

## Repository software

### Background

One of the most frequently asked questions at RUBRIC is “Which repository software do you recommend?”

The RUBRIC project is **not in a position to recommend repository solutions**. The project was conceived to assist regional universities in adopting best-practice emerging from a previous round of projects funded by the Australian Government.

The Australian government has funded a number of projects under the *Systemic Infrastructure Initiative* as part of [Backing Australia's Ability - An Innovative Action Plan for the Future](#)<sup>1</sup>

A first round of projects, funded in 2003, are known as FRODO ( Federated Repositories of Digital Objects) projects.

RUBRIC-central took on the role of Matchmaker – assisting our partner institutions in selecting a repository 'mate' from a small set of solutions recommended to us by the FRODO projects.

While we are unable to answer such questions as “Which repository software should we use?” we are able to share some of the results of our investigations.

### RUBRIC Toolkit

The [RUBRIC Toolkit](#)<sup>2</sup> launched at [!IDEA 2007](#)<sup>3</sup> in October includes information in regards to the process of selecting repository software. However, as with any software procurement choosing a repository is a context-dependent process, and most institutions will have their own selection processes for software evaluation.

### About the FRODO projects

The FRODO projects are [summarized on the DEST website](#)<sup>4</sup>. What follows is a simple glossary, with links to the relevant projects.

#### MAMS

Meta Access Management System Project . Lead Institution - Macquarie University.

<http://www.melcoe.mq.edu.au/projects/MAMS/><sup>5</sup>

**ARROW**

Australian Research Repositories Online to the World. Lead Institution - Monash University.

<http://arrow.edu.au/><sup>6</sup>

**ADT**

Australian Digital Theses Program Expansion. Lead Institution - University of New South Wales.

<http://adt.caul.edu.au/adtariic.html><sup>7</sup>

**APSR**

Australian Partnership for Sustainable Repositories. Lead Institution - Australian National University.

<http://www.apsr.edu.au><sup>8</sup>

A second round of projects, of which RUBRIC is one, is known as [MERRI](#)<sup>9</sup>.

## Background: FRODO recommendations

As part of the initial work on RUBRIC; we surveyed the FRODO projects asking about software that has become available as a result of their explorations. There were three candidate institutional repository software solutions that warranted investigation. Other FRODO outcomes will also be of interest but one of the key goals for RUBRIC is to help regional institutions establish research repository infrastructure, so we looked first at repository software that could meet achievable goals.

### Initial scope

One of the most obvious starting points for an institutional repository is to concentrate on documents in two broad categories:

**Theses & Dissertations**

This includes masters and PhD theses and may or may not include undergraduate (Honours) theses, or even major reports and/or essays.

**Research output**

Published research in the form of pre-print, re-prints or author's drafts, and other non-peer reviewed materials such as working papers or reports.

RUBRIC central was involved in evaluating three software products to determine their suitability for use in a Open Access research repository. The [APSR](#)<sup>10</sup> project recommended Dspace and Fez, while the [ARROW](#)<sup>11</sup> project recommended VITAL.

In addition to these products, USQ staff have experience with the Gnu Eprints package. While this is a mature and dependable software package that has been used in Australian

universities for some time it was not recommended by any of the FRODO projects and was consequently not included in our initial investigations.

Finally, there is one more repository solution of interest: The [MAMS project](#)<sup>12</sup>, has created a repository software application as a reference platform for their work on authorization and authentication. Given that the recently announced [Research Quality Framework](#)<sup>13</sup> process will ultimately require federated authentication and authorization, RUBRIC-central staff will be familiarizing themselves with this software as well.

## Issues to consider

There are a number of general issues to consider in choosing a repository. This page is **not intended to deal with all of the them in detail**, rather it is meant to offer some insights gained from the RUBRIC project.

## How many repositories?

It is by no means a given that a single repository needs to be used for everything. While there are benefits to minimizing the number of different packages required, different types of repository may be best served by different software.

Consider whether different software solutions might be appropriate for:

- An Open Access research repository.
- A thesis repository, which may require embargo features and authorization, for example if theses contain third party copyright material, confidential material or information that could compromise patents.
- Image collections.
- Work in progress.
- A preservation repository, containing records form all of the above but without a public portal.

## Configuration management & architecture

Many repositories require some configuration to be done in text files while others are beginning to allow 'point and click' web-based configuration. While the point and click approach is superficially appealing, it is worth considering whether it is really necessary and the potential impact on quality control.

- Will it encourage future repository managers to change things arbitrarily?
- Can it be turned off or locked-down in a production instance?
- Can the results of the configuration be easily version-controlled and managed? Eg can a

configuration change be rolled back? Can a server be reconfigured quickly in a disaster recovery situation?

