Updating the Data Curation Continuum: Not Just Data, Still Focussed on Curation, More Domain-Oriented

Andrew Treloar
Australian Research Data Commons

Jens Klump
CSIRO Mineral Resources

Abstract

The Data Curation Continuum was developed as a way of thinking about data repository infrastructure. Since its original development over a decade ago, a number of things have changed in the data infrastructure domain. This paper revisits the thinking behind the original data curation continuum and updates it to respond to changes in research objects, storage models, and the repository landscape in general.

The Data Curation Continuum

Inception

In 2007 and 2008, Treloar (together with Harboe-Ree and Groenewegen) published a pair of papers describing a way of thinking about data repository infrastructure called the Data Curation Continuum (Treloar, Groenewegen and Harboe-Ree, 2007; Treloar and Harboe-Ree, 2008). This model arose from work done at the time to develop the Monash University Information Management Strategy. This approach was informed by a body of theoretical work developed by researchers in the School of Information Management and Systems at Monash. They developed the notion of an information continuum, based on a multiple-axis analysis of the various characteristics of information in organisations (Schauder, Stillman and Johanson, 2004). These information management dimensions were largely determined to have particular values.

The Role of Continua

An analysis of the research data management space suggested that it was not appropriate to identify specific values along each dimension. Instead, it was decided to have continua that graduated between two endpoints. This analysis was based on user requirements from within Monash University, a literature review, the use-case work undertaken in the DART project (Tsoi, McDonell and Treloar, 2007), and the results of the work undertaken to clarify the role of the Australian National Data Service (DEST, 2007). The reason for calling these curation *continua* was that they all deal with things that the curation domain needs to address: object properties, management decisions and access constraints. The term 'curation' in this paper is used in accordance with the definition used by the Digital Curation Centre. Table 1 summarises the Data Curation Continua identified at the time of the creation of the Data Curation Continuum model.

Table 1. Data Curation Continua.

Object:	Less Metadata	←→	More Metadata
	More Items	←→	Fewer Items
	Larger Objects	←→	Smaller Objects
	Objects Continually Updated	←→	Objects Static
Management:	Researcher Manages	←→	Organisation Manages
	Less Preservation	←→	More Preservation
Access:	Closed Access	←→	Open Access
	Less Exposure	←→	More Exposure

Domains and their Repositories

One way of using these continua is to make a series of choices about where to place a dividing line on each continuum. The sum of these choices serves as a way of defining three different domains within which data stores/repositories might be used. Note that this tripartite division is not the only possible arrangement. Both finer and coarser ways of dividing the space are possible and may be appropriate for particular institutional settings.

The first domain is the private research domain. This is where the immediate research team is working with its data and producing its results. The team may use a Laboratory Information Management System (LIMS) or other research management system (or even something as lightweight as an Excel spreadsheet) to keep track of its data files. The files themselves will live in a research data store. This might be as simple as a file system or something more sophisticated like Fedora/dSpace or an iRODS instance. In terms of the data continua, this domain is characterised by having less metadata, more items, larger objects that are often continually updated, researcher management of the items, less preservation, mostly closed access and less exposure.

The second domain is the shared research or collaboration domain. Here the research team is prepared to open up a subset of its research results to other researchers to access and analyse. Depending on the nature of the collaboration and the size of the data, the data originators may allow remote collaborators to run data analysis jobs using compute cycles located with the data store. Because of the need to structure the collaborative interaction, a collaboration support system (Drupal or one of a number of Virtual Research Environments) can be useful. It allows for blogging, collaborative document editing and content management for non-data objects. The data objects now need to be in a repository that supports greater structuring of the data collections, as well as more sophisticated access controls. Compared to the private domain, this domain is characterised by having more metadata, fewer items, smaller objects that are usually static or derived snapshots (rather than actively updated data), researcher management, possibly more preservation, and less restricted (but not open) access.

The third domain is the publication domain. At this point the research is 'finished' in the sense that the resulting publications (and possibly linked data objects) are available for public viewing. The documents will probably be made available through a traditional (if one can use the term for something that has probably been in existence less than five years) institutional repository. The associated data objects will need to be lodged in a public data repository. This may or may not be the same system as the institutional repository. In terms of the curation continua, the publication domain is characterised by having more metadata than the collaboration domain, fewer items again, smaller objects that are almost certainly static or derived snapshots, organisational management, more preservation, open access and exposure of metadata for harvesting.

