ARROW HERDC Working Group
Interim Report May 2008

Contents:

Introduction

ARROW HERDC Working Group

The ARROW HERDC Working Group was formed in January 2008 to look at the role repositories could play in assisting to meet the university's HERDC requirements. The objectives of the Working Group are to:

- Identify commonalties in research collection practices in ARROW community members.

- Identify good practice HERDC models for integration of research systems and repositories.
- Report to the ARROW community on potential HERDC models.
- Circulate findings with the Australian and New Zealand repository community.

Members

The Working Group was formed from members of the ARROW community who were actively involved in planning for repository and HERDC interaction of some kind within their institution. The members are:
- ARROW – Angela Lang
- LaTrobe University – Michael Wood
- Monash University – Andrew Harrison
- Swinburne University of Technology – Teula Morgan
- University of Newcastle – Vicki Picasso
- University of New South Wales – Tom Ruthven
- University of South Australia – Jenny Quilliam
- University of Sunshine Coast – Kate Watson

Interim report

This report is an interim report on the activities of the Working Group. The interim report aims to stimulate discussion and invite feedback which will inform the final report.

As a first step the Working Group has identified and developed four potential broad models to represent the relationship between the university's research management system and the institutional repository.

The purpose of these models is to suggest possible workflows and processes, to stimulate discussion, and to identify the strengths and weaknesses of each approach. It is assumed that local use of these models will necessitate variants as required by the environment of each institution.

Models

The diagrams below describe four potential broad models to represent the relationship between the research management system and the institutional repository in meeting the HERDC requirements. In these models the output refers to the HERDC reporting output. More general outputs from either the institutional repository or the research management system are out of the scope of these diagrams.

Model 1: Institutional repository to research management system

This model describes an input process where data is captured into the institutional repository, with the data then flowing to the research management system.
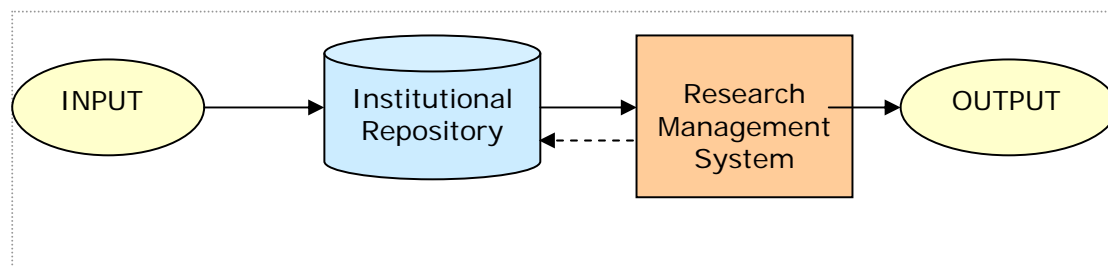
Model 2: Research management system to institutional repository

This model describes an input process where the data is captured into the research management system, with the data then flowing to the institutional repository.
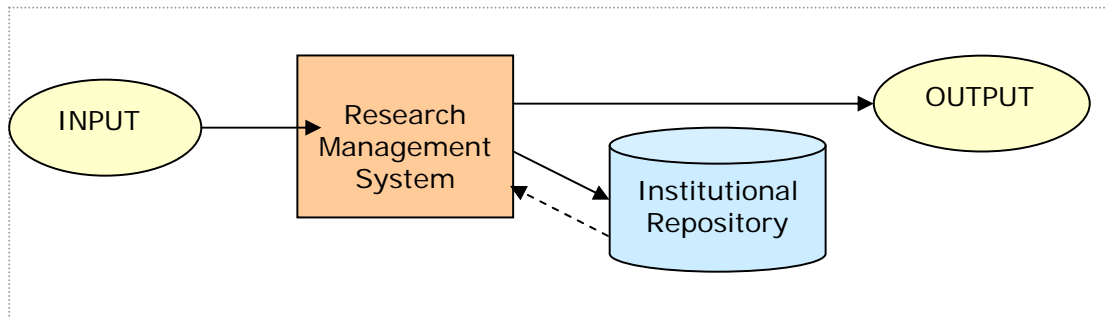
Model 3: Shared input

