The ARROW project after 2 years

David Groenewegen
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

Outline

- What is ARROW?
- What did we set out to achieve?
- What is the state of play now?
- What do we still need to achieve?
- What have we learnt so far?
- Where to from here?

What is ARROW?

- Funded by the Australian Commonwealth Department of Education, Science and Training, under the Research Information Infrastructure Framework for Australian Higher Education.
- Consortium comprises Monash University (lead institution), National Library of Australia, the University of New South Wales, and Swinburne University of Technology.

"The ARROW project will identify and test software or solutions to support best practice institutional digital repositories comprising e-prints, digital theses and electronic publishing."

Why have a repository?

- Provides a platform for promoting research output in the ARROW context
- Safeguards digital information
- Gathers an institution's research output into one place
- Provides consistent ways of finding similar objects
- Allows information to be preserved over the long term
- Allows information from many repositories to be gathered and searched in one step
- Enables resources to be shared, while respecting access constraints (when software allows access controls)
- Enables effective communication and collaboration between researchers

What did we set out to achieve?

- A generalised institutional repository solution for research information management
- Not about Open Access – although that is possible too
- Initial focus on managing and exposing traditional “print equivalent” research outputs
- Expanded to managing other digital research outputs
- Design decisions accommodate management of other digital objects such as learning objects and research inputs such as large data sets
- DEST Research reporting and audit, and Research Quality Framework likely to drive deposit of content by academics and research managers in ARROW universities
- Employing Open Standards where possible

Repositories : Open Source Software and Sustainability issues

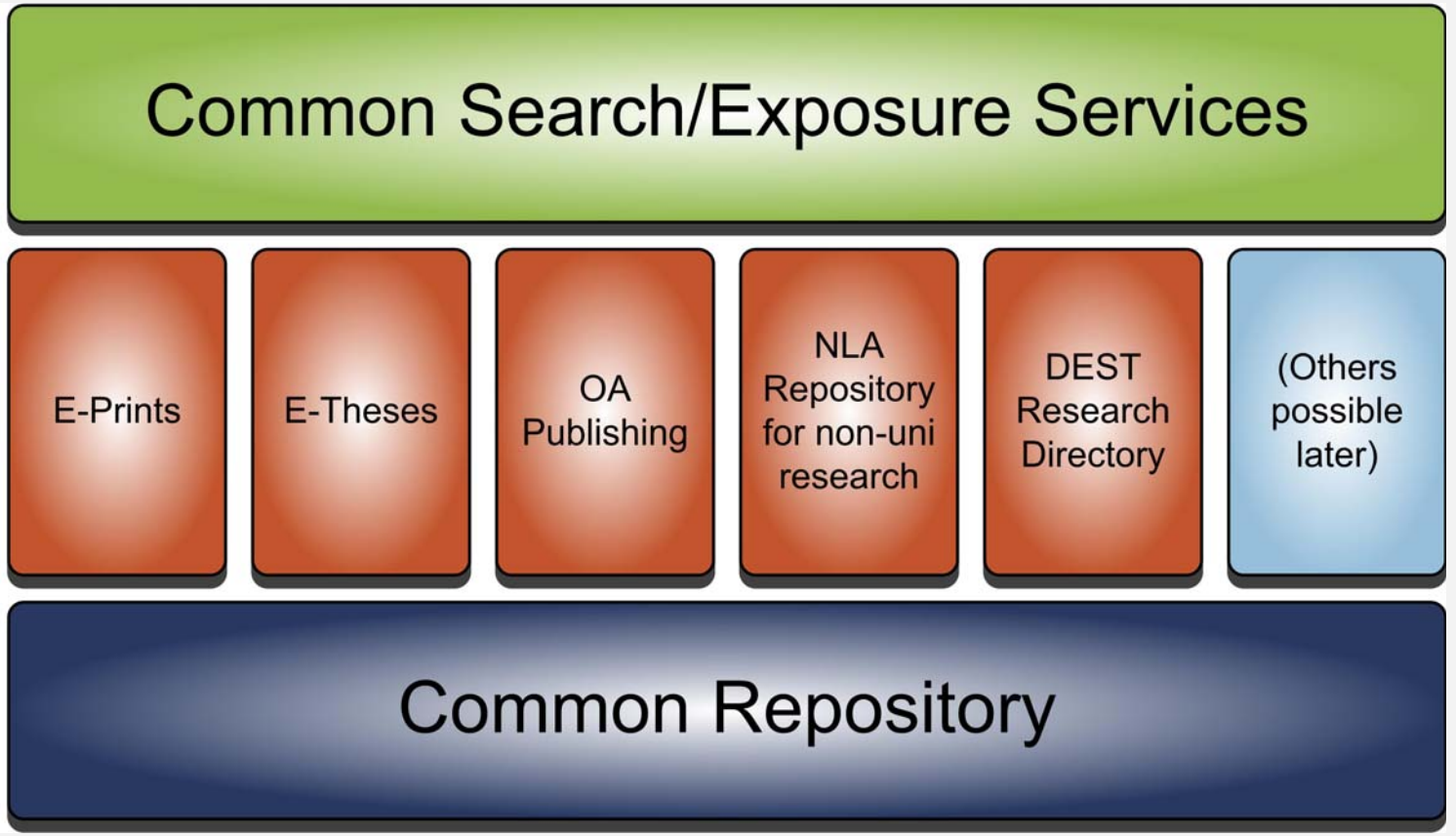
- The business case for open source software is not clear cut
  - Red Hat model - “manageable” open source software for fee
  - Need for reasonable level of technical expertise
  - Complete self reliance
  - Reliance on a consortium of users of a particular product
  - Total cost of ownership is difficult to calculate
  - 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
  - Proprietary software has advantage of support

Original vision and implementation

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

Fedora and VITAL

Fedora



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

Why Fedora?

- **ARROW needed something as a** platform **to develop its own application(s)**
- **ARROW wanted a** flexible object-oriented **data model**
- **ARROW wanted to be able to have** persistent identifiers **down to the level of** individual datastreams**, accommodating its compound content model**
- **ARROW wanted to be able to** version **both** content **and** disseminators **(which can be thought of as software behaviours for content)**
- **ARROW required clean and open exposure of** APIs **with well-documented SOAP/REST web services.**

Fedora satisfied these requirements

Why VTLS?

