Data management and the curation continuum: how the Monash experience is informing repository relationships

Andrew Treloar
Director, Australian National Data Service Establishment Project
Monash University
andrew.treloar@its.monash.edu.au

Cathrine Harboe-Ree
University Librarian
Monash University
cathrine.harboe-ree@lib.monash.edu.au

Abstract:
Repositories are evolving in response to a growing understanding of institutional and research community data and object management needs. This paper (building on work already published in DLib, September, 2007) explores how one institution has responded to the need to provide management solutions that accommodate different object types, uses and users. It introduces three key concepts. The first is the curation continuum, which identifies a number of characteristics of data objects and the repositories that contain them. The second divides the overall repository environment based on these characteristics into three domains (research, collaboration and public), each with associated repository/data store environments. The third is the curation boundary, which separates each of the three domain types.

1. Introduction

The work being done at Monash University on rethinking the role of repositories in supporting data management has grown out of a number of local factors that distinguish Monash University from other institutions: its size, a focus on e-Research, an Information Management Strategy, a range of inter-related and innovative projects and a whole-of-organisation approach, both strategically and practically.

Monash University is Australia's largest and most internationalised university. In addition to its six Australian campuses, it has campuses in Malaysia and South Africa as well as centres in London and Prato (Italy). Monash University has over 50,000 students and 6,000 equivalent full time staff. Its current strategic direction is focussed on achieving excellence through a cross-disciplinary, multi-campus, international approach. Consistent with this the university has invested strongly in information technology to support research, teaching and administration.

One manifestation of this has been a recent strategic investment in e-Research. This is what the UK would call e-Science, and the US calls cyberinfrastructure (Atkins, 2003). This has taken the form of the establishment of the Monash e-Research Centre, as well as significant investment in network, storage and grid computing infrastructure. The Information Technology Services Division is also creating a new department to support research and e-Research activities.

In 2003, a group was formed to develop an Information Management Strategy for the university. The initial membership (later augmented) included the Executive Director (ITS), the University Librarian, the Head of the Centre for Learning and Teaching Support, the Manager of Records and Archives and discipline experts from the School of Information Management and Systems. The resulting strategy (Monash 2005) took an explicitly holistic approach that did not see information as belonging to only one part of the university. The strategy was adopted as one of the University's key priorities for 2006, and it retains an ongoing importance. The strategy articulated a set of principles, including statements about information being of corporate importance and available to anybody, anytime, anywhere and anyhow, as appropriate. These principles have informed a range of initiatives since then, including the ARROW, DART and ARCHER projects (see below). For more details about the strategy, see Treloar (2004, 2005a, 2006a), and Palmer (2007).

Monash has also been a leader in the establishment and operation of Australian government funded research projects in this area, such as the institutional repository project (ARROW) and two projects on researcher workflow and data management (DART and ARCHER).

Australian Research Repositories Online to the World (ARROW – http://arrow.edu.au/) is a consortium consisting of Monash University (lead institution), together with the University of New South Wales, Swinburne University of Technology, and the National Library of Australia. This project aimed to identify and test software or solutions to support best practice institutional digital repositories that would contain e-prints, electronic theses, e-research and electronic journals. The project has partnered with a commercial software developer (VTLS Inc.) to develop a number of open source software modules and VITAL, a licensed commercial offering built on top of these modules. The current offering provides a rich product on top of the Fedora open source repository platform (Lagoze et. al. 2006). Fifteen of the thirty-nine universities in Australia have licensed the VITAL software solution to support an institutional repository. For more details about ARROW, see Payne and Treloar (2006) and Treloar and Groenewegen (2007).

In August 2005, the Dataset Acquisition, Accessibility and Annotations e-Research Technologies (DART – http://dart.edu.au/) project came into existence. The DART request for funding built on the work already done in the ARROW project in establishing the basis for institutional research publication repositories, as well as antecedent activity at each of the DART partners (Monash University – lead institution, James Cook University and the University of Queensland). It did this by extending its scope into the challenges arising from the management of large datasets and sensor networks, as well as annotation technologies and collaborative, composite documents. In particular, the DART project investigated the most appropriate response to the challenges inherent in new forms and producers of raw data, new forms of collaborative research activity, new forms of publication, and new forms of research validation. The DART project completed its ambitious work program of 28 separate workpackages in June 2007. The DART website provides consolidated access to the outputs, including reports and source code. For more details about DART, see Treloar (2006b, 2007).

