DART: Building the new collaborative e-research infrastructure

Dr Andrew Treloar, {Director Information Management and Strategic Planning, ITS, Monash University | ARROW Technical Architect | DART and ARCHER Project Architect }

E: Andrew.Treloar@its.monash.edu.au W: http://andrew.treloar.net/

Abstract

In early 2005, the Australian Government called for proposals for collaborative projects that brought together consortia to improve accessibility to Australian research. The call for proposals identified four areas of interest:

- maximising access to digital resources in Australian universities, especially regional universities;
- creating new types of digital libraries to manage extremely large data sets;
- adopting a national approach to improving open access to the results of publicly funded research;
- providing effective linkages between sets of research information to enable seamless access by researchers.

The call for proposals also identified a number of key trends that are changing the way in which research is conducted and its outputs consumed. These included new technologies, such as computer simulations, synchrotrons and sensor networks, the expanding size of the datasets on which research is based, increasing volumes of information generated through research, and a recognition of the need to work across traditional disciplinary, institutional and national borders.

The Dataset Acquisition, Accessibility, and Annotation e-Research Technologies (DART 2006) Project was successful in receiving funding under this program. DART is responding to these trends and changes by developing and assessing new e-research collaboration tools and infrastructure. The project will be nearing completion by the time of the Educause AustralAsia 2007 conference, and so a late-stage progress report will be delivered as part of the presentation.

This paper begins by describing the context in which the DART bid was submitted, both environmental (described above) and theoretical. The paper then goes on to describe the overall architecture for DART and its various work packages. Each group of work packages is placed in the context of the domain requirements and the theoretical model. Finally, the paper concludes by examining progress to date on the project, as well as describing links to related activity.

Context

In early 2005, the Australian Government called for proposals for collaborative projects that brought together consortia to improve accessibility to Australian research. The call for proposals identified four areas of interest:

- maximising access to digital resources in Australian universities, especially regional universities;
- creating new types of digital libraries to manage extremely large data sets;
- adopting a national approach to improving open access to the results of publicly funded research;
- providing effective linkages between sets of research information to enable seamless access by researchers.

The call for proposals also identified a number of key trends that are changing the way in which research is conducted and its outputs consumed. These included:

- new technologies, such as computer simulations, synchrotrons and sensor networks
- the expanding size of the datasets on which research is based
- increasing volumes of information generated through research
- greater complexity
- recognition of the need to work across traditional disciplinary, institutional and national borders.

To this one might add a growth in research practices that are producing a paradigm change in the types of research that this new large-scale computing/data management environment can support. These emerging research practices:

- are intensely collaborative (often involving trans-national teams)
- require high-quality network access; and
- are data and simulation-intensive.

These changes first became evident in high-energy physics, science and engineering (Atkins, et. al. 2003) but are now also becoming apparent in the social sciences and humanities (Waters 2003). Some disciplines have good practices around, and support for, lodgement of datasets as part of publication while other disciplines are just starting to explore this area. The role of datasets (historical, sensor-produced and simulation-derived) is also becoming increasingly important to a wide range of disciplines.

These changes in, and pressures on, research practices, are occurring at the same time as changes in the communication of research results. Communication through scholarly journals and the archiving of those journals has been the mainstay of a range of research communities for the last three centuries. The

advent of the World Wide Web has made possible a whole range of new forms of publishing (Treloar 1999), and the past decade in particular has seen a great deal of experimentation with new journal forms and new publishing models.

More recently, the open access movement has been particularly vigorous in proposing solutions to a number of concerns that have become obvious in the current system of scholarly communication: the serials crisis (the increasing subscription costs of scholarly journals), the relative inaccessibility of paper-based archival journals (the need to physically examine the paper based publication), and the permissions crisis (the way in which publishers impose restrictions on the use of material published under their imprints). These solutions can be seen as a response to the potentials latent in the advent of the World Wide Web.

In May 2005 a joint CNI-JISC-SURF invitational conference was held in Amsterdam with the title "Making the strategic case for institutional repositories". This conference emphasised the potential for repositories to move beyond the kinds of traditional publications that have been the concern of the open access movement to support innovative new forms of research and research output exposure. Some of the possibilities discussed were:

- life cycle management of research (from lab book to formal outputs to teaching)
- smart publications that link experiments, results, and a range of documents
- the ability to validate not only research conclusions, but also research results, by replication and comparison
- the ability to allow other researchers access to original raw data which might result in quite different discoveries (and possibly more important discoveries) by someone other than the generator of the original research; this can be seen as a form of post processing and knowledge mining
- the potential to shorten the publication cycle (time to release information about new research)
- environments that provide stronger support for authenticity, authority, and integrity of research.