- Wanted to be a development partner
- Willing to create a combination of Open Source and proprietary solutions
- Familiar with library sector

Contracts and licensing

- Monash is the lead institution, and has a Head License Agreement with VTLS
- VTLS are required to make a certain amount of work Open Source
- Monash arranges a sublicense to participating institutions
- Sub-licensees enter a separate maintenance agreement directly with VTLS for support of the Software
- Monash coordinates software development
- VTLS installs, trains and supports
- System and training documentation are made available by VTLS
- ARROW to establish an ARROW community

Some key areas ARROW has tackled

- Descriptive metadata
- Persistent identifiers
- Open versus Controlled access
- Compound vs. atomistic object model
- Use cases and content modelling

Descriptive metadata

- Decided to support the metadata generated by communities of practice to accompany their digital objects
- Developed content models using MARCXML for seven commonly used research objects
- VITAL 2.0 can transform MARCXML and ETD-MS into Dublin Core for OAI-PMH and internal purposes
- Investigating possibility of using OCLC's interoperable core
- May need to write something ourselves

Persistent identifiers

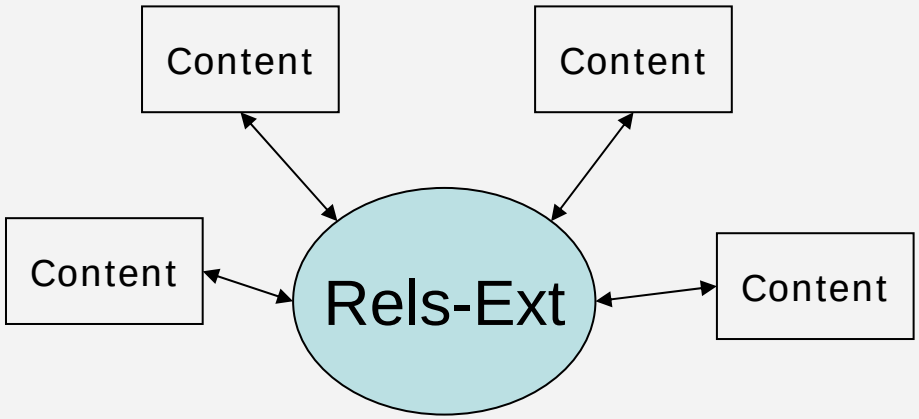
- After **long** discussions decided to adopt CNRI handles
- Wanted to be able to persistently cite both objects and components of objects
- ARROW software assigns handles to:
  - each entire ARROW object (such as a thesis)
  - every component or datastream of an ARROW object (such as the metadata, the thesis abstract, the thesis body, and the reference list)
- Repository managers can disaggregate and re-aggregate objects as required in the future without user being aware of it.
- Minimum persistently citeable unit can be made as granular as is required

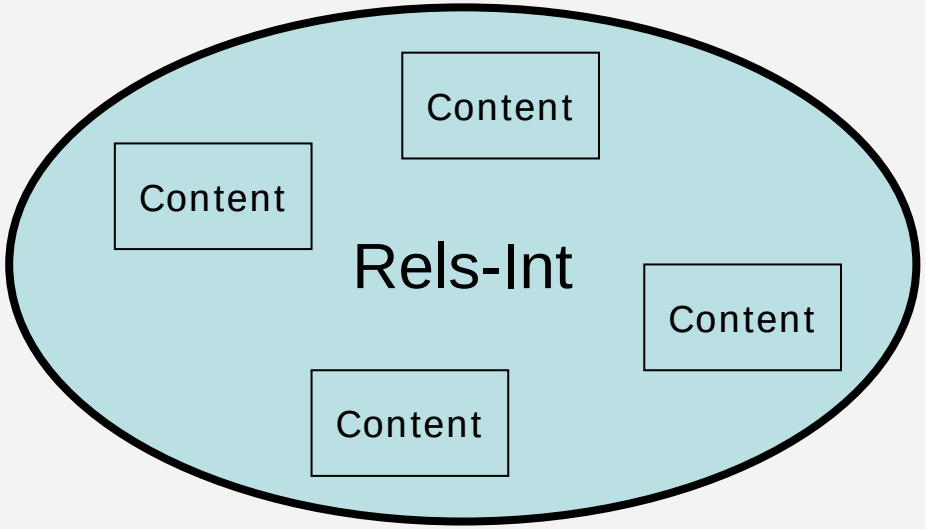
Open versus Controlled access

- Open Access in the classical sense is not always appropriate
  - Requires access controls in some areas
  - Planning to use XACML delivered in Fedora 2.1
  - Co-ordinating work to develop standard XACML policy vocabulary

Atomistic vs. Compound object model

Atomistic – a data object with one or more content datastreams that are all considered primary to the object.



