ANDS Guide

Curation continuum

Level: Awareness

Last updated: 3 February 2016

Web link: http://www.ands.org.au/guides/curation-continuum

This guide introduces the concepts of the data curation continuum and curation boundary, and describes how they impact on the management of research data for discovery and reuse. It is likely to be of interest to researchers who create data, to research data stewards and to data management professionals.

Scholarly communications and curation

The scholarly communications cycle includes all the activities associated with research, including the collection and analysis of data as well as its dissemination and reuse. An increasing range of things are being considered to be first class research outputs: publications, data, models, workflows and software. For the purposes of this guide, unless otherwise noted data is assumed to cover all of these except publications.

The management and preservation of publications are well catered for by established infrastructure such as libraries, archives, electronic print repositories, and academic and commercial publishers. With traditional publication, most curation activities occur at the end of the research cycle. In contrast, digital curation of data and other non-publication outputs is more demanding, with activities needing to occur throughout the scholarly communications and research process.

Digital curation is defined by the UK's Digital Curation Centre as "The activity of managing the use of data from its point of creation to ensure it is available for discovery and reuse in the future". This management process includes storage and security of the data, its quality management, recording important information about the data, including its source, analysis methods and changes to the data, and preserving the data so that it can be accessed and reused in the future. The same considerations apply to other non-publication outputs.

Data curation continuum

The data curation continuum begins in the private domain, with the capture or creation of research data by a researcher (see Figure 1). There may be a large number of data objects which are updated frequently. At this stage, researchers typically manage their own data, possibly within small teams. Preservation and metadata may not be needed, and access to the data is limited.

