Digital Curation 101

PRESERVATION PLANNING

About *Preservation Planning*

Topics:

- Preservation planning
- Digital preservation and digital curation
- Aims of digital preservation
- Achieving the aims
    - Ensuring the integrity of the bit stream
    - Maintaining accessibility
- Planning for preservation
    - Data lifecycles
    - The main planning stages
- The next action in the curation lifecycle

Preservation planning

Preservation planning is a full lifecycle action in the data curation lifecycle. Specifically, it encompasses:

- Planning for preservation throughout the curation lifecycle of digital material
- Developing and applying plans for management and administration of all curation lifecycle actions.

This module introduces digital preservation and describes why it must be considered at all stages during the life-cycle of digital data, noting the need for planning for preservation throughout the curation lifecycle of digital material.

Digital preservation and digital curation

Digital preservation refers to activities aimed at ensuring that we can access data (in the form of digital objects and databases) in the future – for longer than the lifespan of the software and hardware.

Many definitions of digital preservation exist.

Digital Curation 101

The key points of these definitions are that digital preservation:

- Is a set of *managed activities*
- Aims at ensuring the *bit-stream* is maintained
- Aims at ensuring that data are *accessible*
- Is concerned with maintaining bit streams and ensuring accessibility for a *definable period of time*.

But digital preservation is not the same thing as digital curation. Think of it as a subset of digital curation. Digital preservation is a *necessary* part of curation, but by itself is not sufficient. Just preserving the data, for example by copying the bitstream onto new forms of data storage as old forms become obsolete, does not by itself ensure that we can use these data in the future. More is needed.

This is where digital curation comes into play. Its requires the active management and appraisal of data over the life-cycle of scholarly and scientific materials so that their integrity is protected and their value is enhanced with the aim of making them useful and usable in the future. To do this, we have to actively manage data over the whole of their life.

Aims of digital preservation

Digital preservation aims to produce data with three characteristics:

- *Longevity*: the data will be available for the period of time their current and future users (the *designated community*) requires. The lifespan of data is short unless action is taken. How long the data need to be maintained varies, but a minimum is for a period greater than the life expectancy of the access system (the hardware and software).

- *Integrity:* the data are authentic[1] – they not been manipulated, forged or substituted. Because digital preservation techniques such as migration inevitably alter the data, authenticity has to be demonstrated by paying attention to characteristics of the data such as provenance[2] and context[3].

- *Accessibility:* we can locate and use the data in the future in a way that is acceptable to its designated community. For example, an image (such as a pdf)

[1] http://www.archivists.org/glossary/term_details.asp?DefinitionKey=9
[2] http://www.archivists.org/glossary/term_details.asp?DefinitionKey=196
[3] http://www.archivists.org/glossary/term_details.asp?DefinitionKey=103

may be acceptable for some digital objects (such as documentation), but for other objects (a database, for example) the ability to manipulate or interrogate that object is required by its designated community in the future.

Achieving the aims

To meet the aims of ensuring the integrity of bit streams over time and maintaining access to them, we need to:

- record enough representation information
- manage intellectual property and other rights
- maintain the ability to locate the digital materials reliably
- monitor technology for changes that affect accessibility

The specific techniques commonly applied to meet these aims are listed next. (This section is based closely on the National Library of Australia's *Recommended Practices for Digital Preservation*[4].)

.

Ensuring the integrity of the bit stream

The main digital preservation practices applied to ensure that bit streams are authentic are:

- Copying data to a reliable digital storage system
- Managing ongoing data protection in accordance with good IT practices for data security, backups, error checking
- Refreshing (moving to a newer version of the same storage media, or to different storage media, with no changes to the bit stream), checking accuracy of the results (for example, checksums) and documenting the process
- Maintaining multiple copies of the bit stream
- Ensuring you have the right to copy and apply preservation processes, which may require negotiation with rights owners.

[4] http://www.nla.gov.au/preserve/digipres/digirespractices.html

Digital Curation 101

Maintaining accessibility

The key practices in current use are:

- Assigning persistent identifiers to the data to ensure they can be found
- Adding sufficient representation information to data (for example, information about file format, operating system, character encoding) so that the bit stream is still meaningful and understandable in the future
- Producing data in open, well-supported standard formats
- Limiting the range of preservation formats to be managed (often by normalising data to standard formats)
- Keeping track of developments (especially obsolescence) in hardware, software, file formats and standards that might have high impact on digital preservation
- Retaining and managing the original bit stream in case future developments mean we can restore access to it.

