

ARROW INSTITUTIONAL REPOSITORIES: A REPORT ON THE DECISIONS AND EXPERIENCES OF THE ARROW PROJECT

Presentation to the ALIA Information Online Conference, Sydney, 1 February 2005 by Geoff Payne, ARROW Project Manager

Abstract

This paper provides an overview of institutional repositories, and describes the first twelve months work of the Australian Research Repositories Online to the World (ARROW) project. ARROW has chosen the Fedora open source software as the storage layer software for the repositories to be established at the four ARROW partner sites. Repository workflows and searching are supported by the VITAL software from VTLS Inc, as enhanced in partnership with the ARROW project. Open Journal Systems (OJS) software has been selected to support open access journal publishing. Metadata practices, data modelling and strategies for attracting content to the repositories are discussed. An ARROW research resource discovery service built on metadata harvested from project repositories is being tested, and indexing of content through web search engines is planned.

Introduction

In October 2003 the Department of Education Science and Training, through the Australian Government's Backing Australia's Abilityⁱ initiative, allocated funding of AU\$3.66 million over three years to ARROW to identify and test solutions to establish institutional repositories at Monash University (the lead institution for the project), Swinburne University of Technology, The University of New South Wales and the National library of Australia. The project descriptionⁱⁱ based on the funding bid outlines the project objectives. In the first instance the repositories will manage e-prints, electronic publishing and digital theses, however design decisions are being taken with the intention of allowing management of other types of digital objects as the project progresses.

Metadata from the four repositories is being harvested as the basis of a single ARROW research resource discovery service managed by the National Library of Australia.ⁱⁱⁱ

Institutional Repositories

For ARROW's purposes, an institutional repository is a managed collection of digital objects which

- is institutional in scope rather than subject focused
- has consistent data and metadata structures for similar objects
- enables resource discovery by the "community of practice" for whom the objects are of interest
- allows reading, inputting and exporting of objects to facilitate resource sharing
- respects access constraints

- is sustainable over time, and
- facilitates the application of preservation strategies

An institutional repository should be capable of holding any mix of anything that can be represented digitally. Print equivalents such as research papers, books, book chapters, theses, still images and archival records are the initial targets for inclusion in the ARROW repositories. Management of digital audio, moving images, and multimedia objects comprising a mixture of any of the above formats are in scope but lower priority. The project is aware of interest in our managing learning objects and research data sets and is endeavouring to track these fields to ensure decisions we make now do not preclude adding such objects in the future.

Institutional repositories are being explored by many organisations as a means of better managing digital resources, as tools for enabling researchers, and as a means of improving the accessibility and impact of research results.

Managing digital resources effectively is essential to safeguard the growing volume of these assets generated by institutions. At present an individual resource may be developed through reliance on the efforts of a few dedicated individuals undertaken as an adjunct to wider responsibilities. That the hardware and software on which it relies may be unsustainable in the longer term is no impediment to the resource becoming widely used and relied upon. The cumulative investment of effort and funds represented by the resource can be considerable, and for all the above reasons it is essential that a means of ensuring the long term viability of such resources is developed.

Individual researchers with a sufficient understanding of information management techniques to construct a hardware and software environment to facilitate their research are at a distinct advantage to those without this capability. Institutional repositories have the potential make it easier for less technologically independent researchers to construct, exploit and maintain data sets, and to publish their results. Additionally, collaboration between researchers with related interests should be facilitated through access to institutional repositories affording opportunities to discover related research and to share tool kits for data access and manipulation.

Publication in open access repositories will expand the readership of research results beyond the subscribers to traditional academic publications, thus providing a better return on the public funding invested in the research.

The presentation of research results on line provides an opportunity to expose information in ways not achievable in print. For example, detailed imagery for which the cost of printing would be prohibitive can be manipulated interactively online, or a data set can be exposed for further analysis in ways the compiler of the information may not have anticipated.

Reshaping scholarly publishing is possible with institutional repositories. Publication of research for which the readership would be insufficient to cover the costs of printing is possible. Online publication may shorten the time

between the completion of research and its wide availability. Providing an alternative means of publication to expensive print journals, and in the process retaining intellectual property rights over research outcomes is a possibility.

The FRODO Projects

The ARROW project is one of four related projects dubbed the Federated Repositories Of Digital Objects (FRODO) projects by the Department of Education Science and Training.