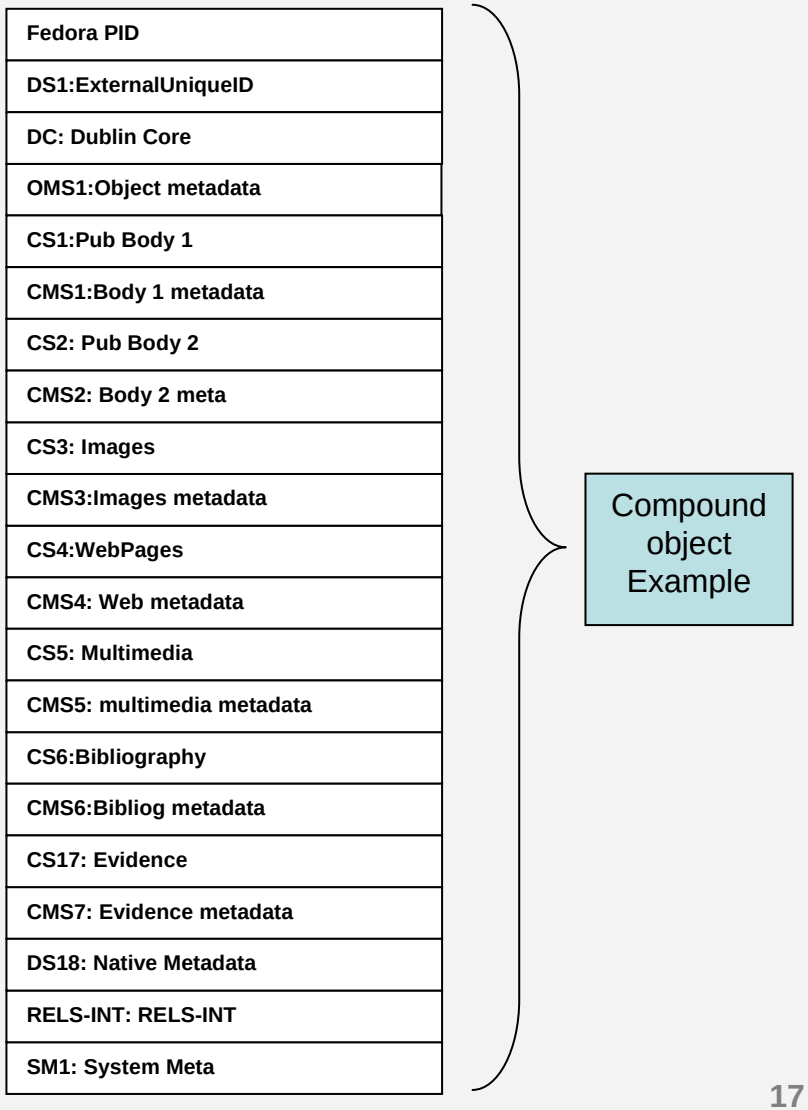


Compound – a data object consisting of multiple content datastreams that are not all primary to the object.

Arrow has chosen the compound object model

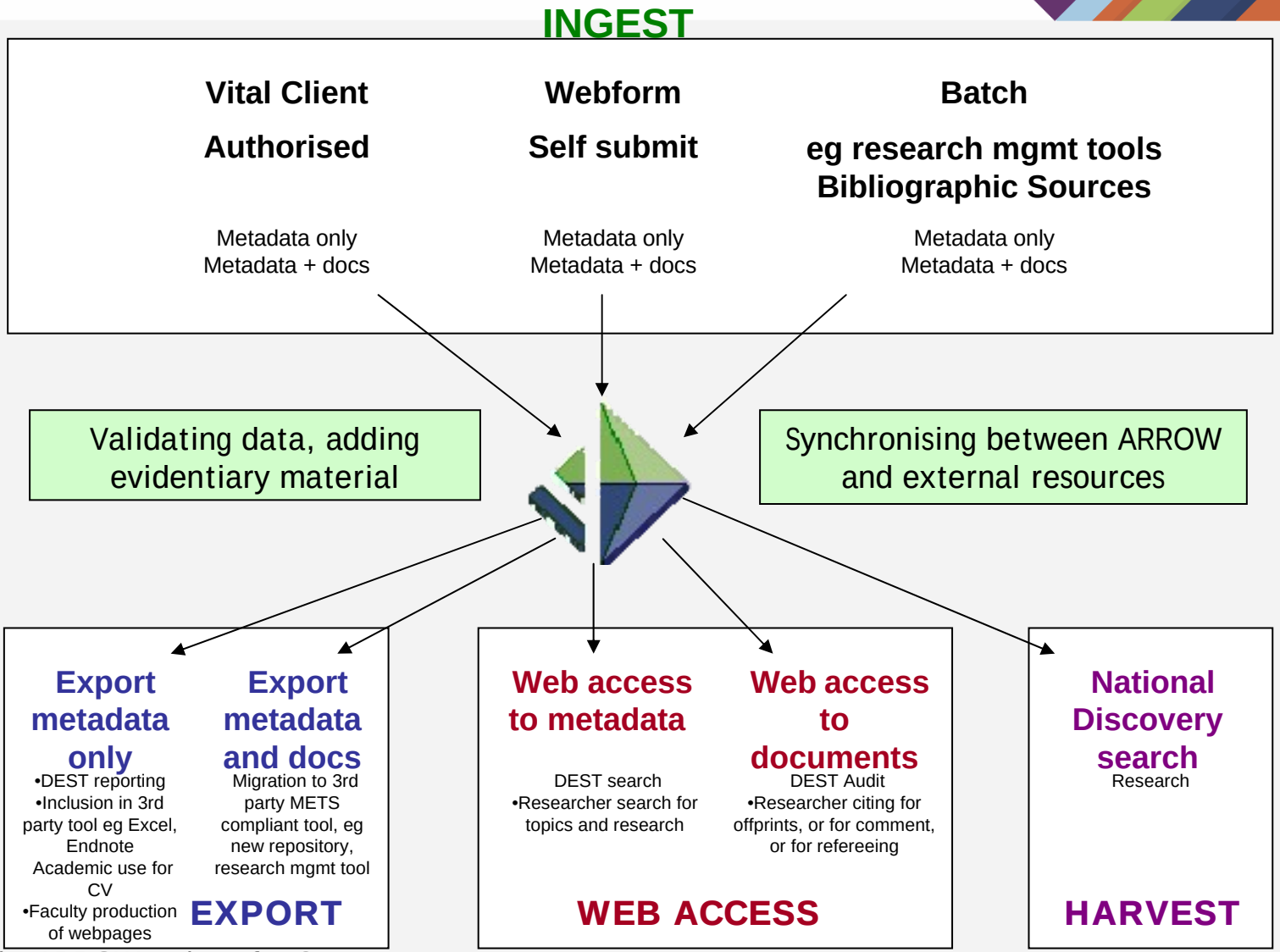
Each object in the repository comprises two or more datastreams.

One object may contain many different kinds of files.



Use cases and content modelling

- Based our work on use-cases
- Selected seven representative types of objects journal articles, conference papers, working papers, books, book chapters, theses and images
- Developed content models using MARCXML
  - more will be needed for RQF


INGEST
Vital Client
Authorised
Webform
Self submit
Batch
eg research mgmt tools
Bibliographic Sources
Metadata only
Metadata + docs
Metadata only
Metadata + docs
Metadata only
Metadata + docs
Validating data, adding evidentiary material
Synchronising between ARROW and external resources
Export metadata only
•DEST reporting
•Inclusion in 3rd party tool eg Excel, Endnote
Academic use for CV
•Faculty production of webpages
Export metadata and docs
Migration to 3rd party METS compliant tool, eg new repository, research mgmt tool
EXPORT
Web access to metadata
DEST search
•Researcher search for topics and research
Web access to documents
DEST Audit
•Researcher citing for offprints, or for comment, or for refereeing
WEB ACCESS
National Discovery search
Research
HARVEST

What have we done so far?

