Institutional Repositories Down Under: The ARROW Project Revealed

Presentation to OCLC, 2 September 2005

Geoff Payne,
ARROW Project Manager

The ARROW Project is funded by the Australian Commonwealth Department of Education, Science and Training, under the Research Information Infrastructure Framework for Australian Higher Education.

arrow.edu.au

The ARROW Consortium comprises Monash University [lead institution], National Library of Australia, The University of New South Wales and Swinburne University of Technology.

arrow
australian research repositories online to the world

MONASH University | NATIONAL LIBRARY OF AUSTRALIA | UNSW | SWINBURNE UNIVERSITY OF TECHNOLOGY

ARROW Institutional Repositories

Presentation structure:

- Why Institutional repositories?
- ARROW and the other FRODO Projects
- ARROW Services
- ARROW Software Strategy
- ARROW Metadata Strategy
- ARROW Content and Advocacy

What is an Institutional Repository?

A **managed** collection of digital objects

- institutional in scope
- with consistent data and metadata structures for similar objects
- enabling resource discovery by the “Communities of Practice” for whom the objects are of interest
- allowing read, input and export of objects to facilitate resource sharing
- respecting access constraints
- sustainable over time
- facilitating application of preservation strategies

Why Institutional Repositories? – As Good Management of resources

- Need to safeguard digital resources generated already by institutions.
- Existing digital resources often:
  - are managed by grace and favour arrangements
  - rely on unsustainable hardware, software or individual support
  - need future-proofing migration strategies
- Yet are widely used and reflect substantial investment in generating their content

Why Institutional Repositories? – As Research Enablers

- Need an enabling environment for other less technologically independent researchers
- Need to facilitate collaboration between researchers with similar interests but located in different faculties or institutions

Why Institutional Repositories? - Research Exposure and Impact

- Greater exposure & impact of institutional research outputs
  - Readership is otherwise limited to subscribers to the journal in which research is published
  - Better return on investment of public funds in research through greater accessibility
  - Can publish online material for which printing is not financially viable
  - Opportunity to expose materials other than the print friendly
  - Opportunity to preserve and expose research data sets for further analysis by others

Why Institutional Repositories? - Reforming Scholarly Publishing

- Potential to reform the scholarly publishing system, by demonstrating we can
  - Facilitate publication of research for which the audience is too small to justify the costs traditional publication mechanisms
  - Provide alternatives to expensive journals
  - Regain intellectual property rights over research outputs
  - Achieve shorter times between research output and its accessibility

Different Types of Repository Content

An Institutional repository may be expected to store any mix of anything that can be represented digitally

- Print equivalents – Research papers, Theses, books, book chapters, archival records
- Audio
- Still and moving images
- Multimedia objects
- Learning Objects
- Research data sets

Repositories - Technical Issues

- Interoperability
- Metadata
- Federated Searching
- Semantic web
- Authentication and Authorisation of users
- Rights Management
- Persistent Identifiers for digital objects

Repositories – Technical Issues – Interoperability

- Few standards are available to assist in the exchange of digital objects between repositories
  - No widely accepted data models for complex objects – cf SCORM for learning objects
  - Few “archival” formats agreed for digital objects
  - Few Metadata standards, but lots of pragmatic Metadata schemata to meet the needs of specific communities of practice

Repositories – Technical Issues – Metadata Exchange

- Dublin Core – insufficiently granular for many purposes
- Learning Object Metadata – not good for “bibliographic” metadata
- Need to preserve metadata relevant to categories of objects as decided by the “community of practice” that produced the object
- Open Archives Initiative Protocol for Metadata Harvesting (OIA-PMH) – can gather Dublin Core metadata to establish resource discovery services

Repositories – Technical Issues – Federated Searching

- eXtensible Access Control Markup Language (XACML)
  - No profile defined as yet to tag repository content to signify who can access it
  - Hence no standard way to allow search software to determine who can access what across a federation of repositories
    - Eg All University staff can access …
    - All enrolled students in “State” can access…
    - All members of “professional association” can access…
    - Research Assessment panel of “discipline” can view…

Repositories – Technical Issues – Semantic Web

- Semantic Web
  - Relies on machine interpretable data to allow application of business rules
  - Hence Metadata standards need to be granular and follow consistent encodings of concepts
  - Example - Machine analysis of citations to link to full text often fails as citations are not consistently expressed

Repositories – Technical Issues – Authentication, Authorisation, Rights Management

- MAMS Project is working in this area
  - Shibboleth as a model
  - XACML as a way of encoding fine grained access control
  - Digital Rights Expression Languages and Patents
- Repositories need access control to honour constraints imposed by copyright owners
  - eg to meet the ROMEO database expressions of publishers permissions policies for depositing previously published content to repositories

Repositories – Technical Issues – Persistent Identifiers

- Repositories need to offer a preferred form of citation for their content
  - Which does not break as URLs do when files are moved or web sites restructured
  - Handles from CNRI seem to be becoming widely adopted
    - DOI (Digital Object Identifier is a Handle)
    - UK Stationery Office adopting Handles
    - DSpace uses Handles

