

Data 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 re-use. 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

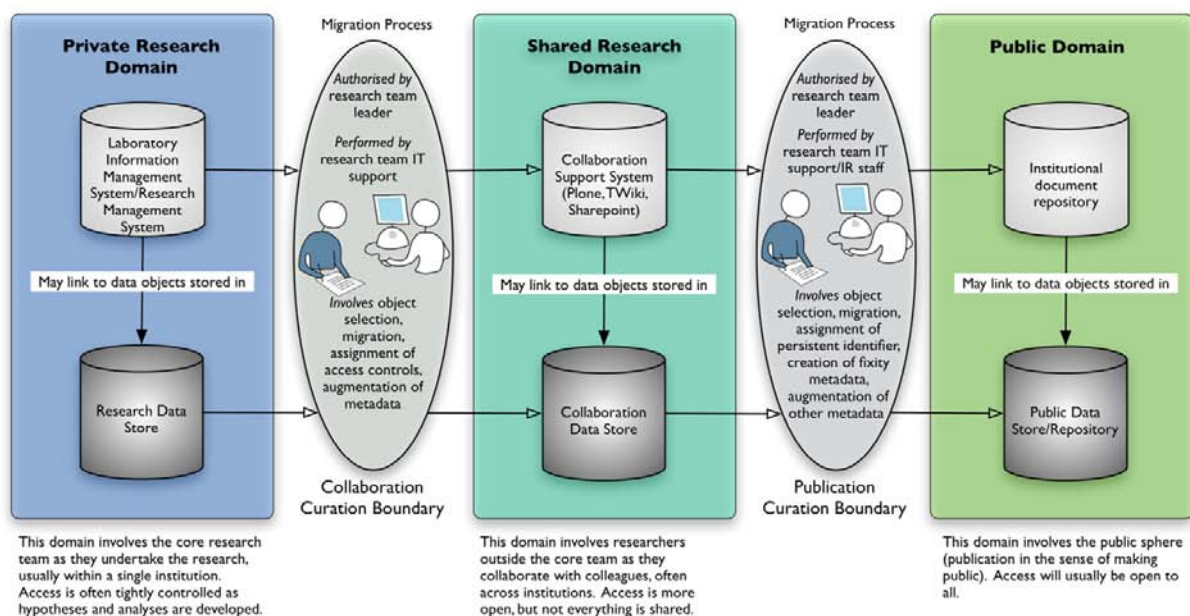
The scholarly communications cycle contains all the activities associated with research, including the collection and analysis of data as well as its dissemination and re-use. Both scholarly publications and data are first class research outputs. The management and preservation of publications are well 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 is more demanding, with activities occurring 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 re-use 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 re-used in the future.

Data Curation Continuum

The data curation continuum begins in the private domain, with the 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. Preservation and metadata may not be needed, and access to the data is limited.



Version 1.4, <http://andrew.treloar.net/>, 07Dec07

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 data objects which have accrued more metadata, and which may be managed and preserved through institutional arrangements such as repositories. This data is more likely to be publicly accessible, possibly in association with print publications.

Curation boundaries

Each research project has a unique pattern of data curation dimensions and transition along those dimensions. The key transitions occur when data moves from the private domain into a collaborative environment, and when data moves 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 moves 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 moves 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 re-use 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 re-use of the data.

Implications

The Australian National Data Service (ANDS) promotes data sharing and re-use 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 re-use. 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

ANDS Guides and other Resources: www.ands.org.au/guides

Digital Curation Centre "Curation lifecycle model" <http://www.dcc.ac.uk/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.

http://www.valaconf.org.au/vala2008/papers2008/111_Treloar_Final.pdf