- 2004
  - ✓ Developed architecture, selected, tested and developed initial software (Fedora, VITAL and OJS)
  - ✓ Demonstrated functionality
- 2004 – end 2005
  - ✓ Deployed at Monash University, Swinburne University, University of NSW, National Library of Australia
  - ✓ Began populating repositories
  - ✓ Began harvesting by the NLA (National Discovery Service)
  - ✓ Continued developing VITAL
- Mid 2005 –end 2006
  - ✓ Enabled others to participate: four additional partners signed by January 2006

Questions?

VITAL software

- VITAL 2.0 was installed on partner servers in December. This provides
  - Searching
  - Support for a range of content types
  - Range of ingest alternatives
  - Management functions
- VITAL 2.1 now in beta. This adds
  - A range of UI enhancements, including better browsing
  - Collections support
  - OpenURL support
  - Endnote export
  - Harvesting via Google and enhanced OAI-PMH support

Searching and harvesting

- The Access Portal offers simple and advanced searching
- The National Discovery Service provides consolidated searching across many repositories
- Picture Australia harvesting (still in test)
- Google and other service providers to be established

Object types: Text plus

- **Text only**
  - Thesis Title:
    To define the ways in which VITAL will support the creation of a range of branded interfaces, showing either all of the repository or particular subsets
    - http://arrowdev.lib.monash.edu.au/hdl/1959.100/486
- **Text and supporting images**
  - Title:
    History Australia, Volume2, No.1, 2004. Ferals and their muddies: Making a home in the bush
    - http://arrowdev.lib.monash.edu.au/hdl/1959.100/418

Object types: Images

- **JPEG Images**
  - http://arrowdev.lib.monash.edu.au/hdl/1959.100/459

    Or http://hdl.handle.net/1959.100/459

    Composite weather image of a tropical cyclone Creator NASA
  - http://arrowdev.lib.monash.edu.au/hdl/1959.100/457
    Satellite image of Victoria and Northern Tasmania Creator NASA
- **MrSid images with navigation**
  - http://arrowdev.lib.monash.edu.au/hdl/1959.100/507
    Victoria Dock, circa 1910 and 1942
  - http://arrowdev.lib.monash.edu.au/hdl/1959.100/516
    Victoria Dock, 1972 and 2002

Object types: more…

- **XML plus images**
  - http://arrowdev.lib.monash.edu.au/hdl/1959.100/283
    Melbourne 2030: Chapter 5 - Residential infill and its threat to Melbourne's liveability
- **MPEG movie**
  - http://arrowdev.lib.monash.edu.au/hdl/1959.100/586
    Medical computer animation #20
- **Quicktime movie**
  - http://arrowdev.lib.monash.edu.au/hdl/1959.100/566
    Medical computer animation #10
- **mp3 audio**
  - http://arrowdev.lib.monash.edu.au/hdl/1959.100/571
    Ash Grunwald, Bakelite Radio and Blues Progression 5Mb
- **AVI movie**
  - http://arrowdev.lib.monash.edu.au/hdl/1959.100/551
    Fantastic Four, movie trailer

Ingest and submission

- Current Web Self Submission allows users to submit and describe theses
- Web Review allows a staged review process
- Batch Ingest allows bulk ingest of objects with their matching metadata
- Forthcoming Web Self Submission upgrade (now in testing):
  - Customisable interface for self or mediated deposit of an extendable range of publications and object types.
  - Metadata elements adjust according to the object type selected.

Management services

- The Access Explorer web interface provides indexing options and management functions.
- The Vital Manager: a 3rd generation (= Windows) repository application for managing and editing metadata, content and data streams for objects held in the repository
  - designed for trusted repository admin staff (not for end-users)

Partners

- ARROW@UNSW
  - School of Biological, Earth & Environmental Sciences (BEES) is trialling the deposit of honours theses by final year students in 2005.
  - School of Mining Engineering Trial: academic and research staff (including postgraduate students) are trialling the deposit of research publications until the end of 2005.
- Swinburne
  - ResearchMaster Metadata
  - Journal of Applied Psychology
    - Peer review management
    - Open Journal System
    - Assembling and publishing journal issues
    - Well liked by academics using the software
- Monash
  - Faculty of Business and Economics Working/Discussion papers
  - Centre for Gippsland Studies picture collection
  - Theses
  - Patents
- NLA
  - Looking at numerous possibilities, including capturing email

What do we still need to achieve?

- Full functionality – VITAL 4.0
- Access and authentication issues
  - XACML
  - MAMS
  - Shibboleth
- Metadata interoperability
- Improved usability for submitters
- Improved services for repository managers
- Better provision and integration of tools and functionality for both submitters and managers
- ARROW community

What have we learnt so far?

- Multiple partners are both good and bad:
  - Good:
    - Sharing of information and experiences
    - Sharing of development work
    - Multiple perspectives on issues
  - Bad:
    - Multiple perspectives on issues
    - Scope creep
    - Managing expectations
    - Pressure on the ARROW office
- Software development *feels* slow, both commercial and open source
- Development with a commercial partner can be tricky

What have we learnt so far? (2)

- Lots about the nature of digital repositories
- That there aren't enough real standards in this area
- Open versus closed repositories, *or* information management versus accessibility is a big issue
- Repositories are only partly about software - advocacy, policy, institutional engagement and grunt work need equal attention
- Constraints of dealing with copyright
- Institutions are hungry for information - hence the ARROW community

You can’t do it alone

- Current partnerships:
  - Government:
    - DEST
    - Other FRODO/MERRI projects e.g.:
      - MAMS
      - DART
  - Open source development
    - Fedora
  - Commercial
    - VTLS
    - Thomson Scientific – Web Citation Index
- Potential partners:
  - OCLC – Metadata Interoperability
  - Oxford – new VITAL customer
  - Publishers?

