Digital Curation 101

CREATE AND/OR RECEIVE

Create and/or Receive is the second sequential action in the curation lifecycle, following *Conceptualise*.

Topics:

- Create or receive data
- Creating data for curation
- Structuring data for use and reuse
    - Open source
    - Significant properties and authenticity
- Structuring data for management
    - Data quality
- Structuring data for discoverability
    - Persistent identifiers
- Receiving data for curation
- The next stage in the curation lifecycle

Create or receive data

Create or Receive is the second sequential stage of the data curation lifecycle. Its activities are:

- Create data including administrative, descriptive, structural and technical metadata. Preservation metadata may also be added at the time of creation.
- Receive data, in accordance with documented collecting policies, from data creators, other archives, repositories or data centres, and if required assign appropriate metadata.

Scientists create data. These data should be created with curation principles in mind, so that they are accessible in the future and can be shared and reused. Archivists, data curators, librarians and others who receive data into an archive are keenly interested in ensuring that these data are curation-ready. *Create and Receive* explains the principles and practices of making data curation-ready.

Digital Curation 101

Creating data for curation

Scientific data is increasingly held in collaborative curated databases which may contain source experimental data, annotations, metadata, and data extracted from other curated databases. It is essential for the scientific record that these databases are preserved. Key ways of ensuring their preservation is to adhere to standards for data structure, data quality, citation, and annotation and provenance.

In the *Conceptualise* section, we asked the questions: What are the characteristics of 'good' data in terms of making effective curation more likely? What do we want data to be able to do, and what do we want to do with that data, in data curation terms? Three curation outcomes were identified:

1) Keep data, and the ability to process it – by creating data in standard data formats and file types that can be processed with open source, well-documented programs
2) Make ownership and allowable uses clear – by keeping documentation about the data, formats, software, agreements about its use, and so on
3) Make it citable – by applying standards for how data are referred to.

Structuring data for use and reuse

To keep data, and the ability to process it, the data needs to be:

- Authentic (it is what it claims to be)
- Accurate (it hasn't been tampered with)
- Renderable (it can be used in the ways for which it was intended, or viewed as originally intended)
- In a form that best ensures its longevity.

One way to achieve these is to use file formats that we stand a good chance of understanding in the future. These file formats are likely to be in widespread use, and very likely to be open source. Criteria we can use to predict the ongoing viability of a file format include:

- Openness – is there an open, publicly available specification for the format; are its specifications in the public domain; is it unencrypted?

Digital Curation 101

- Portability – is the format independent of hardware, operating system, of other software; is it independent of particular institutions, groups, or events; is it in widespread current use; does it contain little or no built-in functionality?
- Quality – is it robust, simple, highly tested, loss-free?[1]

Open source

All file formats and software adopted or developed for data curation should be open source. Open source file formats and software assists curation because curators have control over the software source code. Curation is made difficult or impossible if proprietary file formats are used, because software code is usually inaccessible for preservation purposes. Initiatives such as the Open Source Initiative[2] and SourceForge[3], a repository of open source software, illustrate the increasing popularity of the open source concept, which is a key aspect of data curation.

An example is the use of Joint Photographic Experts Group (JPEG) for images. Defined by an international standard (ISO 10918), it exists in several profiles, of which the lossless version of JPEG is preferred for preservation purposes, and JPEG 2000, *Part 1, Core Coding Version* with lossless compression is also preferred.

Significant properties and authenticity

To keep data and the ability to process and use them, we need to know precisely what it is we are trying to keep. To ensure that *authentic* digital objects or databases remain accessible and useable over time, we need to know which of their properties or characteristics we must maintain. These properties or characteristics are known as *significant properties*.

The InSPECT Project defines significant properties [4] as those aspects of the digital object or database, which *must* be preserved over time in order for the digital object to remain accessible and meaningful. The properties of digital objects and databases can be categorised as:

- Content (i.e., text, image, slides)

---

[1] Based on L. Clausen, *Handling File Formats* (2004), pp.11–12. http://netarchive.dk/publikationer/FileFormats-2004.pdf
[2] http://www.opensource.org/
[3] http://sourceforge.net/
[4] http://www.significantproperties.org.uk/

Digital Curation 101

- Context (i.e., who, when, why)
- Appearance (i.e., font and size, colour, layout)
- Behaviour (i.e., hypertext links, updating calculations, active links)
- Structure (i.e., embedded files, pagination, headings).

The concept of *authenticity* has become central to data curation. Authenticity is defined as 'the quality of being genuine, not a counterfeit, and free from tampering'[5] and is demonstrated from evidence such as the characteristics, structure, content, and context of a digital object or database. Authenticity is closely linked with significant properties – preserving a digital object's significant properties helps to ensure that it retains its authenticity over time.

Significant properties are defined by the requirements of the research community for whom the data are being preserved. Digital preservation usually involves some change to the data. Some change may be acceptable to some communities, but not to others. For example:

- For some categories of word-processed documents, the content (the text) may be the most significant characteristic, with other characteristics (layout or font size) not essential to its use in the future.
- For statistical datasets, the ability to manipulate the data is essential to future users and must be retained.