All of these new possibilities also present new challenges in lifecycle management, attribution and provenance of the full set of research outputs, not just the conventional formal publication.

Theoretical basis

In responding to this changing context and its attendant challenges the DART project is drawing on three bodies of research into scholarly activity and communication. The relationship between these three realms (re-engineering scholarly communication, scholarship as ecology, and the information continuum) is shown in Figure 1 below. So what are the differences between these approaches?

Re-engineering scholarly communication

The first body of research arises out of nearly 15 years of examination of the current system of scholarly communication. Much of this activity has been focussed on the scholarly article as the main form of output. A recent article by Herbert van de Sompel from Los Alamos and others (Van de Sompel, et. al. 2004) argues that current experimentation within the publishing world is both limited and backwards-looking. The authors argue for two radical changes.

The first change is to deliberately engineer a new scholarly communication system that is intertwined with the process of generating new knowledge. Roosendaal and Geurts (1997) have already distinguished the following basic functions required from any system of scholarly communication:

- Registration (to resolve questions of precedence)
- Certification (of the validity of a scholarly claim)
- Awareness (to ensure that scholars are kept aware of new work)
- Archiving (to preserve the scholarly record over time)
- Rewarding (for measured performance according to discipline or system-specific metrics)

The researchers in Van de Sompel, et. al. (2004) build on this work to argue that we should decompose the current system of scholarly communication into a network of scholarly value chains. A repository (serving the registration and possibly certification functions) would be one hub on such a chain. Similar chains are also being developed for datasets in the Grid domains, through network-based services for data sharing and information storage.

The second change is to redefine what constitutes a unit of communication. Instead of just focussing on journal publications, they suggest that a new scholarly communication system should also accept “datasets, simulations, software and dynamic knowledge representations”. These units of communication should be capable of being aggregated into complex documents, which are themselves units of communication. They should also be able to be registered and preserved regardless of their nature or stage of development, thus facilitating collaboration, network-based research and faster discovery.

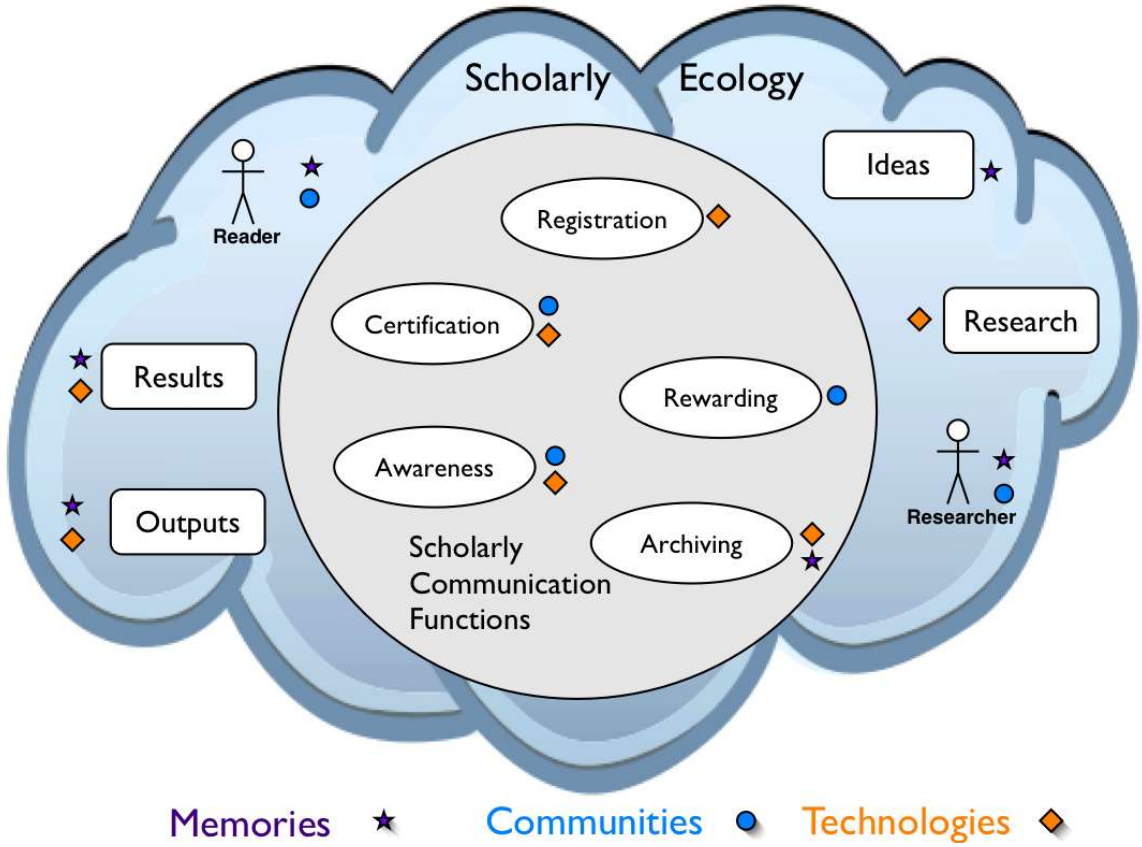
Van de Sompel, et. al. (2004) also identies the need to develop “information models, process models, and related protocols to enable interoperability among existing repositories, information stores and services”.

