

The Dataset Acquisition, Accessibility, and Annotation e-Research Technologies Project (DART)

Lead institution and grant recipient:

Monash University (MU)

Partner institutions:

James Cook University (JCU), University of Queensland (UQ) and DSTC Pty Ltd, operating the CRC for Enterprise Distributed Systems Technology (DSTC¹)

Team leader:

Dr Andrew Treloar, Senior Project Manager (Information Management) and ARROW Technical Architect, Information Technology Services, Monash University. Telephone +61 3 990 51138. Facsimile: +61 3 990 59888. Email: Andrew.Treloar@its.monash.edu.au.

Contact officer:

Dr Andrew Treloar, Senior Project Manager (Information Management) and ARROW Technical Architect, Information Technology Services, Monash University. Telephone +61 3 990 51138. Facsimile: +61 3 990 59888. Email: Andrew.Treloar@its.monash.edu.au.

Summary statement:

The DART project will undertake a co-ordinated programme of eResearch requirements analysis, software development, policy and guideline creation and prototyping to investigate how best to:

- collect, capture and retain large data sets and streams from a range of different sources;
- deal with the infrastructural issues of scale, sustainability and interoperability between repositories;
- support deposit into, access to, and annotation by a range of actors, to a set of digital libraries which include publications, datasets, simulations, software and dynamic knowledge representations;
- assist researchers in dealing with intellectual property issues during the research process;
- adopt next-generation methods for research publication, dissemination and access.

Amount of Funding Sought:

The programme of activities for this project will be carried out in a series of inter-related work packages. Some of these work packages will complete and deliver production services/outcomes prior to 2006. Others will only be able to develop proof of concept implementations. The costs to provide full production services will be sought from NCRIS, with work to commence in 2007.

Total funding sought from this programme: \$3,237,000.

¹ See Appendix F: Project Governance for an explanation of DSTC's involvement post June 2006.

Table of contents

BUDGET	ii
Project Work Package Costs	ii
Project Shared Costs	ii
Budget Summary	ii
In Kind Contributions	ii
TABLE OF CONTENTS	i
1 BACKGROUND AND CONTEXT	1
1.1 Background	1
1.2 Context	2
2 OBJECTIVES	2
3 PROJECT OUTLINE	4
3.1 Data Collection, Monitoring and Quality Assurance	4
3.2 Storage and Interoperability	5
3.3 Content and Rights	6
3.4 Annotation and Assessment	6
3.5 Discovery and Access	7
4 INTENDED OUTCOMES	7
4.1 Project deliverables	7
4.2 Future work	8
4.3 Long-term sustainability	8
5 USEFULNESS OF PROJECT	8
5.1 Maximising access to digital resources in Australian universities, especially regional universities	9
5.2 Creating new types of digital libraries to manage extremely large data sets.	9
5.3 Providing effective linkages between sets of research information to enable seamless access by researchers.	9
5.4 Adopting a national approach to improving open access to the results of publicly funded research.	10
6 INFORMATION DISSEMINATION	10
7 IMPLEMENTATION TIMETABLE	10
APPENDIX A: DETAILED WORK PACKAGES	11
Data Collection, Monitoring and Quality Assurance	11
Storage and Interoperability	14
Content and Rights	18
Annotation and Assessment	21
Discovery and Access	23
APPENDIX B: REFERENCES	25

APPENDIX C: KEY ACRONYMS AND INITIATIVES	27
APPENDIX D: THEORETICAL FRAMEWORK	29
APPENDIX E: CAPABILITY STATEMENT	30
CRC for Enterprise Distributed Systems Technology (DSTC)	31
James Cook University	32
Monash University	32
University of Queensland	33
Governance issues	35
Governance structure	35
Statement regarding role of CRC for Enterprise Distributed Systems Technology	35
APPENDIX G: LETTERS OF SUPPORT	36

1 Background and context

1.1 Background

In order to meet the emerging needs of e-Researchers, the DART project proposes a comprehensive end-to-end approach to developing a new system for managing research activity and publication. This will have far-reaching impacts in many places. The DART proposal seeks to respond to rapid and ongoing changes in the way research is carried out, and the way its results are communicated. The call for proposals 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 a whole range of new forms of publishing possible (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.

Recently (May 10/11, 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 (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.

1.2 Context

The DART proposal builds 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 four DART partners. It does this by extending its concerns into the areas of large datasets and sensors, as well as annotation technologies and collaborative, composite documents. In particular, DART seeks to investigate the most appropriate response to the challenges inherent in:

- new forms and producers of raw data
 - particularly high-rate and large-volume data streams
- new forms of collaborative research activity
 - particularly distributed multi-disciplinary teams
- new forms of publication
 - not just journal articles, but also datasets, simulations, software and dynamic knowledge representations (see Van De Sompel 2004 for more on this)
- new forms of research validation
 - extending peer-review and allowing for annotation of online information resources, including all of the above forms of publication
 - access to the original data for appropriately authorised readers and researchers.

In doing this, DART is able to draw on the differing needs and discipline perspectives of a range of research communities, including the following disciplinary and interdisciplinary groups:

- climate research at Monash University²
- environmental/water research at the University of Queensland
- tropical and marine sciences at James Cook University (the reef monitoring sensor network)
- biomedical imaging at Monash University and the University of Queensland
- protein crystallography/structural biochemistry at Monash University
- historians and social science researchers at Monash University and the University of Queensland.

The Dataset Acquisition, Accessibility, and Annotation e-Research Technologies (DART) project will draw on the embedded expertise and unique attributes of each of its four partners:

- JCU brings a strong regional focus and expertise in marine sciences and sensor grids to the project
- DSTC brings expertise in distributed systems, international web and metadata standards, large-scale digital libraries and annotation and knowledge mining technologies for the cultural, scientific, health and educational domains
- UQ brings expertise in workflow technologies, database research and large dataset management
- Monash brings a growing focus on e-Research, expertise in high-bandwidth networking, grid computing research, large-scale infrastructure, the location of the technical base for the ARROW project, and legal expertise in intellectual property, information security and privacy issues.

The DART team (see Appendix E for Capability Statements) has been carefully chosen to include people who have maturity in leading research efforts in related and highly relevant fields. There are a number of other groups worldwide working on the same or similar problem spaces. The DART team is very well connected into this community, and can contribute to the worldwide effort in a major way. This will assist in eliminating duplication and help Australia remain a first-class member of global research and e-research communities in a wide range of fields.

2 Objectives

This proposal, drawing on the theoretical work described in Appendix D, seeks to support the evolving new paradigm for e-Research 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 does so, within a holistic framework, by tackling issues related to the various new forms (of data, publication, etc.) discussed above. It also recognises the need for data curation, which is defined as follows by the UK Digital Curation Centre:

² led by Professor Amanda Lynch (a Federation Fellow)

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. (<http://www.dcc.ac.uk/what.html>)

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 1 shows the rationale behind the DART project. This figure draws on the work described in Van de Sompel (2004), which re-conceptualises the processes that take place in scholarly communication, and adds to this model the research process itself, as well as the process of annotation. Figure 1 shows the current situation, the situation with the innovations that will be delivered by the DART project, the benefits for researchers and the benefits for the general public.

Scholarly Processes	Research	Registration	Certification	Awareness	Archiving	Annotation	Rewarding
Process Outputs							
Without DART	<ul style="list-style-type: none"> Poor curation Fragmented collaboration Poor support for sensors, large datasets 	<ul style="list-style-type: none"> Reliant on journal processes Rarely possible for datasets 	<ul style="list-style-type: none"> Based on journal quality as proxy for article Datasets problematic 	<ul style="list-style-type: none"> Hard to discover datasets and other digital objects 	<ul style="list-style-type: none"> Reliant on journals Poor support for datasets 	<ul style="list-style-type: none"> No ability for annotation of publications or datasets 	<ul style="list-style-type: none"> Largely based on publications Based on peer evaluations
With DART	<ul style="list-style-type: none"> Data curation Collaboration support Support for eResearch 	<ul style="list-style-type: none"> Immediate registration Datasets and other digital objects accepted 	<ul style="list-style-type: none"> Other quality measures possible Digital objects rateable 	<ul style="list-style-type: none"> Datasets now treated in same way as publications 	<ul style="list-style-type: none"> Datasets now treated in same way as publications Secure archive 	<ul style="list-style-type: none"> Annotation of publications or datasets by researchers and readers 	<ul style="list-style-type: none"> Now based on datasets and annotations Visible to wider group
Benefits for Researchers	<ul style="list-style-type: none"> More effective research No data loss 	<ul style="list-style-type: none"> Guaranteed priority Range of digital objects 	<ul style="list-style-type: none"> Better assessment of all research outputs 	<ul style="list-style-type: none"> Easier to locate and build on existing work 	<ul style="list-style-type: none"> Ability to locate archival datasets No data loss 	<ul style="list-style-type: none"> Improved collaboration and validation Faster communication 	<ul style="list-style-type: none"> More immediate feedback
Benefits for Public	<ul style="list-style-type: none"> Better use of taxpayer funds Improved research outcomes 	<ul style="list-style-type: none"> Improved efficiency Visibility of priority claims 	<ul style="list-style-type: none"> Better visibility of quality measures for range of outputs 	<ul style="list-style-type: none"> More efficient research Improved research outcomes 	<ul style="list-style-type: none"> Access to archived data Improved visibility over research outputs 	<ul style="list-style-type: none"> Visibility into research process Ability to annotate! 	<ul style="list-style-type: none"> Ability to influence rewards Improved research outcomes

Figure 1: Improvements to Scholarly Processes arising from the DART Project

In this DART proposal, dynamic publications are one of a number of digital objects to be included in the consideration, because they represent a significant emerging form of communication among the collaborating researchers. This aspect was not considered in the original ARROW project as the focus at that time was more on the deposit of, and access to, published articles. In addition, the DART project also deals with the requirements for large-scale data collection and curation, which was completely outside the ARROW project scope. Here the DART project can also build on the work being undertaken in the UK by the eBank UK project (www.ukoln.ac.uk/projects/ebank-uk/) which is investigating the issues surrounding the provenance, use and reuse of original data for research and learning purposes as well as DSTC's PANIC and FUSION projects.

Figure 1 answers the question of *why* DART is important. Figure 2 shows *how* the project will 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 with the work packages that are involved for each component. For instance, the process of editing dynamic collaborative documents is described in work package AA4. Details of the work packages can be found in Appendix A.