The most recently updated version of the original Data Curation Continuum is shown as Figure 1. It depicts three distinct domains within which people work with data: the private, collaboration and publication domains. Each domain shows a data location and a location for associated publications. As data moves across the boundaries between the three domains, it undergoes a migration process involving a mixture of manual and machine processes.

There are two main changes as one reads from left to right.

Firstly, there is a quantitative change in object numbers. This is a consequence of a progressive selection decision. The owner(s) of the private domain select a subset of all available objects to move to the collaboration domain, and from the collaboration

domain a further selection for publication. While we are unaware of any published data on the reduction steps, a reasonable estimate might be an order of magnitude drop for each transition.

Secondly, there is a qualitative and quantitative change in the metadata. This is a consequence of the need to take the implicit context for the objects and make it explicit via associated metadata. For instance, a laboratory might have particular protocols for encoding experiment numbers or sample details in the names of files. These would need to be made explicit in order to enable a different laboratory to collaborate over these data. Similarly, the assumptions about particular kinds of data that might be made within a discipline would need to be made explicit if the data is to be used by another discipline.

Visualising this Model

Figure 1 shows the result of this way of viewing the space (as of its original incarnation).
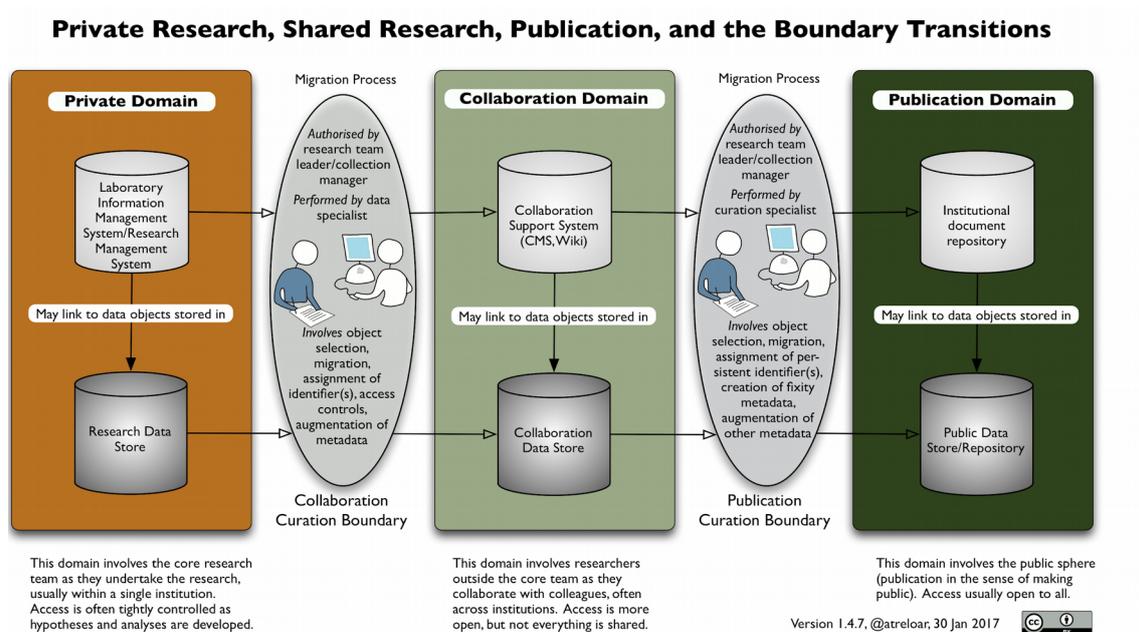

Figure 1. Data Curation Continuum (latest version in this form, prior to rework).

Boundary Transitions

In the three domain model, there are two transitions that need to be negotiated for stored content: from the private to the shared domain (the collaboration curation boundary), and from the shared to the publication domain (the publication curation boundary). The process of ongoing curation in the public domain relies on provenance metadata that should have been captured during the research process. However, the ongoing work of active curation will largely take place on the publication side of the boundary. Researchers are not, in general, focussed on curating their data. This is a task more suited to the professionals who will take responsibility for the data in the publication domain.

There needs, therefore, to be a process to migrate objects from the research to the collaboration, and the collaboration to the publication, domains. In some cases the movement will be in name only, due to storage or other limitations. That is, an object may stay in a research or collaboration repository but be exposed in the publication domain. Obviously, this has security implications for the underlying repository infrastructure. Often this migration process will involve a mixture of human and computer actions. In practice, humans will need to make selection decisions and then use automated assistance to modify and augment the objects as they cross the curation boundary.