Repositories - Open Source Software and Sustainability

- The business case for open source software is not necessarily clear cut
  - Red Hat model - “manageable” open source software for fee
  - Complete self reliance
  - Reliance on a consortium of users of a particular product
  - Total cost of ownership is difficult to calculate
- But open source software is suited as an environment for preservation – no software features are buried in proprietary encodings which compromise the ability to extract content

ARROW Project

ARROW Consortium Partners

- Monash University (Lead Institution)
- University of New South Wales
- Swinburne University of Technology
- National Library of Australia

- October 2003 funding granted
- AU$3.66 Million over three years to identify and test solutions to establish institutional repositories at the ARROW partners

ARROW stages

- Demonstration (2004)
  - Developing architecture, selecting, testing and developing software
- Deployment (late 2004 – end 2005)
  - Populating the ARROW Partners' repositories
- Distribution (mid 2005 – end 2006)
  - Enabling others to participate
    - Under review for earlier participation by others

The FRODO Projects

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

ARROW FRODO Partnerships

- MAMS
  - Access control through eXtensible Access Control Markup Language (XACML) metadata
  - Needs development of a FRODO profile of XACML for access control interoperability
- APSR
  - Interoperability through consistent metadata for similar objects
  - Needs FRODO Metadata schemata for object exchange, export and ingest into new repository environments as part of sustainability and preservation initiatives
- ADT
  - Interoperability through harvestable Dublin Core metadata
  - Supporting e-theses online which are pointed to from ADT
- Web services strategy?

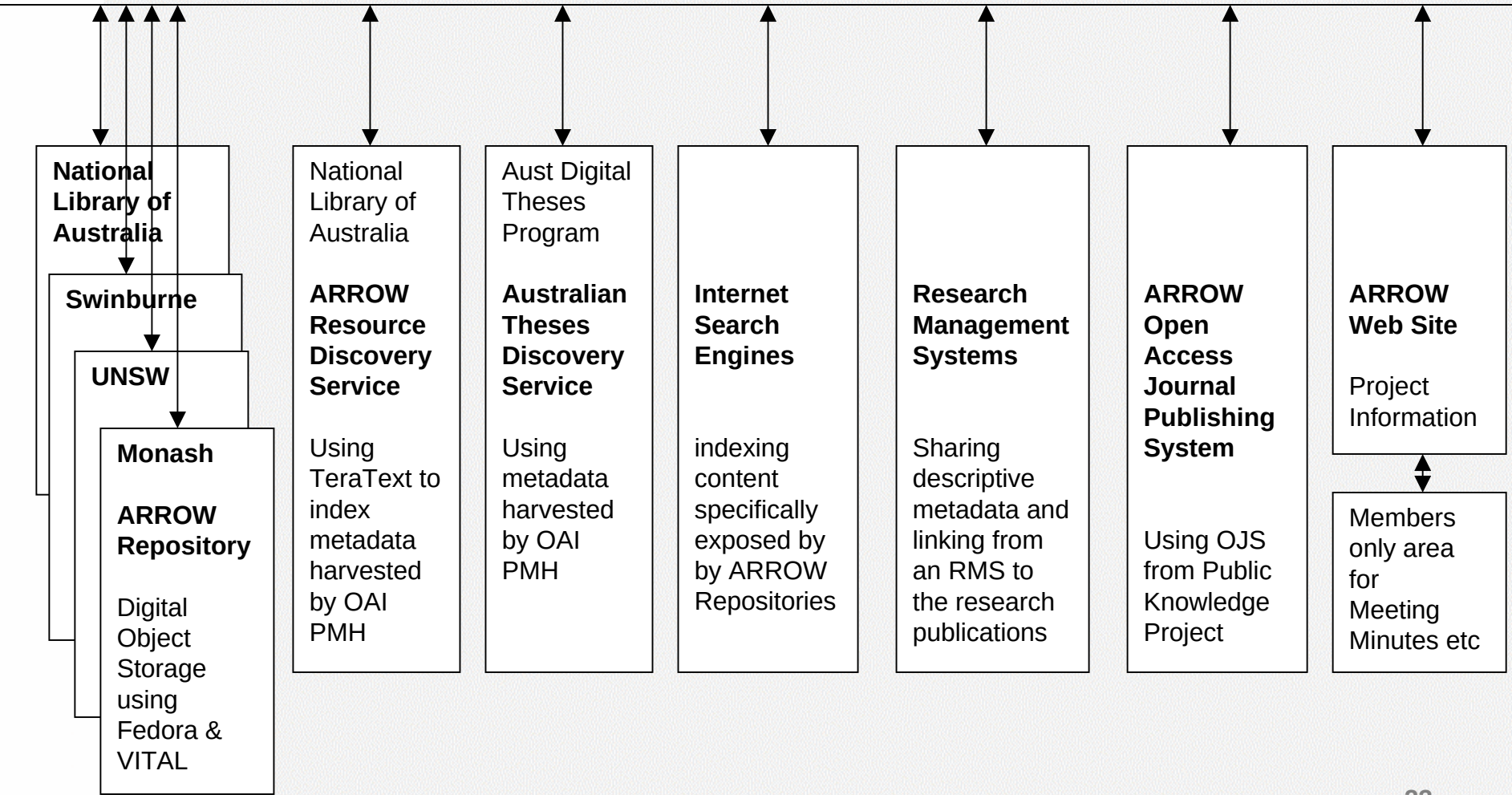
Australian Partnership for Sustainable Repositories