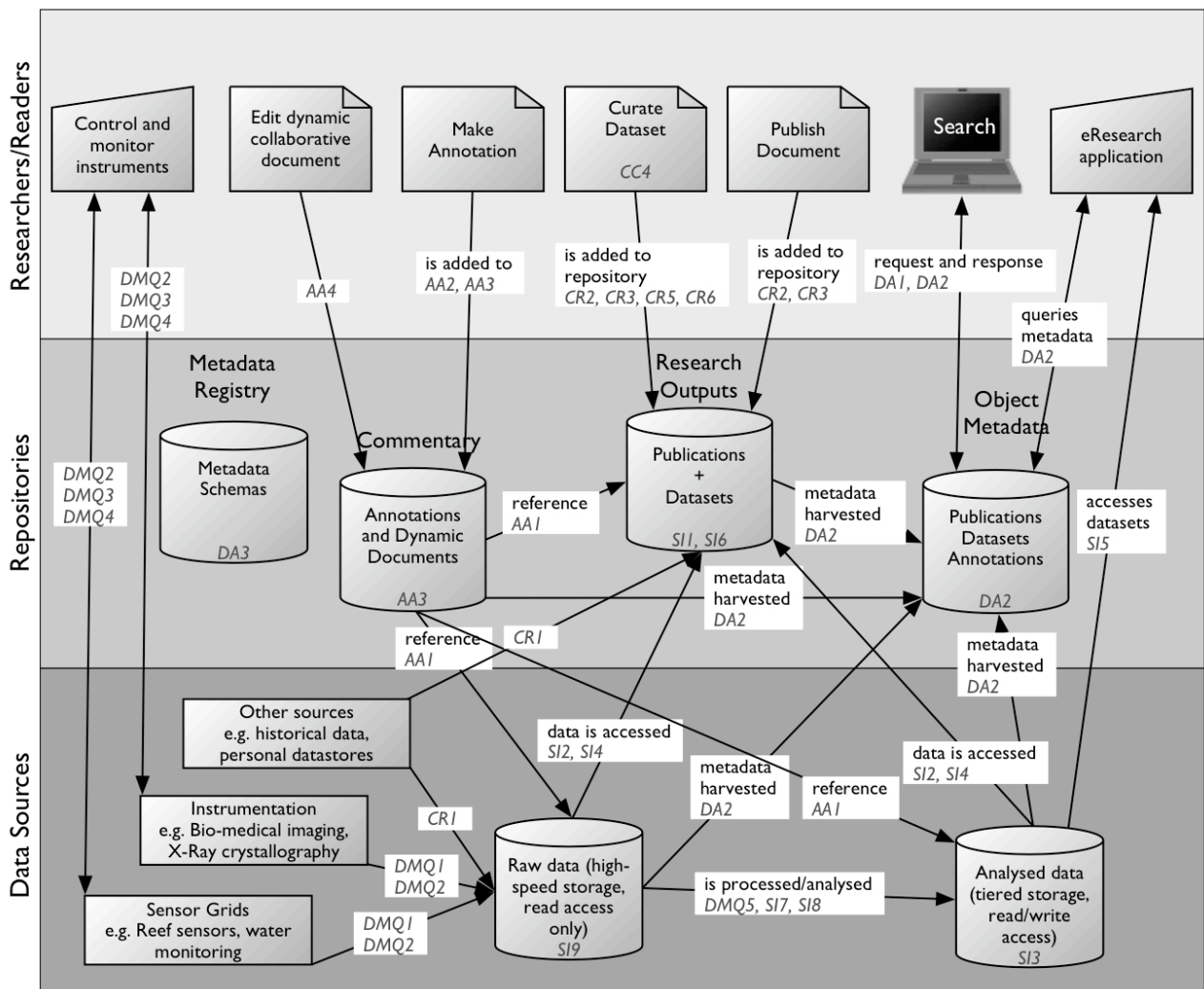


Figure 2: DART high-level architecture (codes indicate relevant DART work packages)

The proposed project thus substantially extends and enhances the focus of the FRODO projects on research outputs to include the needs of dataset creation, acquisition, management, and curation, as well as providing support for collaborative research practices. This aspect of research information management is an opportunity afforded by the advent of the World Wide Web to transform in a fundamental manner the traditional scholarly publication environment to include pre-publication collaboration, as well as access to data sets, software, and/or commentaries, and annotations, thus providing a much richer research collaborative fabric than was imaginable even 10 years ago.

3 Project outline

The DART project will be structured as a number of inter-related thematically-grouped sets of work packages. For each set of work packages (details of each package in Appendix A), this submission provides below an overview of the functionality to be developed, and the rationale for its inclusion,

3.1 Data Collection, Monitoring and Quality Assurance

In this group of work packages, DART proposes to tackle 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:

- Two way communication with the instruments and sensors 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. This 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. Unauthorized access could lead to tampering with the data at its source.

The DART project has chosen to base this group of work packages on the Common Instrument Middleware Architecture (CIMA). This architecture emerged from work supported by the National Middleware Initiative specifically for connecting instruments and sensors to the Internet. 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.

3.2 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 will be storing need to be managed, preserved, persistently identified, aggregated and disseminated in flexible ways in order to deliver the improvements outlined in Figure 1. 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. DSTC staff also have experience working with Fedora and have also been collaborating with the Fedora developers at Cornell (Doerr, et. al. 2003).

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/whatisrb.html>). In order to extend Fedora to work with large datasets, the DART project will need 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 order to build more advanced knowledge mining services in the future, MCAT also needs to be semantically augmented using the Resource Description Framework (RDF).

It will regardless of whether the datasets are stored in Fedora or SRB, it will be advisable to be able to monitor changes to datasets through something like an audit log. This would provide greater assurance of the validity and quality of the data.

3.3 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 will investigate methods, incentives and technologies to motivate researchers to submit their research data and results into institutional repositories. This will include 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 will be 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 section 3.4) that support qualitative and quantitative assessment of research deposited within institutional repositories will also provide additional incentives for researchers to deposit their results.

3.4 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 will be 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 will concentrate 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 will concentrate on tools to support collaborative annotations, thus enabling research communities to document shared practices and assessments. This will involve the refinement and deployment 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 will focus 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 will allow 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 will involve piloting the use of hosted wikis (see below) linked to research data repositories to facilitate interaction between researchers and research groups.

A wiki is a relatively new web collaboration tool that can be a genuinely innovative and useful tool for research collaboration and information management. A key concept of the wiki is that it makes it very easy for any user to easily add content, files and other digital objects to a wiki environment without any need to understand web authoring technologies. Essentially wikis start as empty repositories, and are built, dynamically by the users of the community. Because rights management (groups etc.) can be easily added to the wiki, and they are web based, they also form the basis of an effective information dissemination system. A case study may be instructive. The predictive mineral discovery Cooperative Research Centre (pmd*CRC) is one of the most geographically distributed and diverse CRCs within the program. It also has a very direct industry/outcome focus. For the last 18 months the pmd*CRC has very effectively augmented its conventional face-to-face (and more recently AccessGrid facilitated) meetings with what is now a very sophisticated wiki framework. Users from any part of the CRC can read about the work/plans/outputs of any part of their program within the CRC. However, they can also look at raw data / images / calculations / analysis that have been conducted and uploaded into the format. Any user (with appropriate access permission) can add content, comment or provide feedback and generally interact and be informed of the progress within the entire CRC. JCU has a nascent program to model a series of discipline specific wikis as group/department level tools. The key feature is that a wiki that supports a single research group can – with

almost no effort – be cloned and support an entire school. Ultimately this process can continue without limit so that almost any type of scale of collaboration can be supported. DART will build on this and integrate wiki technology with the Storage Resource Broker (SRB). This combination will allow a wiki to act as an interface to, and commentary on, distributed file systems, data grid management systems, digital libraries and semantic webs.

At present, the final publication is seen as the only 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.

3.5 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 will involve the development of portals that provide seamless search interfaces across distributed archives implemented in SRB and Fedora. Ontologies and the semantically-augmented MCAT RDF data store will be developed to provide semantic interoperability across heterogeneous metadata schemas. Shibboleth and PKI will provide the authentication and access control.

In addition, one work package will develop and provide 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 software tools being 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.

4 Intended outcomes

4.1 Project deliverables

The DART project will deliver, as proof of concepts, a wide range of benefits for the Australian research community and for users of research outputs. Details for all work packages can be found in Appendix A.

The Data Collection, Monitoring and Quality (DMQ) Assurance work packages will:

- Connect instruments and sensors effectively to the network (work package DMQ1) – details for this and all other work packages at Appendix A)
- Ensure effective and reliable connection of selected instruments and sensors to storage repositories (DMQ2)
- Ensure that data from instruments and sensors is of sufficient quality to warrant curation (DMQ3)
- Provide online, remote access to pilot working instruments/sensors (DMQ4)
- Improve the intelligence of the storage framework (DMQ5).

The Storage and Interoperability (SI) work packages will:

- Facilitate distributed data management using Fedora (SI1)
- Investigate the appropriateness of SRB for dataset management and improve interoperability between Storage Resource Broker (SRB) based environments and Fedora repositories (SI2)
- Support richer descriptive and preservation metadata for dataset objects to enable more effective discovery (SI3)
- Provide a secure service for transferring data from sensors/instruments to repositories using Grid security (SI4)
- Develop an abstraction layer that supports a range of data replication systems (SI5)
- Allow simulation data to be retrieved from repositories or regenerated dynamically using computational services (SI6)
- Develop a cost-effective data pre-processing system for the secondary storage (SI7)
- Pilot long-distance high speed and secure data transfer between repositories (SI8)

- Scope and pilot storage infrastructure requirements (SI9).

The Content and Rights (CR) work packages will:

- Move data from personal data stores to secure trusted alternatives (CR1)
- Reduce barriers to content acquisition by providing more rights-assignment options for non-science researchers (CR2)
- Reduce barriers to content acquisition by providing more rights options for science researchers (CR3)
- Improve information management practice in research communities (CR4)
- Make it as easy as possible for researchers to deposit datasets and other digital objects into institutional repositories (CR5)
- Improve content deposit rates by clarifying legal issues around IP, information security and privacy (CR6).