Adopting the Data Curation Continuum

The original data curation continuum concept has been well accepted and used across a range of settings: institutional repositories (Salo, 2008) environmental data (Baker, 2009) the sciences (Kowalczyk and Shanker, 2011), geosciences (Klump, Huber and Diepenbroek, 2016; Klump, Ulbricht and Conze, 2015), data staging repositories (Dietrich, 2010), and research data system design (Wehle, Wiebelt and Suchodoletz, 2017).

Why the Need for an Update?

The design of the original data curation continuum concept was very much a creature of its time. It occurred when the authors were involved in a series of early Australian projects: the Australian Research Repositories Online to the World (ARROW) institutional repository project (Treloar and Groenewegen, 2008), the Dataset Acquisition, Accessibility and Annotations e-Research Technologies (DART) data project (Treloar, 2007), and the Australian ResearCH Enabling EnviRonment (ARCHER) project (Atkinson, et al., 2008). It was also informed by what was then early thinking about the role of the Institutional Repository in a university (Lynch, 2003) and early work in the UK on Virtual Research Environments (VREs) as well as the UK eScience program of work. Finally, it was conditioned by experience of using the eResearch infrastructure that was available in Australia prior to the implementation of solutions through the Platforms for Collaboration program under the Australian National Collaborative Research Infrastructure Strategy (NCRIS).

So, what has changed in the last ten years to require a redesign? In short, a large number of things relating to infrastructure availability and research practices.

More Storage Options

The original diagram assumed both a small number of discrete storage offerings in each domain, and an implied process of data movement/copying across the curation boundaries.

Researchers now have access to a much greater range of available storage solutions, in particular through a range of cloud storage offerings. These might be offered by an institution for use by its staff, as a commercial offering (Dropbox, AWS S3), or nationally (Australia's Research Data Services, AARNet's CloudStor+).

Because this storage is often used from the point of data creation, the need to copy the data from the private to the collaboration domain in order to share it is potentially

removed. The data effectively starts life on shareable storage. More granular access control mechanisms mean that it is possible to start capturing data using one solution and provide wider access to the collaboration or public domains by just changing the access restrictions.

Wider Range of Research Outputs

When the original Data Curation Continuum was conceived, the primary research output was the publication – everything else was viewed as being in service to this, and data was seen by many researchers primarily as a way to generate figures or tables. The idea that data could be a first-class research output was only being discussed in particular contexts. Since then we have seen a greater diversity in research data outputs (including large reference datasets relied on by disciplines), the addition of models/workflows/code as an additional category of output, and a growing number of ways to present *n*-dimensional datasets through techniques such as video, simulations, virtual reality and augmented reality. This means that describing the model as just applying to data is no longer sufficient.

For an example in the software domain, a researcher might start out writing code on their own computer, move that code to Github to enable others to collaborate on it, and then publish it to Zenodo (Smith, et al., 2016).

Consequences of Increasing Data Volumes

Within the data domain, the size of the data objects and their number are increasing rapidly. Some disciplines work with large numbers of files (each of which might itself be large) and others with complex multi-Terabyte databases (both relational and noSQL). Even with high-speed networks, it may not be practical to move such volumes between storage solutions and leaving them in situ will be required. It is also often the case that these large volumes will need to stay located close to the HPC software that is needed to process or visualise them. This requires a curation-in-place approach. The data is not moved from one storage subsystem to another, because it can't be (at least, not on timescales that are acceptable). Instead, the access permissions are changed to allow greater access and more and more context metadata are added. This is effectively a process of curation by addition (Shotton, 2011).