Scholarship as ecology

The second body of research draws on systems theory and sees scholarly communication as an ecology (Kaufer and Carley 1995). In this ecology, a communicative transaction is a cyclic process of interaction, communication and adaptation between actors and entities. Like any ecology, each member affects the others, and complex behaviours emerge in unpredictable ways. Kaufer and Carley then applied this model to the published output of a piece of research. But what happens if one considers the ecology around the entire process of

research? Instead of a fairly linear model leading from idea -> research -> publication output -> reader, one ends up with an ecology consisting of Actors (researchers and readers, who may well be the same person) and Entities (Ideas/Problems, Experiments/Research Activities, Results, Outputs). These actors and entities are all co-evolving, co-adapting and influencing one another. As in real ecologies, these influences occur in a very non-linear way, and changes emerge in ways that are difficult to predict in advance.

Figure 1: Inter-related bodies of theory



Information Continuum

The third body of work has been developed by a group of researchers in what was at the time known as the School of Information Management and Systems (now the Caulfield School of Information Technology) at Monash University. They have 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). One of the axes in the model relates to the processes that are applied to information:

- Create – the original idea
- Capture – in some fixed form
- Organise – in some storage and retrieval system

- Pluralise – extend the usefulness of the information object across time, space and community

This research team is now building on the information continuum work as part of a project called “Memories, Communities and Technologies”. This project is already looking at how knowledge is created and communicated in research communities. The memories are what the research community knows, encoded in tangible publications, memories and intangible practices. The technologies themselves support the storage and transmission of these memories, as well as the creation, use and re-use of knowledge. As the communities develop and transform themselves along the memory axis of the information continuum model, technology can be applied to enable this transformation. A very successful conference to explore these ideas, and how new e-research technologies (such as those being created by DART) might influence research practices (and vice-versa), was held at the Monash campus in Prato, Italy in early October 2006.

DART objectives

The DART project, drawing on the theoretical work described above, is undertaking development and integration work to support the evolving new paradigm for e-Research. It is doing so by addressing issues across the entire research continuum from the creation of the original research problem through to the pluralisation of the resulting work, its annotation by others, and its reuse in new research. It also recognises the need for data curation, which is defined as follows by the UK Digital Curation Centre:

> the actions needed to maintain digital research data and other digital materials over their entire life-cycle and over time for current and future generations of users … [including] all the processes needed for good data creation and management, and the capacity to add value to data to generate new sources of information and knowledge. (DCC 2005)

The specific objectives of the DART project are:

> to *support and enable researchers, end-users, and appropriate computer systems* to manage the creation and collection of data and to gain greater access to data and documents
>
> by *gathering, managing and archiving data and documents* and managing their access
>
> so that *researchers are more easily able to perform their work* and do so at a much higher level of insight and productivity than was previously possible,
>
> and so that the *Australian public has greater visibility of, and access to, publicly funded research*.

Figure 2 shows the rationale for the DART project. This figure draws on the work described in Van de Sompel (2004) and adds to this model the research process

itself, as well as the process of annotation. Figure 2 shows the current situation, the situation with the innovations that are being developed by the DART project, the benefits for researchers and the benefits for the general public.

