DCC JISC National e-Science Centre

Digital Curation 101

APPRAISE AND SELECT

Appraise and Select is the third sequential stage in the curation lifecycle, following *Create and/or Receive*.

Topics:

- Appraise and select
- What data do we want to keep?
- How long do we want to keep that data?
- Appraisal and selection policies
- Reappraisal and disposal

Appraise and select

Appraise and Select is the third sequential action of the data curation lifecycle. Its activities are:

- Evaluate data and select for long-term curation and preservation
- Adhere to documented guidance, policies or legal requirements.

What data do we want to keep for the future? How do we decide what is likely to be useful? How long should we plan to keep them? Do we want them to be fully functional (for example, all linked data is also available), and to what extent, in the future?

Developing and applying policies for appraisal and selection can address these and other questions: deciding what data is worth keeping, why, and for how long.

What data do we want to keep?

The question of what data we want to keep is based on the assumption that it is not feasible to keep all of it: this is too expensive, and there isn’t sufficient organisational capacity to do so. The question has two dimensions:

1) Which datasets or digital resources do we want to keep?

2) Which characteristics or elements of those datasets or resources do we want to keep?

Digital Curation 101

Answers to these questions become increasingly necessary as the rate of data production continues to outstrip the rate at which resources become available for data curation.

To answer both questions requires knowledge of the designated community – the people who will understand it and use it in the future. It is not just the data itself that is selected, but also representation information – the information about the data that is needed to make it understandable in the future. It also requires thinking about what the designated community will consider sufficient in the future: for instance, will it be sufficient to keep just the information content of a database, but not the ability to search and manipulate its contents?

How long do we want to keep that data?

Another question is: how long do we want to keep the data? Answers to this question help us determine the resourcing we will need to effectively carry out digital curation. The answer is often phrased in different ways in terms of:

- changes of technology (for example, through several generations of hardware)
- the mission of the organisation curating it (for example, to meet specific business requirements)
- user requirements (for example, as evidence to verify conclusions derived from research).

Increasingly, the benefits and risks of keeping/not keeping are being recognised:

- What are the consequences of not keeping the data?
- How much would it cost to recreate it in the future?
- Is it even possible to recreate it in the future?

Appraisal and selection policies

Appraisal and selection policies are developed to assist in answering these questions and ensuring that the right data are kept for valid reasons. Policies are important because they allow informed consistent decisions to be made that can be defended if challenged.

There may be legal requirements for keeping data, or legal restrictions that mean data can’t be curated. For example:

- Copying data for preservation purposes (copying is the basis of the digital preservation strategies of refreshing, migration, and emulation) without specific approval from the copyright owner may not be covered by copyright legislation
- The data owner may not allow reuse of that data. Intellectual property rights for some material may be restrictive to the extent that there is no real possibility of access to data being made available in the future. In this case, it is probably pointless to expend resources on its curation.

Reappraisal

Revisiting appraisal decisions (reappraisal) may be required. An appraisal decision about data is not a decision made once and for all. As requirements and needs change, so too do appraisal decisions. The appraisal policy should include a statement of reappraisal principles and a reappraisal schedule.

For example, for research datasets:

- Initial appraisal could result in most datasets being kept
- Criteria for reappraisal are developed
- Reappraisal occurs at defined intervals to test the dataset against these agreed-upon criteria to decide whether it still meets the conditions for applying resources to its long-term retention.

The appraisal and reappraisal processes may result in a decision not to commit further resources to curation of specific data. In this case these data could be offered to another repository, or disposed of.

Disposal of data

Dispose is an occasional action in the data curation lifecycle. Its action is:

- Dispose (by transfer elsewhere, or destruction) of data which has not been selected for long-term curation and preservation in accordance with documented policies, guidance or legal requirements.

The decision to dispose of data is an outcome of the appraisal and selection process or the reappraisal process. Disposal refers to one of two actions:

- Transfer of data to another archive, repository, data centre or custodian
- Destruction of the data.

The decision to dispose of data is made by assessing the data against appraisal criteria. These are developed by archives, repositories and data centres, and used to determine if data are relevant to their aims and if resources should be committed to maintaining them in the long term. These appraisal criteria should be fully documented.

The decision to dispose of data may also result from reappraisal (testing the dataset against agreed-upon criteria to decide whether it still meets the conditions for applying resources to its long-term retention).

For some data there may be legal reasons why they must be destroyed in a secure manner. For example, many countries have legislation to protect the rights of individuals with regard to their personal data. For the U.K. this is the Data Protection Act 1998. If datasets contain personal data that has not been anonymised (meaning individuals cannot be identified from the data) then it is likely that they will need to be destroyed in a secure manner within a stated timeframe.

Transfer of the data

If data are determined not to be relevant to one archive, repository or data centre, they may be transferred to another that is interested in them. For example, a dataset may be transferred from its creator at the end of a research project, to a data archive whose mission is to maintain datasets for long-term use. Examples of such data archives include:

- The UK Data Archive[1]: contains data in the social sciences and humanities in the U.K.
- ICPSR[2] (the Inter-University Consortium for Political and Social Research) is the world's largest archive of digital social science data, based at the University of Michigan.
- The Life Sciences Data Archive[3] provides information and data from space flight experiments funded by the National Aeronautics and Space Administration (NASA).

[1] http://www.data-archive.ac.uk/

[2] http://www.icpsr.umich.edu/

[3] http://lsda.jsc.nasa.gov/

Digital Curation 101

In addition to the data, also transferred should be appropriate and adequate metadata and representation information and documentation about the data. These are essential to ensure that the data can be curated by the receiving organisation.

Destruction of the data

If data are to be destroyed, this should be carried out in a secure manner: that is, in such a way that the data cannot be reused or reconstructed. Erasing or deleting files is not sufficient to achieve this. Software tools that remove all data so that they cannot be reconstructed are widely available. For example, a ‘wiping’ program deletes the data and also overwrites it with random data several times. Standards that define approved methods include:

- DoD 5520.22-M Standard: Chapter 8 of the DOD 5220.22-M National Industrial
- Security Program Operating Manual (NISPOM)
- the Gutmann Method[4]: this is relevant for extremely sensitive data.

The next stage in the curation lifecycle

The next sequential action in the curation lifecycle is ***Ingest*** which investigates the transfer of data to an archive, repository, data centre or other custodian.

[4] http://www.usenix.org/publications/library/proceedings/sec96/full_papers/gutmann/