- Audit of sustainability of existing collections of digital objects (some of which are repositories by another name)
  - While ARROW is focused on “greenfield” repositories
- Dspace in use at several partner sites, including Australian National University

ARROW Project Governance

- **ARROW Management Committee**, Advised by
  - **ARROW Technical Committee**
    - Developing a vehicle for content management
  - **ARROW Content Committee**
    - Content issues
    - Advocacy to achieve cultural changes to ensure content capture

ARROW Branded Services Profile



ARROW Technology – Software

Needed a repository system early in the project

- To learn what works and what does not work
- To manage content as a demonstration system
- But all repository software is immature at present

Commitment to open source software in the ARROW Funding Agreement

- Evaluation of DSpace, Fedora, other software

ARROW commitment to Open Source Software

- Open Society Institute “A Guide to Institutional Repository Software” 3d ed August 2004
  - Software systems criteria for inclusion:
    - Freely available as open source software
    - Compliant with the latest version of the Open Archives Initiative Protocol for Metadata Harvesting
    - Currently released and publicly available

- ARROW Internal review of open source repository software

ARROW Technology – Software Selected

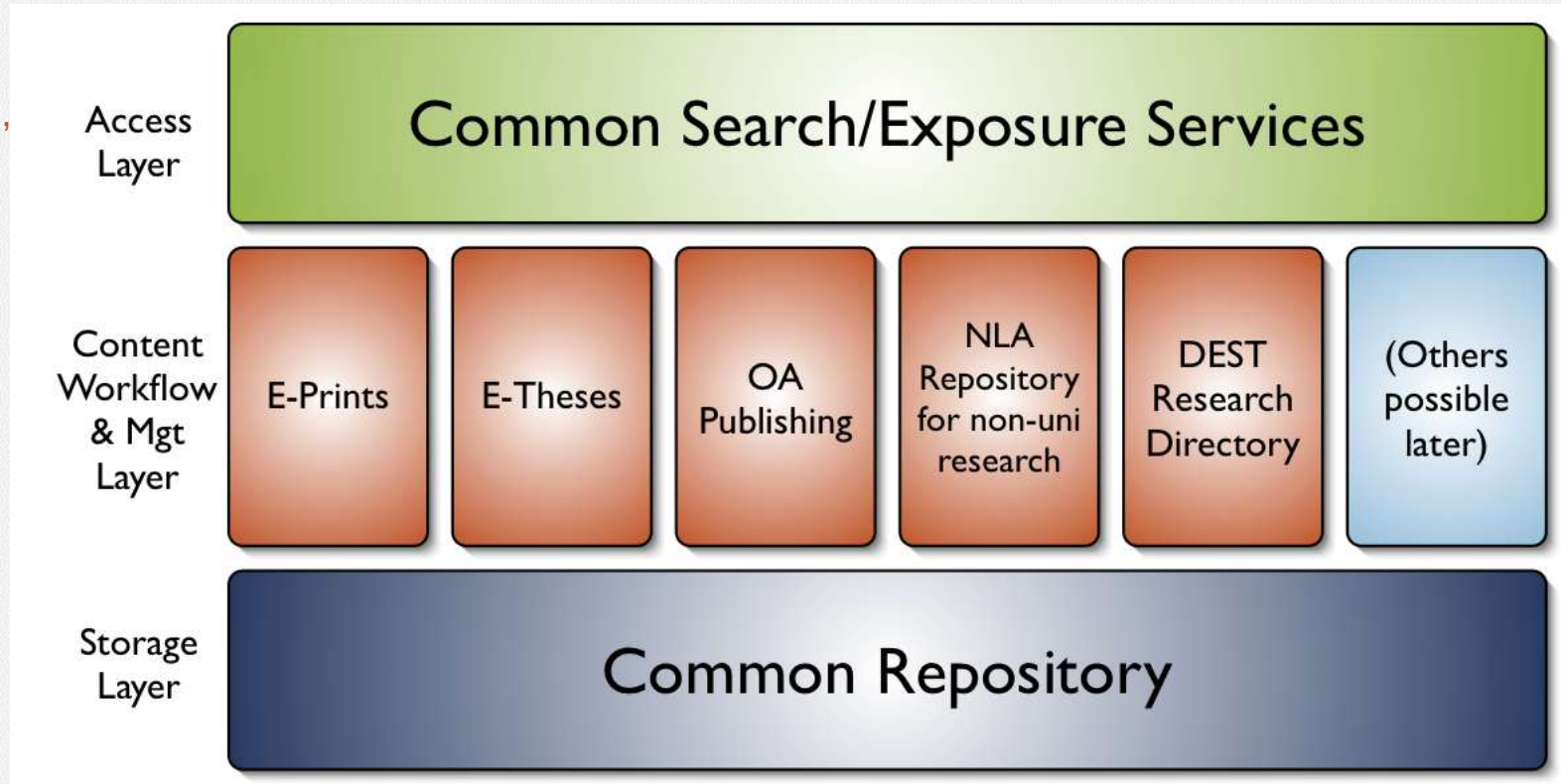
- **Flexible Extensible Digital Object Repository Architecture - Fedora™** http://fedora.info
  - Cornell and University of Virginia
  - ARROW a founding member of the Fedora Development Consortium
- **VITAL** from VTLS Inc http://www.vtls.com
  - ARROW / VTLS partnership to take the Fedora "engine" and construct a working repository to meet ARROW's functional requirements using VITAL and open source web services
  - Sustainability through vendor support
- **Open Journal Systems** (OJS) from Public Knowledge Project (University of British Columbia) http://www.pkp.ubc.ca/ojs/
  - for open access journal publishing