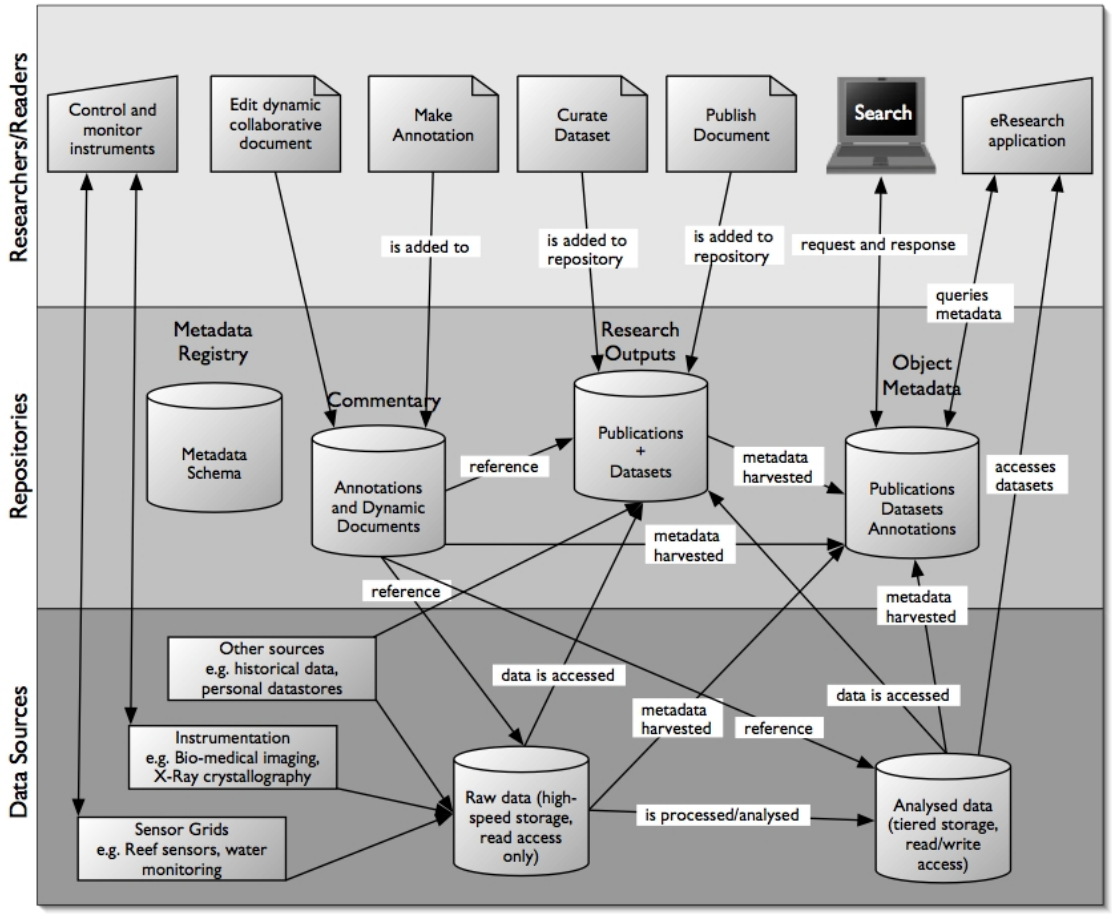
Figure 2: Potential benefits arising from the DART project

Scholarly Processes	Research	Registration	Certification	Awareness	Archiving	Annotation	Rewarding
Process Outputs	Data-sets; Publication	Timestamped Deposit	Quality rating	Discovery/ Alerting Database	Repository	Annotation Server	Tangible/ Intangible rewards
Without DART	• Poor curation • Fragmented collaboration • Poor support for sensors, large datasets	• Reliant on journal processes • Rarely possible for datasets	• Based on journal quality as proxy for article • Datasets problematic	• Hard to discover datasets and other digital objects	• Reliant on journals • Poor support for datasets	• No ability for annotation of publications or datasets	• Largely based on publications • Based on peer evaluations
With DART	• Data curation • Collaboration support • Support for eResearch	• Immediate registration • Datasets and other digital objects accepted	• Other quality measures possible • Digital objects rateable	• Datasets now treated in same way as publications	• Datasets now treated in same way as publications • Secure archive	• Annotation of publications or datasets by researchers and readers	• Now based on datasets and annotations • Visible to wider group
Benefits for Researchers	• More effective research • No data loss	• Guaranteed priority • Range of digital objects	• Better assessment of all research outputs	• Easier to locate and build on existing work	• Ability to locate archival datasets • No data loss	• Improved collaboration and validation • Faster communication	• More immediate feedback
Benefits for Public	• Better use of taxpayer funds • Improved research outcomes	• Improved efficiency • Visibility of priority claims	• Better visibility of quality measures for range of outputs	• More efficient research • Improved research outcomes	• Access to archived data • Improved visibility over research outputs	• Visibility into research process • Ability to annotate!	• Ability to influence rewards • Improved research outcomes