The Meta Access Management System^{iv} (MAMS) project is charged with developing solutions to manage authentication, authorisation and identities, together with common services for digital rights, search services and metadata management.

The Australian Partnership for Sustainable Repositories^v (APSR) project is focussed on the critical issues of access continuity and the sustainability of digital collections.

The Australian Digital Theses^{vi} (ADT) project is creating a national collaborative distributed database of digitised theses produced at Australian Universities, by harvesting metadata about theses into a single resource discovery service for theses. This project is redeveloping and expanding the existing ADT database.

The ARROW, APSR, MAMS and ADT projects are working together to ensure that requirements and proposed solutions in all these areas are explored and shared to mutual benefit. For example agreement on consistent application of metadata standards across the four projects will be required to allow the MAMS authorisation and digital rights management strategies to be adopted by the other projects. Easy resource discovery across the set of repositories fostered by the four projects will be facilitated by the adoption of agreed metadata standards for access management, data elements and vocabularies within data elements.

Standards, Profiles, and in the meantime...

The field of institutional repositories is relatively new. DSpace was first released in late 2002^{vii}, and Fedora was first made available for downloading in May 2003^{viii}. Many aspects of institutional repositories are yet to be standardised, including data models and metadata schemas for describing and managing various types of objects. There is no widely accepted standard set of data models for various types of objects that would allow their export and import between disparate repositories without manual intervention. For example a thesis may be stored as a set of pages with one file per page, or as a set of chapters with a file per chapter, and separate files for the abstract and the bibliography. Suppressing selected parts of the thesis from viewing needs to be achieved in ways that are respected by search engines and a variety of potential viewing client software. Standardising the way other types of objects

are presented and managed presents a wide range of challenges requiring analysis and wide discussion.

Experience with the application of the Fedora architecture is limited, and Thus ARROW is making decisions in the abstract about how to structure objects in the Fedora repository architecture. Once we obtain more experience, and as the field evolves, some of these decisions may need to be revisited.

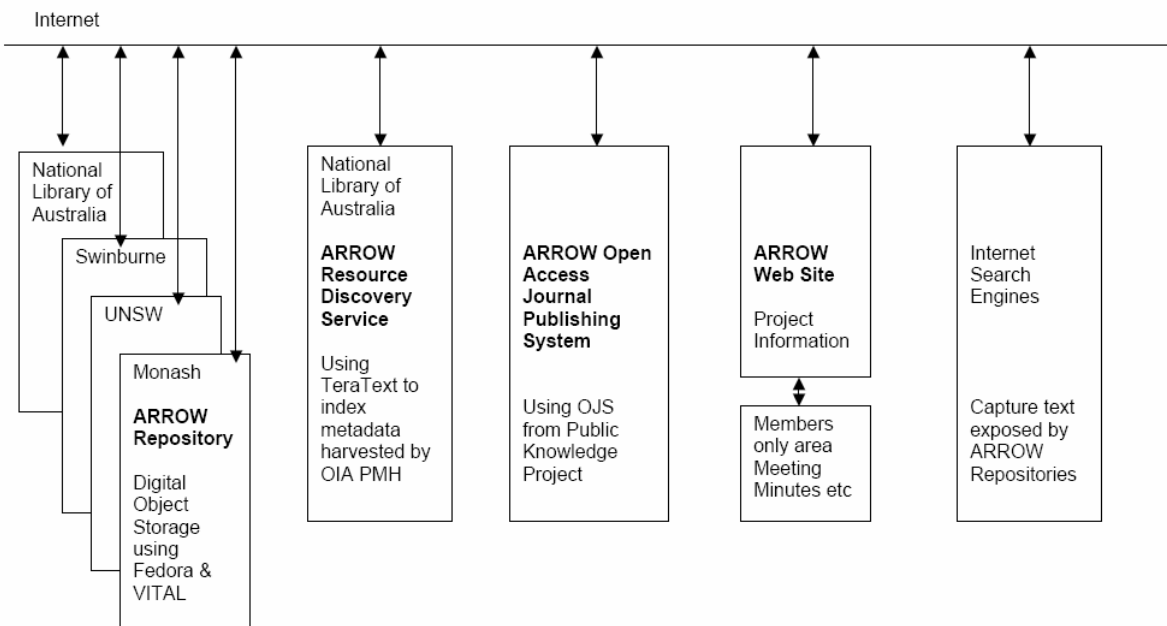
ARROW Project Governance

The ARROW project is overseen by the ARROW Management Committee, which is advised by the ARROW Technical and Content Committees.