The Annotation and Assessment (AA) work packages will:

- Allow researchers to annotate each other's works (including datasets) (AA1)
- Improve rates of annotation by allowing end-user control over who can annotate what and who can access the annotations (AA2)
- Help collaborative annotation services contribute to the life and productivity of research communities (AA3)
- Foster emerging wiki-based collaborative work practices in research teams (AA4).

The Discovery and Access (DA) work packages will:

- Improve repository deposit rates, sharing and reuse by allowing end-user control over who can access what (DA1)
- Improve repository deposit rates, sharing and re-use by improving discoverability (DA2)
- Reduce wasted effort in creating metadata schemas and improve interoperability of metadata schemas (DA3).

4.2 Future work

Some of the above work funded by DART will deliver proof-of-concept implementations only. This will need to be further developed and tested in a wider context to ensure that it supports the range of disciplines across the Australian research community as well as anticipated load expansion. The DART project therefore anticipates seeking NCRIS funding to further develop its activities beyond the end of the ARIIC funding in 2006. In the detailed work packages (see Appendix A), such future work is clearly marked with the letters NCRIS.

4.3 Long-term sustainability

Beyond the grant period, the sustainability of the DART initiatives will be determined by whether they are seen as useful by the communities they serve, by how much they are embedded into the ongoing activities of Australian research agencies (universities, CSIRO, CRCs), and by the extent to which they are built into existing repository initiatives. If, as expected, the demonstration project can show that concepts expounded in the previous sections are integrable and may be served by the combination of Fedora/SRB, then it is likely that such an architecture, and its associated mechanisms will be adopted as the mainstay of institutional repositories. Monash has already committed to implementing an institutional repository as one way to support research outcome dissemination, facilitate public access to research outcomes, and retain institutional knowledge. Hence, the long term sustainability of such a proposal is not in question as far as the institutional repository components are concerned. The National Library of Australia has committed itself to continuing with the national research discovery service available at <http://search.arrow.edu.au/>, and has agreed to absorb some extensions to this service as envisaged in this bid, with some additional resourcing. The main new activity requiring an ongoing home is the proposed annotation service which will be prototyped in the DART timelines. NCRIS funds will be sought to put it into production, but without this support the prototype will still provide useful information for the development of e-Research in Australia.

5 Usefulness of project

The DART project will comply with the requirements outlined in Call for Proposals as detailed below.

5.1 Maximising access to digital resources in Australian universities, especially regional universities

Text from Call for Proposals	DART Compliance Statement
“Regional universities”	DART will increase the accessibility to data and documents for researchers working in regional universities. JCU’s involvement draws on its significant expertise and acts a demonstrator for other regional universities.
“Exchange of information between research groups”	DART will enable better discovery from, and controlled access into, repositories and databases.
“Management of research resources, data and results”	DART will support deposit into well managed open access institutional repositories, thus providing effective long term curation and management of resources, data and results, as well as their associated metadata.
“Storage and preservation of research output”	DART will facilitate secure management of research outputs rather than existing ad hoc/individual curation practices.
“Accessibility and re-use of research results”	DART will provide support for new forms of scholarly collaboration and re-use through linked publications and distributed annotation, and through making data and datasets widely available and discoverable.

5.2 Creating new types of digital libraries to manage extremely large data sets.

Text from Call for Proposals	DART Compliance Statement
“Enable institutions to link data from diverse and large multi-centred national and international data sources more efficaciously, economically and efficiently than is possible to date”	DART will develop a range of integration mechanisms that enhance interoperability between repositories and data sources and assist easier linking.
“Provide solutions whereby the research outcome becomes an interim document that bundles primary data, statistical processing methods, figures and derived data together with the textual documentation and conclusions”	DART will enable the integration of non-research outputs, including sensor data, with more traditional outputs (publications). By storing these outputs, together with their associated semantic descriptors and annotations, in secure repositories with persistent identifiers, it will be much easier to bundle research outcomes into the interactive media-rich documents envisaged in the Call for Proposals. DART will also investigate ways to combine interim results and research outcomes into interactive documents based on wiki technology.

5.3 Providing effective linkages between sets of research information to enable seamless access by researchers.

Text from Call for Proposals	DART Compliance Statement
“Providing effective linkages between research information”	<p>DART will enable this by providing secure storage, persistent identifiers, standardized semantically-rich metadata and controlled access (where required).</p> <p>DART will also make it easier to discover research information to link to in the first place through an enhanced national research discovery service.</p> <p>DART will enable linkages between the heterogenous objects</p>

	related to a set of research activities: datasets, publications, annotations, etc through semantic inferencing.
Enabling “seamless access by researchers in discovering information quickly and easily”	DART will enhance existing research discovery services to enable this, and will implement a system of persistent identifiers for deposited research results to reduce the likelihood of broken links.
Implementing “web-visible, predominantly open-access [repositories], well-structured and conforming to common international standards and protocols”	The proposed DART repositories will comply with all of these criteria for published output, and all standards criteria for restricted access research data.
“Explore the complexity of authorisation models to encourage a more flexible approach to providing researcher access while applying high levels of security”	DART will do this by exploiting the power of XACML and building on the Shibboleth work already being undertaken by the MAMS project.
Provide “innovative ways to manage intellectual property in the digital environment”	This is at the heart of the DART project. In addition, DART will look at the applicability of the Creative Commons and Science Commons work to the identified problems of intellectual property assignment and copyright in the dataset and research results domains.

5.4 Adopting a national approach to improving open access to the results of publicly funded research.

Text from Call for Proposals	DART Compliance Statement
“Expose research output through open access repositories”	DART will provide support for adding research output to open access repositories based on ARROW technology.
“a national platform for acquiring, sharing and integrating research data”	DART will contribute to this by demonstrating the viability of a distributed, open-standards based approach.

6 Information dissemination

DART seeks to disseminate its results as widely as possible, both within Australia and internationally. This dissemination will be ongoing throughout the project, guided by a Communication Plan. Specific communication channels will include:

- DART website: the website will be regularly updated by the project Manager and project Management Partners to provide news of project developments.
- Announcements: important milestones will be announced on the project web site, relevant email discussion lists and via RSS feeds.
- Publications: publications about the technology, theoretical approaches, guidelines and lessons learned will be submitted to relevant conferences, workshops and journals.
- Conference and Meeting presentations: key conferences and international standards meetings in relevant areas will be used as dissemination opportunities.
- Wider community: findings will be disseminated through close collaboration with the communities identified in Section 1.2, as well as other research communities
- Road show: at the mid-point and end of the project, the project team will present its findings nationally through a road show.
- Collaboration with the ASK-OSS project (if funded) to disseminate findings and advice to potential adoptees.

7 Implementation Timetable

The timetable for the implementation of the work packages in this project is shown in the Milestones section for each work package.

Appendix A: Detailed work packages

This appendix provides the detail for the work packages. The format of presentation is as follows:

- an identifier
- the high-level objectives
- who is the lead institution
- the activity statement
- the deliverables
- a list of possible issues that may need consideration
- partners who will be involved with this work package
- and milestones.

Because the time when funding will be allocated is not yet known, dates have been provided in terms of S+X or E-X, where S = project Start time, E = project End time, and X is a number of months.

Data Collection, Monitoring and Quality Assurance

Work package DMQ1

Objective	Connect instruments and sensors effectively to the network.	
Responsible	JCU	
Partners	Monash	
Deliverables	A set of policies that govern the connection of instruments and sensors to the network. Note that this work package requires extensive discussions with the instrument and sensor owners in various research communities. Considerable time may be taken before a policy framework can be ratified.	
Activity Statement	Identification and formalisation of policies. A security model will be developed that supports user authentication and authorization requirements at storage and compute facilities and role-based authorization requirements for instrument facilities. A mapping facility similar to the X.509 certificate/Globus gatekeeper/grid map file arrangement used to control access to Globus compute sites will be developed to perform user to role mapping and authorization for instruments and sensors on the grid.	
Issues	Availability of instruments, and sensors. Sharing of resources.	
Milestones	Policies formulated.	S + 6
	Security layer mapped and operational.	S + 18

Work package DMQ2

Objective	Ensure effective and reliable connection of selected instruments and sensors to storage repositories (SRB) via CIMA middleware, and efficient use via changed work practices.	
Responsible	JCU	
Partners	JCU, Monash, UQ, CIMA group (Indiana University)	
Activity Statement	Scoping of initial instruments/sensors; target groups marine sensors (JCU), crystallography (Monash, UQ), water (UQ).	

	Determine details of how users expect to interact with selected facilities as regards remote interaction, data handling and overall experiment design and execution. These usecases will drive the development of the instrument's grid interface and details of the security implementation. Design and implement CIMA/Web services for the instruments.	
Deliverables	Instruments connected to the network, and operating correctly as web services.	
Issues	Physical connections, proprietary software, power supply compatibility. Note that this work package needs to work with various selected research communities. Each community will have their own instrument/sensor and work case requirements. Hence considerable time may be taken to connect these instruments online.	
Milestones	Scoping of instruments and work case assessment.	S+3
	An initial selection of instruments (HF surface ocean radar) and sensors (ocean temperature and weather) CIMA enabled and connected to the data storage via SRB.	S+12
	All intended instruments and sensors are connected to the data repositories .	S + 18

Work package DMQ3

This work package will assess the quality of the digital objects obtained from instruments and sensors, check the calibration of instruments and sensors, validate the instruments and data they generate on ingestion.

Objective	Ensure that data from instruments and sensors is of sufficient quality to warrant curation.	
Responsible	JCU	
Partners	Monash, UQ Core Instrument developers (including JCU: Peter Ridd, Mal Heron, AIMS), CIMA group	
Deliverables	The instruments and sensors will be properly calibrated to work in their operating range. A set of procedures and metadata framework to describe instrument health/calibration and data status.	
Activity Statement	Develop a metadata framework that supports instrument health and calibration status. Generate workflows or procedures that can assess via historical comparison, or metadata the quality of the incoming data stream and tag the data appropriately.	
Issues	Two-way communications require reliability, response of the instrument, proprietary data formats from manufactures. Requires triggers built into SRB. Processing requirement for very high rate data streams.	
Milestones	All instruments and sensors report calibration status in metadata tags associated with the data stream.	S + 12

Work package DMQ4

A significant benefit of grid-enabling the instruments comes from the ability to remotely access and in certain cases control these instruments using the network, sometimes called tele-instrumentation. This is a separate issue from accessing the data streams received from sensors/instruments addressed in packages DMQ1-3 above.

