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
Challenges of digital curation: 
*Nature of digital objects*

Some of my old files: how to open them?
Challenges of digital curation: Nature of digital objects
Challenges of digital curation: Reproducing authentic digital objects

Are these the same?

---

**DOC version: not searchable**

**CSV version: searchable**
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


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.

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
7. Optimize data for long-term storage

Harvey Digital Curation I: 5th EABH Summer School, 25 June 2013
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?
Example 2: What planning could mitigate this?
Example 2: What planning could mitigate this?
<table>
<thead>
<tr>
<th></th>
<th>Checklist for conceptualisation</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>Get into the habit of equating data curation with good research.</td>
</tr>
<tr>
<td>✓</td>
<td>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?).</td>
</tr>
<tr>
<td>✓</td>
<td>Determine intellectual property rights from the outset and ensure they are documented.</td>
</tr>
<tr>
<td>✓</td>
<td>Identify any anticipated publication requirements (embargoes, restrictions on publishing over multiple sites).</td>
</tr>
<tr>
<td>✓</td>
<td>Identify and document specific roles and responsibilities as early as possible.</td>
</tr>
</tbody>
</table>
# Checklist for create and/or receive

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>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.</td>
</tr>
<tr>
<td>✓</td>
<td>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.</td>
</tr>
<tr>
<td>✓</td>
<td>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.</td>
</tr>
<tr>
<td>✓</td>
<td>Identify data quality metrics as soon as possible and ensure that these are communicated and monitored.</td>
</tr>
<tr>
<td>✓</td>
<td>Work together - researchers and information managers need to communicate regularly. Neither can do their job in isolation.</td>
</tr>
<tr>
<td>✓</td>
<td>Be realistic – strike a balance between what is sufficient and what is ideal based on your practical realities.</td>
</tr>
</tbody>
</table>
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
<table>
<thead>
<tr>
<th>Broad Function</th>
<th>Type</th>
<th>Specific Function</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Describes data and their location</td>
<td>Descriptive metadata</td>
<td>Allows data to be identified so they can be linked with requests</td>
<td>Name of the creator of the data set 名称数据集的创建者</td>
</tr>
<tr>
<td></td>
<td>Structural metadata</td>
<td>Describes how compound digital objects are organized</td>
<td></td>
</tr>
<tr>
<td>Provides the technical information needed to use data</td>
<td>Technical metadata</td>
<td>Provides the technical information needed to use data</td>
<td>Format 格式</td>
</tr>
<tr>
<td></td>
<td>Administrative metadata</td>
<td>Provides information about the use, management, and encoding processes of digital objects over a period of time</td>
<td>Information about data creation, subsequent updates, transformation, versioning, summarization 数据创建,后续更新,转换,版本化,总结信息</td>
</tr>
<tr>
<td></td>
<td>Preservation metadata</td>
<td>Records the preservation actions that have been applied to data over time</td>
<td>File format 文件格式</td>
</tr>
</tbody>
</table>

Ross Harvey (2010) *Digital Curation: A How-to-do-it Manual* Figure 6.1
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

<table>
<thead>
<tr>
<th>Files</th>
<th>Size</th>
<th>Format</th>
<th>View</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSLIS_Guide_2010.pdf</td>
<td>836.3Kb</td>
<td>PDF</td>
<td></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Field</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dc.contributor.author</td>
<td>Graduate School of Library and Information Science</td>
</tr>
<tr>
<td>dc.date.accessioned</td>
<td>2010-10-27T12:47:56Z</td>
</tr>
<tr>
<td>dc.date.available</td>
<td>2010-10-27T12:47:56Z</td>
</tr>
<tr>
<td>dc.date.issued</td>
<td>2010</td>
</tr>
<tr>
<td>dc.identifier.uri</td>
<td><a href="http://hdl.handle.net/10090/17896">http://hdl.handle.net/10090/17896</a></td>
</tr>
<tr>
<td>dc.description.provenance</td>
<td>Submitted by Stephanie Satalino (<a href="mailto:stephanie.satalino@simmons.edu">stephanie.satalino@simmons.edu</a>) on 2010-10-27T12:47:56Z No. of bitstreams: 1 GSLIS_Guide_2010.pdf: 836334 bytes, checksum: 9d2f8097bb105c3ad5ece00b3efa0973 (MD5)</td>
</tr>
<tr>
<td>dc.language.iso</td>
<td>en_US</td>
</tr>
<tr>
<td>dc.publisher</td>
<td>Simmons College</td>
</tr>
<tr>
<td>dc.title</td>
<td>The Guide For GSLIS Students</td>
</tr>
<tr>
<td>dc.type</td>
<td>Other</td>
</tr>
</tbody>
</table>
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?

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

<table>
<thead>
<tr>
<th></th>
<th>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).</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>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?</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Conversely, know what you need to dispose of. Destruction is often vital to ensure compliance with legal requirements.</td>
</tr>
<tr>
<td></td>
<td>Ensure that your data meets minimum quality assurance metrics (based on intended use).</td>
</tr>
<tr>
<td></td>
<td>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.</td>
</tr>
<tr>
<td></td>
<td>Work with researchers and information managers to develop policies and to identify realistic and implementable workflows.</td>
</tr>
<tr>
<td></td>
<td>Appraise for the here and now but with an eye to the future.</td>
</tr>
</tbody>
</table>
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)

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)