A Digital Curation Centre and Australian National Data Service 'working level' guide

How to Appraise & Select Research Data for Curation

Angus Whyte (DCC) and **Andrew Wilson** (ANDS)

How to Appraise & Select Research Data for Curation

Introduction

This guide will help you develop a managed approach to appraising and selecting datasets for curation. It provides working knowledge of current approaches, issues and challenges, and of the roles of research groups and institutional data services in addressing these. This guide should interest researchers responsible for managing data or who work in data-intensive fields, and those supporting them at research group level, or in institutional repositories, data centres or archives.

Why select and appraise

It is not possible for all digital data to be kept forever but outside the archive and library communities there is no widespread recognition of the need to select data for curation. Instead there is a view that "storage is cheap so why don't we just decide to keep everything". While that may in theory be technologically possible in practice there are four main objections to this view[1]:

1. Digital content expands. And "…if the growth of content (per byte or per object) keeps pace with the declining cost [of storage], then the real cost of keeping everything may actually be the same as it is now, or higher"[2].
2. Backup and mirroring increases costs. No digital preservation approach can survive without appropriate mirroring and backup systems. This instantly increases the storage cost by at least a factor of two.
3. Discovery gets harder. Keeping everything means that the noise to signal ratio of searches will be high, requiring additional individual effort to ascertain which data is the intended target of a search.
4. Managing and preserving is expensive. We must consider the cost of creating and managing preservation metadata, and the cost of preservation actions on data that does need to be retained.

The decision to be selective may raise a difficult question. Does the cost of selection outweigh the combined cost of creating and managing metadata, and undertaking preservation? Although no-one really knows the answer to this question there is some evidence that the answer is no, given the extremely large volumes involved and the absolute necessity to keep adequate metadata to ensure the data is findable, understandable and useable over time.

Beyond all this is the inescapable fact that long-term retention and curation of data requires a commitment to incur future costs; this necessarily imposes on any community a need for careful consideration of what should be retained. Expenditure on data curation will have to be justified and data creators and managers will not be able to escape the necessity of making selection decisions.

Appraisal Concepts

"Appraisal is the noblest function, the central core of contemporary archival practice." [3]

What archivists call 'appraisal' is often referred to outside the archival profession as 'selection' or 'acquisition', and is closely linked to a repository or institutional policy on collection development. Appraisal/selection is a primary activity whose importance cannot be overstated. Given that it is not feasible to retain everything, repository and data managers must be prepared to decide what will be retained. In the DCC Curation Lifecycle, the "appraise and select" activity requires data managers to "evaluate data and select for long-term curation and preservation". The Australian Records Management

[1] See "The Diverse and Exploding Digital Universe: An Updated Forecast of Worldwide Information Growth Through 2011", *IDC White Paper*, March 2008, which reports that the amount of digital data created in 2007 exceeded (for the first time) the amount of storage available (new and existing). IDC estimate that by 2011 half the digital information created will not be able to be stored.

[2] Paradigm Project: *Workbook on Digital Private Papers, Section 04: Appraising digital records: a worthwhile exercise?* Retrieved Feb 17 2010 from: http://www.paradigm.ac.uk/workbook/appraisal/digital-appraisal.html

[3] C. Couture, "Archival Appraisal: A Status Report", Archivaria 59, 2005, p. 107

Standard, AS 4390 (basis for ISO 15489), defined appraisal as evaluating records to determine which are to be retained, which are to be kept for specified periods, and which will be destroyed.[4] The UK National Archives (TNA) defines appraisal as: "the process of distinguishing records of continuing value from those of no further value so that the latter may be eliminated".[5] The similarities are obvious, and the intent is clear: appraisal is the process whereby some records are selected for retention, others (the great majority) are deemed of insufficient value to justify permanent retention. Although these references refer to records the principle of appraisal holds good for any other type of information that needs to be managed over time.

Selection is not an *ad hoc* process; it must be guided by local and community policies and legal requirements. The process used to make data selection decisions must be transparent and accountable. It cannot be based on individual views about possible future research needs. Research communities and institutions need to develop and agree on a set of objective criteria for assessing the long-term significance of research data sets. These must be widely disseminated and understood so that researchers and institutional information managers can make justifiable and rigorous decisions.

