Populating a new repository for Open access, HERDC and ERA.
Riding on the back of the wave ...

*Patricia Scott – Deakin University Library*

**The preparation**

Deakin University set up an open access repository to fulfil all the requirements of the ASHER grant. It had to be able to describe the University’s research output appropriately, support open access for as much of that research as possible and be sustainable into the future.

It was purposely designed therefore to be able to store and report material for a retrospective government audit, such as ERA, to be ongoing in order to inform future annual HERD reports, as well as provide to our researchers the benefits from increased awareness of and access to their research.

As one of the later Universities to establish an open access repository, we were able to use the OakLaw manual as a blueprint. Information was also gathered from various Australian Universities, and even Loughborough University in the U.K. We also surveyed Science Commons, SPARC, JISC and SURF and attended various APSR sessions before we finally put together a plan of action and the documents we needed to offer.

We are using Fez/Fedora, from the University of Queensland, and have been greatly assisted by the UQ team as well as the Fez community globally.

This project from the outset has been a joint project between the Library and Research Services and is project managed by ITSD.

**The aims – fill the repository for open access, HERDC, ERA**

The task to fill the repository was three fold.

1. We needed to be sure that the descriptions of the items were accurate
2. We needed to provide access to the items wherever they were.
3. We needed to provide a copy within the repository and make it available if we could.

1. It is inevitable that errors creep into any database. Checking the data to source again provides a more accurate result.
2. By providing a link to the published version we would ensure that anyone could verify the data we were reporting and make use of the item, even if they or their institution had to pay for it.
3. In order to get content into the repository, copyright ownership needs to be known and permissions discovered or obtained.
Providing background data – how we knew what to collect

The repository was primed with data and objects gathered for the RQF exercise. ERA was expected to be built on the RQF collection. As some data was unchecked and documents for the most part couldn’t be made public, all was added into dark collections set up within Fez/Fedora. Fez supports security for groups or individuals, or a mix of both at the community, collection, record or object level.

- **Record source** - Using the Research Services database as the only comprehensive source of data about outputs, a file of records for outputs 2001-2007 was provided. These included the Research Services record ID number.
- **Reference document source** – Documents scanned by the Library in preparation for RQF review under the copyright exemption provided for the purposes of the Crown may be used for strictly administrative purposes i.e. for validating the information in the database. The scanned images were named by the Research Services ID number so they were collected from their server and matched with the records created. Both record and image were then loaded into “dark collections” within Fez. Useful for validating the information in the records, they also help to find the published version where errors had crept into the description. More recently, some have been extremely useful where we have permission to make an item available and the published version no longer exists e.g. some conference papers.
- **Search guide** - As part of the load process the systems team created open URLs for the records to enable us to use our resolver to more easily find the online version of the items. This has proven invaluable, allowing us to use more junior staff and much greater speed to find the published version.
- **Researcher link** - Research Services keeps an accurate file of current Deakin Researchers. Fez is one repository package that provides an author record structure, which is one of the attractions of choosing it. We loaded in records created from another spreadsheet supplied by Research Services with their author data. Fez allows you to link all relevant items by the same author to this record, even if you chose to use a different form of name on the item record. Once linked the display of author heading changes to “green” as an easy identifier of our authors. This is, of course essential when reporting works of a particular researcher.

* A good start, but now action was required.*

Aligning with the strategic plan and publicity

The repository development coincided with a change in strategic policy for research at Deakin University. Raising the awareness of the research record from Deakin Research Online and providing a profile of their research to each researcher fitted neatly to a step change in research output that is now required from every researcher as part of Deakin University’s strategic plan.

We gained support at the most senior level of the University, both in enthusiasm and with funds. Two permanent positions were funded as a Library strategic initiative by the University and supported by the Deputy Vice Chancellor Research to ensure the establishment of an ongoing repository. He also officially launched Deakin Research Online (DRO) jointly with the University Librarian at the Deakin University Research Conference on the 30th of September 2008.

This gave confidence that we were here to stay, and we received much applause from researchers present at the conference for the promise that DRO offered.

We then had to get the attention of the rest of the research community, and action from all.
Researchers are busy people. We only wanted to bother them once, cause them the least amount of work and let them see the maximum benefit for their cooperation.

It was possible to check and process the existing records independently of any contact with researchers. But that is only part of the story

- Some academics had failed to report some material, and as ERA required data about research for any census author between 2002-2007, some material had been reported to another institution if the researcher wasn’t at Deakin at the time.
- Permission was required to add in content, and to follow up on publications that had disappeared.
- In addition, as extra material that for the most part has not been collected for HERDC is required for ERA we need full cooperation from specific researchers.

It was essential to contact all census authors, do as much work as we could for them and get their support for describing new material and getting in content.