Increased Importance of Capturing the Process of Research

Van de Sompel and Treloar (2014) have argued that there is an increasing need to move from archiving the outputs of the research process to capturing the process whereby research takes place. The research process itself is in transition from being hidden in the system of journals towards being visible in the web of objects. The increased use of commodity networked technologies, such as on demand cloud computing infrastructure and collaboration/sharing platforms for a variety of objects including software and workflows, make sharing objects that are created during that process not only possible but also attractive. MyExperiment, GitHub, Dropbox, networked lab notebooks, scientific wikis and blogs stand out as obvious examples of this.
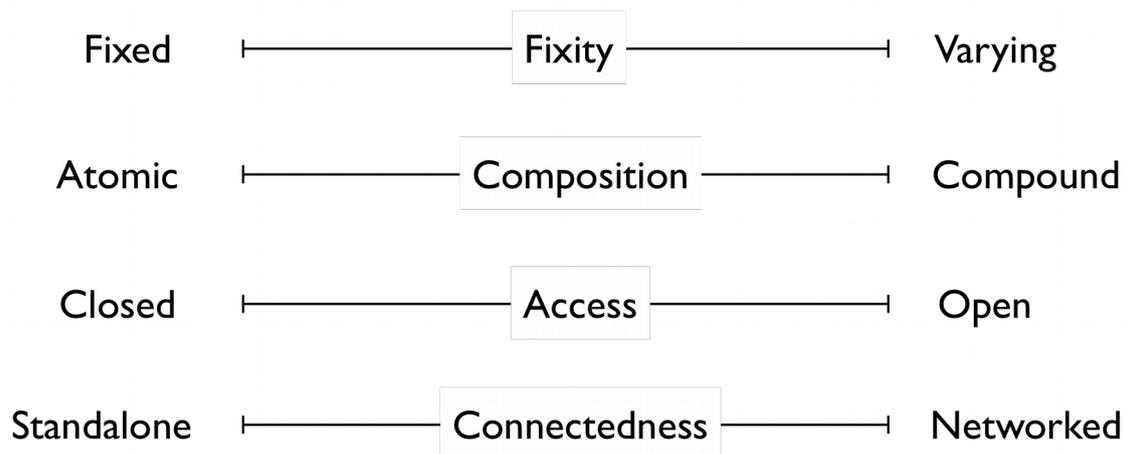
Figure 2. Changes in scholarly objects.

Figure 2 observes the changing nature of the objects that are communicated in the scholarly communication system, confirming the evolution from fixed to varying, from atomic to compound, from uniform to diverse, and from standalone to inter-related or networked that was anticipated in Van de Sompel et al. (2004). In addition, it observes the evolution from journal articles that exhibit a clear sense of fixity towards dynamic objects that (at least during part of their visible life cycle) are continuously changing (for example, as they are being collaboratively edited on the aforementioned commodity platforms). The ongoing evolution from restricted to unconstrained access to scholarly objects catalysed by the Open Access and Open Science movements is also depicted. The data repository landscape also needs to respond to this series of changes.

Looking at the infrastructure requirements from this perspective can be visualised as shown in Figure 3 (in some ways this can be seen as a 90 degree rotation of Figure 1). Here the private infrastructure stores the ephemeral results of research activity, the recording infrastructure (corresponding to the Collaboration domain) captures the transitory process of shared research, and the archiving infrastructure provides persistent storage for the public outputs of research.
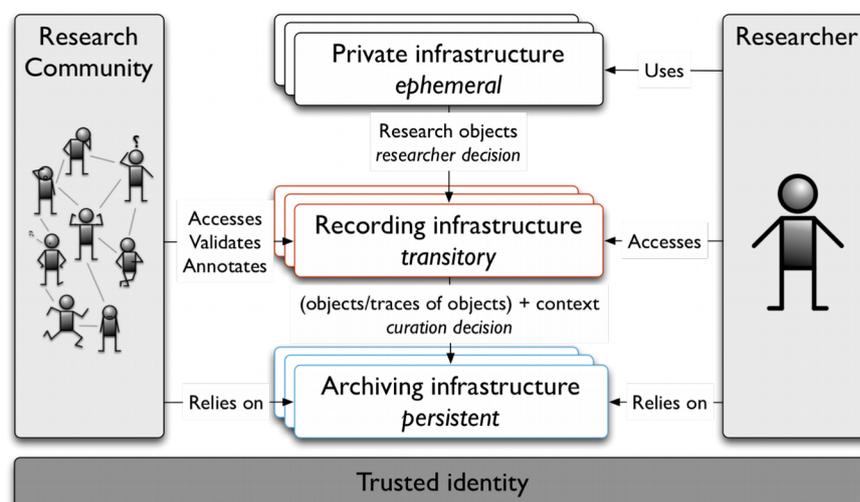

Figure 3. An archival perspective on capturing research.