This model describes a separate input process that feeds into both the research management system and the institutional repository.
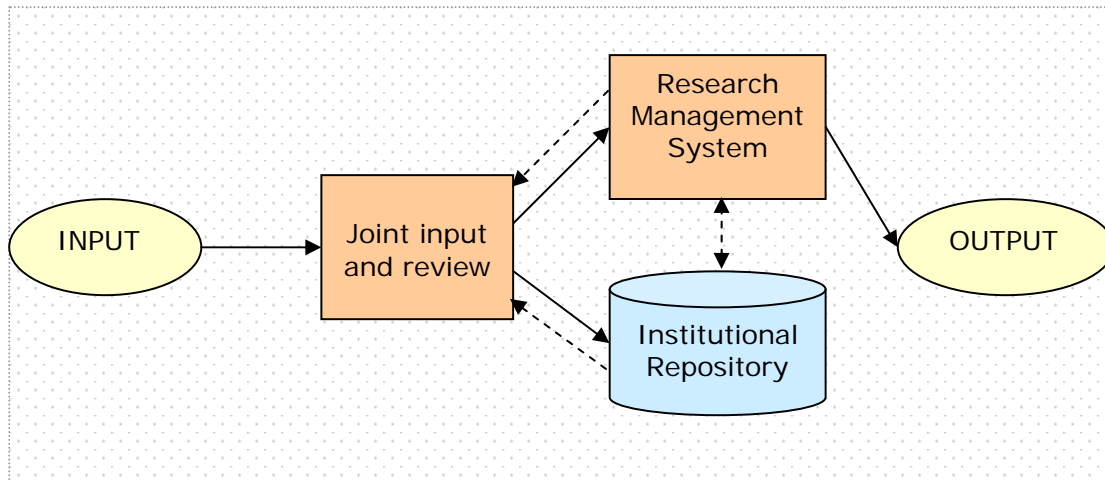
Model 4: Combined

This model describes a single system that captures, manages and exposes publications for both the institutional repository and the research management system.

Considerations - across all models

The selection of appropriate systems and workflows is always a balance between what is ideal, and what is practical.

The Working Group has attempted to outline a number of issues that will need to be considered when determining how best to integrate HERDC and institutional repository processes. These have been broken down into five broad areas - the organisation, the collection, the data, the content, and the software.

The organisation

Management issues

The identified models all require varying levels of communication and collaboration between organisational units. The existing environment and culture will need to be taken into account when selecting the most appropriate model.

Some institutions have instituted committees or working groups to bring organisational units together, or have tapped into existing collaborative networks.

Organisational change

The amount of organisational change required, and the readiness of the institution to make that change, will need to be considered.

The various models have attempted to build on existing university processes. Models 1 and 2 attempt to minimise the change required, while Models 3 and 4 accept more extensive changes in order to maximise potential benefits.

Resources available

As with all new workflows or systems the resources available will impact on the potential solutions. Possible improvements in efficiencies as well as valuable outcomes could be used to justify expense in implementing change.

The collection

Timing of collection process

The methods and timing of collection of HERDC publications varies at each institution. However, it is not uncommon for the Research Office to open and close collection at specific dates. Outside this period the research management system may be closed for input.

An alternative model is to maintain the system open all year round for collection, avoiding the creation of peak periods of data collection activities.

The process used locally will help determine the best mechanism and timing for data collection at your institution.

Individual submission versus central identification

The main processes by which data is entered into a repository are individual submission or central identification of content. Individual submission is generally associated with author involvement, whereas central identification usually involves the library managing automated processes for data identification such as customised alert services or database interrogation.

Workloads across organisational areas

As the HERDC / institutional repository interaction is likely to involve new workflows, the workload spread across organisational units will need to be monitored. The timing of the collection, and the choice of individual or central submission, will influence these workloads.

Differing priorities, time lags

All of the models to some degree introduce dependencies across organisational units. This creates the need to identify where differing priorities could potentially cause problems with the workflow.

In most cases libraries have not previously had a large role to play in the HERDC. If the new workflows rely in some part on the institutional repository they will need to be aware of HERDC requirements and deadlines they must meet. From the other perspective the research office has usually not been involved in institutional repository activities, and must be aware of the deadlines and requirements in that area.

The data

Input

Input may be done by any or all of the following groups of users:
- academic staff
- researchers
- staff in faculty
- publication officer
- administrative staff
- library staff
- research office staff

Processes for input can be:
- paper based form/template (requiring more effort and resources)
- online/web submission (requiring less effort and resources)

Good design of the data input mechanism is extremely important to ensure that there is no negative impact on the quality of the data being captured. If data is being entered by a range of staff, as opposed to smaller core groups within the library or research office, considerable care will be needed to make the data entry process as simple and clear as possible to reduce input errors and subsequent data quality issues.

Good design must be supported with consistent standards and training for all data entry staff.

Review processes

Based on the choice of input user groups, the need to implement a review process for the data should be considered. Data for review may include bibliographic citations, DEST categories, affiliation, etc.

If you require review processes, you will also need to determine:
- who will be responsible for the review?
- when will the review occur?
- does the input interface facilitate the necessary review processes?
- does the completion of the review process for individual data records need to be flagged in any way?

A review process could be divided across the library and research office according to areas of expertise, e.g. reviewing a DEST category versus verification of bibliographic information post input.