The broad configuration of a repository is usually done once. Arbitrary changes in metadata schemas or repository organization may break downstream services such as federated repositories, so it is worth considering whether easy configuration is really of high value.

It is preferable to use a multi-tier architecture with development done on desktop computers, testing on a server, with all configuration kept in a source-code management system.

RUBRIC and partners use the Subversion version control system to develop, manage and share configuration. This is a commonly used system: many other options are explored in [this summary](#)<sup>14</sup>.

When configuration is stored in a database (as is the case with the Fez repository, for example), there should be a procedure developed for exporting and importing configuration.

RUBRIC [uses virtual infrastructure](#)<sup>15</sup> to good effect. Our technical team are well-practiced at creating, configuring, populating and then destroying repositories in a order to ensure there are comprehensive disaster recovery processes in place.

## Metadata schemas

RUBRIC recommendations for the most appropriate metadata schema have been based on standards of granularity, interoperability, extensibility, support and common practice, and accordingly work with Dublin Core (both simple and qualified), MARCXML and MODS metadata schema. The Dublin Core schema is supported by the [Dublin Core Metadata Initiative](#)<sup>16</sup> (DCMI); and MARCXML and MODS both have the institutional support of the [Library of Congress](#)<sup>17</sup>. These schema are compliant with [SRU/SRW](#)<sup>18</sup> protocols.

Dublin Core is the standard metadata schema, supported by the [OAI-PMH](#)<sup>19</sup> (Open Archives Initiative Protocol for Metadata Harvesting) for harvesting and the semantic web. It is also endorsed by [OCLC](#)<sup>20</sup> (Online Computer Literacy Centre) as a library application. The MARC schema is relevant for management and harvesting of metadata in the library world. MODS, an evolutionary development of MARC, has an applicability beyond library service providers, and is more easily applied by non-cataloguers than MARC. RUBRIC testing has shown that any loss of MARC data in the process of transferring data to MODS is negligible and has no impact on the needs and functions of the repositories.

RUBRIC has done work on preparing and applying crosswalks between different metadata schema for records with specific “in house” features and display and interoperability requirements. Documentation of standards of data entry and formats are also being prepared.

In addition to the above, RUBRIC is exploring the most appropriate metadata schema for nonstandard repository resources such as images and video files. Other types of metadata for preservation, validation and rights information management are specialist fields that are under development by various standards organizations (e.g. OCLC, NISO). RUBRIC is maintaining close contact with developments in these areas as they occur.

Preliminary studies have been made of metadata requirements for electronic theses and

dissertations to ensure various theses types will be compatible with both [ADT](#)<sup>21</sup> and [NDLTD](#)<sup>22</sup> requirements for national and international searching and access.

**DSpace** uses qualified Dublin Core metadata. Qualified DC enables more specific information to be managed and harvested. MODS application profile is also available for migration of DSpace records.

**Eprints** uses Dublin Core metadata, although not always consistently with DCMI standards (e.g. Eprints dc.identifier applies to the Eprints metadata record and not the target resource.)

**VITAL** uses MARCXML as the default data entry schema from which it generates an additional Dublin Core datastream. Data can also be entered in the MODS schema which is also used to generate a Dublin Core datastream. (VITAL can additionally generate a Dublin Core datastream from imported EAD files – Encoded Archival Description -- in place of MARXML. But EAD schema is designed for specialist archive collections and RUBRIC does not recommended its use in normal library repositories.

**Fez** now uses MODS metadata by default. Dublin Core is also generated.

Eprints and DSpace use variants of the Dublin Core standard. This has proven to be adequate for open access research repositories, but has some limitations where detailed querying is required.

## Collections

Collections are content aggregations that may be considered metaphors for the way libraries and archives group materials. In an electronic environment 'collections' may be more fluid, and can be thought of as the results of dynamic queries on a repository against metadata. Eg “All the theses from the Faculty of Engineering”.

So far on the RUBRIC project we have not encountered a need for collections other than those that can be created dynamically by a query: these have been dubbed *Virtual Collections*.

Some repository software mandates collections, and some allows it, while Eprints avoids the concept altogether. There are some management issues with collections which have been noted in the pages on individual solutions.

## Preservation

We have made notes on each of the repository solutions regarding their support for preservation metadata, designed to help with long-term archiving.

Even if the repository under consideration does not deal with preservation metadata it may be worth considering setting up a preservation repository. For example data could be exported from an Eprints repository and automatically ingested into a fedora repository complete with automatically generated preservation metadata. Please contact the RUBRIC

team if this idea appeals and we will look into it.

One of the keys to preservation is to make sure that data is clean and technical staff understand how to extract it from the repository and import into other systems.