Planning for preservation

Data need to be managed from their point of creation until they are determined not to be useful. During this process, ensuring the data's long-term accessibility, authenticity and integrity is essential. The key is planning for active management of data over the whole of their lifecycle – providing 'constant maintenance and elaborate "lifesupport" systems' (NSF Workshop on Research Challenges in Digital Archiving and Long-Term Preservation 2003).

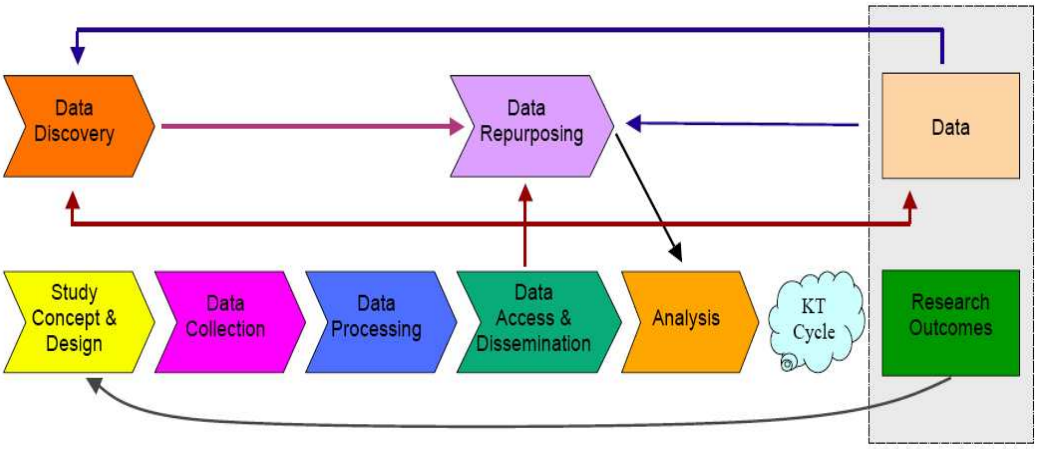
Planning covers:

- planning of the data (for example, how they are structured, what data about them –representation information, metadata – needs to be created, when, and by whom)
- planning of data storage and long-term management (for example, where is the data kept, who provides the continuing funding to maintain the data at the end of the research project).

Digital Curation 101

Data lifecycles

Data-driven research (e-Science) focuses on the data, rather than the results, with reuse of data as a key component.



The Life Cycle of Research

(Source: Chuck Humphrey. The Role of Academic Libraries in the Digital Data Universe[5], ARL Workshop on New Collaborative Relationships 26-27 September 2006)

Data curation requires that data can be searched, accessed, manipulated and mined during their lifetime. These require that the primary data are annotated with relevant representation information (metadata) about their provenance and content, and how the data were produced.

The main planning steps

Planning for each main step of the data lifecycle is essential:

- *Study concept and design* – plan to design data and representation information so that they can be preserved and reused
- *Data collection* – plan to collect data and representation information in ways that ensure their accuracy
- *Data processing* – plan to add representation information describing the data and recording their storage and manipulation (to provide the basis of validation of scholarly output, accountability and

[5] http://datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc

recordkeeping); plan to store data so that it retains its authenticity; plan to develop sustainable models and long-term institutional commitments to preserving data

- *Data access and dissemination* – plan to ensure that data and accessible to users and reusers, usually in the form of publicly available published information
- *Analysis* – plan to allow analysis of data, for example by ensuring they are consistent and cited according to relevant standards
- *Data discovery* – plan to make data discoverable, for example by adding representation information describing the data
- *Data repurposing* – plan to provide multiple access methods to data, to heighten their visibility and enable reuse (for example, by adding value to data sets so they can be reused or re-purposed, by adding metadata that assists in its discovery

The next action in the curation lifecycle

The next full lifecycle action in the curation lifecycle is *Description & Representation Information* which investigates description and representation information and indicates why they are essential to ensure that data curation is effective.