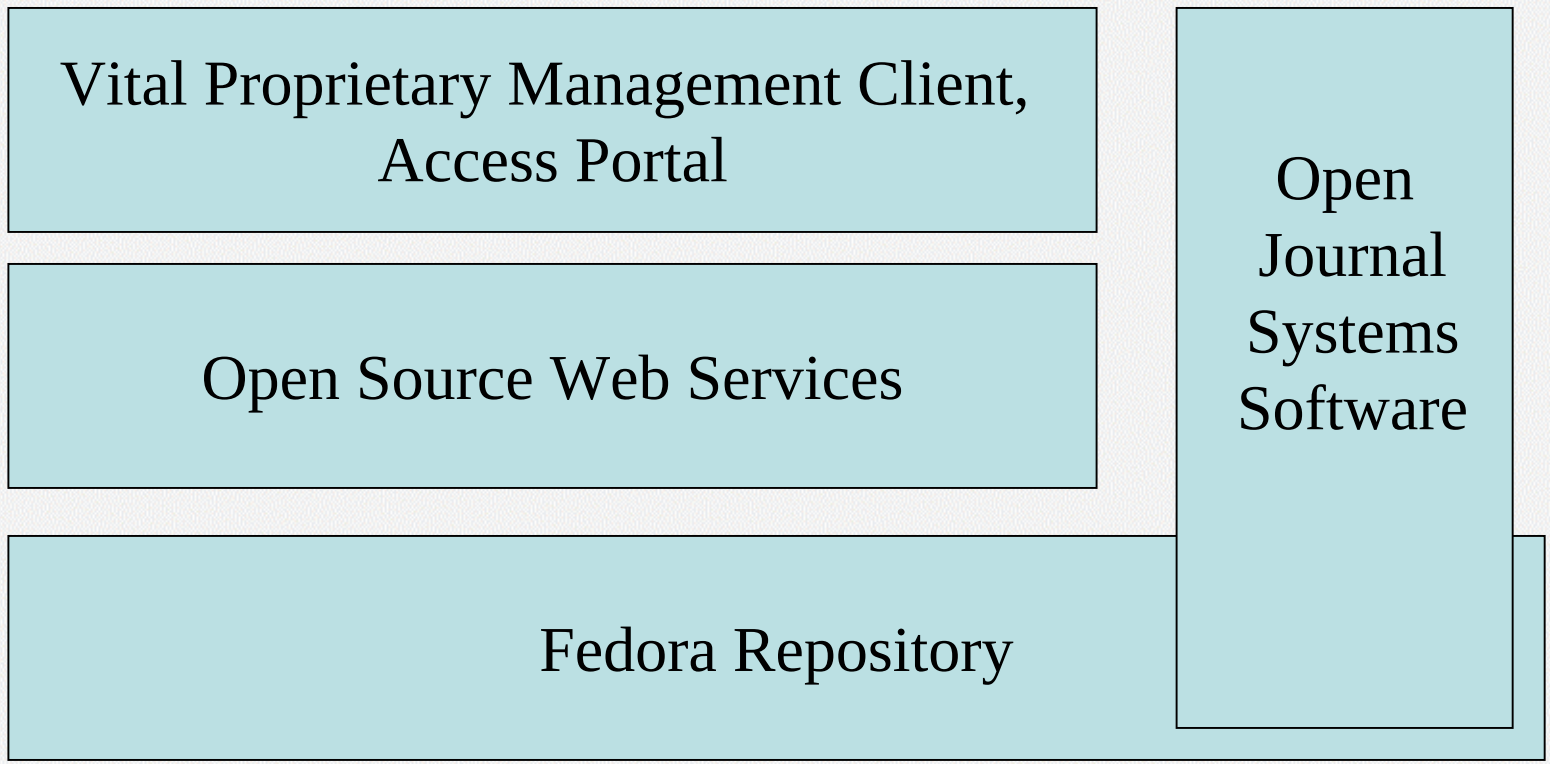
ARROW Architecture & software components

