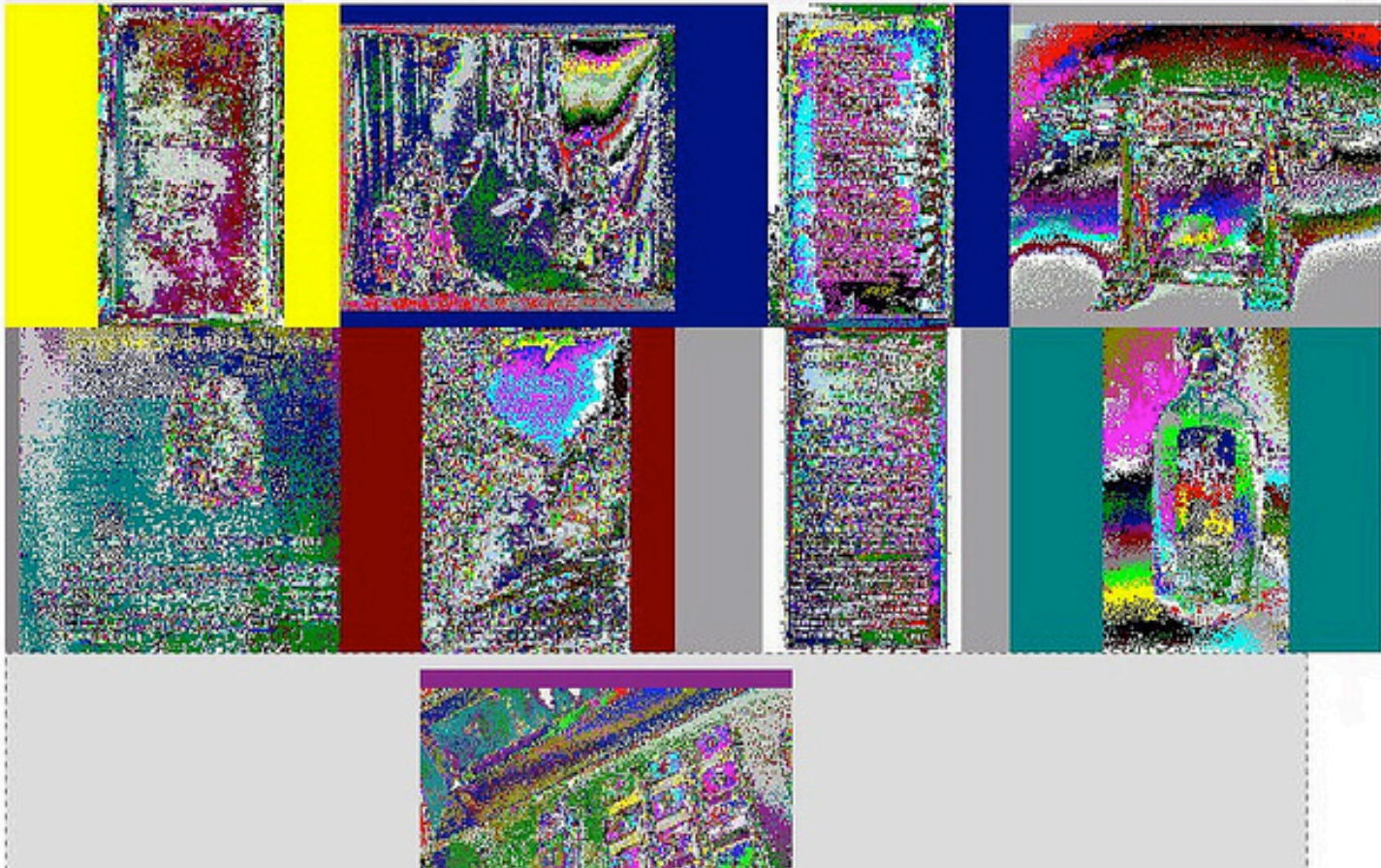
Example 1: What planning could mitigate this?



Employee turns pages (quickly).

From p. 16-17 of *The Whist Reference Book* by William Mill Butler (1898). Original from the New York Public Library. Digitized September 7, 2005. Cited in *The Art of Google Books*, by Krissy Wilson. books.google.com/books?id=fLmgrpR2e9fYC&dq=win&pg=...

Example 2: What planning could mitigate this?



TIFF to PNG thumbnail migration fail

Example 2: What planning could mitigate this?



g i t c h . 0x0c

another_glich

0x0c, One of many glitched PSD files that were recovered from a hard disk failure by Data Rescue 3. From www.flickr.com/photos/warzauwynn/7256819760/



Checklist for conceptualisation

<input checked="" type="checkbox"/>	Get into the habit of equating data curation with good research.
<input checked="" type="checkbox"/>	Know what your funding body expects you to do with your data and for how long. Assess your ability to be able to meet these expectations (i.e., do you need additional funding or staff?).
<input checked="" type="checkbox"/>	Determine intellectual property rights from the outset and ensure they are documented.
<input checked="" type="checkbox"/>	Identify any anticipated publication requirements (embargoes, restrictions on publishing over multiple sites).
<input checked="" type="checkbox"/>	Identify and document specific roles and responsibilities as early as possible.