## Harvesting

RUBRIC follows the Open Archives Initiative Protocol for Metadata Harvesting ([OAI-PMH](#)<sup>23</sup>) which has become an international standard for metadata interoperability among open access and institutional repository movements. The minimum standard for interoperability across service providers is the simple Dublin Core schema.

Specialist providers will also seek more granular metadata for their fields of interest and OAI-PMH strongly encourages the use of more granular and specialist metadata schema for this more fine-tuned searching. For this reason qualified Dublin Core, and other metadata schema – in particular [MARCXML](#)<sup>24</sup> and [MODS](#)<sup>25</sup> -- are also strongly encouraged by the OAI-PMH. These have the advantage of making more granular data available in a recognized format for harvesters in the library and educational communities.

**DSpace** uses qualified Dublin Core. This complies with minimum OAI harvesting standards but also allows for the more detailed searching undertaken by some service providers.

**Eprints** uses simple Dublin Core which is adequate for basic OAI harvesting standards and interoperability.

**VITAL** generates a simple Dublin Core datastream for basic OAI harvesting standards and interoperability. It also houses a MARCXML or MODS datastream for more granular harvesting.

**Fez** handles multiple metadata schema, including Dublin Core and MODS.

## Specific repositories

### **DSpace**

DSpace is a mature open source repository solution which was recommended to RUBRIC by the APSR project.

Read our summary

### **Eprints**

Eprints was not recommended by any of the FRODO projects, but it has been used at the University of Southern Queensland, which hosts the RUBRIC project.

Read our summary.

### **Fedora based**

Fedora is a back-end repository solution, which requires front-end software.

**Fez**

This open source software produced at the University of Queensland was recommended as a best practice repository by the APSR project.

Read our summary.

**VITAL**

VITAL is a commercial software package developed with some input from the Australian ARROW project.

Read our summary.

**Name TBA**

This is a repository developed as a bi-product of the MAMS project, demonstrating best-practice for repository architecture. RUBRIC has not tested this in any detail, but we will be looking at it as part of our ongoing investigations into repository technology.

[Summary forthcoming]

- 1 <http://backingaus.innovation.gov.au/>
- 2 [http://www.rubric.edu.au/packages/RUBRIC\\_Toolkit/default.htm](http://www.rubric.edu.au/packages/RUBRIC_Toolkit/default.htm)
- 3 <http://www.linkaffiliates.net.au/idea2007/>
- 4 [http://www.dest.gov.au/sectors/research\\_sector/policies\\_issues\\_reviews/key\\_issues/australian\\_research\\_information\\_infrastructure\\_committee/ariic\\_projects.htm#2003\\_Federated\\_Repositories\\_of\\_Online\\_Digital\\_Objects\\_\(FRODO\)\\_Projects](http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/australian_research_information_infrastructure_committee/ariic_projects.htm#2003_Federated_Repositories_of_Online_Digital_Objects_(FRODO)_Projects)
- 5 <http://www.melcoe.mq.edu.au/projects/MAMS/>
- 6 <http://arrow.edu.au/>
- 7 <http://adt.caul.edu.au/adtariic.html>
- 8 <http://www.apsr.edu.au/>
- 9 [http://www.dest.gov.au/sectors/research\\_sector/policies\\_issues\\_reviews/key\\_issues/australian\\_research\\_information\\_infrastructure\\_committee/ariic\\_projects.htm#2005\\_Managed\\_Environment\\_for\\_Research\\_Repository\\_Infrastructure\\_\(MERRI\)\\_Projects](http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/australian_research_information_infrastructure_committee/ariic_projects.htm#2005_Managed_Environment_for_Research_Repository_Infrastructure_(MERRI)_Projects)
- 10 <http://apsr.edu.au/>
- 11 <http://arrow.edu.au/>
- 12 <http://mams.melcoe.mq.edu.au/>
- 13 [http://www.dest.gov.au/sectors/research\\_sector/policies\\_issues\\_reviews/key\\_issues/research\\_quality\\_framework/default.htm](http://www.dest.gov.au/sectors/research_sector/policies_issues_reviews/key_issues/research_quality_framework/default.htm)
- 14 <http://linuxmafia.com/faq/Apps/scm.html>
- 15 [http://rubric.edu.au/techreports/VMware\\_At\\_RUBRIC.htm](http://rubric.edu.au/techreports/VMware_At_RUBRIC.htm)
- 16 <http://dublincore.org/>
- 17 <http://www.loc.gov/standards/>
- 18 <http://www.loc.gov/standards/sru/>
- 19 <http://www.openarchives.org/>
- 20 <http://dublincore.org/>
- 21 <http://adt.caul.edu.au/>
- 22 <http://www.ndltd.org/>
- 23 <http://www.openarchives.org/>
- 24 <http://www.loc.gov/standards/marcxml/>
- 25 <http://www.loc.gov/standards/mods/>