Objectives	Provide online, remote access to pilot working instruments/sensors.
-------------------	---

Responsible	UQ	
Partners	JCU, Monash	
Deliverables	Online access / remote operation to pilot instruments and sensors from remote locations.	
Activity Statement	CIMA supports streaming data from sensors to multiple consumers on the network, creating virtual, composite instruments by aggregating control and data functions into a single interface. For high bandwidth data significant issues result from the inability of databases to cope with the volume of data - it is either lost or received out of order. Use needs to be made of high throughput data stores in order to approach near-realtime data processing. Such processing is required in order to process realtime data from low-latency instruments or put in place real-time data-casting.	
Issues	Complex software engineering is necessary in order to put in place the real-time manipulation of data streams. Pilot systems and demonstrators will only be developed in the life-time of this project. To deliver a production service from the work on this package is likely to take more than 2 years for a basic level of production service with support for tele-presence and virtual laboratory framework.	
Milestones	Pilot demonstrator of a tele-presence / tele-instrumentation system.	S+18
	Establishment of remote access and control for coral reef network.	NCRIS
	Access Grid for tele-presence and virtual laboratory services.	NCRIS
	Establishment of remote control for the synchrotron beam-lines.	NCRIS

Work package DMQ5

Triggers are a concept whereby the system responds to object manipulation events by executing a script or other code (and is used in database applications). However, triggers are not available within SRB, but adding this ability to respond to file events will allow for the evolution of the data grid from a static storage medium, to an automatic processing framework. System level triggers could monitor object events such as file ingestion, renaming, addition of meta data, change of permissions, etc, and respond by executing an arbitrary workflow, automatically keeping the system in a consistent state, and performing required batch mode processing automatically.

Objectives	Improve the intelligence of the storage framework.	
Responsible	JCU	
Partners	JCU, Monash, UQ, SRB group: San Diego Supercomputing Centre	
Deliverables	Event triggers built into SRB to trigger 'events' on arrival of desired data objects.	
Activity Statement	Enhance SRB by building event triggers into SRB. Triggers are a concept where by the system responds to object manipulation events by executing a script or other code; Triggers will allow for the evolution of the data grid from a static storage medium, to an automatic processing framework. System level triggers can monitor object events such as file ingestion, renaming, addition of meta data, change of permissions, etc, and respond by executing an arbitrary workflow, automatically keeping the system in a consistent state, and performing required batch mode processing automatically. These triggers can also act as a mechanism to derive and capture dataset change audit information.	
Issues	This package will be able to deliver proof of concept implementations only.	
Milestones	MCAT Triggers built into SRB source stream.	S+12
	Production implementation.	NCRIS

Storage and Interoperability

Work Package SI1

The initial integration approach between the Fedora repository software and SRB is at the storage level of the repository software. This would enable Fedora to natively handle large datasets.

Objectives	Facilitate distributed data management using Fedora.	
Partners	SDSC, ARROW project, Fedora developers	
Responsible	UQ	
Activity Statement	Investigate the appropriateness of SRB (relative to contenders such as GFarm or OPeNDAP), as a candidate for integration (as the underlying data grid) into Fedora, and if deemed appropriate implement: <ul style="list-style-type: none"> • Data access integration • Data model integration for archival context management. 	
Deliverables	Working integration.	
Issues	Identifying most effective way to do this, open-source licensing, metadata compatibility.	
Milestones	Integration complete and available to research community.	S+9

Work Package SI2

Not all large dataset environments should be, or will want to be, integrated into Fedora. It will therefore also be necessary to ensure interoperability between these non-Fedora repositories and existing institutional repositories.

Objectives	Improve interoperability between Storage Resource Broker (SRB) based environments and Fedora.	
Responsible	Monash	
Partners	SDSC, UQ, ARROW project, Fedora developers	
Activity Statement	Interoperability here means ability to link from one to the other/move objects between one and the other <ul style="list-style-type: none"> • Investigate use of Metadata Encoding and Transmission Standard (METS) to exchange archived resources between systems • Develop authenticity management between federated systems • Implement ability for Fedora to reference SRB-based stores as externally-managed datastreams. 	
Deliverables	Documented and working interoperability solution.	
Issues	Identifying possible partners, selecting appropriate standards, determining how best to match metadata across repositories, liaising with DSpace regarding work already underway in that domain.	
Milestones	Solution tested and ready for deployment by others.	S+12

Work Package SI3

Closer integration between Fedora and SRB will require the MCAT metadata catalogue to be extended and to be semantically augmented by extensions to the schema and possible replacement by a Resource Description Framework (RDF) data store (e.g., Kowari). DSTC have extensive experience with RDF, and the Fedora developers have just integrated an RDF triple store into the latest version of Fedora.

Objectives	Support richer descriptive and preservation metadata for dataset objects to enable more effective discovery.	
Responsible	DSTC	
Partners	NPACI, SDSC, ARROW	
Activities	Extend the SRB/MCAT metadata schema. This will involve semantic augmentation of MCAT and possible replacement with an RDF data store (e.g., Kowari) and links to domain-specific descriptors: <ul style="list-style-type: none"> • extend MCAT metadata schema to support richer descriptive and preservation metadata (e.g., METS) • replace MCAT with Kowari RDF data store, linked to domain-specific ontologies and association rules • integrate with SRB and Fedora. 	
Deliverables	Prototype RDF-OWL-based implementation of MCAT.	
Issues	Metadata schema extensions; RDF representations; OWL ontologies	
Milestones	Prototype to be completed.	S+12
	Solution evaluated and ready for deployment.	S+18

Work package SI4

Secure access using Grid security for transferring data between the instrument and sensor data layer and the higher layers.

Objectives	Provide a secure service for transferring data from sensors/instruments to repositories using Grid security.	
Responsible	Monash	
Activity Statement	Develop and test lateral integration of Globus GT4 with other project specific middleware – CIMA and SDSC SRB <ul style="list-style-type: none"> • Implement a secure data transfer mechanism using Globus security for SRB transfers • Identify and bridge the gaps within Globus GT4 and web services, in context of CIMA and possibly SRB. 	
Deliverables	Secure data transfer mechanism.	
Issues	Emerging standards within middleware for the project may lead to changes in the way these may interoperate in future.	
Partners	JCU, DSTC	
Milestones	Data transfer service.	S+15

Work package SI5

There are a number of different middleware proposals for managing collections of replicated data. The SRB is certainly one of the most mature, but Globus provides its own RLS, and there are other competitors like Gfarm from Japan. This work package will build on an existing ARC grant for the GriddLeS system.

Objectives	Develop an abstraction layer that supports a range of data replication systems, such as SDSC SRB, Globus RLS and GFarm.	
Partners	Nil	
Responsible	Monash	
Activity Statement	<ul style="list-style-type: none"> • Build an abstraction to SRB, Globus and GFarm to provide portability for applications whilst also minimizing the changes required to applications. • Design and implement a higher level service to better match application requirements. 	
Deliverables	An implementation that provides access to data regardless of the underlying replica management system.	
Issues	Need to track developments in Globus, Web Services and related technologies.	
Milestones	Design complete.	S+9
	Production system.	S+18

Work package SI6

Much of this proposal targets data from instruments and static data in repositories. Some of the papers in this area talk about capturing the data from simulations, and imply that it is simply put into the repository. It might be more efficient to store the input parameters for such simulations and rerun them when required. The user/software can be shielded from the underlying implementation and think it is accessing an actual repository. The system might instead choose to recompute the data and present it to the user. This would leverage Monash experience with Nimrod/G (DSTC funded) and GriddLeS (as above).

Objectives	Allow simulation data to be retrieved from repositories or regenerated dynamically using computational services.	
Responsible	Monash	
Partners	Nil	
Activity Statement	Build on Nimrod/G and GriddLeS to provide a software implementation that allows data to be retrieved or regenerated using standard metadata for describing the data.	
Deliverables	Working software.	
Issues	Need to track developments in Globus, Web Services and related technologies.	
Milestones	Proof of concept.	S+18

Work package SI7

Objectives	Develop a cost-effective data pre-processing system for the secondary storage	
Responsible	Monash	
Partners	QPSF, JCU, UQ	
Activity Statement	Establishment of a data pre-processing service for refining, integrating, and storing real-time data streams from instruments and sensors into the secondary storage for data analyses and processing by higher layers.	

Deliverables	Data pre-processing service for moving data from primary storage into secondary storage.	
Issues	Different pre-processing requirements, time constraints at clearing the primary data store, error recovery – recovering discarded data sets.	
Milestones	Proof of Concept.	S+6
	Pilot Service.	S+12

Work package S18

Moving large volumes of data over longer distances would be required between the secondary storages at Monash and JCU/UQ.

Objectives	Pilot long-distance high speed and secure data transfer between repositories.	
Responsible	Monash	
Partners	QPSF, JCU	
Activity Statement	Design and implementation of a high-speed and secure data transfer mechanism between JCU and Monash.	
Deliverables	File Transfer service between the two sites over AARNet3/Grangenet.	
Issues	Access rights, emerging use patterns, growing rate at which data gets produced.	
Milestones	Data transfer service.	S+12

Work package S19

A very large storage infrastructure will be required to store outputs from the various instruments and sensors in the identified research communities. This infrastructure will need to be built using a hierarchical storage management approach comprising fast online and relatively slower nearline storage systems. This work package will look at the engineering and configuration requirements (including data audit) for such a system, and implement a proof of concept using Monash's SAN and HSM infrastructure. The full production hardware will be sought via a separate LIEF grant in which Monash is a participant, and future NCRIS funding.

Objectives	Scope and pilot storage infrastructure requirements.	
Responsible	Monash	
Partners	QPSF, UQ, JCU, VPAC, ANU	
Activity Statement	Develop specifications and configuration for a suitable large storage infrastructure <ul style="list-style-type: none"> • Work with research communities to identify requirements • Design a system to meet these requirements. 	
Deliverables	System design.	
Issues	Difficulties in accurately estimating requirements, selection of appropriate data migration middleware.	
Milestones	System Design.	S+6
	Pilot Service.	S+15

Content and Rights

Work Package CR1

This work package will look at the range of issues acting as a barrier to migrating personal repositories.