What is captured?

The data that is captured at each point in the process will be dependant on the chosen model. This is one of the most critical issues impacting on integration of HERDC and repository processes. There will need to be agreement at the onset as to which data elements will be captured and stored.

There is some specific data required by HERDC which is beyond the normal repository bibliographic data and whether, and how, to capture and store this data has been the subject of discussions between the Working Group and MACAR. Some elements may be captured in standard MARC fields while other data may be better captured in a separate metadata stream. No conclusion on these issues has been reached as yet. This data includes:

- DEST/DEEWR category code
- Linking of affiliation to author name
- Total number of authors for the work
- Order of the author's name on the publication
- Number of chapters in a book
- RFCD (ANZSRC) code

As an example, the requirement to capture this additional HERDC data may impact on the repository by requiring further customisations to its input forms and mechanism, which in turn may also require additional resources.

There may also be different standards in use across the institutional repository and the research management system. An example is the preferred data format for description of publication type, where practices commonly vary between systems.

Duplicates

Variable input methods and sources of input will likely result in a significant number of duplicate records flowing into the institutional repository.

To ensure data quality and integrity a de-duplication method is needed. This could be automated identification of duplicates supported by manual verification. The lack of definitive publication identifiers precludes automatic deletion.

There may also be individual practices that require duplicate records. Some research management systems capture multiple records for the same publication where there are multiple authors at the same institution.

Synchronisation

The need for synchronisation of data across systems will need to be identified. If synchronisation is required processes will need to be developed to identify which data requires synchronisation, when changes can be made, how changes will flow through, and how conflicts in the data will be managed.

An example of where data synchronisation may be desired is where a correction to bibliographic information is made in the institutional repository which should also be reflected in the research management system.

There may also be a need for data to be 'locked' after a certain point in the process, for example after formal reporting has taken place.

Author Names

Decisions on the format for capture of author names need to be made locally. The research management system and the institutional repository may well have different needs for the author name format. Examples of this may be the repository records referring to the author's name as per the publication details while the research management system may refer to the author's full name as derived from the university's identity management or HR system.

The requirement of specific author name format may also be dependant on relationships with other university systems where data is also captured or stored for the author or researcher.

The content/evidence

Processes for collection of evidence

The collection and storage of verification material for the HERDC will need to be incorporated into the workflow.

Collection and storage processes may be paper based or electronic. An electronic process has the potential to be more streamlined and eliminate the long term storage of paper based material. This will need to be balanced against existing workflows and the potential need to digitise some material.

Some institutions have a verification and signoff step in the collection process and will need to consider how this would be implemented in an online workflow.

Collection of published version vs. postprint

Collection of the verification material for the HERDC incorporates sourcing a copy of the published versions of the research output, however sourcing the postprint for open access is an extremely important part of the workflow for the repository. There are therefore two versions of the research output to be collected.

The most appropriate point in the workflow to collect both the published version and the postprint must be determined. Simultaneous collection has the advantage of simplicity but also has the potential to confuse authors as to which version is desired. Collection of the postprint at a later time, separate to the HERDC process, is typical of current practice for many universities at this time.

It also needs to be determined whether the publisher's copy itself is collected or a permanent link to the publisher's copy will meet verification requirements.

The software

Existing systems?

Some of the identified models require more systems development than others. Models 1 and 2 have attempted to build on existing software as much as possible. Models 3 and 4 accept that more systems development may be required. The availability of technical resources will influence the most appropriate model for an institution.

Interaction between systems

Interaction between the institutional repository and the research management system will require technical resources. The level required will be dependent on:
- the amount of synchronisation required between systems
- the number and type of data fields identified for synchronisation
- the level of complexity associated with the flow of data between each system
- expected rate of change of synchronisation requirements resulting in the need for programming changes

Interface design

The design of the input interface is an important issue for the designated user groups. Consider the following points in determining who is the appropriate person or people to input data:
- ease of inputting data into the system
- ease of navigating the system
- level of support users need to complete the process, including training and contextual help
- the amount of work required in the data review process, e.g. will additional work be created later in the process as a result of limitations in interface design?

Conclusions and next steps

The ARROW HERDC Working Group invites feedback on this interim report. All feedback on this interim report submitted to the Working Group will be considered and incorporated into a final report. The final report will also attempt to provide more information on issues specific to each model, together with a mapping and analysis of repository metadata fields vs. required HERDC data collection elements to assist with the selection of an appropriate model.

Comment and feedback on the following questions is particularly welcome:
- What other information should be provided to assist institutions in considering future workflows?
- Advantages and disadvantages of the specific models
- Capture and storage of HERDC data elements

Feedback can be provided via email to arrow@arrow.edu.au .