A successor project, called the Australian ResearCH Enabling EnviRonment (ARCHER – http://archer.edu.au/), is currently building on selected DART deliverables (collaborative workspace environments, data acquisition and management, and frameworks), as well as integrating other open source components, to provide a robust and comprehensive end-to-end set of data acquisition and management tools. These tools are progressively being aligned with the development priorities and needs of the interoperation and Collaboration Infrastructure (ICI) and the Australian National Data Service (ANDS) programs within the Platforms for Collaboration (PfC) capability under the National Collaborative Research Information Strategy (NCRIS). These two programs will make a significant contribution to support for innovative Australian research over the next four years.

2. Data Curation Continua

The Monash University Information Management Strategy adopted a multi-dimensional view of information. 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, et al. 2004). These information management dimensions were largely determined to have particular values.

An analysis of the research data management space suggested that it is 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, and the results of the work undertaken to clarify ANDS (DEST 2007b). 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: http://www.dcc.ac.uk/. Table 1 summarises the Data Curation Continua identified so far.

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

Metadata

At one end of the continuum, objects will contain the minimum metadata needed by the object's creators and users. This will often be a mix of simple descriptive metadata (filename, creator) and discipline-specific technical metadata. The drivers for minimal metadata are the combination of significant numbers of objects (see below), insufficient time to provide extensive metadata for each object, automatic generation/capture of the objects, and no business case for providing more comprehensive metadata. At the other end of the continuum, objects will contain much richer metadata. This might include provenance metadata (indicating what operations have been performed on the object), more detailed descriptive metadata and preservation metadata to facilitate curation.

Item Count

One end of the item count continuum describes repositories with lots of items. These may be different versions of data objects, the results of failed or inconclusive experiments, or objects that should have been purged but have not been as yet. For large projects or institutions, the object count could run into the millions. The other end of the continuum describes repositories with many fewer items. This is because the objects have been winnowed and selected. This selection may be on the basis of an institutional data management policy or because the objects have been referenced in a publication.

Object Size

Object size is another possible continuum. Of course, repositories will contain objects of many different sizes – this continuum is based on the most common object size. Many e-research projects now routinely work with very large (i.e., multi-gigabyte) objects. These may either be single files, containers of files, or complex databases. On the other hand, some types of repositories (typically publication-focussed institutional repositories) are designed to work with smaller (megabyte-sized) objects. The size continuum is actually quite critical – a number of designers of repository software have made implementation decisions on the basis of object size or type, and repository managers are now having to revisit these decisions to support larger/different objects.

Object 'Fixity'

Fixity refers here to whether the objects change once ingested into the repository. Researchers often want data objects that are continually changed or updated as the research project progresses. A good example is a record of climate data that grows each time new values are collected. Such a changing data object is not appropriate if it is linked to a publication as part of the permanent scholarly record. For this purpose it is preferable to have a snapshot of the data object or an extracted subset that will not change post publication.

Management Responsibility/Control

Another continuum that can be used to characterise data curation practices defines who is responsible for the management of the data (or who has control over it). Research data objects in the collaboration space are often managed by the researcher, members of the research team, or local IT staff. These managers may not have the skills or the commitment needed for long-term curation. At the other end of this continuum is management by a dedicated group within the university. Such a group, which might be a virtual team, would ideally include a range of specialists: records and archives, library, information technology, and e-Research.

Preservation

One of the processes that can be applied to data objects is evaluating the degree to which preservation occurs. In the case of a research group, the preservation horizon is unlikely to be more distant than the end of the current project or the requirements of the grant that is funding the work. An institution should have a longer-term focus; one that is concerned with the long-term scholarly record, national codes of conduct, and institutional obligations enshrined in legislation.

Access

The work of investigators in the DART project has indicated that many researchers are conservative when it comes to granting access to research data. This appears to be associated with increasing competition in attracting research funds and having articles accepted by high-value publications. The recent move in Australia towards assessment of institutional research performance based on quality metrics (the Research Quality Framework – DEST 2007a, being re-evaluated at the time of writing after the change of government in November 2007) is likely to intensify this. As a result, many researchers want tightly controlled access prior to publication. It is theoretically possible to provide the levels of access control demanded by researchers in a repository that also hosts open-access content, but separation of the two types of repositories may be a preferred solution. Post publication, there is some evidence that open access leads to increased accessibility and increased citation rates. This may, over time, encourage more researcher openness.