VITAL Access Portal, OAI/PMH, SRU/SRW, Web Exposure

VITAL, Fedora, OJS

Fedora

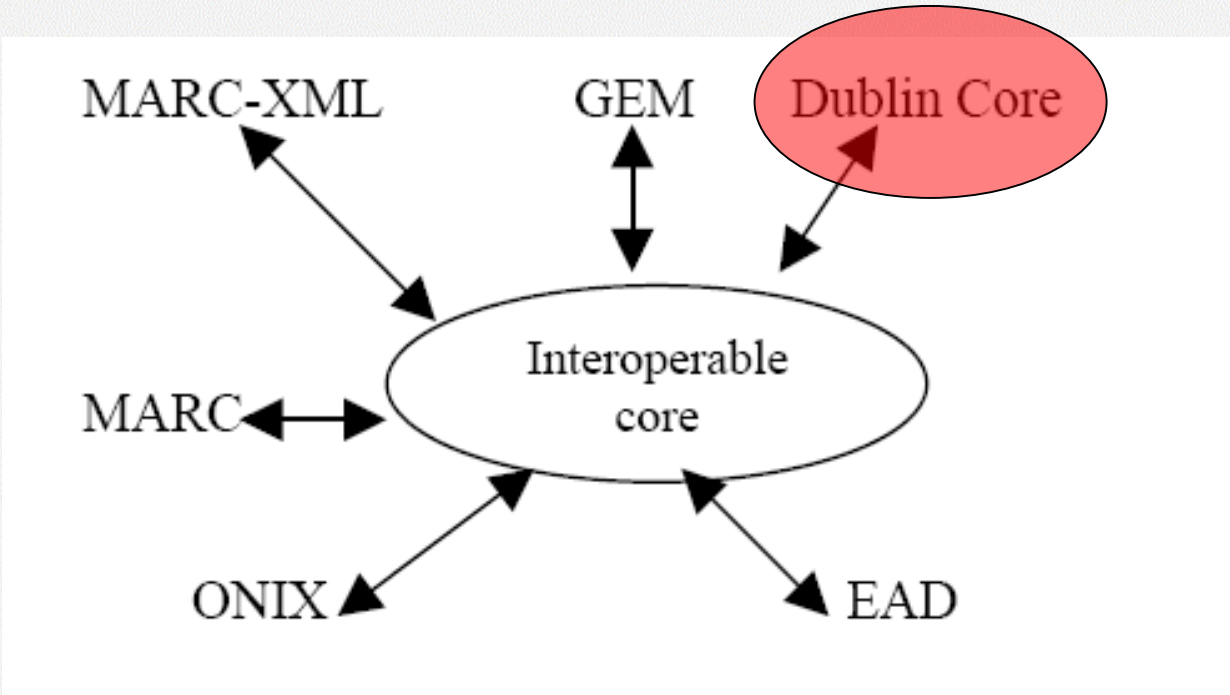



Vital Proprietary Management Client, Access Portal
Open Source Web Services
Fedora Repository
Open Journal Systems Software

ARROW Metadata Strategy

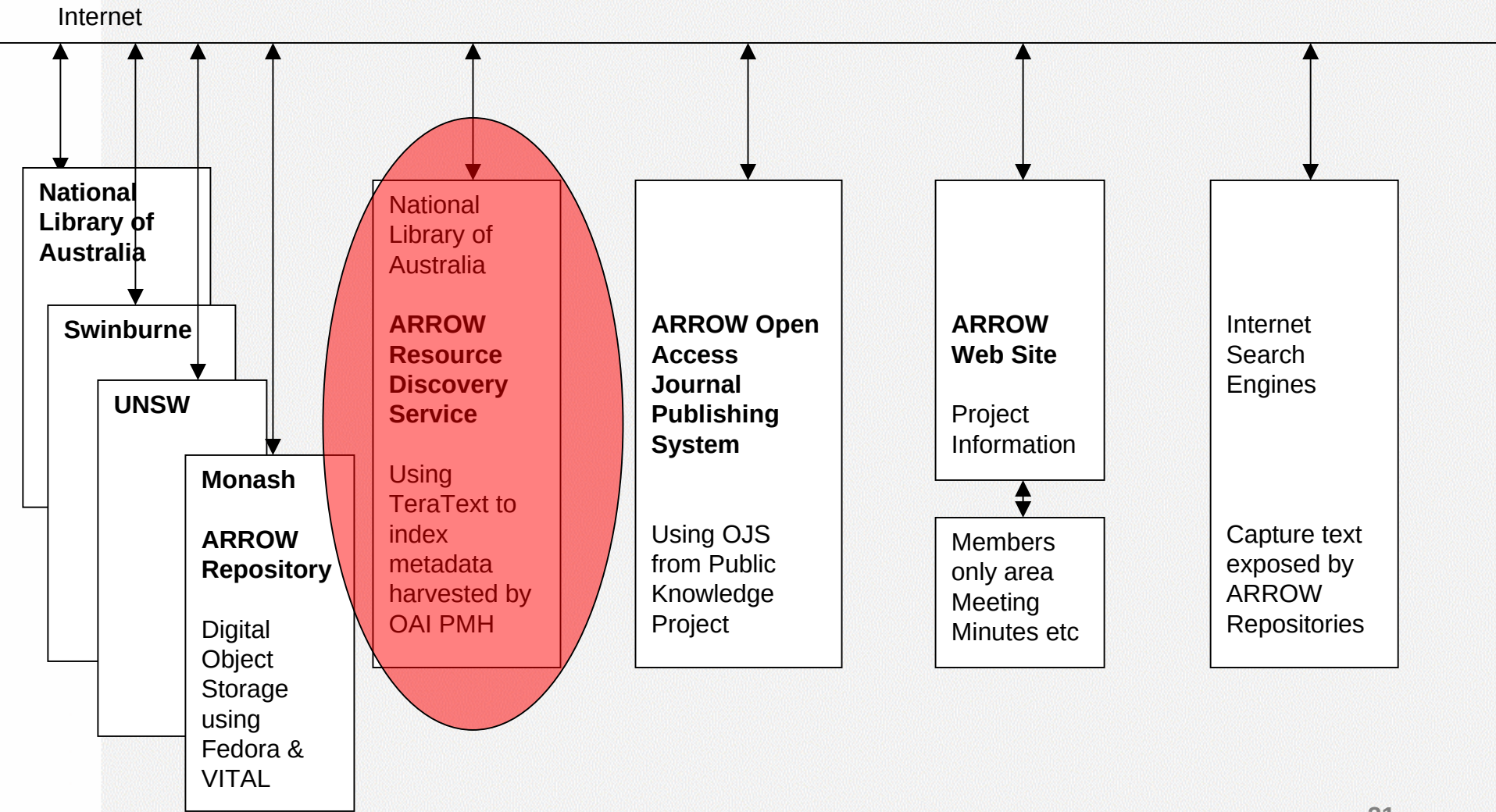
- Supports metadata schemata to suit individual data models
  - No requirement to shoehorn all metadata into one schema
  - Each stored object can retain metadata developed for it by the community of practice which generated the object
  - Maintains flexibility to store many types of digital objects in the repository
  - No need to anticipate every object type now