Structuring data for management

Data need to be managed to ensure that we have the ability to process, access, and reuse them over time. Some of the processes involved in managing data are data cleaning, production tracking, storage, and maintenance. Selecting viable file formats (noted above) is essential. Characteristics of file formats that are particularly helpful to the management of data include:

- Metadata support
- Interoperability
- Viability.

Metadata support. Description and representation information are essential for curation. Some software applications generate description and representation

[5] http://www.archivists.org/glossary/term_details.asp?DefinitionKey=9

information automatically, while other description and representation information has to be provided by data creators or data managers. Some file formats accommodate metadata. An example is metadata fields in a Tagged Image File Format (TIFF) file record details about the make and model of scanner, software and operating system, creator's name, and a description of the image.

Interoperability. Managing data over time will almost certainly require its migration from one technical environment to another. File formats that are platform-independent and/or are supported by wide range of software are easier to migrate.

Viability. File formats that have built-in mechanisms for error checking can assist data management by indicating when files have become corrupted. An example is the Portable Network Graphics (PNG) format, which allows PNG decoders to detect errors.

Data quality

Data cleaning is an important early step in the curation process. The best results when managing data over time and reusing data come when data are of high quality. In data cleaning incomplete, noisy, and inconsistent data are identified and actions applied to rectify these issues by filling in missing values, smoothing out noisy data, and detecting outliers.

An example from bio-informatics explains this:

> Data quality is an essential aspect of databases. It generally refers to the 'fitness' of the data in the databases. Data quality can be improved using data cleaning, the process of detecting and removing errors and discrepancies. Bioinformatics databases are usually cleaned manually, or with the use of proprietary programs … Manual curation of the data is commonly used in many biological databases to improve the quality of data originated from other public domain databases or directly submitted by individual researchers … using data analysis and visualisation tools, curators inspect and correct the data for consistency, accuracy, completeness, correctness, timeliness, relevance, and uniqueness (Koh, J L Y and Brusic, V (2005) 'Database Warehousing in Bioinformatics' in Chen, Y-P B (ed) *Bioinformatic Technologies* (Berlin: Springer) p.58)

Data quality is usually achieved by a combination of automated and manual processes, which calibrate data created by scientific instruments, validate, verify, and clean the data. For example:

- Post-measurement calibration of the instrumentation generating the data, to check characteristics of the measurements such as precision and bias
- Validation by checking for equipment errors and transcription errors

- Verification by checking the veracity of the data, for example by taking multiple samples.

Structuring data for discoverability

Data curation requires that data is discoverable: that is, they can be located. Standardised methods of identifying data are applied, such as a persistent identifier – a standardised method of identification that does not change, even if the locations of the data or digital objects change.

Standardised methods of identifying resources are essential for:

- *Citing data, databases and digital objects*. These are increasingly available online and often cite other online data, databases and digital objects. If the location of these changes the link is broken and the data, database or digital object cannot be located and referred to. This raises several questions: Does the data exist? If so, where are they? If data that looks similar is located, can we be sure they are the same?
- *Archiving of digital materials.* Reliable long-term access is based on persistent identification of data, databases and digital objects. If these cannot be reliably identified and located, they are effectively lost and hence any curation and preservation activities applied to them will have been in vain
- *Rights management and access management*. Automated transactions such as allowing access to only authorised users requires unambiguous identification of data, database or digital object for computer processing.

Persistent identifiers

Reliable identification of data is essential for providing long-term access to them and ensuring their reliability and authenticity. A persistent identifier – 'a name for a resource which will remain the same regardless of where the resource is located'[6] – provides this reliable identification. Applying persistent identifiers to data, databases and digital objects is an important aspect of the curation process. The section on *Description and Representation Information* provides more information about persistent identifiers.

[6] National Library of Australia *Managing Web Resources for Persistent Access* , 2002. http://www.nla.gov.au/guidelines/persistence.html

Digital Curation 101

Although much of the discussion about persistent identifiers has been about their application to web resources, their use is not limited to web material. It is equally essential for linking and citation of primary research to datasets.

Receiving data for curation

Archivists, data curators, librarians and others who receive data, databases and digital objects into an archive are keenly interested in ensuring that these are curation-ready. If they have been created using the criteria noted above, the curator's task of ensuring long-term access is more likely to be successful.

The data, databases and digital objects considered for ingesting (taking into an archive) may come directly from data creators such as scientists, or may come from other archives, repositories or data centres. Whatever their source, they will be considered for ingest according to documented policies for what that particular archive collects.

Ideally the material received is of high quality, created in curation-friendly open source formats using open source software, able to run on a variety of hardware platforms, and fully documented with administrative, descriptive, structural and technical metadata. If required, the curator will assign appropriate metadata to ensure that the data can be successfully maintained and accessed.

The next stage in the curation lifecycle

The next sequential action in the curation lifecycle is *Appraise and Select* which investigates the evaluation and selection of data for long-term curation and preservation.