Objectives	Move data from personal data repositories to secure trusted alternatives.	
Responsible	Monash	
Partners	ARROW project, APSR project, SIMS	
Activity Statement	Investigate issues around personal research data repositories <ul style="list-style-type: none"> • Locate willing representative researchers with personal data repositories • Identify barriers to deposit into secure repositories • Develop strategies to overcome these barriers. 	
Deliverables	Recommendations, identified migration solution.	
Issues	Wide range of existing problems to be determined and solved – trust, security, IP etc.	
Milestones	Recommendations and available solution.	S+12

Work Package CR2

The existing Creative Commons work may be applicable for non-science researchers.

Objectives	Reduce barriers to content acquisition by providing more rights-assignment options for non-science researchers.	
Responsible	UQ	
Partners	QUT (Legal Protocols for Copyright Management Project), ARROW, Science Commons, School of Information Management and Systems (Monash), KCL	
Activity Statement	Apply Creative Commons work to non-science research data and results, and integrate this into software <ul style="list-style-type: none"> • Investigate current state of work at QUT on contextualising Creative Commons for Australian conditions • In collaboration with QUT (Legal Protocols for Copyright Management Project), develop guidelines for application of Creative Commons to non-science research data and results • Integrate the results of this work into repository submission and display software. • Provide access to disciplinary testbeds for QUT (Legal Protocols for Copyright Management Project) in application of Creative Commons to non-science research data and results 	
Deliverables	Guidelines, working software.	
Issues	Identifying user requirements, interoperability with other related work.	
Milestones	Working software.	S+12

Work Package CR3

Science Commons (<http://sciencecommons.org/>) is a new project exploring legal and technical mechanisms to remove the barriers that inhibit the sharing of scientific information. The Science Commons work package will involve developing software to enable scientific researchers to attach standardized licenses that define attribution, commercialization, derivative works and re-distribution rights.

Objectives	Reduce barriers to content acquisition by providing more rights options for science researchers.	
Responsible	DSTC	
Partners	QUT (Legal Protocols for Copyright Management), Science Commons, ARROW	
Activity Statement	Apply Science Commons work to scientific research data and results, and integrate this through software that enables the attachment of standardized licenses <ul style="list-style-type: none"> • Work collaboratively with QUT (Legal Protocols for Copyright Management Project) to develop set of standardized licenses suitable for scientific data and publications • Develop software tools to enable the easy selection and attachment of licenses recognized by Science Commons. • Provide access to disciplinary testbeds for QUT (Legal Protocols for Copyright Management Project) to evaluate the application of Science Commons to-science research data and results 	
Deliverables	Guidelines, working software.	
Issues	Identifying user requirements within Australia, adoption within Science Commons developments.	
Milestones	Working software.	S+15

Work Package CR4

In order to have the best chances of getting quality content, and at the earliest stage in the research process, it may be useful to seed research teams with information management professionals. They can assist researchers with their information management needs and also develop guidelines for particular disciplines.

Objectives	Improve information management practice in research communities.	
Responsible	Monash	
Partners	APSR, JISC-funded eBank UK projects, SIMS	
Activity Statement	Develop guidelines for best information management practice for research data set issues <ul style="list-style-type: none"> • Locate representative research teams with data set information management issues • Embed information management professionals into these teams • Identify areas requiring improved information management practices • Develop strategies and guidelines to improve these barriers. 	
Deliverables	Recommendations as to best practice.	
Issues	New nature of field; difficulties in predicting future.	
Milestones	Interim recommendations.	S+9
	Final recommendations.	S+15

Work Package CR5

Whatever software is provided to researchers will need to be easy to use, fit with their work practices and workflows, and provide them with a persistent identifier once an item is lodged.

Objectives	Make it as easy as possible for researchers to deposit datasets and other digital objects into institutional repositories.	
Responsible	Monash	
Partners	ARROW, VTLs, Fedora	
Activity Statement	Develop workflow steps and a user interface (UI) for software to enable researchers to easily ingest/deposit communication units into repository and receive a persistent identifier <ul style="list-style-type: none"> • Determine required workflow • Determine required metadata to be captured/machine-generated • Prototype data-entry screens • Implement into ARROW software. 	
Deliverables	Working software for self-deposit of communication units and management of these units once deposited.	
Issues	Need to ensure that development of software is informed by user-centred design and usability principles. Need to ensure that we work with a range of researchers from different disciplines to ensure a general solution.	
Milestones	Initial requirements.	S+3
	Initial software release.	S+9
	Final software release.	S+12

Work Package CR6

This work package will address the legal issues (separate from the granting of rights dealt with in the two work packages dealing with Science Commons and Creative Commons).

Objectives	Improve content deposit rates by clarifying legal issues around IP, information security and privacy.	
Responsible	Monash	
Partners	ARROW, SIMS	
Activity Statement	Comprehensively identify and examine legal issues, including intellectual property issues, privacy issues and issues concerning information security, that arise from information deposit into institutional repositories, and its subsequent storage, use and dissemination.	
Deliverables	Identification of legal issues/issue mapping & recommendations.	
Issues	Intellectual property, information security and privacy.	
Milestones	Identification of legal issues and recommendations .	S+12

Annotation and Assessment

Work Package AA1

This work package will develop tools and services that enable peers to attach reviews, opinions, comments or assessments to research data, reports, publications etc. It will concentrate on extending and refining existing annotation tools to enable annotation of digital objects held within research repositories such as SRB, DSpace or Fedora.

Objectives	Allow researchers to annotate each other's works (including datasets).	
Responsible	DSTC	
Partners	W3C, NSDL	
Activity Statement	Develop software to enable the attachment of annotations to resources, either as part of validating them, assessing them or in order to add richness <ul style="list-style-type: none"> • Extend Annotea to support annotation of multimedia objects in SRB, Fedora or DSpace • Build and index annotation servers • Capture annotation changes as a record of collaborative activity. 	
Deliverables	Pilot annotation software and associated infrastructure.	
Issues	Identifying user requirements, interoperability/integration with related work.	
Milestones	Working software.	S+12

Work Package AA2

This work package will focus on the development of secure authenticated access to annotation servers by trusted team members through the development of a Shibboleth-based interface to the W3C's open source Annotea server. DART aims to use XACML to implement this fine-grained access control and will need to build on work done by the MAMS project.

Objectives	Improve annotation and deposit rates by allowing end-user control over who can annotate what. and who can access the annotations.	
Responsible	DSTC	
Partners	W3C, Shibboleth Internet2	
Activity Statement	Develop software to enable authentication and access constraints on annotations, based on Shibboleth (and MAMS outcomes) <ul style="list-style-type: none"> • Modify Annotea to enable Shibboleth integration • Build tools to define XACML access constraints associated with annotations. 	
Deliverables	<ul style="list-style-type: none"> • Secure annotation servers. • Secure, access controlled search, browse and retrieval tools for annotations. 	
Issues	Identifying user requirements, integration of related existing middleware.	
Milestones	Prototype software.	S+1815
	Tested and Working software.	NCRIS

Work Package AA3

This work package will concentrate on tools to support collaborative annotations by groups of users either asynchronously or in real-time through application sharing.

Objectives	Help collaborative annotation services contribute to the life and productivity of research communities.	
Responsible	DSTC	
Partners	UK eScience VidGrid and CancerGrid projects, Monash, KCL	
Activity Statement	Develop software to enable collaborative attachment of annotations to resources during group application sharing within videoconferencing or access grid sessions. Develop collaborative asynchronous annotation tools <ul style="list-style-type: none"> • Develop searchable annotation server • Implement application sharing environment • Implement real-time collaborative annotation tools • Capture annotation intent and annotation changes as a record of collaborative activity. 	
Deliverables	Pilot collaborative annotation software and associated infrastructure.	
Issues	Identifying user requirements, integration and extension of existing tools, interoperability with complementary work packages.	
Milestones	Prototype software. Production level software.	S+15 NCRIS

Work Package AA4

Wikis can be used to support collaborative research practices, act as an interface to disparate digital objects, and comment on/annotate these objects.

Objectives	Foster emerging wiki-based collaborative work practices in research teams.	
Responsible	JCU	
Partners	Monash, DSTC, NSDL	
Activity Statement	Integrate Wiki and SRB technologies together and pilot with a number of discipline groups by: <ul style="list-style-type: none"> • identifying the most suitable form of wiki, and placing the back end file system under federated SRB control • integrating wiki with chosen directory services and rights management. Building – where possible - a semantic web map of the wiki and its content. • developing wiki templates/structures and plug-ins to support board adoption (this last step will involve working with different user communities) • supporting capture of annotation changes as a record of collaborative activity. 	
Deliverables	Working wiki/SRB integration. Annotated wiki content and a semantic web/grid content flowing from the wiki content.	
Issues	Identifying the right flavour of wiki, user education, software integration/engineering issues.	
Milestones	Working software with rights management flowing from chosen directory services.	S+6

	Wiki content and metadata used to generate semantic web maps of content.	S+9
	Series of discipline specific templates and wiki structures.	S+12

Discovery and Access

Work Package DA1

DART will need to provide depositors with control over access to their contributions. This might be by a range of attributes, including user, role, time or location.

Objectives	Improve repository deposit rates, sharing and reuse by allowing end-user control over who can access what.	
Responsible	Monash	
Partners	ARROW, VTLS	
Activity Statement	Enable controlled access to distributed archives through the RDF-data store – to resources and data stored both within SRB, DSpace and Fedora repositories <ul style="list-style-type: none"> • Identify generic range of access control requirements • Develop XACML statements that encode these • Implement XACML statements in ARROW software. 	
Deliverables	Working software that allows depositor control over access both within an institution and across institutions.	
Issues	Need to allow for a range of access types.	
Milestones	Working software.	S+9

Work Package DA2

Materials in the repositories will need to be discoverable (unless access restrictions prevent this).

Objectives	Improve repository deposit rates, sharing and re-use by improving discoverability.	
Responsible	Monash	
Deliverables	Enable public discovery of available communication units through working integration of DART repositories, and other dataset repositories, with national research discovery service hosted by NLA at http://search.arrow.edu.au/ discipline-based discovery services as appropriate.	
Activity Statement	<ul style="list-style-type: none"> • Extend NLA RDS to harvest metadata from dataset repositories (both OAI-PMH and SRB-MCAT). • Implement search interface for dataset metadata. 	
Issues	Defining what metadata schema to apply, what schema to harvest, how to allow for discovery across different repositories with different ontologies.	
Partners	NLA, Monash, KCL.	
Milestones	Working integration.	S+12

Work package DA3

This work package will develop and provide access to a centralized repository/registry of metadata schemas, building on the open source software tools being developed within the JISC IE Metadata Schema Registry Project (IEMSR) by UKOLN and ILRT (<http://www.ukoln.ac.uk/projects/iemsr/>).