Appraisal is perhaps the most contentious and certainly one of the most difficult undertakings of the professional archivist. It determines what records will be preserved for posterity. In a very real way, archivists, certainly those working in jurisdictional archives, shape national narratives about the past. Appraisal is only a recent addition to the archivist's role. Before the mid-twentieth century, the archivist's role was traditionally just to accept and care for whatever records of the administration had survived.[6] In the mid-twentieth century, however, with the proliferation of records and the development of multiple technological processes for copying, more and more paper records lasted in organisations beyond their original administrative purpose. This gave rise to the need for archivists to make decisions about which records should be kept and become part of the archives' collections, and the consequent development of theories of and policies for appraisal applicable to all kinds of records including datasets.

Roles and Responsibilities

A data librarian or archivist will be mainly responsible for setting a selection and appraisal policy, developing criteria with input from other stakeholders. The communities producing and re-using the data need to be consulted, especially local data managers, as they are best placed to judge what makes the data valuable. Researchers will also benefit from knowing in advance how their own data will be assessed, and what they should plan to do in order to increase the chance of their research having an enduring impact.

Overall responsibility is shared between individual researchers and their organisations. These may provide guidelines to researchers as well as to any institutional data repository. In Australia, all universities are bound by the Australian Code for the Responsible Conduct of Research[7]. The Research Councils UK (RCUK) *Policy and Code of Conduct on the Governance of Good Research Conduct*[8] outlines general responsibilities:

- Make relevant primary data and research evidence accessible to others for reasonable periods after the completion of the research: data should normally be preserved and accessible for ten years, but for projects of clinical or major social, environmental or heritage importance, for 20 years or longer;
- Manage data according to the research funder's data policy and all relevant legislation;
- Wherever possible, deposit data permanently within a national collection.

[4] Adapted from Standards Australia, AS 4390-1996, Records Management, Part 1, Clause 4.3. AS 4390 has been superseded by the International Standard on Records Management, ISO 15489-2002, which, unfortunately, does not define appraisal.

[5] The National Archives, Appraisal Policy, August 2004. Retrieved Feb 19 2010 from: http://www.nationalarchives.gov.uk/documents/appraisal_policy.pdf

[6] Sue McKemmish, Barbara Reed, Michael Piggott, "The archives" in Sue McKemmish, Michael Piggott, Barbara Reed, Frank Upward (eds.), Archives: Recordkeeping in Society, Wagga Wagga, 2005, p. 174.

[7] Australian Code for the Responsible Conduct of Research. Retrieved Oct 4 2010 from: http://www.nhmrc.gov.au/publications/synopses/r39syn.htm RCUK Policy and Code of Conduct on the Governance of Good Research

[8] Conduct. Retrieved Oct 4 2010 from: http://www.rcuk.ac.uk/review/grc/default.htm

Relevant roles include the following.[9]

Researcher ('data creator')

- Provide enough information for others to assess the research data's scientific and scholarly quality and compliance with disciplinary or ethical norms.
- Provide relevant information for the repository to identify who will use the data and how i.e. the 'designated community', and any specific access requirements or constraints.
- Provide the research data in formats recommended by the data repository.
- Provide the metadata requested by the repository.

Data centre or repository

- Make explicit its mission in the area of digital archiving, and its selection policy for digital objects.
- Ensure compliance with legal regulations and contracts.
- Ensure the authenticity and integrity of the digital objects and the metadata.
- Assume responsibility from the data producer for ensuring the digital objects are accessible and available to a defined 'designated community'.
- Plan for long-term preservation of the digital assets.

Appraisal and Selection Policy

A policy needs to ensure consistent, transparent and accountable decision-making, so that commitments can be tracked and accounted for. The policy must fit legal requirements, e.g. relating to privacy and Intellectual Property Rights. It may also need to comply with relevant legislation for the jurisdiction, e.g. Public Records Acts, as well as national data policies and codes of conduct adopted by the host institution or funder, and any information governance policies relating to the discipline.