DART Design

Figure 2 answers the question of *why* DART is important. Figure 3 shows *how* the project has been designed to produce these benefits. In the uppermost layer are researchers, readers and computers programs. The middle layer shows the proposed repositories (including traditional publications as research outputs, and raw data) and the data flows between them and the datasets in the lowest layer. The lowest layer shows the data sources and their associated storage. The figure has been annotated to indicate some of the more significant data flows between major components.

Figure 3: DART Technology Architecture



For the purposes of implementation, the DART project has been structured as a number of inter-related thematically-grouped sets of work packages.

Data Collection, Monitoring and Quality Assurance

In this group of work packages, DART is tackling the issues surrounding high-rate and large-volume data streams, particularly those generated by instruments and sensors. There are a number of requirements that are unique to the challenges inherent in dealing with digital objects generated by and derived from instruments and sensors. They include:

- Two way communication with the instruments and sensors is desirable so that their status and information can be probed and monitored remotely.
- Quality assurance processes need to be transparent to the user despite variations between different instruments and sensors. A standard approach for detecting faulty or poor quality data early in the experiment can then be implemented.
- Triggering the download of data contained in the temporary data storage (data cache) into the permanent data storage is a non-trivial process, requiring the automation of metadata creation and data labelling and

indexing, in large volumes, sustainably, and without human intervention.

- Security and access to the instruments and sensors must be carefully managed. Unauthorized access could lead to tampering with the data at its source.

The DART project is basing this group of work packages on the Common Instrument Middleware Architecture (CIMA) (McMullen and Chiu 2005). This architecture emerged from work supported by the National Middleware Initiative. The CIMA architecture allows the connection of instruments and sensors to the internet, and makes them discoverable and their results publishable using web services or the open grid services architecture. In other words, this architecture allows DART to leverage the work being done by the National Middleware Initiative in the US and the Open Middleware Institute Initiative in the UK in using their core middleware for security, access, file transport, etc. Because the CIMA is based on international standards, middleware produced using this architecture will be re-usable in other projects. DART is working with the CIMA developers to extend the software and apply it to a wider range of instruments.

Storage and Interoperability

This group of work packages relates to the need to work with documents, datasets, simulations, software and dynamic knowledge representations in a secure way with controlled access. This includes collection from a range of devices, secure transfer across networks, storage on high-capacity devices, management and preservation in repositories, and maintaining the integrity of the datasets.

The digital objects that DART is storing need to be managed, preserved, persistently identified, aggregated and disseminated in flexible ways in order to deliver the improvements outlined in Figure 2. Of the available pieces of widely used repository software (Open Society Institute 2004), Fedora has been found by a range of projects to be the best match for these requirements. Fedora (Lagoze, et. al. 2005) is “an open source, digital object repository system using public APIs exposed as web services.” (Staples, Wayland and Payette 2003). Its architecture is very flexible, and provides significant advantages as a platform on which to build other applications. In particular, in a DART context it provides the ability to store and manage complex objects and the relationships within and between complex objects. The ARROW team at Monash University have been using Fedora for over 18 months now and are one of a small number of projects which are collaborating as part of the Fedora Developer Consortium.

The pre-eminent technology for working with large datasets is the Storage Resource Broker (SRB), developed at the San Diego Supercomputing Center (SDSC) (Moore 2004a, 2004b). SRB can be described as “client-server middleware that provides a uniform interface for connecting to heterogeneous data resources over a network and accessing replicated data sets. SRB is widely deployed across the world in demanding, large-capacity and high-bit rate environments. In conjunction with the Metadata Catalog (MCAT), it provides a way to access data sets and resources based on their attributes and/or logical

names rather than their names or physical locations. The SDSC SRB system is a comprehensive distributed data management solution, with features to support the management, collaborative (and controlled) sharing, publication, and preservation of distributed data collections. The SRB also serves as middleware via a rich set of APIs available to higher-level applications and by providing a management layer on top of a wide variety of storage systems." (http://www.sdsc.edu/srb/whatissrb.html).