Exposure

The access continuum on its own is not enough to ensure the benefits of open access. The contents of the repository also need to be exposed and discoverable. This can be via a range of techniques: search engine spidering, the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), RSS (Really Simple Syndication – often used for podcast and blog alerting services) feeds, and/or SRU/SRW (a search protocol designed as the successor to Z39.50) access through federated search techniques. At one end of this continuum, there is limited or no exposure to such harvesting software, meaning that even data objects with no access restrictions are unlikely to be discovered. At the other end of this continuum, the contents of repositories (such as the ARROW repositories) are exposed using a range of technologies to provide the maximum accessibility.

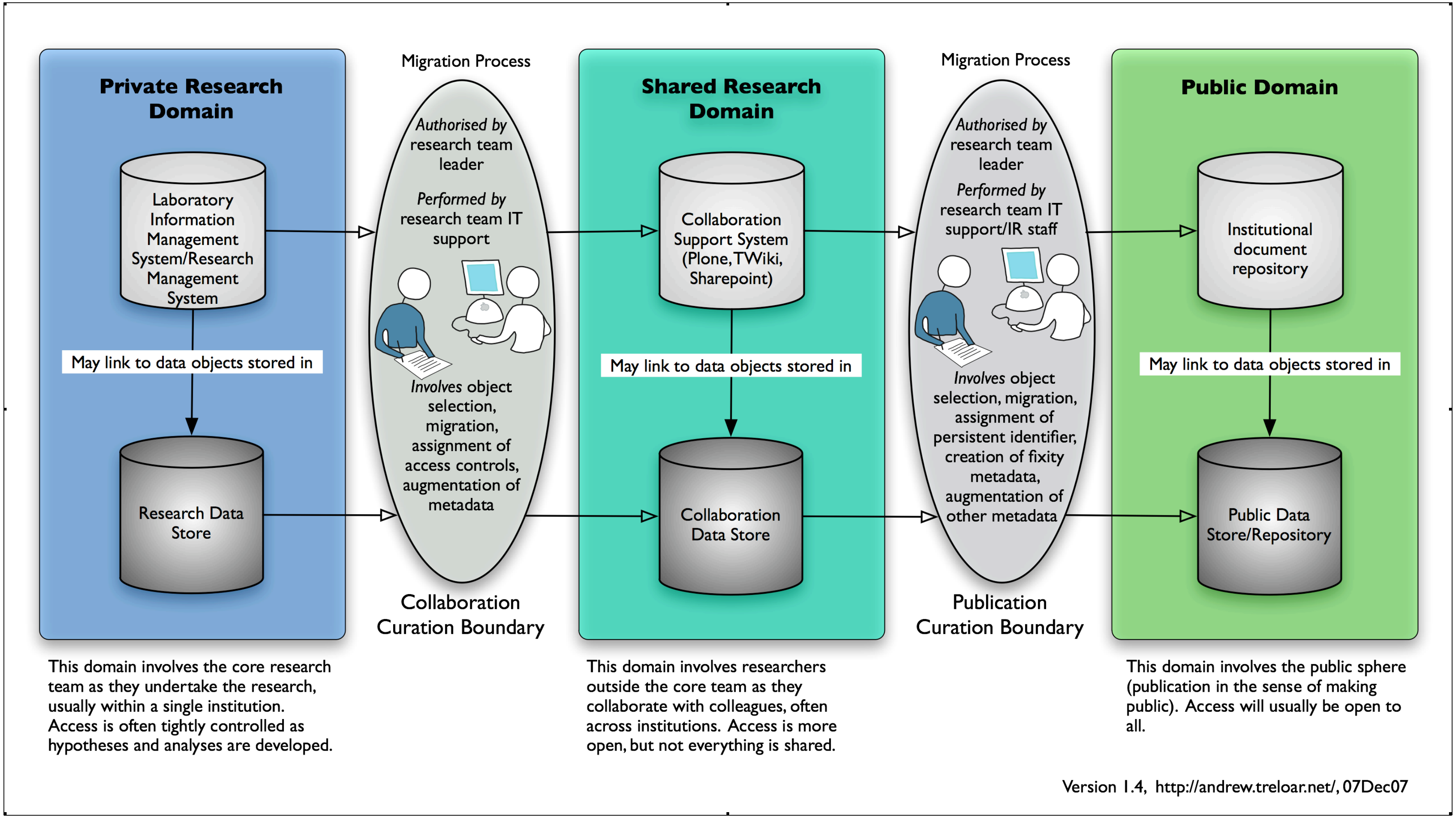
3. Domains and their repositories

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. Figure 1 (over page) shows a technology-neutral version of these three domains and their associated repositories. 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 or a Storage Resource Broker (SRB) 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 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 (such as Plone or TWiki) 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. Both SRB and Fedora (as well as a range of other technologies) support this. Compared to the research 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.