Where to from here – 2006

- Rollout to new partners
- Ongoing development of: VITAL 2.1, 3.0 and 4.0
- Taking advantage of new features in Fedora 2.1
- Providing enhanced repository management tools
- Working on access issues
- Roadshows
- Preparing for RQF
- Planning for a world after ARROW
  - Knowledge
  - Communication
  - Support

Where to from here – 2007 and beyond

- RQF
- DEST requirements?
- ARROW community
- Developing VITAL in conjunction with VTLS

Questions?

Research Quality Framework

- Working with Research Management Systems (initially Research Master)
  - Objects with Metadata
    - Open Supported Standards
      - METS and future XML based formats (MPEG21 DIDL etc)
  - Store Research Objects (with associated Metadata)
  - Provide Persistent Links (HANDLES)
  - Provide Secure Access (XACML)
  - Expose Research Digital Objects (Google, National Discovery Service etc)

RQF support

- ARROW proposed response to RQF
  - All “publications” entered into repository
  - Each “publication” has specific metadata fields
    - RFCD code(s)
    - RFCD panel
    - RQF tag (yes, no)
  - Reports in date range generated for each panel
  - Citation and persistent identifier retrieved and exported to panel members
  - Research office attaches factual statements
  - Associated “document” sent to DEST (or repository)
- See http://www.caul.edu.au/caul-doc/repositories2005harboe-ree.pdf for more

Working with Research Management Systems
(Part 1. Gathering information)