OCLC Metadata Interoperability Core



From: Godby, Smith and Childress. 2003. “Two paths to interoperable metadata” p. 3 at http://www.oclc.org/research/publications/archive/2003/godby-dc2003.pdf

ARROW Branded Services Profile

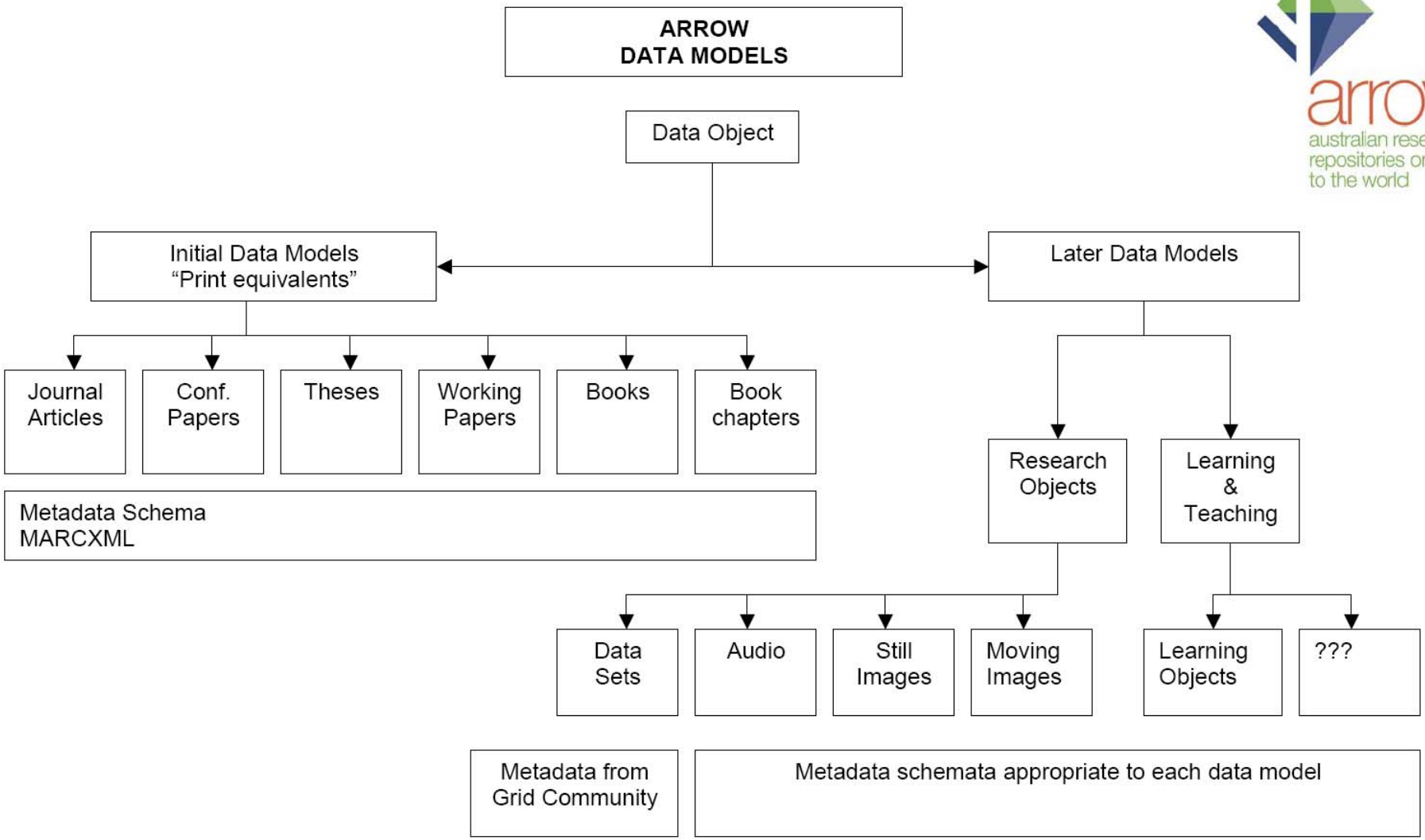


Fedora™ - Flexibility at the expense of implementation design effort

- Allows storage of any number of different types of digital objects
- But extra effort required
  - Data Modelling
    - How any given type of digital object will be stored can be tailored to suit
  - Metadata schemata for each data model (or even every object!) are allowed
  - Persistent Identifiers – Multiple identifiers from different schemes can be used