In order to extend Fedora to work with large datasets, the DART project originally intended to integrate SRB with Fedora, both as a replacement storage layer for Fedora itself, and as a location for content outside a Fedora repository but managed by it. In practice, this work was undertaken by the Fedora developers and DART concentrated on complementary activities. In order to build more advanced knowledge mining services in the future, MCAT also needs to be semantically augmented using the Resource Description Framework (RDF). Core DART developers at JCU are also working to extend SRB by providing support for off-line data collection and by improving the user interface.

Content and Rights

An enormous amount of research data is currently stored within personal or private archives, either on researcher desktops or departmental/institutional servers. In these locations it is largely inaccessible to, and undiscoverable by, other researchers or the public. This group of work packages is investigating methods, incentives and technologies to motivate researchers to submit their research data and results into institutional repositories. This includes the development of:

- Simple user interfaces and workflows to enable researchers to easily deposit documents, research data and results into institutional repositories.
- Tools and services to enable researchers to easily select and attach standardized licenses defining access and re-use rights to their data and research results. These tools are based on the outcomes of both the Creative Commons and the more recent Science Commons initiative (http://sciencecommons.org/).
- Guidelines for information management best practice in research teams, arising from embedding information professionals into such teams as research partners.

Note that assessment technologies (see below) that support qualitative and quantitative assessment of research deposited within institutional repositories will also provide additional incentives for researchers to deposit their results.

Annotation and Assessment

This group of work packages relates to tools and services that enable peers to attach reviews, opinions, comments or assessments to research data, reports, publications etc. These annotation and assessment services can serve either as an alternative, or addition, to existing peer-review mechanisms. This can be seen

as a completely new certification function made possible through this new distributed networked environment. Two annotation approaches are being trialled. One builds on annotation research carried out at DSTC, looking at annotations that are managed and stored external to the digital objects. The other approach builds on work underway at JCU to create collaborative documents including annotations.

The first work package is concentrating on extending and refining existing annotation tools to enable annotation of digital objects held within the Fedora and SRB research repositories such as SRB, DSpace or Fedora. The second work package is concentrating on tools to support collaborative annotations, thus enabling research communities to document shared practices and assessments. This involves the refinement and deployment of a subset of the Vannotea software developed within DSTC's FilmEd project. Vannotea is designed to enable real-time annotation of complex digital objects (images, video, 3D objects) by geographically distributed groups within a videoconferencing environment (Schroeter et. al., 2003). A third work package is focussing on the development of secure authenticated access to annotation servers through the development of a Shibboleth-based interface to the W3C's open source Annotea server (Barstow et. al. 2001). This allows different groups who might want to annotate resources for different purposes (such as referees, grant committees, researchers) different levels of access.

A fourth work package was intended to involve piloting the use of hosted wikis linked to research data repositories to facilitate interaction between researchers and research groups. In fact, the decision was made early on to instead use a content management system based on Plone instead. This provides richer management and organization of content. As a proof of concept, the DART project is using itself as a testbed for this approach (eating our own dogfood, as the phrase has it).

It should be noted that the final publication is often seen as the only official research record worthy of capture and curation. Both the annotation and wiki technologies described above will allow for the capture of a record of some of the collaborative activity around the datasets and other research outputs.

Discovery and Access

This group of work packages relates to tools and services that enable researchers and readers to search, browse and discover resources within the repository and access them, either under controlled conditions or in an unrestricted way.

It involves the development of portals, based on Gridsphere, that will provide seamless search interfaces across distributed archives that are implemented in SRB and Fedora. Ontologies and the semantically-augmented MCAT RDF data store are being developed to provide semantic interoperability across heterogeneous metadata schemas. Shibboleth and PKI provides the authentication and access control.

In addition, one work package is developing and providing access to a centralized repository/registry of metadata schemas and ontologies. Metadata schema registries enable the publication, navigation and sharing of information about metadata. This registry will act as the primary source for authoritative information about recommended metadata schemas. It will enable the sharing and re-use of existing metadata schemas and application profiles – thus enhancing interoperability and reducing costs and effort. This work package will build on the open source work that has been developed within the JISC IE Metadata Schema Registry Project (IEMSR) by UKOLN and ILRT (http://www.ukoln.ac.uk/projects/iemsr/). DART will also work with other related projects towards ensuring that metadata in other repositories is managed and exposed in standards-compliant ways. This will enable later federation through work outside the scope of this bid.

