Digital Curation 101

CURATE AND PRESERVE

About *Curate and Preserve*

Topics:

- Why there's a need for digital curation
- What data?
- Create and preserve
- Why should you be interested in digital curation?
- What does digital curation involve?
- Who does digital curation?
    - Data creators and data users/reusers
    - Data curators
- Summary: main characteristics of data curation
- The next action in the curation lifecycle

Why there's a need for digital curation

Digital curation is necessary because:

- Immense quantities of data (any information in binary digital form) are being generated in all walks of life
- The quantities are increasing
- The scientific, scholarly and research communities increasingly rely on networked computing, as trends such as the move from *in vitro* to *in silico* science becoming dominant
- This places heavy responsibility on data, which are at risk from:
    - technology obsolescence
    - their fragility
    - lack of understanding about what constitutes good practice
    - insufficient resources
    - inappropriate organisational infrastructure.

Digital curation is a set of techniques that address these issues, emphasising the maintenance of data and adding value to these data for current and future use.

Digital Curation 101

What data?

First, some definitions.

Data is any information in binary digital form. It includes *Digital objects* and *Databases.*

Digital objects can be simple or complex.

- *Simple digital objects* are discrete digital items, such as textual files, images or sound files, along with their related identifiers and metadata
- *Complex digital objects* are discrete digital objects, made by combining a number of other digital objects, such as websites.

Databases are structured collections of records or data stored in a computer system.

Curate and preserve

Curate and Preserve is a full lifecycle action in the data curation lifecycle. Its activities comprise:

- Being aware of management and administrative actions planned to promote curation and preservation throughout the curation lifecycle
- Undertaking management and administrative actions planned to promote curation and preservation throughout the curation lifecycle.

Why should you be interested in digital curation?

Digital curation has become urgent for a range of reasons. Most of these will be familiar to you. Seamus Ross (Associate Director, DCC) vividly describes the reasons why digital objects and digital data become unusable:

> Digital objects break. Digital materials occur in a rich array of types and representations. They are bound to varying degrees to the specific application packages (or hardware) that were used to create or manage them. They are prone to corruption. They are easily misidentified. They are generally poorly described or annotated, that is they often have insufficient metadata attached to them to avoid their gradual susceptibility to syntactical and semantic glaucoma and where they do have sufficient ancillary data these data are frequently time constrained. **(**Seamus Ross, '*Digital Preservation, Archival Science and Methodological Foundations for Digital Libraries* ECDL2007).

Digital Curation 101

The following list of threats to digital continuity lists the most significant reasons why digital curation is an urgent imperative.

Threats to Digital Continuity[1]

- The carriers used to store digital materials are usually unstable and deteriorate within a few years or decades at most
- Use of digital materials depends on means of access that work in particular ways: often complex combinations of tools including hardware and software, which typically become obsolete within a few years and are replaced with new tools that work differently
- Materials may be lost in the event of disasters such as fire, flood, equipment failure, or virus or direct attack that disables stored data and operating systems
- Access barriers such as password protection, encryption, security devices, or hard-coded access paths may prevent ongoing access beyond the very limited circumstances for which they were designed
- Those taking responsibility for the material may not have adequate knowledge or facilities
- There may be insufficient resources available to sustain preservation action over the required period
- It may not be possible to negotiate legal permissions needed for preservation
- The digital materials may be well protected but so poorly identified and described that potential users cannot find them
- So much contextual information may be lost that the materials themselves are unintelligible or not trusted even when they can be accessed
- Critical aspects of functionality, such as formatting of documents or the rules by which databases operate, may not be recognised and may be discarded or damaged in preservation processing.

Probably the most commonly recognised of these threats is obsolescence. Our ability to maintain data and render them in a usable form over time is challenged by the wide range of formats, software and hardware, and the rapid speed at which they change.

Another threat is the increasing quantity of data produced in digital form. This is challenging our ability to maintain them and render them in a usable form over time. In addition, the increasingly dynamic nature of digital resources is producing major challenges.