Figure 1: Domains, Data Stores and Curation Boundaries


Private Research Domain
Laboratory Information Management System/Research Management System
May link to data objects stored in
Research Data Store
Migration Process
Authorised by research team leader
Performed by research team IT support
Involves object selection, migration, assignment of access controls, augmentation of metadata
Collaboration Curation Boundary
Shared Research Domain
Collaboration Support System (Plone, TWiki, Sharepoint)
May link to data objects stored in
Collaboration Data Store
Migration Process
Authorised by research team leader
Performed by research team IT support/IR staff
Involves object selection, migration, assignment of persistent identifier, creation of fixity metadata, augmentation of other metadata
Publication Curation Boundary
Public Domain
Institutional document repository
May link to data objects stored in
Public Data Store/Repository

This domain involves the core research team as they undertake the research, usually within a single institution. Access is often tightly controlled as hypotheses and analyses are developed.

This domain involves researchers outside the core team as they collaborate with colleagues, often across institutions. Access is more open, but not everything is shared.

This domain involves the public sphere (publication in the sense of making public). Access will usually be open to all.

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

The third domain is the public 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.

It should be noted that there is no necessary one-one relationship between domains and repositories. It would be theoretically possible to support all three domains from a single repository instance, but in practice the requirements of the different curation continua would make this extremely difficult. This is because the differing sorts of objects that might be stored in a repository can vary widely. In addition, there are characteristics of the management of and access to these objects that also differ. These differences cannot easily be accommodated from a common repository infrastructure. This is particularly true once one moves from a repository of publications and discrete objects to a repository that might also contain data (of widely varying sizes) generated by e-research.

4. Boundaries and migration

Figure 1 shows two boundaries: the collaboration curation boundary and the publication curation boundary. The use of the word curation is deliberate and reflects the fact that 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.

In this model, there is 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. As Figure 1 shows, this migration process involves 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. The University of Hull is currently exploring the role of workflows in facilitating this sort of processing in their Repository Metadata Management (RepoMMan) project (Hull 2007). The process of crossing the collaboration curation boundary will probably be more lightweight than the process of crossing the publication curation boundary.

5. Putting the Data Curation Continuum into Practice

Implementing this at Monash

The way in which Monash University is currently applying the data curation continuum approach is to focus at present on the research and public domains. It is anticipated that the collaboration domain (as defined above) will be more relevant and easier to implement in the context of the nascent Australian Access Federation and ICI and ANDS programs within the Platforms for Collaboration capability of the National Collaborative Research Information Strategy.

The research domain at Monash University (and no doubt other universities) is populated by a variety of different repository solutions. The DART and ARCHER projects have been working with two different Protein Crystallography groups to support their research processes. The DART project (Treloar, 2007) has provided software to assist with data capture from instruments, storage of the data in an SRB repository, and analysis of the data using grid software tools. There is also a managed storage solution (a service centrally provided by Information Technology Services from January 2008) called the Large Research Data Store (LaRDS). This is made available through a number of different software interfaces, including as a mapped R (for research, of course!) drive on the desktop.

The public domain at Monash University uses the ARROW software solution (Treloar and Groenewegen, 2007). This was designed for document objects, and ingesting of large datasets is currently being trialled. It is possible that a separate public data repository may be needed, but at present the existing institutional repository appears able to fulfil both functions.

Two examples from different domains will serve to illustrate how we are currently applying the Data Curation Continuum at Monash.

The first example comes from the domain of Protein Crystallography. The end result of applying this model and the associated migration process is a paper in the prestigious *Science* journal (Rosado et al. 2007), where the final published version points to a dataset that has been migrated across the curation boundary into the ARROW Repository (see http://arrow.monash.edu.au/hdl/1959.1/5863). This process was somewhat ad-hoc and involved a lot of manual work and creative problem-solving by Andrew Harrison, the Monash ARROW Librarian. This was in part caused by the size of the datasets involved. The entire repository object totalled 36 GB in size, (after compression!) with many datastreams being 2 GB in size. This is significantly larger than the software was initially designed for, although it is being reconfigured to support larger file sizes. Procedures are also being put in place that will allow the researchers themselves to undertake much of the work of lodging the dataset objects, with the ARROW Librarian performing more of a quality control and authorisation function. Under this approach, the researchers will provide the quality control over the technical metadata and the library staff will review (and augment) the descriptive metadata.