Increased Automation in the Research Process

The original Data Curation Continuum envisaged the process of migration of objects across the curation boundaries as a primarily manual process. A combination of two developments requires a re-evaluation of this. Firstly, the volume of data objects (and thus curation decisions) means that unassisted manual activity will not scale – most data will only ever be viewed by machines. Secondly, there is now a greater number of automated or semi-automated tools and workflow systems to reduce the need for manual intervention or assist when it is required.

A Greater Focus on Data Re-use

The original Data Curation Continuum had as an unstated assumption that data was captured/created in the private domain and then only moved through a series of transitions from left to right. The increase in the availability of reference datasets and a greater awareness of the value of data reuse, in part driven by the successes of the FAIR movement, mean that this assumption needs to be replaced with a more nuanced view. This means that the continuum needs to accommodate data in the publication domain being combined with new data in the private domain to generate new findings.

The Arrival of FAIRness

The original Data Curation Continuum was created long before the FAIR (Wilkinson, et al., 2016) approach to data was developed and enthusiastically adopted. In the same way that everything now needs to demonstrate how it relates to FAIR, any update to the Data Curation Continuum needs to demonstrate how it relates to each of the FAIR elements. In particular, the publication domain contributes to Findability and Accessibility, and the enrichment of context as data moves from left to right contributes to Interoperability and Reusability.

What is Still Relevant?

At the same time as the above factors have changed, a number of characteristics of the original diagram have remained valid. This argues for the value of revisiting the diagram and updating it.

Validity of the Three Domains

The original three domains were derived from a wide exposure to e-research solutions, as well as the experience of running a number of projects in each of the domains. A decade of reuse of the Data Curation Continuum has not invalidated the choice of these three domains as distinct entities. Indeed, the GFZ case study described below speaks to its continuing relevance.

Value of the Model for Infrastructure Planning

Most research institutions nowadays operate a research data repository. Submitting research data to these repositories is fraught with all the data and metadata challenges

that lead to the development of the data curation continuum concept in the first place. At the same time, research projects and their research data infrastructures come and go. Often, research data management systems start out as a project specific, monolithic application. To keep data accessible, these systems would need to be maintained well beyond the end of the project they were originally built for. In addition, the project-specific functionality to add further data is no longer needed, but difficult and expensive to maintain beyond the end of the project. To keep data accessible, it becomes necessary to transfer the data into other, persistent institutional or disciplinary repositories. Here, the Data Curation Continuum model helps to define the migration path of the data through the project life cycle and the enrichment and transformation of associated metadata to enable future discoverability and reuse. The model also helps to outline the domains of responsibility of different stakeholders involved in the data curation process.

For designing its institutional research data infrastructure, the Helmholtz Centre Potsdam German Research Centre for Geosciences (GFZ) in Potsdam, Germany, adopted a variation of the data curation continuum model (Klump, Ulbricht and Conze, 2015). The model was used to delineate domains and functions of the project specific data management portals and the generic institutional data access portal, which all used the same institutional data storage infrastructure (Ulbricht, Elger, Bertelmann and Klump, 2016). This model separated data curation project portals in the collaboration domain from persistent access to data in the publication domain. This separation of functions and responsibilities allowed project portals to be decommissioned some time after the end of the project without jeopardising access to the already published data. Additionally, the model developed at GFZ allowed for various metadata schemas to be used alongside each other. Only at the point of transferring data from the collaborative domain to the publication domain was it necessary to produce metadata compliant with the DataCite metadata schema.

The application of the Data Curation Continuum model at GFZ also helped to outline the domains of responsibility of the stakeholders involved. In this application of the model, the project specific components that were used for data curation in the collaboration domain were built and maintained by the project and supported by the GFZ Centre for Geo Information Technology, while the publication domain was operated and maintained by GFZ Literature and Information Services through its data publishing service.

Usefulness of Model as a Way to Engage with Stakeholders

The range of ways in which the Data Curation Continuum have been applied, as well as anecdotal comments on its usefulness in explaining research data management to researchers, demonstrates its value as a way of conceptualising the space. The model has proven useful in discussions between stakeholders operating in different parts of the continuum to outline their respective domains of responsibility and clarify who is responsible for what. It has also been used to inform the design of data management systems (Wehle, Wiebelt and Suchodoletz, 2017).