The policy will set out criteria for assessing a dataset or a resource's value, and what should be done with it accordingly i.e. how long it should be kept for, or when it can be destroyed. Criteria will vary depending, for example, on whether the remit includes preservation. In any case the policy will give the basis for further assessment of the datasets.
That will also be influenced by discipline-specific factors and based around general criteria such as the seven listed below, which are drawn from various sources (footnotes [10,11,12])

1. ***Relevance to Mission:*** The resource content fits the centre's remit and any priorities stated in the research institution or funding body's current strategy, including any legal requirement to retain the data beyond its immediate use.
2. ***Scientific or Historical Value:*** Is the data scientifically, socially, or culturally significant? Assessing this involves inferring anticipated future use, from evidence of current research and educational value.
3. ***Uniqueness:*** The extent to which the resource is the only or most complete source of the information that can be derived from it, and whether it is at risk of loss if not accepted, or may be preserved elsewhere.
4. ***Potential for Redistribution:*** The reliability, integrity, and usability of the data files may be determined; these are received in formats that meet designated technical criteria; and Intellectual Property[13] or human subjects issues are addressed.
5. ***Non-Replicability:*** It would not be feasible to replicate the data/resource or doing so would not be financially viable.
6. ***Economic Case:*** Costs may be estimated for managing and preserving the resource, and are justifiable when assessed against evidence of potential future benefits; funding has been secured where appropriate.
7. ***Full Documentation:*** the information necessary to facilitate future discovery, access, and reuse is comprehensive and correct; including metadata on the resource's provenance and the context of its creation and use.

[9] Adapted from the international Data Seal of Approval Guidelines. Retrieved Oct 4 2010 from: http://www.datasealofapproval.org/

[10] NASA Socioeconomic Data and Applications Center (SEDAC) Long-Term Archive. (n.d.). Appraisal for Accesion to the SEDAC LTA. Retrieved June 24, 2010, from http://sedac.ciesin.columbia.edu/lta/Appraisal.html

[11] Faundeen, J. (2010). Appraising U.S. Geological Survey Science Records. Archival Issues, 32(1), 7 -22.

[12] NARA – U.S. National Archives and Records Administration. (2007). Strategic Directions: Appraisal Policy. Retrieved June 24, 2010, from http://www.archives.gov/records-mgmt/initiatives/appraisal.html

[13] Copying data for preservation purposes without specific approval from the copyright owner may not be covered by copyright legislation; or the 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.

Developing the Appraisal Process

The appraisal process will apply the criteria set out in the policy. This process must be transparent and accountable to justify selection decisions to current and future users, so it must follow clear, unambiguous, and objective criteria. It is worth noting that appraisal typically results in only a very small number of records being retained permanently. Most records are destroyed at some point after they have ceased to be of immediate use. In Australia, the National Archives has historically kept around 7-8% of the total federal government output of records. Quantities in the US and Europe are generally much smaller, for example only some 4-5% of UK government records are kept permanently.

High data volumes rule out appraisal at the record level, and may even do so at the data set level. The appraisal/selection process should be undertaken at as high a level of data aggregation as will ensure justifiable outcomes and allow cost effective decision making. The challenge is to identify a set of high-level criteria that can be applied using evidence it is practical to obtain, yet is sensitive to possibly wide variations at the individual item level.

The data repository or archive's management should preferably take selection decisions, as this is the appropriate level of responsibility. The views of discipline specialists will often be essential, especially the research team who created and used the data. All decisions must be recorded, with justifications, so that future researchers can understand why particular data sets were kept or destroyed. Decision records are metadata, to be held in whatever archive/asset/digital object management system is used to manage and control the data, and these metadata must be retained permanently.

Detailed workflows will depend on how the criteria are ordered in terms of priority, and the dependencies between the sources of information drawn on to make decisions. Below we take the general criteria listed in the Policy section and consider some more detailed questions and evidence relevant to appraisal on those criteria.