The second example comes from the domain of musicology. We have some researchers who are working with archival recordings of Jewish music performance. They currently have about 400 GB of digitised audio content up on their LaRDS space, being used for their own private research within their research team. They are now migrating a subset of this content into ARROW for publication. This will be a progressive migration as copyright is gradually sorted through. The estimate is that approx 10% (40 GB) will eventually be published. From there it will be further harvested by and exposed through the National Library of Australia's MusicAustralia service.

Further work

As Monash tests the use of the data curation continuum and the curation boundary ideas, it will accumulate a body of knowledge about how best to make the decision to move data or objects across the boundaries. Part of the migration process over boundaries will need to focus on the object metadata, which will need to be modified and augmented at each transition. In a research repository, the descriptive metadata will be assumed, or encoded within the object name or its directory. In a collaboration repository, it will be the minimum needed for location and management within a collaboration context. In a public repository, the metadata will need to be quality-controlled and significantly augmented so as to improve exposure and accessibility. New metadata, such as PREMIS (preservation) metadata to assist with long-term curation, will also have to be added in a publication/preservation context.

Another part of the migration process shown in Figure 1 is the assignment of a persistent identifier (such as a handle) to the object to facilitate persistent access. The current approach is to only assign a handle once the object is in the publication/preservation repository. An alternative approach is to assign the handle to every object in the research repository and then just update the handle as the object is migrated (Sefton, 2007). This approach facilitates the migration of publications that are linked to data objects prior to publication. As researchers become more comfortable with the new repository-based collaboration environments, it may be worth considering moving to this approach, although this has to be offset against the associated ongoing handle management overhead.

6. Conclusions

When the ARROW philosophy was initially conceived it was thought that a single institutional repository that was integrated, interoperable and flexible would provide the best platform to support teaching and research at Monash.

The single repository approach, while initially attractive, has been found to suffer from a range of implementation challenges and fails to provide adequate management solutions for data generated by researchers over the entire research lifecycle. These challenges can be best addressed when considered in terms of the data curation continua. The ARROW, DART and ARCHER projects have seen the evolution of this concept into a more nuanced understanding of the different types of content that would need to be managed, and the different audiences and uses for that content. This has led to an acceptance that multiple, albeit interoperable, repositories would be better.

One set of decisions about what to do for each of the continua leads to three different sorts of repository domains. Monash University is calling these research, collaboration and public repositories respectively. A further management concept, the curation boundary, provides a mechanism for determining when and how objects can be moved between the domains.

As knowledge about institutional and data management repositories evolves over the next few years, these ideas will be further explored, by Monash and many other institutions.

7. Acknowledgements

The ARROW, DART and ARCHER projects have been funded by the Australian Commonwealth Department of Education, Science and Training. The funding has been provided through the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovation Action Plan for the Future.

We would also like to acknowledge our colleagues. Much of the intellectual work described in this paper arises from intense discussions within the ARROW and ARCHER teams, and their ongoing contributions to activities across the entire data curation continuum. Without their work, Monash would not now be able to tackle the sorts of challenges presented here.

8. References

Atkins, D. et al., 2003, *National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, Revolutionizing Science and Engineering through Cyber-infrastructure*. Available at http://www.communitytechnology.org/nsf_ci_report/

Blinco, K. and McLean, N., 2004, *The Wheel of Fortune: A "Cosmic" View of the Repositories Space*. Available as a Flash animation at http://www.rubric.edu.au/extrafiles/wheel/main.swf

DEST, 2007a. *Research Quality Framework.* Available at http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/research_quality_framework/

DEST, 2007b. *Towards the Australian Data Commons*. Available at http://www.pfc.org.au/twiki/pub/Main/Data/TowardstheAustralianDataCommons.pdf

Jacobs, N., 2006, 'Digital Repositories in UK universities and colleges', *FreePint*, Issue 200. Available at http://www.freepint.com/issues/160206.htm#feature

The Joint Information Systems Committee, 2004. *The Data Deluge: Preparing for the explosion in data*. Available at http://www.jisc.ac.uk/index.cfm?name=pub_datadeluge