DART Dimensions

At the time of writing this paper, the project has been under way for a bit over half its anticipated duration (the project did not effectively commence until December 2005, and has received permission to expend project funds through to the middle of 2007). DART is now a large and complex project with:

- DEST funding of A$3.235 million
- 3 partners:
    - Monash University (host) in Melbourne
    - University of Queensland in Brisbane
    - James Cook University in Townsville
- 5 technical areas of focus within the DART work packages (WPs)
- 7 Chief Investigators
- 27 Separate DART work packages
- 40+ project team members
- 18 months in which to do it all!

Lessons learned so far

So, what are the lessons that have been learned so far?

Demonstrators

When the project bid was submitted, it was conceived as focussing on the technologies, and only engaging with researchers for specific work packages. It soon became clear that this was not going to act as a sufficient testbed for the end-to-end lifecycle approach that DART was trying to exemplify. Nor was it going to ensure that the outputs of DART sufficiently met the needs of real researchers. Accordingly, the DART Board of Management decided in early 2006

to establish a number of demonstrator projects. The areas of activity were chosen across different disciplines, selecting researchers for each demonstrator at each of the three DART partners. The three research areas chosen as DART demonstrators are:

- X-Ray Crystallography
- Climatology
- Digital History

The use of demonstrators is designed to show the value of an end-to-end lifecycle approach and to test **proof-of-concept** outcomes from DART.

The DART demonstrator tasks are to:

- Engage with suitable researchers at each partner university for each of the three selected research areas
- Define the research activities applicable to each area
- Embed information management specialists into research teams
- Construct a custom designed prototype DART portal, incorporating software applications specific to each research discipline
- Progressively refine the model as the DART project adds new features and services

Experience with some of the research teams involved on the demonstrators to date has been somewhat mixed. Researchers are very time-poor (this will surprise no-one, I am sure!). They need to have services delivered that are of use to them, and they are quite careful about how they invest their time and attention. This means that using them as guinea-pigs without a good deal of careful thought will be neither a pleasant experience for them nor productive for the project. The creation of embedded information management specialists as a free resource for the teams was an attempt to ameliorate this. In some cases this has worked very well.

It is also the case that some of the tools and services that DART is working on are in areas where some of the research teams feel that they have things under control. As a result, they are sometimes reluctant to adopt new ways of doing things that they feel offer little or no additional benefit. This is, of course, an issue for many projects delivering e-research tools. Such tools will need to be delivered in a modular way to complement existing practices and/or offer significant advantages over those existing practices.

With the benefit of hindsight, the role of DART demonstrators should have been written into the bid from the beginning.

Integration

The second area where DART has needed to adapt the way it operates is in the area of integration across the various work packages. It was originally envisaged

that this could be split between the full-time project director (Dr Jeff McDonell) and the notionally half-time project architect (the author of this paper). In practice what happened was that the complexity of the project and its co-ordination took up all of the project manager's time. In addition, the project architect was diverted onto preparing a bid for a followup DEST funding round (this bid was funded as the Australian ResearCH Enabling environment, or ARCHER), working on the Australian Research Repositories Online to the World (ARROW) project and undertaking responsibilities in his day job (setting up a new Department of Information Management and Strategic Planning at Monash, completely revising the Monash ICT Strategic Plan for the next five years, and steering the implementation of the Monash University Information Management Strategy). Accordingly, it was decided to appoint a staff member from JCU (Frank Eilert) to act as Integration Architect for one day per week from the beginning of 2006. This provided to be a definite improvement, with a number of important events (including three co-ordination workshops) being organised. However, by the middle of 2006 it was clear that making progress on some of the technical integration (and making sure that the individual DART staff didn't become too narrowly focussed) would require a full-time staff member. A search was undertaken, and the successful applicant, Anthony Beitz, started in September 2006.

Once again, with the benefit of hindsight, this full-time role should have been included in the bid document and costing calculations.