Objectives	Reduce wasted effort on creating metadata schemas and improve interoperability of metadata schemas.	
Responsible	DSTC	
Partners	JISC, UKOLN, ILRT	
Activity Statement	Develop software (or use existing open source software) to enable users to create new schemas, submit schemas to the registry and search and browse the registry <ul style="list-style-type: none"> • Understand approach, technologies being developed by JISC IEMSR • Work on extensions, refinements, gaps • Build prototype tools and preliminary registry. 	
Deliverables	Centralized repository of metadata schemas and application profiles.	
Issues	Metadata schemas to include, moderation and support, federated access to global registries.	
Milestones	Prototype tools.	S+9
	Production system.	S+15
	Centralized working registry.	NCRIS

Appendix B: References

- Abramson, D. 2005. "Software Development for the Computational Grid", *Remote Access and Automation Workshop*, Sydney, 2005. Abstract online at http://mmsn.chem.usyd.edu.au/events/mmsn_ws05_abstracts.html#Abramson
- Atkins, D. et al. 2003. *National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure, Revolutionizing Science and Engineering through Cyber-infrastructure*. Available online at http://www.communitytechnology.org/nsf_ci_report/.
- Barstow, A, Kahan, José, Koivunen, M-R, Swick, R. "Annotea: A Generic Annotation Environment using RDF/XML", *WWW10 Developers Day*, Hong Kong, May 2001
- Clapham, N. T., Green, D. G. and Kirley, M. (2001). Emergent information systems - the role of adaptive agents. *Australian Journal of Intelligent Information Processing Systems* 7 (3/4), 96-101.
- Davis, P. 2005. "Grid Enabling of the Australian Synchrotron", *Remote Access and Automation Workshop*, Sydney, 2005. Abstract online at http://mmsn.chem.usyd.edu.au/events/mmsn_ws05_abstracts.html#Davis
- Doerr, M., Hunter, J., Lagoze, C. 2003. "[Towards a Core Ontology for Information Integration](#)", *Journal of Digital Information*, Volume 4 Issue 1, April 2003 <http://jodi.tamu.edu/Articles/v04/i01/Doerr/>
- Green, D.G. and Bossomaier, T.R.J. (2002). *Online Geographic Information Systems and Spatial Metadata*. Taylor & Francis, London.
- The Joint Information Systems Committee, 2004. *The Data Deluge: Preparing for the explosion in data*, Available online at http://www.jisc.ac.uk/index.cfm?name=pub_datadeluge
- 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". Submitted to *International Journal of Digital Libraries: Special Issue on Complex Objects*. Available online at <http://www.arxiv.org/abs/cs.DL/0501012>
- Lyon, L. "eBank UK: Building the links between research data, scholarly communication and learning", *Ariadne*, Issue 36. Available online at
- McDermott, G. 2005. "Crystallography and Tomography Automation and Remote Access at the Advanced Light Source", *Remote Access and Automation Workshop*, Sydney, 2005. Abstract online at http://mmsn.chem.usyd.edu.au/events/mmsn_ws05_abstracts.html#McDermott
- McMullen, R. and Chiu, K. 2005. "CIMA: Scientific Instruments as First Class Members of the Grid", *Remote Access and Automation Workshop*, Sydney, 2005. Abstract online at http://mmsn.chem.usyd.edu.au/events/mmsn_ws05_abstracts.html#McMullen
- Moore, R. 2004a "Integrating Data and Information Management", *International Supercomputer Conference*, June. Available online at <http://www.sdsc.edu/dice/Pubs/ISC2004.doc>.
- Moore, R. 2004b. "Evolution of Data Grid Concepts", *Global Grid Forum Data Area Workshop*, January. Available online at <http://www.sdsc.edu/dice/Pubs/Grid-evolution.doc>.
- Open Society Institute, 2004. *A Guide to Institutional Repository Software v 3.0*. Available online at <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 1997, Oldenburg, Germany. Available online at <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. Paper presented at *CIRN 2004: Sustainability and Community Technology*, Monash University, Prato, Tuscany, Italy. Available online at <http://www.ciresearch.net/conferences/viewabstract.php?id=68&cf=4>.

Schroeter, R., Hunter, J., Kosovic, D. "Vannotea - A Collaborative Video Indexing, Annotation and Discussion System for Broadband Networks", *Knowledge Markup and Semantic Annotation Workshop*, K-CAP 2003, Sanibel, Florida Oct 2003 <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", in *D-lib Magazine*, April. Available online at <http://dlib.org/dlib/april03/staples/04staples.htm>

Treloar, A, 1999. *Hypermedia Scholarly Publishing: the Transformation of the Scholarly Journal*. PhD Thesis, Monash University 1999. Available online at <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. Available online at <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. Available online at <http://www.cni.org/tfms/2003b.fall/handouts/Fall2003Handouts/H-Watersplenary.doc>.

Appendix C: Key Acronyms and Initiatives

ADT – Australian Digital Theses Program: <http://adt.caul.edu.au/>

API – Applications Programming Interface

APS – American Physical Society: <http://www.aps.org/>

APSR - Australian Partnership for Sustainable Repositories: <http://www.apsr.edu.au/>

ARROW - The Australian Research Repositories Online to the World: <http://arrow.edu.au/>

arXiv: <http://arxiv.org/>

BIOME: <http://biome.ac.uk/>

CIMA – Common Instrument Middleware Architecture:

<http://www.instrumentmiddleware.org/metadot/index.pl?iid=2119&isa=Category>

CNI-JISC-SURF conference - The Coalition for Networked Information (CNI), the Joint Information Systems Committee in the UK (JISC) and SURF - the Dutch higher education and research partnership organisation for network services and information and communications technology (ICT):

<http://www.surf.nl/en/bijeenkomsten/index2.php?oid=6>

Dublin Core: <http://dublincore.org/>

eprints.org: <http://www.eprints.org/>

FRODO - Federated Repositories Of Digital Objects (FRODO) Projects funded by DEST under the Commonwealth Government's *Backing Australia's Ability* Initiative: 1. Meta Access Management System; 2. Towards an Australian Partnership for Sustainable Repositories; 3. Australian Research Repositories Online to the World (ARROW); 4. Australian Digital Theses Program Expansion and Redevelopment.

FUSION - Fuel cell Understanding through Semantic Inferencing, Ontologies and Nanotechnology:

<http://metadata.net/sunago/fusion.htm>

GFarm: http://www.aist.go.jp/aist_e/latest_research/2003/20031205/20031205.html

Globus GT: <http://www.globus.org/>

Griddles: <http://www.csse.monash.edu.au/~davida/griddles/overview.htm>

IPR – Intellectual Property Rights

KCL – Kings College London

MAMS - Meta Access Management System: <http://www.melcoe.mq.edu.au/projects/MAMS/index.htm>

MCAT - SRB Metadata Catalogue

MMSF – Morgan-Monroe State Forest Automated Observatory:

<http://www.instrumentmiddleware.org/metadot/index.pl?iid=2174&isa=Category>

Nimrod: <http://www.csse.monash.edu.au/~davida/nimrod/>

NLA – National Library of Australia: <http://www.nla.gov.au/>

NSDL – National Science Digital Library: <http://nsdl.org/>

OAI – Open Archives Initiative: <http://www.openarchives.org/>

OAIS – Open Archival Information Systems: <http://ssdoo.gsfc.nasa.gov/nost/isoas/>

OPeNDAP – Open-source Project for a Network Data Access Protocol: <http://opendap.org/>

PADI – Preserving Access to Digital Information: <http://www.nla.gov.au/padi>

PANDORA – Preserving and Accessing Networked Documentary Resources of Australia:

<http://www.nla.gov.au/pandora>

PANIC – Preservation and Archival of New media and Interactive Collections: <http://metadata.net/panic/>

ARIIC DART BID - Updated

PMH – Protocol for Metadata Harvesting: <http://www.openarchives.org/OAI/openarchivesprotocol.html>

RSS – Really Simple Syndication: <http://www.xml.com/pub/a/2002/12/18/dive-into-xml.html>

SDSC – San Diego Supercomputing Center: <http://www.sdsc.edu>

SIMS – School of Information Management and Systems, Monash University

SPARC – Scholarly Publishing and Academic Resources Coalition: <http://www.arl.org/sparc/>

SRB – Storage Resource Broker: <http://www.sdsc.edu/srb/>

VTLS – Visionary Technology for Library Systems, Inc. One of two groups of ARROW software developers: <http://www.vtls.com/>

W3C – World Wide Web Consortium

XACML – extensible Access Control Markup Language: http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=xacml

Appendix D: Theoretical framework

The Call for Proposals identifies a number of key trends that are changing the paradigm for how research is carried out and consumed. These include:

- new technologies, such as computer simulations, synchrotrons and sensor networks
- expanding size of datasets
- increasing volumes of information
- greater complexity
- recognition of the need to work across traditional discipline, national and institutional borders

To this one might add a growth in research practices that are:

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

The DART project proposal recognises the significance of these trends and their impact, and draws on three bodies of research into scholarly activity and communication in order to frame an appropriate response. The relationship between these three bodies of theory (Re-engineering scholarly communication, scholarship as ecology, and the information continuum) is shown in Figure 3 below.

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, 2004) argue that current experimentation within the publishing world is both limited and backwards-looking. They argue for two radical changes.

The first is to deliberately engineer a new scholarly communication system that is intertwined with the process of generating new knowledge. Roosendaal and Geurts (1997) have 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 (2004) 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 developing 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 should be capable of being aggregated into complex documents, which are themselves units of communication
- these units should be capable of being registered and preserved regardless of their nature or stage of development, thus facilitating collaboration, network-based research and faster discovery.

They also identify 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, the 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 applied this model to the

published research output. 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.