Relevance to Mission

Does the dataset or resource fall within the repository's scope?

- Refer to the remit or mandate set by the host institution or other funders, their broader data policies, and codes of conduct for research e.g. for retention periods.
- Consult the strategic priorities of the host institution or other funders.

Are there other relevant legal requirements or guidelines?

- These may for example include legislation relating to Public Records, Copyright and Patents, Freedom of Information, Data Protection, Health and Safety, and Equality.
- Also check any discipline-specific information governance guidelines and codes of practice, e.g. professional associations' ethical codes.

Scientific or Historical Value

Does the dataset reflect the interests of contemporary society?

- Consider how the research questions relate to trends in research awards by national funding bodies, and assign a value (rating) based e.g. on the number of projects funded or the amount provided for the relevant research topic.

Is there authoritative evidence of current value to the research field?

- This may be available from citations to publications the data has been used in, or other authoritative sources such as research assessments, indicating whether the data should be retained as part of the research record, considering the findings based on them.

Uniqueness

Is the dataset the only source of its content and will it be preserved elsewhere?

- Check whether the dataset(s) duplicates existing work, is new or unique.
- Try to find out if other copies of the data exist and are accessible and useable. If other copies exist, where is the most comprehensive or up-to-date version?
- Are any other copies at risk of loss, or will they be preserved where they are?

Potential for Redistribution

Are Intellectual Property Rights (IPR) issues addressed?

- Check the institution's policy on IPR and sharing, access to and re-use of data.
- Check whether the funder or project consortium have IPR policies affecting the work, and whether these have been adhered to.
- Identify any contractual or licence terms affecting the dataset, e.g. has the copyright owner given permission for archiving?

- If a Creative Commons or similar 'copyleft' licence has been used, with what conditions, or has a public domain waiver been given?
- Are database rights applicable and if so have these been obtained?

Are human subjects issues addressed?

- Was informed consent obtained from the research subjects for archiving and re-use of personal data, on what terms, and is it feasible for the archive to adhere to them? E.g. can the data be effectively anonymised, and any keys curated?
- Was approval by an Ethics Committee required to collect the data and if so is there evidence of this?
- Are there any other restrictions on sharing, access and re-use if the research involved human subjects (e.g. sensitive health or political data)?

What is the reliability and usability of the dataset?

- Is the dataset in a format that allows others to use it without costs or other restrictions?
- Is software available to access, view and query the data, and if so will any costs or terms apply to users?
- Is there enough metadata and documentation for the dataset to be readily used and understood away from its original context of creation?

Has the data been stored in a way that ensures its integrity has not been compromised?

- Whoever has kept and stored the dataset needs to ensure that the data cannot be tampered with or inadvertently changed.
- Backups must have been kept safely to ensure corrupted data can be replaced.

Does the dataset meet technical criteria that allow its easy redistribution?

- Has the data been created, or kept, in an open, machine-independent or easily accessible format?
- Can the data be easily migrated to other formats that might be more accessible to external users?

Non-Replicability

Can the data be easily replicated, recreated or re-measured?

- Are the data records transient or one-off events that cannot be repeated, such as weather observations, volcanic eruptions or rainfall records?
- Is the event/project which caused the data to be created easily reproducible?

Is the cost of replicating or re-measuring the data financially viable?

- Would another body be prepared to fund the future reproduction of the data?
- Even if the data can be recreated or re-measured it may be so expensive to do so that it is preferable to retain the original.

Economic Case

Has the total cost of retaining the data been considered?

- Keeping data for long periods involves more than storage. Data must be kept accessible, backups kept, and sharing and access implemented. All this adds to the cost of keeping data. The total cost must be considered and estimated to check whether it is financially viable to keep the data.
- The JISC *Keeping Research Data Safe* (KRDS) Phase 2 Project produced a cost model for digital preservation[14] including 'acquisition' costs. The British Library's LIFE (Life Cycle Information for E-Literature) projects[15] have also developed a lifecycle model and a predictive costing tool that can help to determine costs.

If the cost is acceptable who is going to pay for data retention?