Value of Context Augmentation as a Way of Thinking about Metadata

A core part of the Data Curation Continuum is the notion of augmenting the context around an object as it moves from left to right. In the Private domain, much of the context is tacit – assumed by the researchers based on a combination of local practice

and discipline conventions. Moving to the Shared domain, neither of these are still valid – a range of researchers are involved in collaborative activity around the objects, and the context may now be multi-disciplinary. What is assumed in one discipline needs to be made explicit for another. If one now moves to the Public domain, even more context needs to be encoded – anyone might now need to be able to find, access and re-use the objects, including those outside the research domain altogether (such as citizen scientists).

What does this Mean for an Update?

This paper has argued for the validity of the ideas that underlie the Data Curation Continuum, but also demonstrated a number of developments that require it to be updated. Thinking about possible changes, what should be the re-design elements?

Simplify

One is to remove elements that are no longer central to how it is used and that confuse the message. Chief among these is the collaboratory layer (the upper element in each of the domains in the original diagram). At the time, the expectation was that each domain would have one or more environments supporting the creation of documents that connected to data objects. An example of such a setup might be a Twiki- or Drupal-based environment that is used by a laboratory to create publications and document the research process, coupled with a data store to manage the data resulting from that process.

In practice, collaboratories of this form have not been taken up as widely as was originally anticipated, and cloud-based generalist environments, such as Google Docs or osf.io, have proven more popular. Given the increasing awareness described above of the importance of data as a first class research output, an implied mandatory coupling between document and data is also no longer helpful.

Another simplification is to remove some of the detail of activities at the boundaries, and the specific techniques used, to make the model more general.

Amplify

At the same time as simplifying particular elements, it is necessary to amplify others.

The main requirement here is to provide both discrete and contiguous storage layers (the latter based on the increasing prevalence of cloud solutions), to include a wider range of object types (which, by implication, requires a focus broader than just data), and reflect a greater focus on data reuse.

Another is to make clearer the process by which the provenance trail is captured and augmented as the objects move across the domain transitions.

Clarify

The domain transitions also need to be clarified to reflect the ways in which research outputs are now being managed. This needs to clarify what happens at the boundaries, and also distinguish between object context and object provenance.

Result

The result of this process is Figure 4. This is an update of the Data Curation Continuum for a new decade, building on what worked in the past and anticipating the future. The title has been changed to Object Curation Domains to reflect an updated understanding of the core elements of the model. The activities taking place at the boundaries have been broken out into discrete streams of activity within defined layers, to make it clearer where the activities are taking place and what their consequences are.

The **object layer** shows the increased range of research objects now in scope (data, models, workflows, software, publications, documentation), and also the process by which the number of objects decreases as the result of a process of intentional selection as one moves from left to right.

The **storage layer** now distinguishes between discrete storage (different for each domain) and cloud storage (contiguous across each domain). It is also possible to use a combination of local storage and cloud storage, or even three different cloud storage solutions (one for each domain) but adding all the possible option combinations would have had made the model much more complex for little additional benefit.

The **context layer** shows the way in which object context is added as the object(s) transition across the boundaries. This reflects the way in which tacit context needs to be made explicit for audiences broader than the setting in which the objects are being created/used. Note that this layer has deliberately not been called the metadata layer. This is because it is a means for encoding context (and of course other things). The end is the context capture itself. Note also that the step of adding context happens at the boundary transition. There is little value in documenting context of use only outside a domain if that object will never leave that domain. This can be viewed as 'just-in-time' context addition.

The **provenance layer** is subtly different. Provenance can be viewed as just another kind of context, encoded in provenance metadata, but the provenance of an object (and particularly the changes that have occurred to that object) are often used in different ways to (for instance) discovery or description context information. Note that for the provenance information, this is added within the domain (by whatever systems for data management/generation are being used) and then simply migrated across the boundary transition. This is consistent with the model of capturing research activity shown in Figure 3.

The **archival** layer shows the ways in which archival elements can be included in the object lifecycle from the point of creation, rather than being added as an afterthought later on.[1]