Collaboration and Co-ordination

DART was always going to be an ambitious project. The challenges of managing work across 27 work packages, three sites, and 40 staff were never going to make for an easy operation. The bid document was designed to try to overcome this by having a full-time project director and three separate project managers at each site. For a variety of reasons, this was not what ended up happening. In addition, many of the Chief Investigators are also undertaking multiple projects. As a result, the project has had to work harder than expected at keeping groups talking, and working towards an integrated solution (rather than focussing on optimising individual components of the project). It is difficult to see how this can be improved (short of getting less busy people to drive the project, which doesn't seem like a good idea on other grounds!).

Conclusion

The DART project has just commissioned a mid-term review by an external consultant. His interim report was presented to the DART Board of Management in late September, and his final report was presented to the Board in November 2006. By the time of the EDUCAUSE conference itself, the project will be nearing its end, and the ARCHER project will be well underway. For these reasons, it is perhaps too early to write a conclusion. Probably all that can be said for now is the DART project, despite being very ambitious, appears to be progressing well in terms of individual work packages, and delivering useful tools to some of the

researchers it is engaging with. The full story of DART and what it delivered will need to wait for a later date.

References

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

Barstow, A, Kahan, José, Koivunen, M-R, Swick, R 2001, 'Annotea: A Generic Annotation Environment using RDF/XML', *WWW10 Developers Day*, Hong Kong, May.

DART 2006, *The Dataset Acquisition, Accessibility and Annotations e-Research Technologies Project*. Viewed 29 September 2006, http://dart.edu.au/.

DCC 2005, *Digital Curation Centre*, viewed 2005. <http://www.dcc.ac.uk/what.html>

Kaufer, D. S. and Carley, K. M. 1993. *Communication at a Distance: the influence of print on sociocultural organization and change*. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Lagoze, C., Payette, S., Shin, E. and Wilper, C. 2005, 'Fedora: An Architecture for Complex Objects and their Relationships', *International Journal of Digital Libraries: Special Issue on Complex Objects.* <http://www.arxiv.org/abs/cs.DL/0501012>

McMullen, R. and Chiu, K. 2005, 'CIMA: Scientific Instruments as First Class Members of the Grid', *Remote Access and Automation Workshop,* Sydney*.* <http://mmsn.chem.usyd.edu.au/events/mmsn_ws05_abstracts.html#McMullen>

Moore, R. 2004a, 'Integrating Data and Information Management', *International Supercomputer Conference*, June <http://www.sdsc.edu/dice/Pubs/ISC2004.doc>.

Moore, R. 2004b, 'Evolution of Data Grid Concepts', *Global Grid Forum Data Area Workshop*, January, <http://www.sdsc.edu/dice/Pubs/Grid-evolution.doc>.

Open Society Institute, 2004, *A Guide to Institutional Repository Software v 3.0*. <http://www.soros.org/openaccess/software/>

Roosendaal, H., and Geurts, P. 1997, 'Forces and functions in scientific communication: an analysis of their interplay', *Cooperative Research Information Systems in Physics*, August 31—September 4, Oldenburg, Germany. <http://www.physik.uni-oldenburg.de/conferences/crisp97/roosendaal.html>.

Schauder, D., Stillman, L., & Johanson, G. 2004, 'Sustaining and transforming a community network. The Information Continuum Model and the Case of VICNET', *CIRN 2004: Sustainability and Community Technology*, Monash University, Prato, Tuscany, Italy. <http://www.ciresearch.net/conferences/viewabstract.php?id=68&cf=4>.

Schroeter, R., Hunter, J., Kosovic, D. 203, 'Vannotea - A Collaborative Video Indexing, Annotation and Discussion System for Broadband Networks', *Knowledge Markup and Semantic Annotation Workshop*, K-CAP 2003, Sanibel, Florida October <http://metadata.net/filmed/pub/Vannotea.pdf>.

Staples, Thornton, Wayland, Ross & Payette, Sandra 2003, 'The Fedora Project: an open-source digital object repository management system, *D-lib Magazine*, April, <http://dlib.org/dlib/april03/staples/04staples.htm>.

Treloar, A, 1999. *Hypermedia Scholarly Publishing: the Transformation of the Scholarly Journal*. PhD Thesis, Monash University. <http://andrew.treloar.net/research/theses/phd/index.shtml>

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

Waters, D. 2003. *Cyberinfrastructure and the Humanities*. Fall Task Force Meeting of the Coalition for Networked Information. >http://www.cni.org/tfms/2003b.fall/handouts/Fall2003Handouts/H-Watersplenary.doc.>