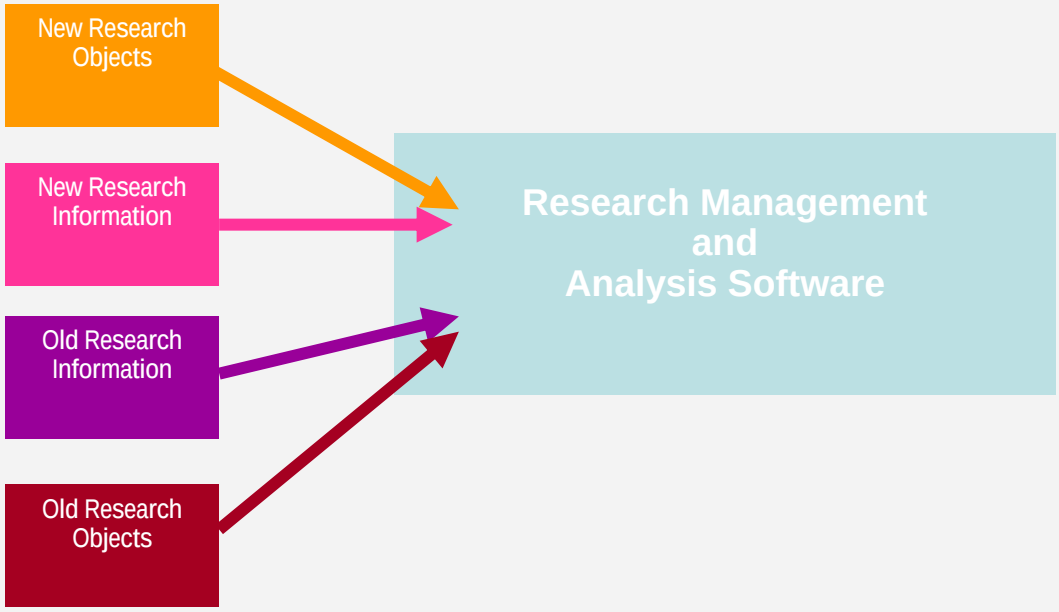
New Research Objects

New Research Information

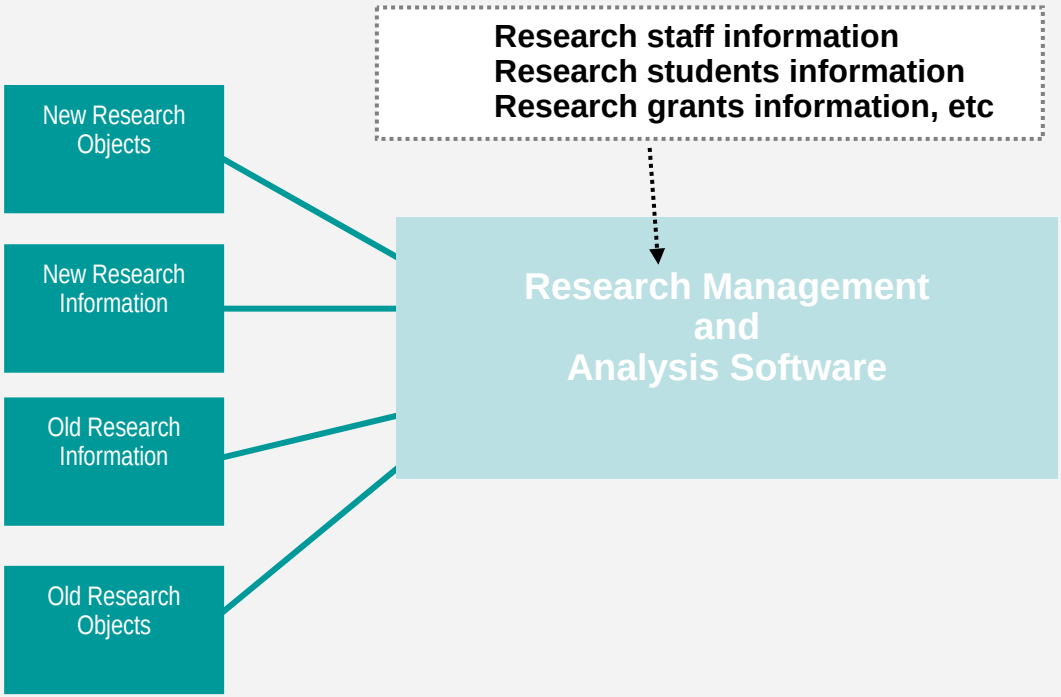
Old Research Information

Old Research Objects

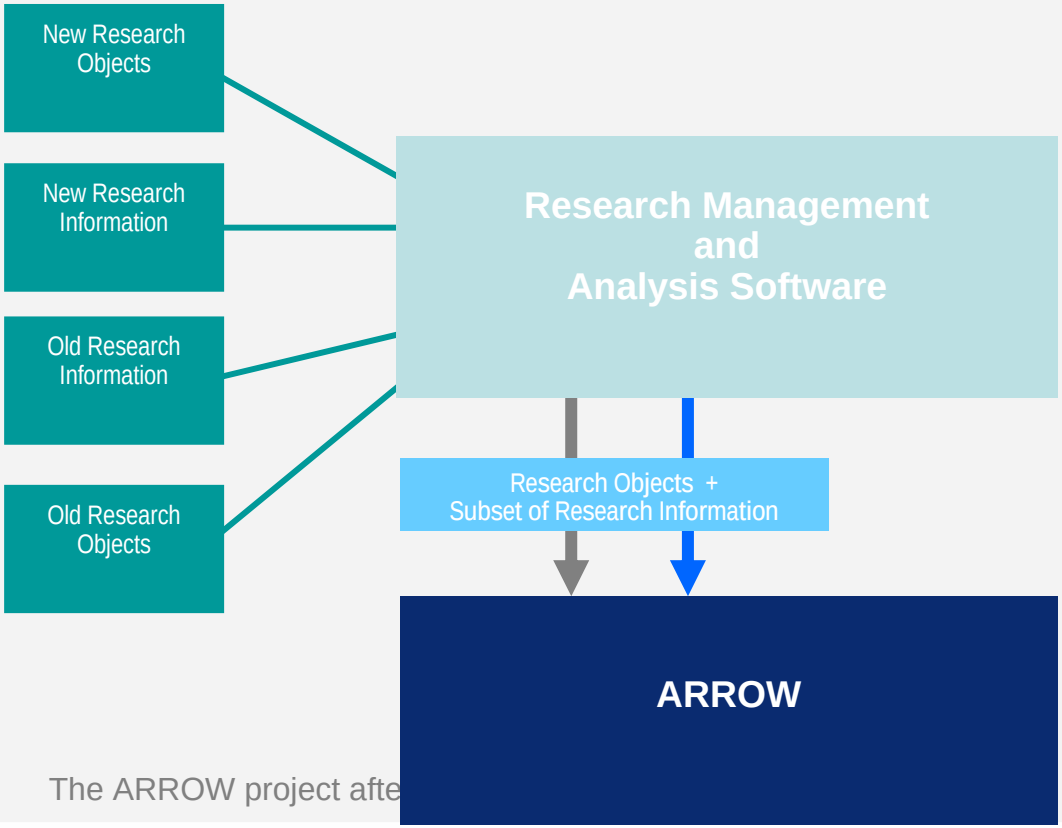
Working with Research Management Systems
(Part 2. Record input)



Working with Research Management Systems

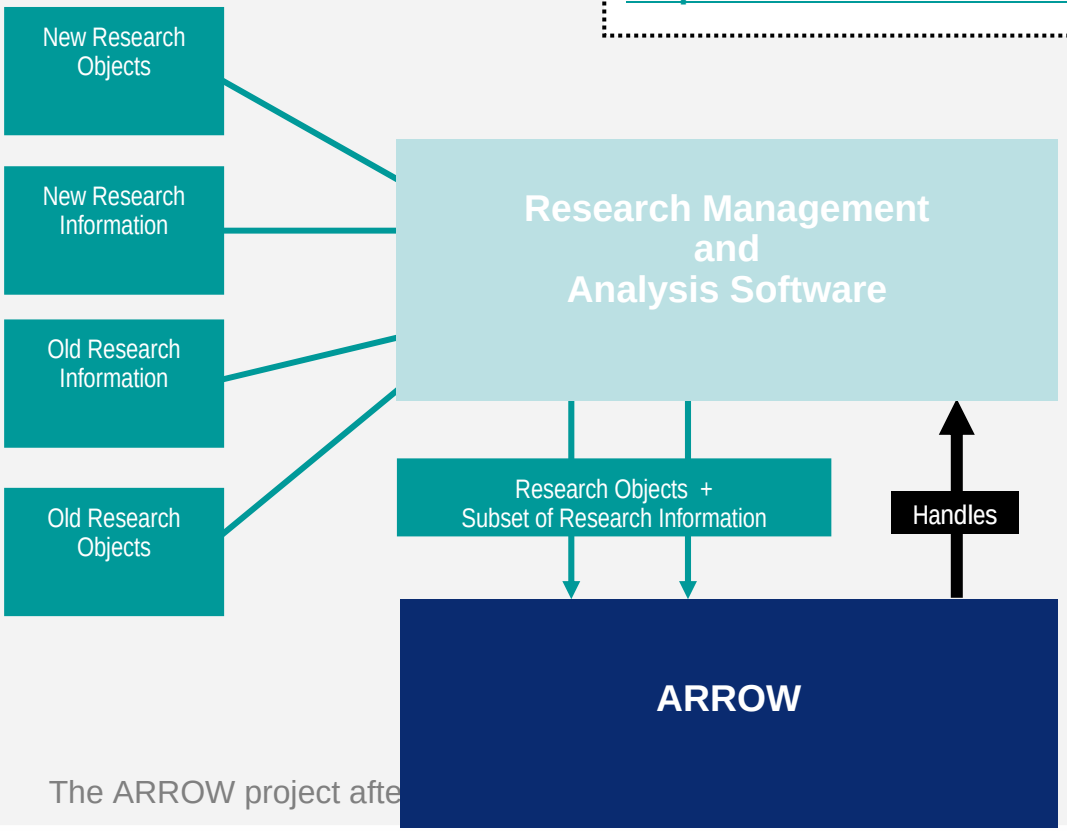


Working with Research Management Systems
(Part 3. Deposit)