Lagoze, C., Payette, S., Shin, E. and Wilper, C., 2006, 'Fedora: An Architecture for Complex Objects and their Relationships', *International Journal of Digital Libraries: Special Issue on Complex Objects*, Volume 6, Issue 2, April. Available at http://www.arxiv.org/abs/cs.DL/0501012

Monash University, 2005, *Information Management Strategy.* Available at http://www.monash.edu.au/staff/information-management/

Open Archives Initiative, 2007, *Open Archives Initiative Object Reuse and Exchange.* Available at http://www.openarchives.org/ore/

Palmer, H., 2007, 'From strategy to action: the information management initiative at Monash University', *Proceedings of EduCause AustralAsia 2007*, Melbourne, April. Available at http://www.caudit.edu.au/educauseaustralasia07/authors_papers/Palmer-237.pdf

Payne, G and Treloar, A., 2006, 'The ARROW Project after two years: are we hitting our targets?, *Proceedings of VALA 2006*, Melbourne. Available a t http://www.valaconf.org.au/vala2006/papers2006/57_Treloar_Final.pdf

Research Information Network, 2007, *Stewardship of digital research data: a framework of principles and guidelines* (Consultation Draft). Available at http://www.rin.ac.uk/data-principles/

Rosado, C. J. et al. (2007), “A Common Fold Mediates Vertebrate Defense and Bacterial Attack”, *Science Express*, August 23 2007. Science DOI: 10.1126/science.1144706

Schauder, D., Stillman, L., and 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. Available at http://www.ciresearch.net/conferences/viewabstract.php?id=68&cf=4

Sefton, P., 2007, *Another good use for handles: identifying items in the ICE content management system throughout their lifecycle* (blog entry). Available at http://ptsefton.com/blog/2007/05/16/handles_curation_boundary

Treloar, A., 2005a, 'Developing an Information Management Strategy for Monash University', *Proceedings of Educause AustralAsia 2005*, Auckland, April. Available at http://andrew.treloar.net/research/publications/educause05/A19.PDF

Treloar, A., 2005b, 'ARROW Targets: Institutional Repositories, Open-Source, and Web Services', *Proceedings of AusWeb05, the Eleventh Australian World Wide Web Conference*, Southern Cross University Press, Southern Cross University, July. Available at http://ausweb.scu.edu.au/aw05/papers/refereed/treloar/

Treloar, A., 2006a, 'The Monash University Information Management Strategy: from development to implementation', *Proceedings of VALA 2006*, Melbourne, January. Available at http://www.valaconf.org.au/vala2006/papers2006/56_Treloar_Final.pdf

Treloar, A., 2006b, 'The Dataset Acquisition, Accessibility, and Annotation e-Research Technologies (DART) Project: building the new collaborative e-research infrastructure', *Proceedings of AusWeb06, the Twelfth Australian World Wide Web Conference*, Southern Cross University Press, Southern Cross University, July. Available at http://ausweb.scu.edu.au/aw06/papers/refereed/treloar/

Treloar, A., 2007, 'DART: Building the new collaborative e-research infrastructure', *Proceedings of Educause Australasia 2007*, Melbourne, April. Available at http://www.caudit.edu.au/educauseaustralasia07/authors_papers/Treloar-183.pdf

Treloar, A. and Groenewegen, D., 2007, 'The ARROW Project: A consortial institutional repository solution, combining open source and proprietary software', *OCLC Systems & Services: International Digital Library Perspectives* (in press)

Treloar, A., Groenewegen, D. and Harboe-Ree, C., 2007, ‘The Data Curation Continuum: managing data objects in institutional repositories’**,** *Dlib,* September/October September/October. doi:10.1045/september2007-treloar. Available at http://www.dlib.org/dlib/september07/treloar/09treloar.html

University of Hull, 2007, *RepoMMan Project Aims*. Available at http://www.hull.ac.uk/esig/repomman/project_aims/index.html

Van de Sompel, H, et al., 2004, 'Rethinking Scholarly Communication: Building the System that Scholars Deserve'. *D-Lib Magazine*, September. doi:10.1045/september2004-vandesompel. Available at http://www.dlib.org/dlib/september04/vandesompel/09vandesompel.html

Van de Sompel, H, et al., 2006, ' An Interoperable Fabric for Scholarly Value Chains'. *D-Lib Magazine*, October. doi:10.1045/october2006-vandesompel. Available at http://www.dlib.org/dlib/october06/vandesompel/10vandesompel.html