ARROW Project Services Profile

At the end of its first twelve months work the ARROW project has adopted the structure shown in Figure 1 below to deliver the service profile envisaged in the funding bid.

Figure 1. ARROW SERVICES



ARROW Software

The ARROW project has at its heart a commitment to fostering the development of open source software to support institutional repositories. Once the ARROW repository solution is proven, and subject to a feasibility study, ARROW is committed to offering its repository solution to other Australian universities. The utilisation of open source software is one strategy intended to contain the costs of implementing the ARROW solution more widely.

Following a review of open source repository software in February 2004 ARROW decided to adopt the Fedora software as the basis of the ARROW repositories. This decision was based on an internal review and the published comparisons contained in the Guide^{ix} published by the Open Society Institute.

To be eligible for inclusion in the Guide, software systems must be:

1. freely available as open source software,
2. compliant with the latest version of the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH), and
3. currently released and publicly available.

The second criterion is fundamental to the development of the ARROW Resource Discovery Service, which harvests metadata from each of the ARROW repositories using the OAI-PMH protocol, then builds the discovery service using TeraText^x software. This service builds on the National Library's experience with its Picture Australia service, which harvests metadata to construct an index linking searchers back to the digital images housed at their owning organisations.

The Fedora software utilised by the ARROW project has been developed jointly by Cornell University and the University of Virginia. The name is an acronym for the Flexible Extensible Digital Object Repository Architecture^{xi} and this Fedora is not to be confused with the unrelated Red Hat sponsored Fedora project^{xii} to work with the Linux community to develop a general purpose operating system entirely from free software.

Because institutional repository software is an emerging field, ARROW decided to review annually its repository software choice to take account of any new developments. The first of these reviews in late 2004 concluded in a commitment to continue working with Fedora during 2005.

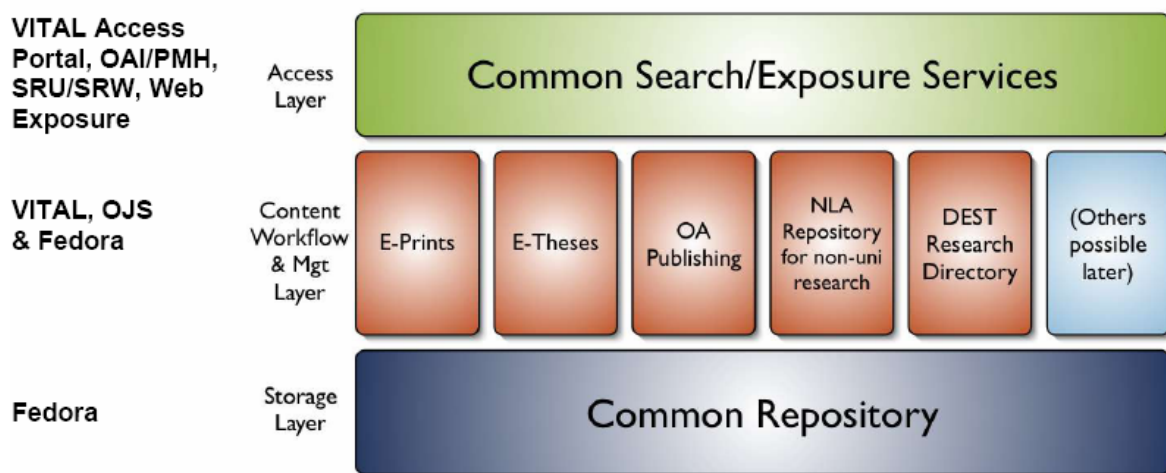
The Fedora software as it exists now is not a complete working repository system. Just as a relational data base software package needs complementary applications software, Fedora provides the core functionality required to operate a repository but needs an applications layer on top to perform functions such as managing workflows, performing object validation and managing metadata collection in schemas other than Dublin Core.