---

1 See http://www.dcc.ac.uk/resources/curation-lifecycle-model/lifecycle-model-faqs# for an archival perspective on the DCC curation lifecycle model
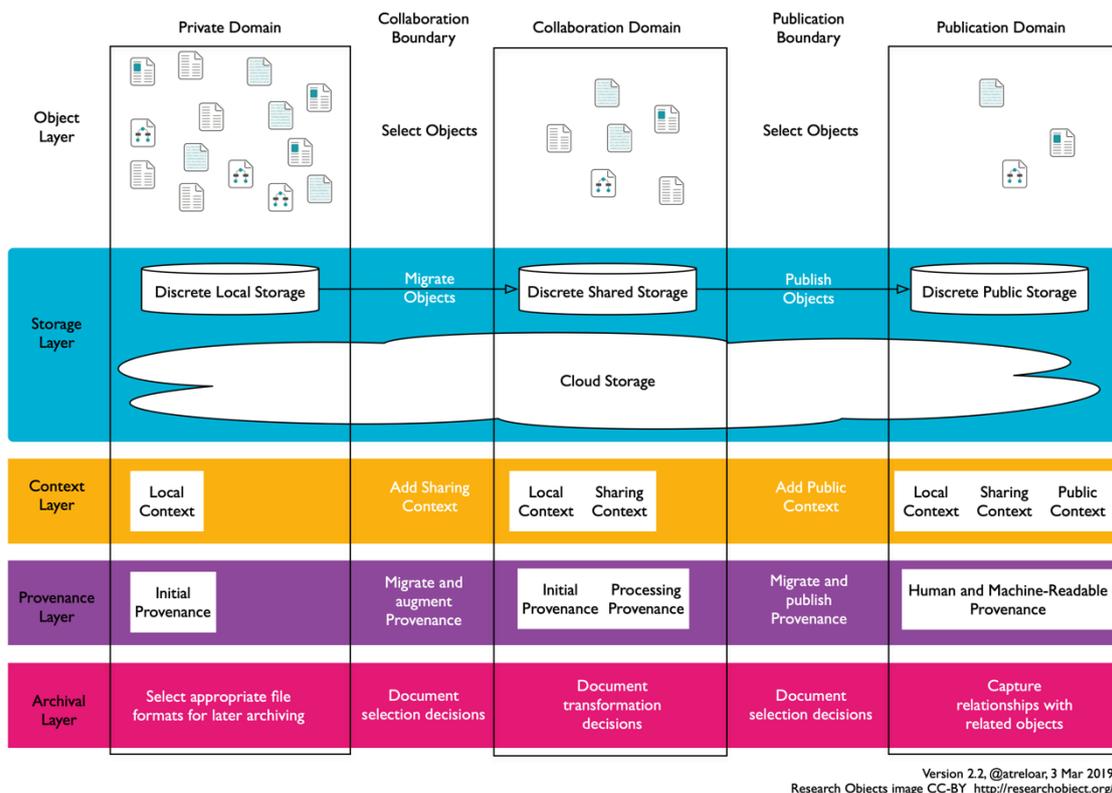
Figure 4. Object Curation Domains.

Applications

The updated Object Curation Continuum model has application in a number of domains relevant to scholarly communication, and implications for practitioners in those domains.

For researchers creating objects in the Private Domain, they need to be aware that some of these objects will make the transition to the Collaboration and Publication Domains. This requires them to be explicit about capturing enough context early in the life of the object to enable this to be translated (perhaps automatically at the boundary transitions) and presented for a wider audience. They also should be aware that only a subset of the objects they are creating will be selected for this transition, and thus not to overinvest in manually adding contextual information too early (automated capture of context costs much less and thus is less of a concern).

For system designers, they should aim to capture as much context automatically as possible, as early in the process as possible (Treloar and Wilkinson, 2008). They should also be aware of the need to capture provenance information early and migrate it across the boundaries in a form that can be presented as human-readable.

For data librarians, they should assist researchers to understand the kinds of sharing context and public context that will most assist in making the research objects most findable, accessible, interoperable and reusable.

Conclusion

As we hope is clear from the section on adoption and the GFZ case study, the Data Curation Continuum model has been adopted by a range of domains over the last decade and demonstrated its value in informing infrastructure planning. This update of the model retains its value, simplifies the core concepts, reflects changes in the environment it describes, and prepares it for use over another decade. The authors look forward to continuing to engage with research infrastructure practitioners to refine the model and further enhance its relevance and applicability.

Acknowledgements

References

Atkinson, A., et al. (2008). ARCHER – e-Research tools for research data management. Proceedings of IEEE e-Science 2008, December, Indianapolis.

Baker, K.S. (2009). Data stewardship: Environmental data curation and a web of repositories. *Digital Discourse: The e-volution of Scholarly Communication, 1*(1).