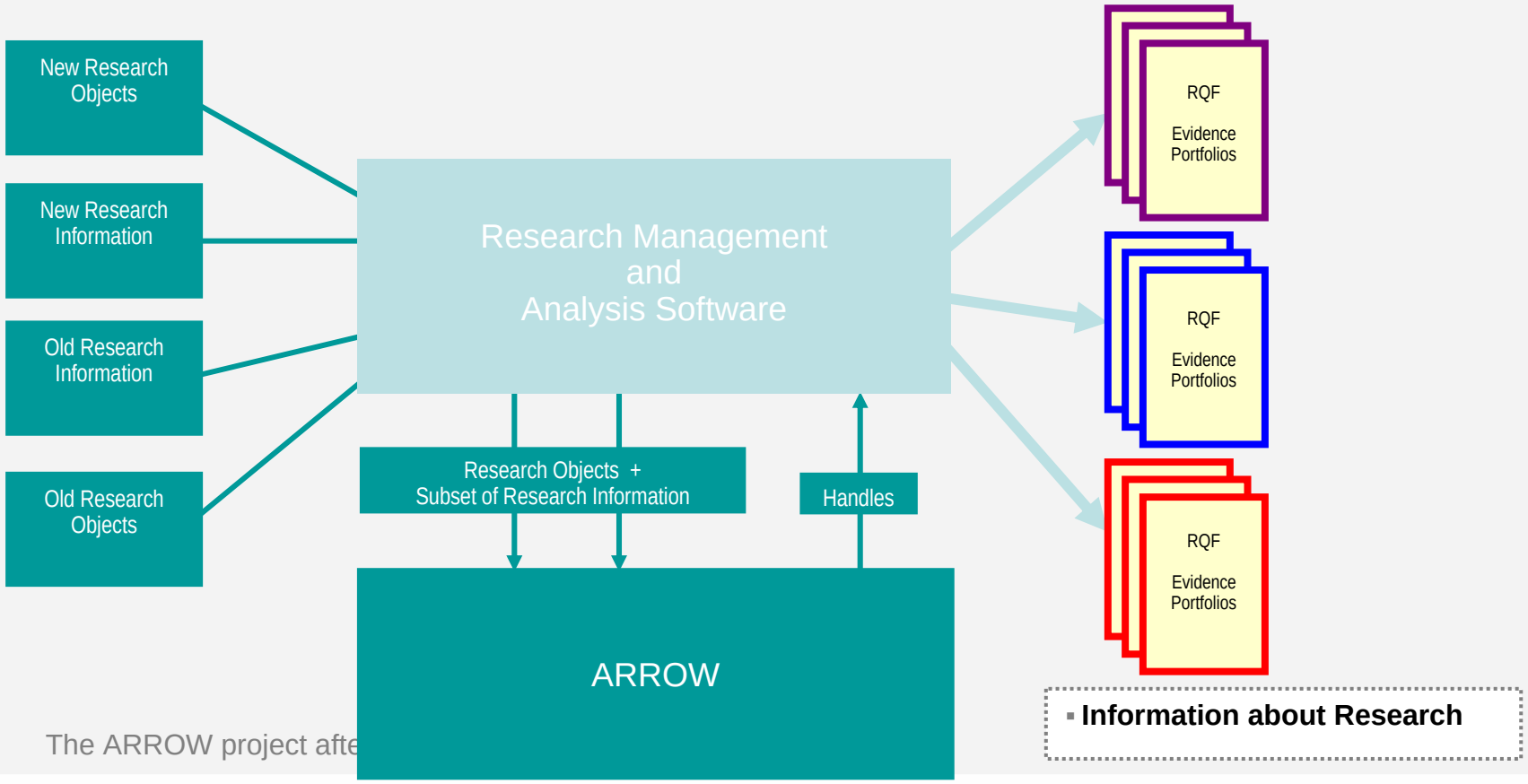


Working with Research Management Systems
(Part 4. Handles)