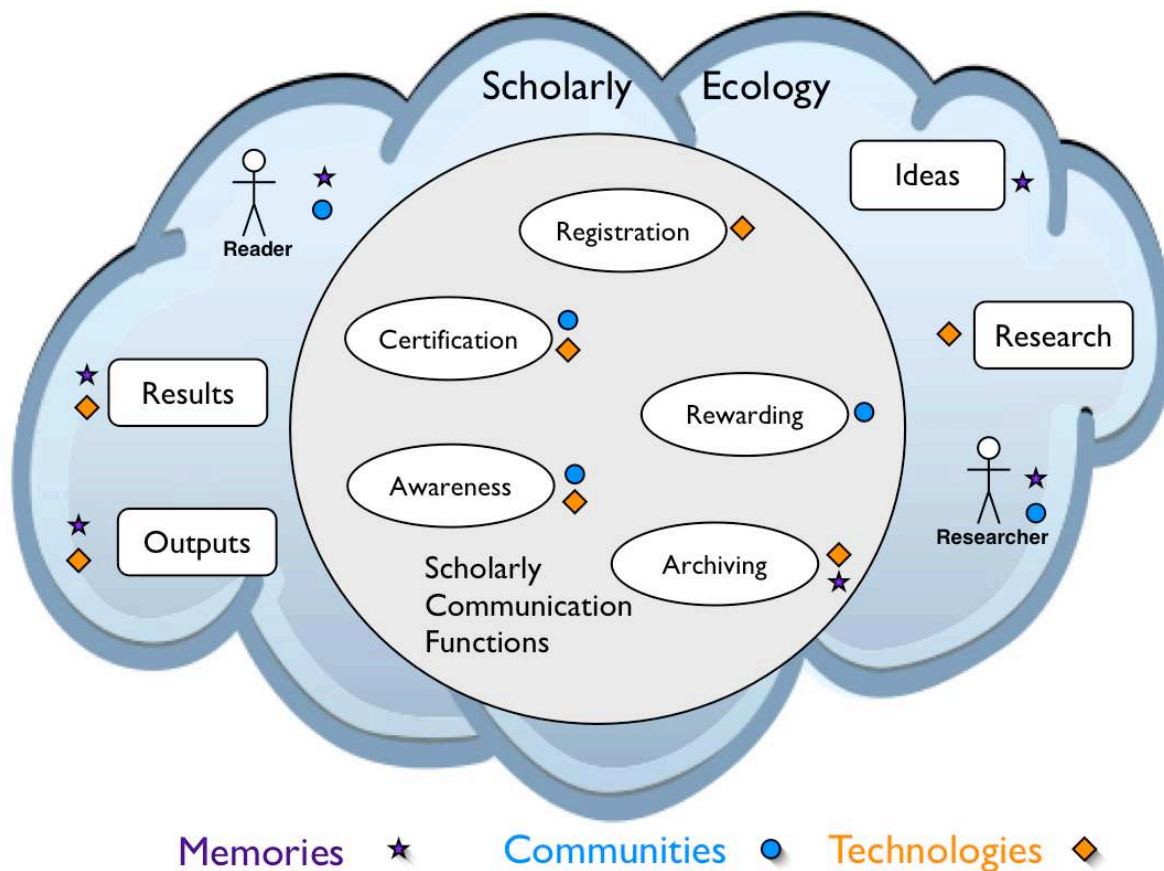


Figure 3: Inter-related bodies of theory

Information Continuum

The third body of work has been developed by a group of researchers in the School of Information Management and Systems 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 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 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.

Appendix E: Capability statement

CRC for Enterprise Distributed Systems Technology (DSTC)

Established in 1991, DSTC Pty Ltd has operated four Cooperative Research Centres including the current CRC, the CRC for Enterprise Distributed Systems Technology. DSTC participants include the Queensland, South Australian and Federal Governments and a portfolio of participating organisations including Telstra, Boeing, Mincom, Sun Microsystems and Fujitsu and a number of Australian Universities. DSTC's collaborative research is focussed on the specification, interoperability, and management of enterprise information systems both within and across organisations. Our applied research is primarily focused on providing technology services to the Government, Defence, Healthcare, Finance and Education sectors.

Indicators of DSTC's success in establishing its international standing as a centre of research excellence in the field of distributed systems include:

- Three consecutive IT&T Awards for Excellence in Research
- Over 500 international conference papers, journal articles and books
- Linkages with leading Australian and international organisations, industry representatives, researchers and government officers
- Attraction of leaders in IT from around the world.
- Representation on international standards bodies, such as OMG, W3C, ISO, MPEG7 & IETF
- World wide technology deployments and license deals
- Commercial revenues exceeding \$10 million per annum
- 4 internationally successful spin-off companies

DSTC also hosts the Australian office of the World Wide Web Consortium (W3C) and as such has organized numerous conferences, workshops, seminars and training programs to increase information technology awareness and capabilities of Australian industry, government and research organizations.

DSTC's Knowledge Management research program concentrates on developing data models, ontologies, metadata schemas, workflow technologies, software tools and query languages to enable efficient indexing, archival, discovery, analysis, integration, management and preservation of large multimedia collections and the mining of new knowledge - within the educational, cultural and scientific domains. Projects include:

- **Harmony** International Digital Library project – a collaboration between Cornell, DSTC and ILRT (jointly funded by the NSF, JISC and DEST) which developed the ABC ontology adopted by MIT's DSpace system. <http://metadata.net/harmony/>
- **The Australian Global Grid project** – a DEST-funded collaborative e-Research program involving UQ, DSTC and Monash and the UK eScience program, investigating innovative middleware technologies in the areas of Integrative Biology and e-Health. <http://www.acmc.uq.edu.au/eScienceProjects.htm>
- **FUSION** – a collaborative project between UQ, DSTC and Ceramic Fuel Cells Limited which is developing a knowledge management system comprising web-services, workflows, secure distributed data repositories, tele-presence sessions, data exploration and visualization tools, and statistical-driven data capture tools to enable the optimization of nano-materials. <http://metadata.net/sunago/fusion/>
- **PANIC** – An integrated preservation system for digital objects based on METS, software and format registries and semantic web services. DSTC is about to begin a project with NLA and UK DCC which will implement the obsolescence detection and notification component of PANIC.
- **FilmEd/Vannotea** – a GrangeNet funded project which is developing software tools to enable collaborative real-time annotation of digital objects by geographically distributed groups of users. <http://metadata.net/filmed/>

Dr Jane Hunter, is a Distinguished Research Fellow at DSTC, internationally renowned for her work in eScience middleware, Semantic Web/Grid, knowledge management and multimedia digital libraries. Over the past five years, she has been Head of Australia's delegation to MPEG, a member of the Dublin Core Advisory Board, the W3C's XML Schema and Web Ontology Working Groups and the Internet2 Middleware VidMid/ViDe working group. She is currently the liaison between MPEG and the W3C, a member of the DELOS Network of Excellence on Digital Libraries Working Party (WP5) and co-chair of APAN's eScience working group. She is on the Editorial Board of IEEE Multimedia and the Elsevier Journal of Web Semantics and on the programme committees of six international conferences. She has also carried out numerous consultancies for industry and government organizations both nationally and internationally (including the ABC, Learning Federation, Boeing and ACMI). She is currently collaborating with the Smithsonian, the Library of Congress and a number of UK eScience Centres.

James Cook University

James Cook University (JCU) along with various partners, the Australian Institute of Marine Science (AIMS), UQ, QPSF and the CRC-Reef research are engaged in a series of pilot and demonstration projects with the goal of designing and deploying arrays of remote environmental monitoring sensors. These new 'sensor networks' are differentiated in that they support two-way communication with the remote sensors from a central data centre / control room. This means that not only can the sensor packages send data in real time back to the data stores, but that equipment health can continuously be assessed. These environmental sensor network projects are being worked on from every angle; sensor design and deployment, sensor communications and are of particular importance data collection/archiving, meta-data processing and data synthesis.

JCU has the most active high performance computing (HPC) group in regional Australia and has a long-standing history in HPC that dates back more than a decade. JCU has invested heavily (with the support of several ARC-LIEF projects) in developing and supporting HPC and mass-data facilities, and more recently, long-haul multi-gigabit connectivity. At present, the HPC unit at JCU supports a range of facilities including a 84 cpu SGI Origin, a newly installed 155 cpu cluster, and mass data store (40 TB StoTek tape silo), running under hierarchical storage management (HSM). Significantly this mass storage system is coupled to the UQ/QPSF mass storage silo at UQ (1400km south of Townsville) via the recently commissioned AREN funded GigaBit network. Files written to the JCU tape silo are automatically mirrored to the UQ silo. *As we understand it this is the only running example of this form of high capacity, long distance data mirroring in Australia.*

In addition, JCU has a developing presence in the e-Research community. By virtue of its relative remoteness and its capacities in advanced IT, JCU has implicitly understood the value of "e-Research" in the support for a broad array of its research. JCU has been a major contributor to the APAC project and presently employs a team of 5 full-time staff that are working on APAC programs. This core group of staff are also providing support to a wide range of data grid projects at JCU ranging from digital libraries for indigenous media, marine archaeology and robotic telescopes for high-school education through to proteomic data mining, grid supported bio-discovery and materials science. In addition the JCU e-Research group has been working on developing back-end tools to support digital library implementations.

JCU was an early adopter and innovator in AccessGrid technology. The University built the second AG in Australia in 2001 and has used the technology extensively for teaching and research collaboration across multiple campuses for nearly 4 years. We have build a specific AG installation in the JCU advanced analytical centre in order to develop tele-instrumentation/remote control applications and this system has been successfully used to remotely control remote client-instruments as well as being useful in providing user training.

Monash University

Monash University is one of the largest universities in Australia. It has a large number of world ranking scientists working in various areas. For example, Monash has a world ranking group working in climatic changes led by an ARC Federation Fellow, Amanda Lynch. Monash has a world leading group working in bio-informatics, led by Ross Coppel, a Howard Hughes Fellow (there are only a small number of these fellows worldwide). This is in addition to the world famous IVF program which pioneers much cutting edge human reproductive system research, including stem cell research.

Monash is developing an e-Research Centre which will facilitate, coordinate, and support e-Research in Monash. It is currently actively recruiting a world class researcher who has experience leading e-Research to lead this major initiative. The establishment of the e-Research Centre will build on the existing work in middleware in Monash led by David Abramson. The Abramson group is already famous in the implementation of a novel middleware package, called Nimrod (parametric sweeps) and Griddles (Grid workflows). These are some of the few middleware packages in the world which can provide parameter optimisation of complex simulations using grid resources and provide work flow support for a wide range of research disciplines. The establishment of the e-Research Centre will facilitate and enhance the work of Nimrod to include other aspects of the upper middleware (workflow, scheduling, accounting) so that this will be used more widely around the world. Soon, the synchrotron facilities (co-located next to the Clayton campus of Monash university) will become functional (2007). This will provide a driver for a data intensive initiative at Monash. In preparation for the support of the work which will emerge from the synchrotron, Monash has appointed Rob Lewis, a recognised researcher in this area to lead a group. This group will work

closely with other Monash groups to ensure that maximum benefits from the installation and operation of the synchrotron can be obtained.