In January 2004, VTLS Inc, a well known library services company based in Virginia, announced VITAL, software to manage collections of images which uses the Fedora repository software as the storage layer. ARROW's approach to VTLS to consider further developing the VITAL software to meet ARROW's requirements was well received and by June a formal partnership between ARROW and VTLS had been established.

To support open access journal publishing, the Open Journal Systems^{xiii} (OJS) software from the University of British Columbia's Public Knowledge Project^{xiv} has been selected as a good fit for ARROW's requirements. This software has been used by the University of Technology Sydney to publish "Portal"^{xv}, and has been well received by the academics responsible for that journal. Swinburne University of Technology is taking the lead in the use of the OJS software in the ARROW context.

In relation to the three layer architecture for ARROW proposed in the funding bid, these software components together contribute functionality as shown in Figure 2.

Figure 2. ARROW SOFTWARE



ARROW Metadata

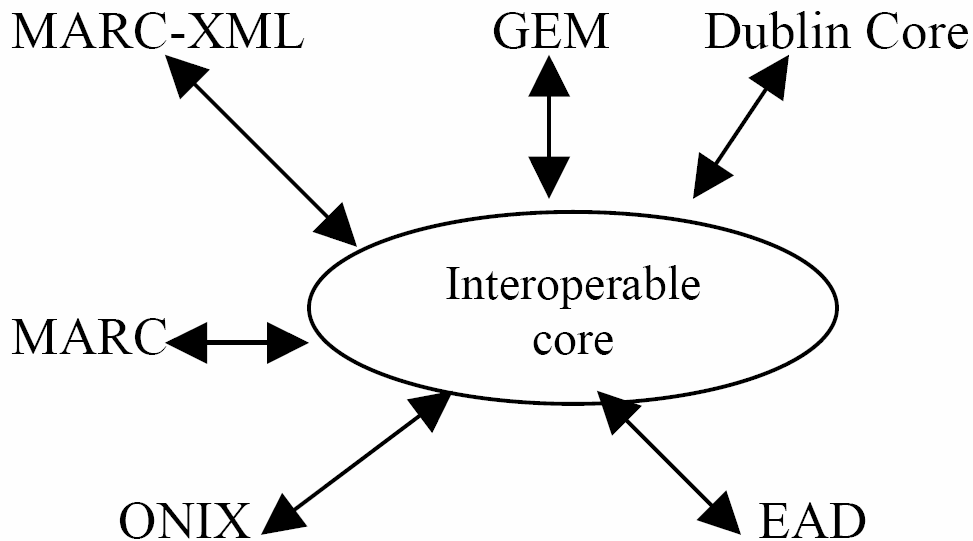
After much discussion it became clear that there was no single metadata schema that could be used effectively for every possible type of digital object we may be expected to accommodate in an institutional repository. Instead ARROW has decided that each type of object should be stored with the metadata developed for it by the community of practice that generated the object. This "native" metadata can then be mapped to Dublin Core for indexing in the ARROW Resource Discovery Service.

To achieve this mapping, ARROW has arranged with OCLC to test its metadata interoperability core. This is designed to perform transformations between individual metadata schemas by establishing inwards and outwards transformations between each individual schema and an interoperability core. The inwards transformation from any schema into the core is intended to be lossless. When this transformation is developed, if a schema has a data element not present in the core, that element will be added to the core. Outwards transformations from the core to other schemas may lose information depending on the characteristics of the target schema. Using the interoperability core thus allows any new schema to be accommodated by

establishing just its inwards and outwards transformations, following which it can be mapped to any other schema via the interoperability core. This is illustrated in Figure 3 below.

It is ARROW's intention to allow searching of the native metadata for any collection of objects stored in the ARROW repositories using Search and Retrieve Web and Search and Retrieve URL (SRW/SRU). SRW/SRU is a web implementation of the functionality delivered by Z39.50 compliant systems, but is much easier to implement. SRW/SRU's explain transaction allows a search client to interrogate a database as to what metadata fields can be searched, and it is this feature that we plan to use to allow native metadata searching for collections of objects in the ARROW repositories. Thus any loss of specificity or granularity in the transformation to Dublin Core can be overcome by accessing the native metadata for fine grained searching.

Figure 3 OCLC Metadata Interoperability Core



From Godby et al, Two paths to metadata interoperability^{xvi}

The ARROW metadata strategy described above allows us to proceed without having to anticipate all possible types of digital objects and their associated metadata at the outset of the project.