Preparing packages

We provided a package of information for each researcher, comprising:

- A list of their publications, sorted by copyright permission, for journal articles and conference papers, with books and book chapters in a section at the end.
- A copy of the non exclusive licence we required the researcher to sign on a once off basis to enable items to be deposited in the repository.
- An introductory letter from the DVCR and University Librarian, also reminding them of their Liaison Librarian who could help them
- An instructions sheet on what we needed from them and how to interpret the list.
- A flyer extolling the virtues of the repository and the benefits of open access

We prioritised researchers by targeting the 30 most prolific authors from the 2007 HERDC report. This became the top 60 authors, by the time influential academics had been added

We then focussed on specific schools, and once the ERA data came out, the ERA cluster 1 and 2 authors took priority.

The packages were given to Liaison Librarians to take on a one by one basis to each researcher.

Giving presentations

At the same time we did a road show of DRO presentations to the different campuses to raise the profile of DRO. It wasn’t well attended, but some key people did attend.

Further presentations at Faculty research days, faculty boards, and at Academic Board have raised the profile of the DRO and what it can do.

More recent publicity about ERA and general panic about the timelines seem to have drawn the attention of researchers.

Research Services Division are working closely with us, as we are with them, they are on our Project Board, and we invite each other to sessions so we speak increasingly with one voice. Now researchers are starting to ask where their package is.
We haven’t completed all packages, but are working systematically through School lists. In addition, some of our early adopters are starting to request our presence at their higher degree training days, as they wish to encourage their students to deposit articles as soon as published as well.

Requests for such presentations are currently filling my diary on a weekly basis.

**Processing records**

At the same time that the packages were produced, the records were processed within DRO, first for our high level authors, then for known cluster 1 and 2 authors, then just by HERDC category, A1, B1, C1 etc. for records that were linked to a current Deakin author record. We are pleased that this coincides with the ERA definitions as well.

An online version of the item or a good catalogue record in the case of print only items was found for almost every output. The best, most permanent link to the online version was added to the record or if not online it was linked to the catalogue entry in Libraries Australia or WorldCat with the “get this item option” of finding in another library or buying from a bookshop. Wherever available we used a DOI, if it worked.

Where we couldn’t verify the item researchers were either emailed directly or the item put into a follow up file. We also ensured that processing the records for a researcher was ahead of completing the package.

As the records were processed they were linked to a collection that was visible, so they came from the dark part of the repository into the light. So it was essential to set security on any RQF scan attached so that it would only be visible to DRO administrators.

The records were linked to the school of each author, the data was corrected and enhanced with a summary, keywords etc, and copyright statements were added.

A single RFCD code had been added to all records in the Research Services database, in many cases by the faculty support staff. FoRs were also added to the records in DRO using the RFCD to FoR translation program offered by James Cook University. It worked better for us than the one offered by ABS. Some manual changes to FoRs were done where there was difficulty in translation. The FoR on the record allowed us to link the record to a new collections created for each ERA cluster. This has identified some additional individuals against clusters who may otherwise not have been considered.

By the end of February this year we have managed processing and linking nearly all of our standard HERDC collection 2002-2007. We have also processed any items recorded in the Research Services data as J category. We are currently processing records for some of the more tricky conference papers, double checking authors not identified as Deakin by a matching author record and bibliographies of creative artists and film makers. We have a growing stream of lists from authors for additional items to add.
Requirements for getting in content.

In order to deposit any item in a repository three things must occur:

1. Permission must be available from the copyright owner
2. The version that is allowed must be available
3. The author must agree to material being deposited, and more importantly to copies of the allowed version being deposited.

1. Permission to store and distribute the item must be obtained from an author if they have retained copyright, or if they have been given back certain rights by their publisher.
   a. First we need to know who owns copyright.
   b. Regardless we need permission from each researcher. Usually the right to post a version in an institutional repository is returned to the author(s). We have devised a one time licence that is valid until revoked to give us this permission. Researchers are reminded that they must not contravene any one else’s rights, i.e. they must ask permission of their co-authors before they deposit. Receiving licences back from researchers is now occurring steadily after a slow start.

2. The version that is allowed is often not available. With new material, the message is filtering down that a post peer review; final author’s version must be kept. For material 2002-2007 most authors have kept the better looking, authoritative published version. in some cases the way in which publishers process outputs makes it very difficult to get the post print version disentangled from the publishers formatting.

3. Many authors are happy to sign a licence and are also happy with the principle of depositing material in the repository. Some are not happy to deposit their post peer reviewed version, even if they have kept it. It doesn’t look professional, or the illustrations are kept out of order are two of the complaints. It also takes authors a long time to find copies of all their publications for that period. From the bibliographic purists I am criticised as we describe the published version, and try to attach something else.

Checking copyright ownership and permissions.

Using a copy of the spreadsheet that had created the records sorted by publisher, then de-duplicated. Most of the publishers had been checked in 2007 for the RQF to the Sherpa RoMEO database. During 2008 a second check using Oaklist was done.

This annotated spreadsheet has continued to be used as we request and receive permissions, as a quick check for staff preparing author lists, and for staff receiving content from authors.
   a. We found that we knew the policy of only 25% of the publishers listed on our spreadsheet. Some have since been clarified as many smaller publishers have been purchased by larger publishers with clear policies that then may be applied to the material published by the smaller publisher.
   b. With the emphasis on processing records requesting publisher policy or permission has taken second place until early 2009, with the addition of an extra team member for just that task.
   c. Processing the publishers hasn’t held us up in getting out the lists to authors. Where we don’t know the policy the items are listed as ‘unknown’ and we request a post-print and any copyright transfer agreement if available

Putting the process into context, we have nearly 1000 unknown publishers. Sherpa RoMEO has just celebrated creating its 500th publisher record!