Checklist for create and/or receive

<input checked="" type="checkbox"/>	Know who you are creating your data for and what you want them to be able to do (and not do) with it. Communicate this with others on the project.
<input checked="" type="checkbox"/>	Identify any data protection requirements that you need to address in the course of your research and ensure that these are communicated to all staff.
<input checked="" type="checkbox"/>	Agree from an early stage any standards you will be making use of for content, syntax, and structure. Once these have been agreed, make sure they are communicated - both to other researchers on the project and to the data/information managers you will be working with. Provide training if necessary.
<input checked="" type="checkbox"/>	Identify data quality metrics as soon as possible and ensure that these are communicated and monitored.
<input checked="" type="checkbox"/>	Work together - researchers and information managers need to communicate regularly. Neither can do their job in isolation.
<input checked="" type="checkbox"/>	Be realistic – strike a balance between what is sufficient and what is ideal based on your practical realities.

Making digital objects preservation-friendly

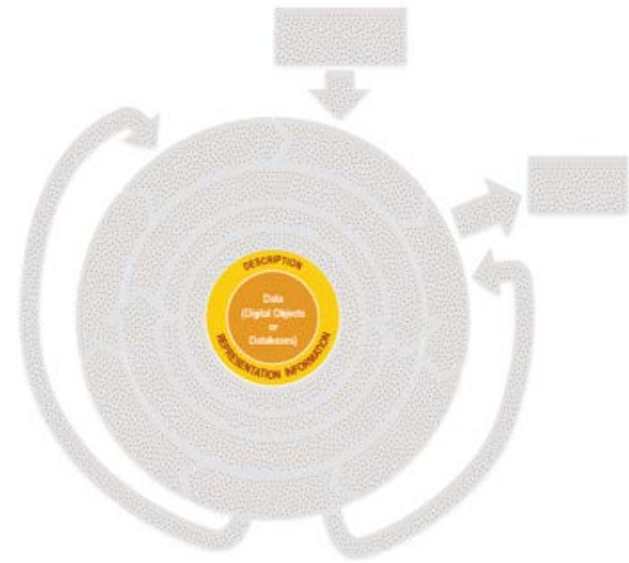
- Capture and store digital objects in preservation-friendly **file formats**
- Keep **documentation** about objects, formats, software, agreements about its use
- Scrupulously **identify** files
- **Store** files on appropriate media

Topic 6: The role of metadata

- *Description & Representation Information* (D&RI)
- What D&RI does
- Examples of D&RI
- Sample repository record

Description & Representation Information (Metadata)

- D&RI is crucial to all aspects of digital stewardship
 - “Assign administrative, descriptive, technical, structural and preservation metadata, using appropriate standards, to ensure adequate description and control over the long-term. Collect and assign representation information required to understand and render both the digital material and the associated metadata”



What Description & Representation Information does

- **Describes** digital objects and where to find them:
 - persistently identifies them
 - clearly describes what they are
 - clearly identifies their technical characteristics
- Gives **technical information** needed to use them:
 - describes what can be done to them
 - describes what is needed to re-present them
- Describes **what happens** to them:
 - Identifies responsibility for their preservation
 - records their history, documents their authenticity

Examples of Description & Representation Information

- **Describes** digital objects and where to find them:
 - Persistent identifier (eg DOI – Digital Object Identifier)
- Gives **technical information** needed to use them:
 - Technical characteristics (eg format, compression or encoding algorithms, encryption and decryption keys, or software - including the release number) used to create
- Describes **what happens** to them:
 - Dates when digital objects created, when updated, when migrated, descriptions of the migration process

Figure 6.1. Description Information and Its Functions

	Broad Function	Type	Specific Function	Examples
Descriptive Information	Describes data and their location	Descriptive metadata	Allows data to be identified so they can be linked with requests	Name of the creator of the data set Name of the author of a document
		Structural metadata	Describes how compound digital objects are organized	
	Provides the technical information needed to use data	Technical metadata	Provides the technical information needed to use data	Format Compression or encoding algorithms Encryption and decryption keys Software (including release number) used to create or update the data
			Provides information about the overall system environment	Hardware, operating systems, application software in which the data were created
	Describes what has happened to data as they move through the curation lifecycle	Administrative metadata	Provides information about the use, management, and encoding processes of digital objects over a period of time	Information about data creation, subsequent updates, transformation, versioning, summarization Descriptions of migration and replication
		Preservation metadata	Records the preservation actions that have been applied to data over time	File format Significant properties Technical environment Fixity information


Sample repository record: how users see it

The Guide For GSLIS Students

[Show full item record](#)

Title:	The Guide For GSLIS Students
Author:	Graduate School of Library and Information Science
URI:	http://hdl.handle.net/10090/17896
Date:	2010

Files in this item

Files	Size	Format	View
GSLIS Guide 2010.pdf	836.3Kb	PDF	

The following license files are associated with this item:

- [Original License](#)

This item appears in the following Collection(s)

- [Student Guide](#) [1]

[Show full item record](#)

Sample repository record: metadata

The Guide For GSLIS Students

[Show simple item record](#)

dc.contributor.author	Graduate School of Library and Information Science	
dc.date.accessioned	2010-10-27T12:47:56Z	
dc.date.available	2010-10-27T12:47:56Z	
dc.date.issued	2010	
dc.identifier.uri	http://hdl.handle.net/10090/17896	
dc.description.provenance	Submitted by Stephanie Satalino (stephanie.satalino@simmons.edu) on 2010-10-27T12:47:56Z No. of bitstreams: 1 GSLIS_Guide_2010.pdf: 836334 bytes, checksum: 9d2f8097bb105c3ad5ece00b3efa0973 (MD5)	en
dc.description.provenance	Made available in DSpace on 2010-10-27T12:47:56Z (GMT). No. of bitstreams: 1 GSLIS_Guide_2010.pdf: 836334 bytes, checksum: 9d2f8097bb105c3ad5ece00b3efa0973 (MD5) Previous issue date: 2010	en
dc.language.iso	en_US	en_US
dc.publisher	Simmons College	en_US
dc.title	The Guide For GSLIS Students	en_US
dc.type	Other	en_US

Topic 7: Selecting digital objects

- Starting out – first steps
- Identify
- Select
- Checklist

Starting out: first steps

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

Identify - What digital content do you have?

You've Got to Walk Before You Can Run:
First Steps for Managing Born-Digital Content
Received on Physical Media

Ricky Erway

Senior Program Officer
OCLC Research

Excellent advice

<http://www.oclc.org/research/publications/library/2012/2012-06.pdf>



A publication of OCLC Research

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

A Digital Curation Centre and Australian National Data Service 'working level' guide



How to Appraise & Select Research Data for Curation

Angus Whyte (DCC) and **Andrew Wilson** (ANDS)

<http://www.dcc.ac.uk/resources/how-guides/appraise-select-data>



Checklist for appraise and select

<input checked="" type="checkbox"/>	Make a start on selection and appraisal from as early a point as possible (e.g., apply the new NERC criteria for identifying valuable data sets at the project plan stage).
<input checked="" type="checkbox"/>	Plan for what you think you'll need to keep to support your research findings. What is the minimum you'll need to support your findings over time?
<input checked="" type="checkbox"/>	Know who you are keeping it the data for and what you want them to be able do with it. This may affect the way you keep it and what you keep.
<input checked="" type="checkbox"/>	Conversely, know what you need to dispose of. Destruction is often vital to ensure compliance with legal requirements.
<input checked="" type="checkbox"/>	Ensure that your data meets minimum quality assurance metrics (based on intended use).
<input checked="" type="checkbox"/>	Re-appraisal can take place before ingest so review what you have and what you need to keep before depositing it to long-term storage.
<input checked="" type="checkbox"/>	Work with researchers and information managers to develop policies and to identify realistic and implementable workflows.
<input checked="" type="checkbox"/>	Appraise for the here and now but with an eye to the future.

Topic 8: Ingest procedures

- Ingest in the DCC Lifecycle Model
- Getting digital objects into the archive:
procedures

Ingest (Sequential Lifecycle Action)

“Transfer data to an archive, repository, data centre or other custodian. Adhere to documented guidance, policies or legal requirements.”



Ingest procedures

- ◊ Establish an accession register listing all submissions and uniquely identifying them
- ◊ Verify file formats (e.g., using JHOVE or PRONOM)
- ◊ Assign unique identifiers
- ◊ Confirm receipt of materials with data creator
- ◊ Copy files submitted on removable media (e.g., CD-ROMs, DVDs) to a secure location
- ◊ Verify that files copied have been transferred properly (e.g., by comparing checksums)
- ◊ Review data for confidentiality issues
 - Remove or recode identifiers if necessary
 - Establish access levels if necessary
- ◊ Convert hardcopy documentation to electronic form
- ◊ Convert software-specific documentation in paper form to PDF/A
- ◊ Generate multiple data formats for dissemination and preservation
- ◊ Create documentation
- ◊ Create a metadata record
- ◊ Assign a Digital Object Identifier (DOI)

More lists at <http://www.neal-schuman.com/curation/>

Summary

- Plan
- Identify
- Select
- Ingest
- For more information

Summary: getting digital objects into the archive

Plan

- Use preservation-friendly file formats
- Keep documentation about the data, formats, software, agreements about its use
- Scrupulously identify files
- Develop file-naming policy
- Identify a safe place for your data (e.g., a trusted archive) and make sure that archive will take your data

Identify - What digital content do you have?

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

Summary: getting digital objects into the archive

Ingest

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

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*

For more information

Web sites

- For tools
 - NDIIPP: 'Partner Tools & Services' section
- For good advice
 - DCC
 - Digital Preservation Europe
- In the U.S.
 - NDIIPP (Library of Congress)