- Even if the cost of keeping the data is acceptable, how the data retention will be funded must be considered. Without this selection cannot be viable.
- Has funding been provided, promised or assured?

Full Documentation

Is there documentation to support sharing, access and re-use of the data?

- Datasets need some way of understanding their structure and the meaning of field names, etc. so anyone not directly involved in creating the data will be able to re-use it. Are there data dictionaries explaining the layout and structure?
- Is there comprehensive information about the context of data creation: the nature of the project; the data collection methodology, post-collection manipulation?
- Have records been kept of any access, copyright/IPR, privacy or ethical restrictions on access and re-use?

[14] KRDS Project factsheet. Retrieved Aug 29 2010, from: http://www.beagrie.com/KRDS_Factsheet_0910.pdf

[15] LIFE Projects. Retrieved Aug 29 2010, from: http://www.life.ac.uk/

New challenges and opportunities

Sources of these are likely to include the need to automate workflows, so as to more cost-effectively manage larger quantities of data and related research material. Another challenge is involving the wider research community in appraisal.

Research funding bodies and institutions are placing increasing demands on researchers to make their research process and results more open and transparent, and to maximise the return on their funding by retaining data for possible re-use, sometimes for many years beyond the research grant. To that end, just as Institutional Repositories are becoming an established means to provide open access to publications, many are called on to keep related datasets, code, and records of the research process safe and re-usable.

Alongside such 'top-down' pressures to curate research data, many fields are thriving around the new capabilities for data-driven analysis, and new tools and approaches to data discovery and management, e.g. linked data technologies. These increase the need for data to be machine-readable, both now and as these technologies change. As broader inter-disciplinary and cross-institution collaborations become more common, so do the needs for shared data facilities that operate to an agreed policy. Economic factors, the changing nature of data and metadata, and the range of actors involved are each each likely to present new challenges and opportunities. We summarise some of these below.

Re-modelling data management workflows

The need to reduce data management costs is driving greater automation of workflows especially in data-intensive research fields. This presents opportunities for automation across the boundary between the originating research group and the data repository. The challenge here is for effective collaboration between the various actors involved; data scientists producing smarter instruments, models and simulations, researchers who add value to their output, and the data curators and research support staff who keep the infrastructure going and make the results accessible and re-usable by others.

More metadata will be generated automatically through tools embedded in everyday research activity. In many fields the research material itself may be marked-up using automated classification techniques, or may be 'linked data' that has been integrated from disparate web sources. These new sources and technologies provide opportunities for data centres to improve data discovery. They also highlight challenges that include:

- Understanding to what extent data has already been 'selected' through pre-processing, sampling or quality control, on what basis, and whether the available metadata allow the process to be followed.
- Documenting the provenance trail for datasets resulting from large-scale collaborations or derived from linked data sources, and dealing with ownership and intellectual property rights issues.
- Assessing the rapidity of technical change and how it affects the case for preservation, e.g. the costs and benefits of migrating to new standards.

Engaging the community in appraisal

There may also be a role for the wider research community in the appraisal process, driven by the need to assess the research value of ever-expanding volumes of data more cost-effectively. According to Fran Berman the "need for community appraisal will push academic disciplines beyond individual stewardship, where project leaders decide which data is valuable, which should be preserved, and how long it should be preserved" [16]. Community curation is an emerging area, and there is widespread expectation that 'community' ratings or comments will become established as a means for peer review of datasets. Open access publisher Public Library of Science (PLoS) has been a leading advocate of 'article-level' metrics, such as usage statistics and reader comments, as a means for readers to assess article quality[17] . Article-level metrics may become accepted as another resource for assessing the value of data linked to publications.

Citations of datasets themselves will likely be a more direct indicator of their value, and may become better established as data citation standards develop, alongside new ways to attribute and reward contributions to datasets. In the biocuration field, for example, research output may not be measurable by links to publications. There have been calls for 'microattribution' systems to enable curators to "unequivocally show reviewers how useful their data content is to the community, by way of accurate citation metrics for datasets" [18]. For a repository or data centre deciding to acquire a database or its contents, dataset-level usage and citation metrics may become an indicator of their current usage, provided of course they are valid measures for the research community concerned. These metrics may also be relevant for re-appraising datasets the archive is already making available.