Our initial choice of metadata schema for the print equivalent objects to be ingested to the ARROW repositories in the first instance is MARCXML. This choice will facilitate the exchange of metadata with traditional library catalogues to permit their maintenance of aggregated metadata for objects such as theses held in print or microform on library shelves, as well as theses held in electronic form in the ARROW repositories.

ARROW Persistent Identifiers

Repositories need to assign persistent identifiers to their content to allow individual digital objects to be cited in ways that do not break if content is relocated from one repository to another, or the internal storage model (the data model) in the repository housing the object is changed for some reason such as performance tuning or better management of access rights. To meet the requirement for a persistent identifier for objects stored in the ARROW repositories, ARROW has selected the handles persistent identifier scheme. The preferred form of citation for an object in an ARROW repository will be as follows: <http://arrow.monash.edu.au/hdl/1959.1/1234>

Any browser can get from the above form of identifier to the Monash handles database which will keep track of the address of the cited item. This form of citation provides a degree of future proofing should http disappear as an Internet protocol at some future date, while avoiding reliance on web browser plug-ins required for browsers to recognise handles at present. Everything after the hdl is a standard Handle, which can be fed into any Handle resolver. Should Handles become unsustainable in future for some unanticipated reason, at the name of the owning institution in the web address gives a starting point for searching in next generation systems to locate the cited object.

ARROW data modelling

As mentioned above there are no agreed standard data models for objects in institutional repositories.

For “simple” objects such as an image collection, the image itself (in a commonly recognised file format) and some metadata in Dublin Core may be a sufficient data model. Depending on the size of the image, it may be desirable to store a thumbnail image for display in summary results screens or as a first response to a user rather than always downloading the high resolution image. Images maintained primarily for preservation purposes may be very large, whereas a low resolution image may be better suited for display on a computer screen where the screen characteristics are the limiting factor for image quality. Immediately a different data model is required to specify which images are to be presented in various circumstances, and additional technical metadata is required to identify the different images and the client software required to render them for viewing.

Only once the use cases for any given class of objects have been identified can the necessary administrative metadata and segmentation or variant versions of the object required for various purposes be determined.

Metadata to support fine grained access control to individual elements of a thesis, for example open access to the abstract but restricted access to the text or certain images, is much more complicated than descriptive metadata required for a simple object such as a low resolution image with no access restrictions.

For a complex object such as a book, thought has to be given to the appropriate level of granularity for the individual components such as chapters, diagrams and images likely to be reused separately from the work as a whole. Identifying an appropriate data model, then providing a persistent identifier and access control metadata for each unique reusable component is a challenge.

ARROW is still working through the use cases for objects to be stored in its repositories. At the initial meeting of the Fedora Development Consortium in Baltimore in October 2004 one of the highest priorities identified by participants was the sharing of data models amongst members, both to learn from what others had done, and to allow the cloning of existing data models where appropriate.

Attracting Content to the ARROW Repositories

In parallel with the focus on technology described above, ARROW has developed advocacy strategies to attract content to the repositories. We know that simply building a repository is not enough to ensure it will be utilised. This is borne out in the PALS Pathfinder^{xvii} research released in January 2004 which reported that the median number of objects in institutional repositories surveyed at that time was a surprisingly low 290 records per institution.

ARROW universities are at various stages along the path to the adoption of mandatory deposit of theses in electronic formats, and storing these in ARROW repositories will expose this previously unpublished research. Monash University is retrospectively digitising some theses from microfiche for inclusion in its ARROW repository.

Project champions have been identified whose research outputs will be incorporated into the repositories early in the project as an example to their colleagues. Research on the relative citation rates for online papers and printed papers is being tracked as a basis for encouraging academics to include their research outputs in the ARROW repositories. For example, in 2001 Lawrence^{xviii} identified that for computer science articles, online articles were cited at a mean rate of 2.6 times more frequently than offline articles.

There is also an administrative advantage to be achieved in streamlining the way in which universities compile the evidence of research outputs required as the audit trail for their statistical reporting to the Department of Education Science and Training. ARROW believes that capturing research outputs into the repositories will make this process easier to manage, and is investigating ways of interfacing ARROW with the software currently used to capture metadata about research publications at the partner universities.