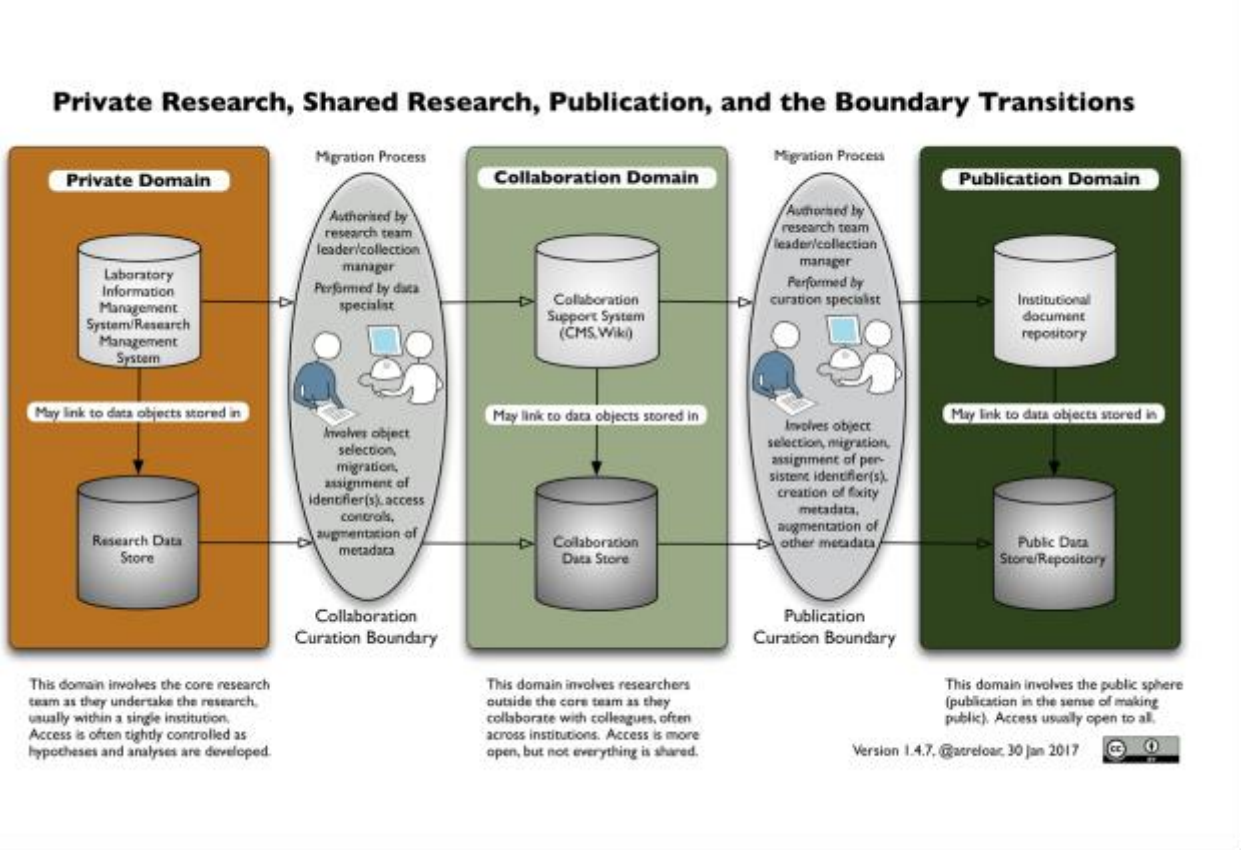

Figure 1: Domains, data stores and curation boundaries

At the other end of the continuum is the public domain. There are likely to be a smaller number of selected static research outputs which have accrued more metadata, and which may be managed and preserved through institutional arrangements such as repositories. These outputs are more likely to be publicly accessible, possibly linked to publications.
In the middle is the shared research domain. Here researchers collaborate on a subset of research outputs, and the need to capture context in metadata is less than in the public domain.
Figure 1 shows the data store layer as three separate stores, one for each domain. This is a conceptual way of viewing the data store - in practice, some combination of the domains could be serviced by a single data store with appropriate access controls.

Curation boundaries

Each research project has a unique pattern of data curation dimensions and transition along those dimensions. The key transitions occur when data move from the private domain into a collaborative environment, and when data move into the public domain with publication.
These transitions are referred to as curation boundaries. They are points in time when curation decisions need to be made and when responsibilities for data management may transfer to others.

Collaboration curation boundary

At the collaboration curation boundary, data move from the private research domain into a shared research domain. The researcher needs to decide what data to be shared, who should be able to access the data, and what additional information (metadata) people will need to collaborate in the project. At this stage the context of the

data is clear, and collaboration is likely to occur mainly with same discipline colleagues, although cross-institutional collaboration is common.

Publication curation boundary

At the publication curation boundary, the data move into the public domain. The researcher again has decisions to make about what data should be shared and who should be able to access it. However, additional constraints come into play at the publication point, and other decision makers may be involved.

There may be legal issues, such as the need to de-identify data to protect privacy, or to protect access to data in accordance with any agreements, contracts or institutional policies. Publication implies continuing availability, so arrangements for storage and preservation need to be made.

Publication also implies a need for the data to be discoverable. As a result, persistent identifiers may need to be assigned to the data, and more complete metadata created. This metadata will need to include descriptions of the data, how it was created and manipulated, access and reuse policies, and copyright statements. Since publication also exposes the data beyond the creating discipline, plain language descriptions of the data may be required to facilitate reuse of the data.

Implications

ANDS promotes data sharing and reuse, both throughout the research process and for future researchers. Decisions made by researchers and others at the curation boundaries strongly influence both the possibility and the extent of this future data sharing and reuse. Particularly critical are choices made about the accessibility of the data and the provision of metadata to provide context to the data and allow its discovery.

Further reading

- Digital Curation Centre Curation lifecycle model
- Treloar, A. and Harboe-Ree, C. Data management and the curation continuum: how the Monash experience is informing repository relationships Proceedings, VALA Conference 2008. NOTE: A publication describing an updated version of the curation continuum is in preparation and should appear in 2017.

Feedback?

We welcome your feedback on this guide. Please email contact@ands.org.au with any comments or questions.

About ANDS

The Australian National Data Service (ANDS) makes Australia's research data assets more valuable for researchers, research institutions and the nation.

ANDS is a partnership led by Monash University in collaboration with the Australian National University (ANU) and the Commonwealth Scientific and Industrial Research Organisation (CSIRO). It is funded by the Australian Government through the National Collaborative Research Infrastructure Strategy (NCRIS).