DEST. (2007). Towards the Australian Data Commons. Retrieved from https://web.archive.org/web/20091015000000*/http://www.pfc.org.au/twiki/pub/Main/Data/TowardstheAustralianDataCommons.pdf

Dietrich, D. (2010). Metadata management in a data staging repository. *Journal of Library Metadata, 10*(2-3), 79-98.

Klump, J., Huber, R., & Diepenbroek, M. (2016). DOI for geoscience data: How early practices shape present perceptions. *Earth Science Informatics, 9*(1), 123-136.

Klump, J., Ulbricht, D., & Conze, R. (2015). Curating the web's deep past – Migration strategies for the German Continental Deep Drilling Program web content. *GeoResJ, 6*, 98-105.

Kowalczyk, S., & Shankar, K. (2011). Data sharing in the sciences. *Annual review of information science and technology, 45*(1), 247-294.

Lynch, C. (2003). Institutional repositories: Essential infrastructure for scholarship in the digital age. *ARL Bimonthly Report 226.* Retrieved from https://www.cni.org/wp-content/uploads/2003/02/arl-br-226-Lynch-IRs-2003.pdf

Salo, D. (2008). Innkeeper at the roach motel. *Library Trends, 57*(2), 98-123.

Schauder, D., Stillman, L., & Johanson, G. (2004). Sustaining and transforming a
community network: The Information Continuum Model and the Case of
VICNET. Paper presented at CIRN 2004: Sustainability and Community
Technology, Monash University, Prato, Tuscany, Italy. Retrieved from
https://research.monash.edu/en/publications/sustaining-and-transforming-a-
community-network-the-information-c

Shotton, D. (2011). IBRG projects to facilitate data publication and data citation.
Retrieved from https://opencitations.wordpress.com/2011/08/04/ibrg-projects-to-
facilitate-data-publication-and-data-citation

Smith et al. (2016). Software citation principles. *PeerJ Computer Science 2:e86*.
doi:10.7717/peerj-cs.86

Treloar, A. (2007). The Data Acquisition, Accessibility, Annotation and e-Research
Technologies (DART) Project: Supporting the complete e-Research Lifecycle.
Proceedings of UK e-Science Programme All Hands Meeting 2007 (AHM2007),
Nottingham, September.

Treloar, A. & Groenewegen, D. (2008). The ARROW Project: A consortial institutional
repository solution, combining open source and proprietary software. *OCLC Systems
and Services, 24*(1), pp. 30-39. doi:10.1108/10650750810847224

Treloar, A., Groenewegen, D. & Harboe-Ree, C. (2007). The Data Curation
Continuum: Managing data objects in institutional repositories. *Dlib, 13*(9-10).
doi:10.1045/september2007-treloar

Treloar, A. & Harboe-Ree, C. (2008). Data management and the curation continuum:
how the Monash experience is informing repository relationships. Proceedings of
VALA 2008, Melbourne, February. Retrieved from
http://www.vala.org.au/vala2008-proceedings/vala2008-session-6-treloar

Treloar, A. & Wilkinson, R. (2008). Rethinking metadata creation and management in a
data-driven research world. Proceedings of IEEE e-Science 2008, December,
Indianapolis. doi:10.1109/eScience.2008.41

Tsoi, A., McDonell, J. & Treloar, A. (2007). Dataset acquisition, accessibility, annotation,
e-research technologies (DART) project: A case study using an X-ray diffractometer.
International Journal of Digital Libraries, 7(1-2). doi:10.1007/s00799-007-0019-4

Ulbricht, D., Elger, K., Bertelmann, R., & Klump, J.F. (2016). panMetaDocs, eSciDoc,
and DOIDB: An infrastructure for the curation and publication of file-based
datasets for GFZ Data Services. *ISPRS International Journal of Geo-Information, 5*(3), 25.
doi:10.3390/ijgi5030025

Van de Sompel, H. et al. (2004). Rethinking scholarly communication: Building the
system that scholars deserve. *Dlib, 10*(9). doi:10.1045/september2004-vandesompel

Van de Sompel, H. & Treloar, A. (2014). A perspective on archiving the scholarly web. Proceedings of iPres 2014, Melbourne, October.

Wehle, D., Wiebelt, B., & von Suchodoletz, D. (2017). Design eines FDM-fähigen Speichersystems. In 10. DFN-Forum Kommunikationstechnologien. Gesellschaft für Informatik eV.

Wilkinson, et al. (2016), The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, Article number: 160018.