ARROW - Data modelling

- Required to define how objects will be stored
  - How many parts are there in any given object that may be cited and repurposed separately
    - For example a diagram may be used in a lecture presentation
  - Do different access controls apply to different component pieces of an object
    - For example a chapter of a thesis with culturally sensitive materials
  - Need to establish use cases, then determine what metadata is required to manage each use case



Notes:

1. Each of the data objects may be simple or complex (ie have one or more data objects as components)
2. Where an Object is complex it may include a mix of bibliographic and/or non-bibliographic data objects
3. An object or components of complex objects correspond logically to a FRBR Expression of a work.

Repository Persistent Identifiers - Recapping

- Repositories need to offer a preferred form of citation for their content
  - Which does not break as URLs do when files are moved or web sites are restructured
  - Handles from CNRI seem to be becoming widely adopted
    - DOI (Digital Object Identifier is a Handle)
    - UK Stationery Office adopting Handles
    - DSpace uses Handles

ARROW Repository Persistent Identifiers

- ARROW Handles* Format adopted:
  - http://arrow.monash.edu.au/hdl/1959.1/nnnn
  - 1959 = ARROW handles naming authority
    - 1959.n – one sub number for each ARROW repository
  - nnnn – running number

  - ARROW will assign a handle to each datastream in a digital object to ensure that individual parts of the digital object can be cited and re-used independently

*http://www.handle.net/index.html

ARROW Content Committee

Unfortunately it is not as simple as build it and they will come…

Publisher and Library/Learning Solutions (PALS) Pathfinder research on web-based repositories , Final Report, January 2004

"We find that IRs are currently rather small, with an average (median) of 290 records per institution (smaller but comparable to the median size of other OAI data providers). (Page 33)"

Incentives are needed for academics to submit their materials to repositories

- Substantial advocacy is required to achieve participation
  - Mandatory deposit of e-Theses
  - Credits towards promotion
  - Funding linkages
  - Demonstrable additional exposure such as in Web Citation indexes and search engines

ARROW Content (Advocacy)

- Advocacy tools prepared and circulated
  - Pro Forma Memorandum of Understanding with a university faculty of department
  - Copyright strategy paper drafted
  - ARROW Frequently Asked Questions
- Pursuing policy changes such as mandatory deposit of e-Theses
- Project champions recruited

ARROW Content (Advocacy) (Continued)

- Design work proceeding on an interface between Research Master (RM) and ARROW for gathering Australia's Department of Education Science and Training (DEST) research evidence
  - Monash, Swinburne, UNSW all use RM v.4, but the solution will be generalised to accommodate other practices
- Migration of content from e-prints repositories planned

DEST Research Evidence

- Around 30-40% of Australian university research funding comes from government
- At present an annual statistical return is required and audit evidence of research outputs is compiled as collections on paper
- ARROW can improve the efficiency of this process as an information management tool
- In Monash context this will capture 4000 publications annually

Research Quality Framework

- Australian Government wanting
  - wider exposure of Australian research
  - Demonstrable value for money, “research impact”
- Moving to a research quality framework similar to the united Kingdom Research Assessment Exercise (RAE)
- RAE 2008 guidelines mandate assembling research outputs in institutional repositories
- RAE utilises approx 70 discipline specific expert panels to rank research outputs from UK universities
  - Around 25% of UK govt research funding goes to the top 4 Universities
- ARROW is envisaged as a tool for the Australian RQF

ARROW Software Development – Current Status August 2005

- VITAL 2.0 alpha over Fedora 2.0
  - Image Management
  - Text Documents
  - VITAL Batch ingest tool for digital objects and/or metadata
  - Handles integration for automatic assignment of persistent identifiers
  - SRU/SRW interface
  - Audio, Moving Pictures and SMIL support
  - Support for Google spidering

ARROW partnerships

- OCLC
  - To test the metadata interoperability core
- Google
  - To test indexing of research materials
- Open Journal System (OJS)
- Thomson ISI Web Citation Index
- VTLS and Fedora
- Research Master

ARROW progress to date – Open Journal Publishing

- Provides full suite of publishing functionality
  - Peer review management
  - Assembling and publishing journal issues
  - Well liked by academics using the software
- Swinburne University is leading ARROW open journal publishing activities
  - E-Journal of Applied Psychology launched 1 July 2005
  - See http://www.swin.edu.au/lib/ir/onlinejournals/ejap/
- Informed by University of Technology Sydney “Portal” e-journal
  - See http://epress.lib.uts.edu.au/journals/portal/