Example:
http://arrowdev.lib.monash.edu.au/hdl/1959.100/630

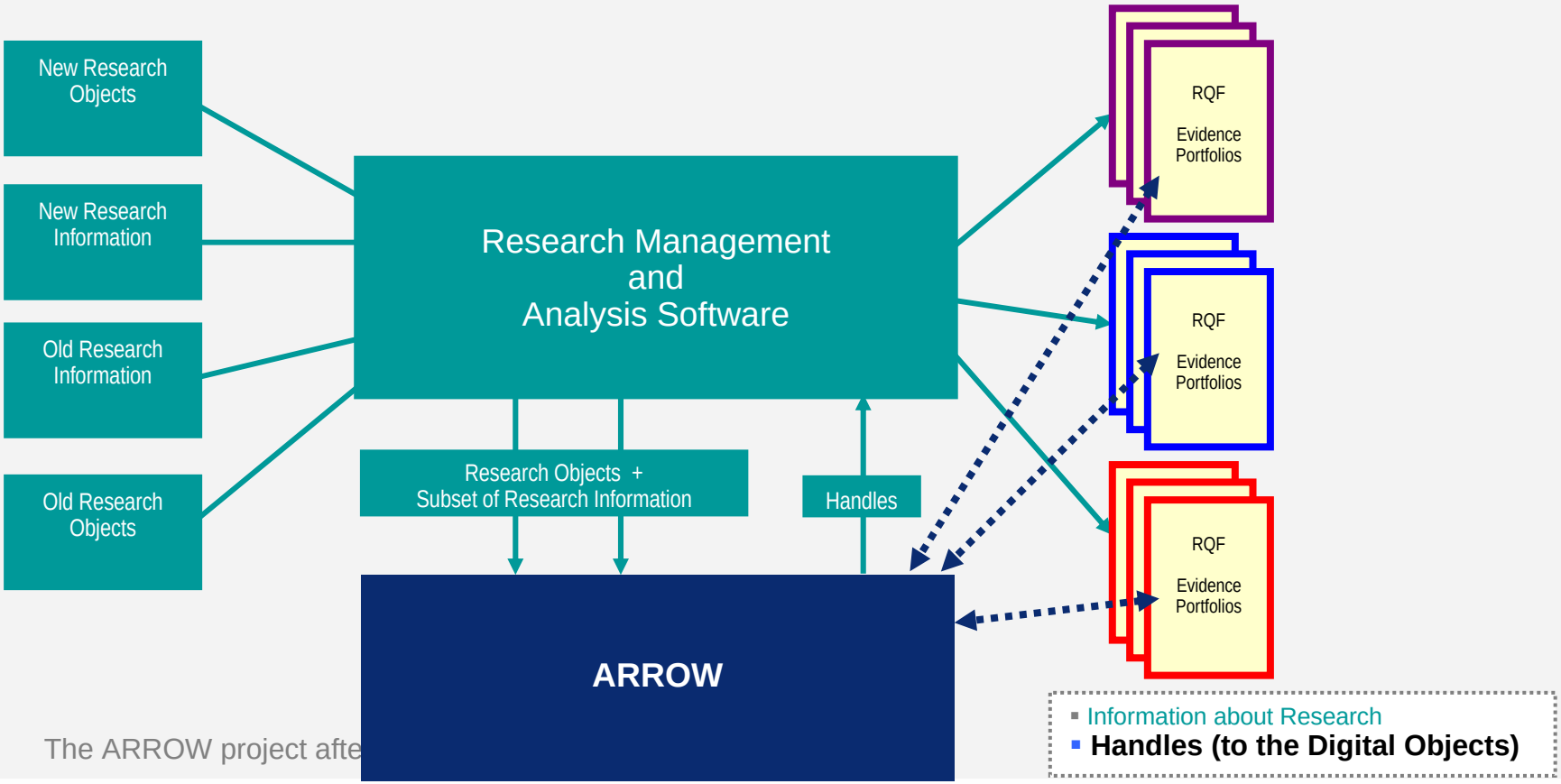


Working with Research Management Systems
(Part 5. Evidence Portfolios)



The ARROW project afte

Working with Research Management Systems
(Part 6. Linking back to Research Objects)



ARROW's role in supporting The RQF

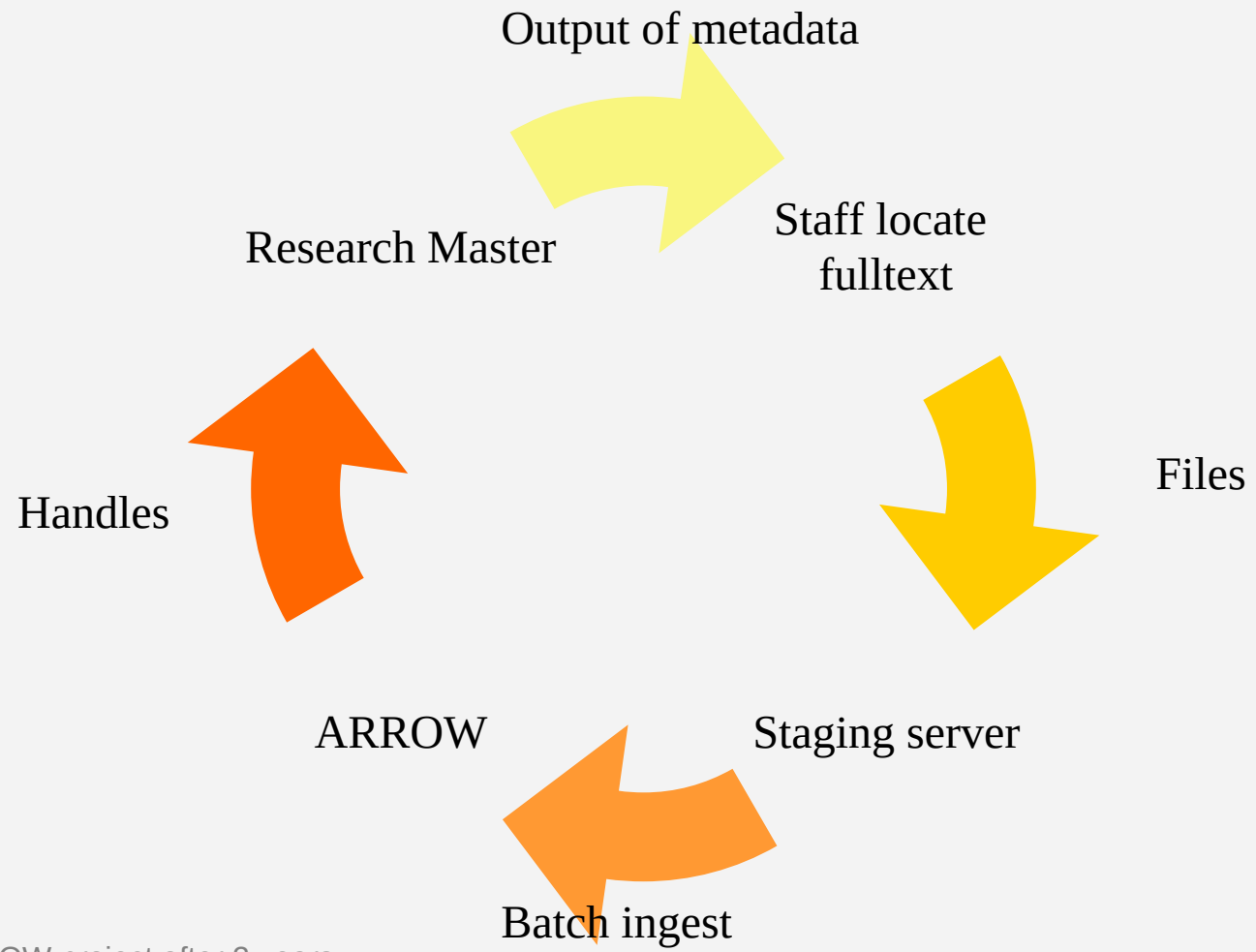
- Without expensive Research Management Systems.

- **ARROW** can:
  - Store Research Digital Objects
  - Provide Persistent Links (HANDLES)
  - Provide Secure Access (XACML)
  - Expose Research Digital Objects (Google, National Discovery Service etc)

RQF and ARROW at Monash

- Monash is testing preparedness for 2007
- Most metadata is already available
- Full text of research objects is not
- Using batch ingest functionality to match metadata to full text and define rules in repository

RQF at Monash



Sample article

- Record – for harvesting, no link to full text

http://hdl.handle.net/1959.100/1230

- Link for use in evidence portfolio

http://ezproxy-test.lib.monash.edu.au/login/ARROW/arrowdev:1230/ATTACHMENT01

- Access to evidence portfolios will be restricted to users with a valid Monash LDAP account

Questions?

- Contact ARROW: arrow@arrow.edu.au
- Contact me: David.Groenewegen@lib.monash.edu.au