The National Library will utilise its ARROW repository to capture research from independent scholars, and to explore managing the addition of materials published in digital format to its collections.

Various pro-forma documents have been produced explaining ARROW's objectives and the advantages which will accrue to academic departments through their contributing their research publications to the repositories.

Conclusion

At the time of writing ARROW has received for testing the first software custom developed by VTLs to our requirements, with further releases planned at intervals through the first half of 2005. Fedora version 2.0, scheduled for release in December 2004, is now expected before the end of January 2005. This has delayed some of the ARROW software developments that are dependent on features expected in the Fedora 2.0 release. Now that we have the first release of the software the implementation of content capture to the ARROW repositories can begin. From this will follow the process of refining and enhancement of the software based on experience in its use.

In summary ARROW is focussed on producing a generalised institutional repository solution, with an initial focus on managing and exposing traditional bibliographic research outputs. Design decisions are being taken with the intention of not precluding management of other digital objects such as learning objects and large research data sets.

ARROW^{xix} is looking forward to continuing to develop expertise in the design and management of institutional repositories in conjunction with the FRODO projects and our colleagues in the Fedora Development Consortium and elsewhere.

References

- ⁱ Backing Australia's Ability, See <http://innovation.gov.au/content/itrinternet/cmscontent.cfm?objectID=523BD2F9-CD5B-F1D4-4CC5FAEC2DF0F22B> visited 4 January 2005
- ⁱⁱ Cathrine Harboe-Ree et al, ARROW: Australian research repositories online to the world, 22 pp December 2003, at <http://eprint.monash.edu.au/archive/00000046/> visited January 10 2005
- ⁱⁱⁱ The betatest version of the Research Resource Discovery Service is available at <http://arrow-test.nla.gov.au/> visited 14 January 2005.
- ^{iv} Meta Access management System, See <http://www.melcoe.mq.edu.au/projects/MAMS/> visited 4 January 2005.
- ^v Australian Partnership for Sustainable Repositories, See <http://www.apsr.edu.au/> visited 4 January 2005.
- ^{vi} Australian Digital Theses Program, see <http://adt.caul.edu.au/> visited 4 January 2005.
- ^{vii} DSpace poised to transform research information storage, (press release) MIT News Office, 6 November 2002, at <http://web.mit.edu/newsoffice/2002/dspace-1106.html> visited 15 January 2005.
- ^{viii} Payette, Sandy, and Thornton Staples, The Fedora Project: An Open-source Digital Object Repository Management System, in D-Lib Magazine, April 2003 Volume 9 Number 4, ISSN 1082-9873 at <http://www.dlib.org/dlib/april03/staples/04staples.html> visited 10 January 2005.
- ^{ix} Open Society Institute "Guide to Institutional Repository Software" 2d ed January 2004, since superseded by the 3rd ed, August 2004 available online at <http://www.soros.org/openaccess/software/> visited 4 January 2005.
- ^x TeraText is described at <http://www.teratext.com> visited 15 January 2005.
- ^{xi} Fedora, see <http://www.fedora.info> visited 10 January 2005.
- ^{xii} Fedora (Red Hat Sponsored project) see <http://fedora.redhat.com/> visited 10 January 2005.
- ^{xiii} Open Journal Systems, see <http://www.pkp.ubc.ca/ojs/> visited 10 January 2005.
- ^{xiv} Public Knowledge Project, see <http://www.pkp.ubc.ca/> visited 10 January 2005.
- ^{xv} Portal, see <http://epress.lib.uts.edu.au/journals/portal/> visited 10 January 2005.
- ^{xvi} Godby, Smith and Childress, 2003 "Two paths to interoperable metadata" p.3, at <http://www.oclc.org/research/projects/mswitch/godby-dc2003.pdf> visited 10 January 2005
- ^{xvii} Publisher and Library/Learning Solutions (PALS) "Pathfinder research on institutional repositories, final report", January 2004 at <http://www.palsgroup.org.uk/palsweb/palsweb.nsf/> visited 15 January 2005.
- ^{xviii} Lawrence, Steve "Online or invisible?" 2001, at <http://ivyspring.com/steveLawrence/SteveLawrence.htm> visited 15 January 2004, an edited version appears in *Nature* 411 (6837): 521.
- ^{xix} Further information about the ARROW project is available at <http://arrow.edu.au>