Also part of this set of challenges are questions of selection:

1. What do we want to keep for the future?
2. Do we keep it all, or only some of it?
3. How do we decide what is likely to be useful?
4. How long should we plan to keep it?
5. Do we want it to be fully usable (e.g. all linked data is also available), and to what extent, in the future?

[1] Based on UNESCO *Guidelines for the Preservation of Digital Heritage*, 2003, p.32 http://unesdoc.unesco.org/images/0013/001300/130071e.pdf

Digital Curation 101

Responses to digital curation that merely apply strategies based on traditional approaches don't work. It is not simply a matter of capturing the data on stable storage media, or copying it onto new storage media when obsolescence threatens the existing media.

Data must be managed from the point that they are created (or – ideally – before they are created) if their survival is to be assured. Active management of data over the whole of their life is also necessary.

What does digital curation involve?

Digital curation encompasses a wide range of tasks and shared responsibilities. An overview is provided in the DCC Briefing Paper *Curating e-Science Data*, which lists some of the practical and technical tasks for scientists and research groups:

- Using Open Source Software and Open Standards to facilitate exchange and persistence of data through and across different systems
- Good annotation and creation of metadata to enable reuse of data
- Ensuring primary, secondary, and tertiary levels of research materials are linked and that links are persistent
- Using unique and persistent identifiers and a consistent citation format
- Identifying and selecting appropriate data for long-term curation and access.

Who does digital curation?

Different roles in the data curation process are played by:

- data creators
- data users/reusers
- data curators.

Data creators and data users/reusers

Scientists, scholars, and researchers are of necessity involved in some of the processes of data curation because they create data and use (and reuse) them.

For data to be usable and reusable, they must be of high quality, well structured, and adequately documented. The best time to ensure that data have these characteristics is

Digital Curation 101

when they are created. If these characteristics are present, then their use and reuse are significantly easier. If they are not present, use and reuse becomes exponentially more difficult, if not impossible.

Data creators, therefore, ensure that the data they bring into being is structured and documented to ensure its longevity and reusability. Data reusers ensure that any annotations they produce are captured and documented to a level that ensures they are understandable to other users of those data.

Data curators

People who have the primary role of managing or 'looking after' data come in a wide range of guises. Their job titles include:

- data curators
- archivists
- librarians
- data librarians
- annotators.

As an example, the tasks of a data curator in the biosciences context include ongoing data management, intensive data description, ensuring data quality, collaborative information infrastructure work, and metadata standards work.

The full range of tasks and responsibilities encompassed by data curation might look something like this:

- Developing and implementing policies and services
- Analysing digital content to determine what services can be provided from it
- Providing advice to data creators and users/reusers
- Ensuring submission of data to a repository
- Negotiating agreements
- Ensuring data quality
- Ensuring that data are structured in the best way to provide access, rendering, storage and maintenance
- Enabling the use and reuse of data
- Enabling data discovery and retrieval
- Preservation planning and implementation (for example, ensuring appropriate storage and backup routine, obsolescence monitoring)

Digital Curation 101

- Ensuring that policies and services are in place to make sure that data is viable, able to be rendered, understandable and authentic
- Promoting interoperability.

Summary: main characteristics of data curation

The main characteristics of digital curation can be summarised as:

- its concern with the range of processes applied to data *over its whole life cycle*, from creation to ultimate disposal; for instance, it places strong emphasis on the importance of designing for curation at the data creation stage
- its emphasis on reproducibility of data as the basis of validation of scholarly output, accountability, and recordkeeping
- its emphasis on adding value to data sets so they can be reused (or repurposed), for example by adding metadata that assists in its discovery, management and retrieval
- the involvement of a wide range of stakeholders cutting across disciplinary boundaries: these include heritage organisations (libraries, archives, museums, art galleries), e-science and e-research groups, scientific researchers, and government bodies who fund e-science, higher education and other activities
- a strong interest in open source solutions
- strong links between research and practice.

The next action in the curation lifecycle

The next full lifecycle action in the curation lifecycle is *Preservation Planning* which introduces digital preservation, describes why it must be considered at all stages during the lifecycle of digital data, and notes the need for planning for preservation throughout the curation lifecycle of digital material.