Dr Asad Khan, who was one of the founding members of Monash data storage initiative, would be associated in a significant manner with this project. Dr Khan has a very strong track record of successful project completions at Monash. He also has a strong research background with 25 refereed publications and being the co-author of one of the first books on parallel finite element computations. He was a co-recipient of over 0.25 million pounds of research contracts in the UK prior to joining Monash. He is the winner of Sun Microsystems 2004 100% equipment grant. He has managed technical teams of up to 15 staff and is currently responsible for 7 PhD and Master by research programs. Monash at present is in the process of establishing a Grid computing service, led by Dr. Asad Khan, which includes a sizable equipment grant from Alphawest-Sun Microsystems, and which was won competitively in 2004. This separate dedicated grid infrastructure will facilitate Monash researchers in exploring and using grid methodologies for their work. This grid network will be connected to the large storage resources indicated previously to create a very large and sophisticated grid infrastructure across various campuses of Monash University.

The University's Information Technology Services (ITS) supports the IT infrastructure of the University, and has been working at the forefront of utilising information technology to facilitate teaching, learning, research and administration of the University. In the past 10 years, ITS has been consistently implementing an aggressive data assurance program. One component of this program is the installation of a number of large capacity data backup services. Currently there are two StorageTek Timberwolf robot tape systems (total capacity > 100 TB) in the University, backing up all user data in a secure and reliable manner. In addition, since 2004, it has installed a state-of-the-art IBM large robotic tape archive system with an online capacity of 2.5PB (Petabyte, being 1000 TB). The University has also recently installed a storage area network (SAN) which has failsafe mechanisms. In the event that one of the fast storage disks fails, the data will be automatically routed via ultra-high speed optical fibre links to hot-standby identical disks. The SAN architecture is ideally suited for the capture of mission-critical data, as the probability of loss of data will be negligibly small as almost non-existent. ITS has built up a very high level of expertise both in installing and supporting these systems. Monash University therefore has potentially one of the largest storage capacities in the higher education sector in Australia.

Monash led the consortium for the establishment of Victorian Educational and Regional Network, as part of the overall Australian Research and Education Network initiative, funded under Systemic Infrastructure Initiative (SII) in 2004 and 2005. This high speed network will allow Monash to experiment with remote instrumentation and tele-presence, initially within the various campuses of Monash, including some located in regional cities.

In terms of its legal expertise, David Lindsay is an expert in intellectual property, privacy, communications and information technology law. He is widely published in these areas. He is completing a book on domain name law and policy. Ann Monotti is the author with Sam Ricketson of *Universities and Intellectual Property: Ownership and Exploitation* (Oxford University Press, UK, 2003), is the recipient of ARC funding and has extensive expertise in the area of intellectual property law. Moira Paterson is the author of *Freedom of Information and Privacy in Australia: Government and Information Access in the Modern State* (LexisNexis Butterworths, 2005), is the recipient of ARC funding and has extensive expertise in the fields of privacy and information security.

University of Queensland

The University of Queensland (UQ) has strong expertise in high performance computing (being a founding participant in and host for the Queensland Parallel Supercomputing Foundation, QPSF), distributed computing (a key participant in CRCEDST), enterprise information infrastructure (hosting the ARC Research Network in Enterprise Information Infrastructure), management of very large and complex scientific data, tera-scale data archives (via QPSF), workflow systems and Web information systems (through its highly successful Data and Knowledge Engineering Research Division).

The University's School of Information Technology and Electrical Engineering (ITEE) has a strong, internationally recognised research base. ITEE staff and students are published widely with more than 70 academic staff and 240 research higher degree students active in key research areas of complex and intelligent systems, data and knowledge engineering, electromagnetics and imaging, embedded systems, and systems and software engineering. The School has forged links with a number of national and international

centres of research excellence at the forefront of technological innovation, in areas such as sustainable energy, computational modelling, human cognition and genomics.

Professor Xiaofang Zhou and the Data and Knowledge Engineering (DKE) research group at UQ have a substantial record in conducting high quality research in databases (in particular, very large scientific databases and spatial databases), information systems interoperability and integration, high performance query processing, Web information systems, distributed and parallel processing, workflow management systems, data mining, and data issues in bioinformatics. Both Zhou and Pailthorpe (below) are CIs in the ARC Centre for Bioinformatics.

Professor Bernard Pailthorpe (jointly School of Physical Sciences and ITEE) is CEO of QPSF and has established vislab.uq in ITEE, which provides one of the nodes of Storage resource Broker (SRB) support in Queensland and collaborates with JCU and AIMs in SensorNet deployment on the Great Barrier Reef. The lab also hosts R&D programs for Access Grid development, within APAC, and high resolution computer displays. Professor Pailthorpe is a CI in the ARC Research Network in ARC Molecular and Materials Structure, which includes synchrotron users.

Appendix F: Proposed project governance

Governance issues

The DART project needs to implement a governance model that responds to a number of specific challenges. These include:

- partners in 3 institutions that are widely geographically separated
- short timelines (at best 18 months total project duration)
- significant programme of work
- wide-ranging coverage within holistic framework

Governance structure

The governance structure we are proposing to respond to these challenges is built around:

- distributed project management
- high-level management committee
- technical committee

Project management

There will need to be a full-time project manager at each geographical location (Townsville, Brisbane, Melbourne). They will be responsible for day-to-day management of work in their areas and co-ordination with the DART project secretariat.

The DART project secretariat located at Monash will need a full-time project director, a part-time technical architect and a part-time assistant. The secretariat will be responsible for overall co-ordination of the project and supporting the work of the two committees.

The local project managers and the DART project manager should teleconference once a week.

Management committee

This will contain two high-level (University Librarian, DVC(Research), CIO or equivalent) representatives from each location, plus the project manager (ex officio), plus a co-opted external with particular expertise in this area. It should be chaired by one of the representatives. This committee is responsible for advocacy within their institutions as well as strategic direction. It should aim to meet every 3 months, although some of the later meetings could be held via video conference/Access Grid.

Technical committee

This will contain two staff from each location who are involved in the actual implementation of the work packages, plus the technical architect, plus the project manager. It should aim to meet every 3 months, although some of the later meetings could be held via video conference/Access Grid. The MAMS project, already funded under the FRODO funding round, will be technical consultants to this committee. MAMS has already committed itself to provide technical guidance and support to the FRODO projects in the area of federated Identity and Access Management. As MAMS has already been funded by DEST to provide this support, they are willing to extend this service to DART as well, as this has been requested by DEST. This support will include advising on requirements, system architecture and design, and implementation; it will also include reviewing requirement documents, system architecture and design documents; but it will not extend to implementation, software development or writing glue-code.

Statement regarding role of CRC for Enterprise Distributed Systems Technology

DSTC will complete the CRC for Enterprise Distributed Systems Technology (CRCEDST)'s seven year lifespan on 30th June 2006, 6 months prior to the scheduled completion of the DART project. At this point, or earlier if this is operationally preferable, the work packages allocated to DSTC will be reassigned. UQ has agreed to undertake the completion of all DSTC work packages, that could not be completed beforehand, using the ARIIC funds available to DSTC and any DSTC staff who work on the project. UQ will complete the work within the team lead by Prof Xiaofang Zhou.

Appendix G: Letters of support

Attached are letters of support from:

- Richard Marchioni, SDSC
- Carl Lagoze, NSDL
- Cathrine Harboe-Ree, ARROW Project Director

Subject: Re: Possible visit in October 2005
Date: Tue, 10 May 2005 14:42:39 -0700 (PDT)
From: Richard Marciano <marciano@sdsc.edu>
To: Sue McKemmish <Sue.McKemmish@sims.monash.edu.au>
CC: rachel.uren@infotech.monash.edu

Dear Sue,

...

I shared your DAART project proposal with Reagan and also received an emphatic endorsement of the project. Please consider us supporters. BTW, Reagan will be traveling in Australia the last week of September. He's there with "OECD"? and will be presenting in Sydney and Brisbane. Some of the groups he's connecting with are using the SRB technology in existing grids in Australia: APAC (Bernard Pailthorpe), CSIRO, are some of the efforts involved, I believe.

...

Best,

-Richard

ARIIC DART BID - Updated

Date: Sun, 15 May 2005 19:42:09 -0400

From: Carl Lagoze <lagoze@cs.cornell.edu>

Subject: DA3RT Bid

To: Andrew Treloar andrew.treloar@its.monash.edu.au

Dear Andrew,

Thank you for giving me the opportunity to review your DA3RT bid. This is extremely exciting research! As you know from our publications and presentations, we at Cornell share the vision that you describe of a distributed e-science environment that allows scientists to share, annotate, and publish new scholarly products. These products are rich aggregations of text, data, images, and computational services – representing science in the 21st century in a way that traditional journal publication can never accomplish. As you know, we are currently engaged in three projects that compliment the DA3RT proposed work:

1. Fedora, in which we continue to refine and develop an open source content management architecture, which you have chosen for ARROW and subsequently for DA3RT. I should note that the idea of integrating Fedora with SRB is one that we have contemplated and would be happy to work with you regarding the details of the implementation.

2. Pathways, in which we are exploring new infrastructure for redefining the process of scholarly communication and providing an environment for more flexible scholarly products and processes.

3. NSDL, in which we are developing the technical and organization infrastructure for digital libraries that target the educational needs of students in science and mathematics. We have chosen Fedora as the basis for this work, for the same reasons that you describe in your bid.

I have shared your bid with my co-investigators on these projects and we are all enthusiastic about the intellectual and perhaps material collaboration that might occur in the context of DA3RT and our projects. Please keep me abreast of your progress.

Best Regards,

Carl Lagoze, Senior Research Associate
Cornell Information Science
301 College Ave.
Ithaca, NY 14850
Phone: 607-255-6046
FAX: 607-255-5196
email: lagoze@cs.cornell.edu
WWW: <http://www.cs.cornell.edu/lagoze>