---

[16] Berman, F. (2008). Got data?: a guide to data preservation in the information age. Communications of the ACM, 51(12), 50–56.

[17] PLoS (2009) Article-level Metrics. Retrieved Aug 27, 2010 http://article-level-metrics.plos.org/.

[18] Gen2Phen (2009) Incentives/rewards for scientific contribution. Retrieved Aug 27, 2010 http://www.gen2phen.org/researcher-identification-primer/incentivesrewards-scientific-contributions

Further Information and Bibliography

Two other DCC guides by Ross Harvey cover this topic:
Awareness *Level: Introduction to Curation: Appraisal and Selection* **(2008)**

Expert Level: *Curation Reference Manual: Appraisal and Selection chapter* **(2006)**

Dallas, C. (n.d.). An agency-oriented approach to digital curation theory and practice. In Proceedings: International Symposium on "Information and Communication Technologies in Cultural Heritage" (p. 49).

Digital Preservation Coalition. (n.d.). Decision Tree for Selection of Digital Materials for Long-term Retention. Retrieved June 24, 2010, from http://www.dpconline.org/advice/decision-tree.html

Downs, R. R., & Chen, R. S. (2009). Designing Submission Services for a Trustworthy Digital Repository of Interdisciplinary Scientific Data. In Earth and Space Science Informatics Workshop: Developing the Next Generation of Earth and Space Science Informatics: Technologies and the People That Will Implement Them. August (pp. 3–5).

Esanu, J., Davidson, J., Ross, S., & Anderson, W. (2004). Selection, Appraisal, and Retention of Digital Scientific Data: Highlights of an ERPANET/CODATA Workshop. Data Science Journal, 3, 226.

Faundeen, J. (2010). Appraising U.S. Geological Survey Science Records. Archival Issues, 32(1), 7 -22.

Gray, J., Szalay, A., Thakar, A., Stoughton, C., vandenBerg, J. (2002). *"Online Scientific Data Curation, Publication, and Archiving"*, Microsoft Research Technical Report MSR-TR-2002-74.

Gutmann, M., Sch\ürer, K., Donakowski, D., & Beedham, H. (2004). The selection, appraisal, and retention of social science data. Data Science Journal, 3(0), 209–221.

NARA - US National Archives and Records Administration. (n.d.). Strategic Directions: Appraisal Policy. Retrieved June 24, 2010, from http://www.archives.gov/records-mgmt/initiatives/appraisal.html

NASA Socioeconomic Data and Applications Center (SEDAC) Long-Term Archive. (n.d.). Appraisal for Accesion to the SEDAC LTA. Retrieved June 24, 2010, from http://sedac.ciesin.columbia.edu/lta/Appraisal.html

Norris, R., Andernach, H., Eichhorn, G., Genova, F., Griffin, E., Hanisch, R., Kembhavi, A., et al. (2006). Astronomical Data Management. Arxiv preprint astro-ph/0611012.

Pearce-Moses, R. (n.d.). SAA: Glossary of Archival Terminology. Retrieved July 6, 2010, from http://www.archivists.org/glossary/

Schade, D. (2009). Data Centre Operations in the Virtual Observatory Age. In Proceedings PV2009. Presented at the Ensuring Long-Term Preservation and Adding Value to Scientific and Technical Data, Madrid, Spain.

Wallis, J. C., Borgman, C. L., Mayernik, M. S., & Pepe, A. (2008). Moving archival practices upstream: An exploration of the life cycle of ecological sensing data in collaborative field research. International Journal of Digital Curation, 3(1).

Witt, M. (2008). Institutional Repositories and Research Data Curation in a Distributed Environment. Library Trends, 57(2), 191–201.

Yakel, E. (2007). Digital curation. Perspectives, 23(4), 335–340.

Thank you to Margaret Henty of ANDS and Mark Thorley of NERC for helpful comments.