ARROW progress to date - Repositories

- Repository software installed at all four partner sites
- Test repository at Monash
- Paper content models penidng software management of content models
  - Consistent metadata and datastreams for like objects
- VITAL 1.3 software release on production repository servers
- VITAL 2.0 alpha over Fedora 2.0 in testing at ARROW and VTLS
- Awaiting Fedora 2.1 software release due any day to incorporate XACML access control
- Work plan for July 2005 to May 2006 agreed between partners and with VTLS
  - Enhanced user interface first priority in VITAL 2.1 due October

VITAL 2.0 capabilities

- Manual or batch ingest of digital objects by partner staff
  - Automatic assignment of Handles persistent identifiers
  - JHOVE content validation
  - MARCXML metadata to Dublin Core transformation
  - Advanced searching
  - SRU/SRW
  - OAI harvesting to populate the ARROW Discovery Service
  - User configured indexing
  - User defined Fedora object structures
  - Web submission tool for end user e-theses deposit

VITAL 2.0 capabilities – Imminent

- Web based deposit for theses
  - Followed by review and manual ingest by staff
- Batch ingest tool
  - Match metadata files and object files on various criteria
- Exposure of content to web search engines
- Enhanced user interface
- VITAL 2.1 release to include integration with Fedora 2.1
  - Support for XACML access controls
  - Improved user interface including browsing

ARROW demo – VITAL 2.0 alpha

http://www.swin.edu.au/lib/ir/onlinejournals/ejap/

http://arrowdev.lib.monash.edu.au:8000/access

http://arrowdev.lib.monash.edu.au:8000/access/explorer.php

http://arrowprod.lib.monash.edu.au:8000/access

http://arrow.edu.au

http://search.arrow.edu.au

ARROW demo – VITAL 2.0 alpha

Monash theses ingested using the web self submit tool:

Http://hdl.handle.net/1959.100/630 Treloar, A.E : Hypermedia online publishing: the transformation of the scholarly journal

http://hdl.handle.net/1959.100/628 9.7 MB Robilliard, Frederick Emile:Studies of hollow-cathode metal vapour ion lasers

Monash theses ingested using the batch ingest tool:

http://hdl.handle.net/1959.100/253 Smith, David Alan, Antecedents and outcomes of multiple dimensions of accountants' organisational commitment

http://hdl.handle.net/1959.100/261 Wilson, Campbell:Visual information retrieval via inference networks

http://hdl.handle.net/1959.100/267 Karmakar, Gour Chandra: An integrated fuzzy rule-based image segmentation framework

ARROW demo – VITAL 2.0 alpha

JPEG http://hdl.handle.net/1959.100/459 Composite weather image of a tropical cyclone Creator NASA

http://hdl.handle.net/1959.100/457 Satellite image of Victoria and Northern Tasmania Creator NASA

MrSid images with navigation

Advanced Search:Title: ag050002 http://hdl.handle.net/1959.100/516 Victoria Dock, 1972 and 2002

Advanced Search:Title: ag050009 http://hdl.handle.net/1959.100/507 Victoria Dock, circa 1910 and 1942

ARROW demo – VITAL 2.0 alpha

Text and supporting images

Advanced Search:

Title: muddies http://hdl.handle.net/1959.100/418
History Australia, Volume2, No.1, 2004. Ferals and their muddies: Making a home in the bush

Text only

RTF text Advanced Search:

Title: requirements http://hdl.handle.net/1959.100/486

ARROW demo – VITAL 2.0 alpha

XML plus images: Advanced Search: Title: residential
http://hdl.handle.net/1959.100/283 Melbourne 2030: Chapter 5 - Residential infill and its threat to Melbourne's liveability.

MPEG movie http://hdl.handle.net/1959.100/586 Medical computer animation #20

Quiktime movie http://hdl.handle.net/1959.100/566 Medical computer animation #10

mp3 audio http://hdl.handle.net/1959.100/571 Ash Grunwald, Bakelite Radio and Blues Progression 5Mb

AVI movie: Advanced Search: Title: fantastic
http://hdl.handle.net/1959.100/551 Fantastic Four, movie trailer **WARNING large file 16Mb** DivX codec must be installed first to view.

Building on ARROW

August 2005 Strategic Infrastructure Initiative funding announced for (among others):

- DART (Monash University as lead institution)
  - Supporting the e-research lifecyle
  - Includes managing large datasets in the ARROW repositories
    - Interfacing Fedora and Storage Resource Broker or similar technologies
- RUBRIC (University of Southern Queensland as lead institution)
  - Implementing ARROW in a further six to eight regional universities in Australia and New Zealand
    - Application to management of learning objects
- IP management in repositories (Queensland University of Technology lead institution)
  - Including Creative Commons Australianisation

ARROW - Summary of design criteria

- A generalised institutional repository solution
- Initial focus on managing and exposing traditional bibliographic research outputs
- Expand to managing non-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
- Research Quality Framework likely to drive deposit of content by academics and research managers in ARROW universities

Questions?

Further information?

Details of the ARROW project can be found at:

arrow.edu.au

The ARROW site includes links to the FRODO projects and a glossary of repository acronyms and projects