In some cases it is clear that we will never be able to find definitively who owns copyright, nor gain permission from the organisation who published the output. This is especially true for conference papers. If after through searching we are unable to find an organiser who claims
copyright we have been advised that we can presume that copyright remains with the author. It is
to some extent an exercise in risk management. If a third party later claims copyright, we will take
the item down.
We follow the Oaklist process for most journals in the Australasian and Asian area and use
Sherpa RoMEO to follow up where the journal is published places that we are unable to contact
during business hours. For conference papers we seek direct permission for a limited list of items,
and future items. This generates a quicker and more positive reply, as the publisher is committing
themselves to a limited agreement, for a specific site. Regardless of the method is a slow
process, but at our current rate of working we should have approached at least 600 publishers by
the end of 2009.

**Setting priorities for getting permission**

On receiving a licence we immediately put up any published versions that are permitted and any
post prints sent to us. We then investigate any ‘unknowns’ on the author’s list.

If an author sends us an electronic copy of publications where permission is unknown, this jumps
to the front of the queue to request permission.

I am concerned that there may be little time to request permission for any of the HCA items
identified as the 20% for peer review if it is not already processed. We are hoping for global
permission to add these in secure access for ERA reviewers. More than half of our HCA output
has an RQF object attached to it. With links to the published version, if not yet stored it would be
simple to collect, if permission to store it for ERA reviewers were obtained.

A new focus has been placed on creative works. We have allocated additional hours of a Liaison
Librarian to assist researchers identify the best formats. To the best of our knowledge the
copyright for most of these items are owned by the creator, however making it openly accessible
is not necessarily desirable in its current form. Our most pressing issue with this material is size.
We believe that we can support our artists, but have filmmakers whose files easily exceed 15 mb.
We are looking creatively at what can be done to make the best of these outputs available.

**ERA requirements have set priorities, but not derailed the strategy**

As stated before, we have almost completed processing the description and links for all higher
quality material required for all ERA clusters. Material that falls outside of these categories has
not yet been processed. We intend to process these in the second half of 2009.

For ERA we
- Can supply records and content for clusters with record and content URLs
- Can hide objects, and allow viewing by authenticated individuals or groups
- Can provide all relevant data
  - still to load in Journal FoRs and adjust clusters
  - Still to load in Scopus IDs
- Research Services will link with other required data and provide SEER files
- Need to complete creative works and new to Deakin author works

We have tried to concurrently focus on both data and open access content for ERA with a view
for preservation for the long term. We only want to ask the researcher for this material once.
We also have sufficient flexibility within DRO to enable us to keep versions on behalf of an author to ensure long term preservation of that item. This is particularly important for conference papers and other outputs that are unlikely to remain available in the longer term.

We expect to be able to roll out self submission in 2009 very soon. We have a number of authors keen to raise their profile and deposit in DRO. Despite our focus on ERA we will need to make self submission available, and are starting to do so. We need to review everything submitted by researchers to ensure that the metadata is complete and accurate and that they haven’t added any objects that shouldn’t be made public. Hence all their records will be only submitted by researchers and we will review them and publish them within the repository.

Encouraging/enforcing researchers to populate the repository

Despite the fact that the University retains a nonexclusive right in the Intellectual property policy to use a copy of each output for teaching, research and promotion of the University, this has not been promoted as a mandate.

What has been used as encouragement for deposit is that DRO has been considered as the place for the researchers to describe their new outputs from 2009 because in one step they can:

- Gain all the benefits of open access to their research e.g. early claim to results, ease and equity of access, greater exposure via Google and Arrow, increased citation effect
- Enable them to fulfill their reporting requirements (for HERDC) into a time more convenient to them
- Enable them to provide the post peer reviewed version when it is at hand, and then clear their files
- Create for themselves an updating, linked bibliography that they can put into their web site (html code provided, other DRO outputs include Word docs for CVs, EndNote files, RSS feeds etc)
- Preserve their outputs and additional related data on a well organised, fully backed up server

There is interest from the Faculty support staff that normally complete the HERD collection. They should find it easy to:

- Get lists of all items produced by their researchers and their school for any collection year (Collection year is one of the fields available)
- Input any missed items and edit any existing records to ensure it is complete
- Add in any scanned additional information if not available online for evidence for example of international or national participation
- Use our links to collect additional information
- Find a print version. As the library processes records we automatically ensure that two copies of any book and a single copy where a book chapter is produced by a Deakin author is purchased by the library. In addition as we chase up permission for conferences, if not available online we purchase a copy of the proceedings and add it to the library.

Additional measures to encourage deposit for us are:

- Fuller information and guidelines on the DRO website
- Brochures to hand out at sessions and by Faculty support staff
- Training of Faculty support staff
- Publishing impacts from the repository