

## RUBRIC Toolkit: Data Management

There are many data decisions to be made when establishing an Institutional Repository (IR). These include:

- data sources to locate any existing data
- evaluation of existing data
- data migration planning to obtain and ingest the data
- adding data about the data (metadata) for management purposes
- data reporting and required outputs

The [Australian Code for the Responsible Conduct of Research](#)<sup>1</sup> outlines best practice guidelines for the management of research data. This document collates the results of a working group established in 2003, comprising representatives from:

- the National Health and Medical Research Council (NHMRC),
- Australian Vice Chancellors' Committee (AVCC)
- the Australian Research Council (ARC).

[Stewardship of Digital Research Data – Principles and Guidelines](#)<sup>2</sup> was released by the Research Information Network for comment in April 2007.

## Data Sources

Consider where data sources may already exist in your organisation.

Citation Data Sets can be purchased from commercial data sources, but IR Managers are cautioned about doing this. Data quality should be carefully reviewed before committing to this course of action.

### **Commercial data case study:**

The RUBRIC Project had funding to purchase citation data for use as a basic set of records in each partner repository.

The University of Newcastle conducted an investigation into the usefulness of the proposed source of citation data and reported the following issues back to DEST:

- the data investigated represented only part of the research output from Universities, e.g. in the source reviewed only 33% of overall research output was available for the University of Newcastle
- there were problems with inaccuracies in the data. For example, search results for University of Newcastle included some publications for University of Newcastle upon Tyne (in the UK)

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

- initial costs and an ongoing commitment to the data provider were also of concern

The RUBRIC Project Board decided that it would be preferable to manage data entry from local sources at each institution rather than obtain it from an external vendor.

Alternative strategies include:

- loading citation data from local research databases
- linking to other systems
- migrating theses and dissertations
- digitisation projects to include theses not currently digitised

Local data may require some cleanup to ensure data quality. Most data needs some formatting before extraction and import into the repository.

Most universities have existing sources of data that can be used to “fast start” the IR. These sources might include (but not be limited to):

- a local database of research reportable data. In Australia, this may be linked to the annual HERDC (Higher Education Research Data Collection) reporting process.
- a digitized thesis collection
- a collected body of work in digital format
- a project or collection archive
- a set of data which can be extracted from an external data source (this may or may not be a commercial entity)

Any data source needs some analysis to determine whether it will be suitable for inclusion in the IR. The evaluation should involve staff with data analysis expertise as well as staff with content or subject knowledge.

The [Technical Reports](#)<sup>3</sup> area of the [RUBRIC website](#)<sup>4</sup> documents the technical aspects of data migration from other sources.

## Evaluation of Existing Citation Data

The IR Manager will need to determine whether existing data sources contain suitable data for inclusion in the IR. Useful questions to ask include:

- what existing data can be extracted?
- is the data suitable? (and consistent with the Collection Development Policy)
- is the data of sufficient quality?
- can it be extracted in a useful format?
- will the extracted data need any modification before being imported to the IR?

- when can the data can be migrated?
- will we still maintain the original data source?
- does the data migration need to be done repeatedly to transfer data between systems?
- will we exclude previously exported data with repeat transfers?

Review of 2005 Research Data in Callista produced by the University of Newcastle is an excellent example of this kind of analysis. This document outlines:

- the purpose of the data review
- the scope of the data under evaluation (including number of records)
- the process of data analysis for quality purposes, including:
  - how to access the data
  - bibliographic verification
  - tools required and steps to take for checking purposes
  - additional information required such as DOIs (Digital Object Identifier)

The University of Newcastle conducted a survey for RUBRIC partners asking them to identify:

- the years of coverage (of citation records) required for their institution
- any issues, problems or anomalies with the identified source of citation data that partners were already aware of
- other identified sources (internal or external) for potential citations data for their university
- other library databases that partners felt have strong representation of citation data for their university and that should be investigated for RUBRIC.

## Data Standards

Certain data standards must be used in an IR to ensure potential data is:

- consistent and structured
- compatible with other systems
- harvestable
- discoverable for retrieval and use
- preserved for sustainability

Metadata is the most common data standard. It means “data about data” and is descriptive data about the records stored in an IR. A very full explanation of how metadata can be

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

applied, what it means and what decisions must be made is available in the section on Metadata.

Z39.50 or the SRW/SRU standards are useful for interoperability if supported by the IR software. These are needed for data linking services.

[AIRway](#)<sup>5</sup> is a project in Japan which uses these standards to enable a researcher to obtain Open Access documents easily through the link resolver.

[OAI-PMH](#)<sup>6</sup> (Open Access Initiative Protocol for Metadata Harvesting) is an important standard to facilitate metadata harvesting from individual IR systems by search engines

Harvesting in the Metadata section provides tools which enable you to test compliance of your IR with OAI-PMH.

The [SPARC Institutional Repository Checklist and Resource Guide](#)<sup>7</sup> provides detailed information on technical and systems issues such as:

- ability to migrate and survive
- document formats and longevity
- scalability
- persistent identifiers
- interoperability
- OAI (Open Access Initiative) compliance

#### **Australian Discovery Services**

- [Arrow Discovery Service](#)<sup>8</sup>
- [Picture Australia](#)<sup>9</sup>
- [Australasian Digital Theses \(ADT\)](#)<sup>10</sup>

#### **International Discovery Services**

[Google](#)<sup>11</sup>

[Google Scholar](#)<sup>12</sup>

[Van De Sompel](#)<sup>13</sup> (2000) and [Hunter](#)<sup>14</sup> (2005) provide useful background understanding of the Open Access Initiative.

## Data Migration

The [Technical Reports](#)<sup>15</sup> area of the RUBRIC website provides access to migration scripts produced for RUBRIC Partners. These are freely available for use (under Creative Commons Licence).

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

ADT Migration Scripts were produced for the ADT Technical Committee in March 2007 explaining the process for migrating Australasian Digital Theses (ADT) data. These are available on the RUBRIC website.

The [System Options](#)<sup>16</sup> section provides further information on the range of data migration activities undertaken by the RUBRIC Technical team.

## Data Reporting and Outputs

Data from a repository is more widely used if integrated with other technologies such as

- statistics
- RSS (Rich Site Summary) feeds
- web services

[Statistics](#)<sup>17</sup> on repository usage are valuable in demonstrating return on investment, advertising the IR and also for forward planning.

[RSS feeds](#)<sup>18</sup> from your repository allows users to keep track of changes of interest to them without visiting the repository. At USQ, the Marketing department expressed interest in using the RSS feed capability of USQ ePrints to promote research strengths to prospective students.

[Webservices](#)<sup>19</sup> can be used for a variety of purposes such as the automatic generation of Author pages and Subject listings among others. At the University of Southern Queensland, one of the Departments worked with the ePrints team to create a web service to automatically generate departmental lists of publications from their deposits in USQ ePrints. This enabled them to pass responsibility for archiving the digital objects to the repository, but still gave them their own customised presentation of the information in their departmental web pages with a minimum of effort.

## References and Further Reading

Refer to the Further Reading section at the end of the Toolkit for bibliographic details of works referenced in this section.

---

“RUBRIC Toolkit: Data Management” produced May 2007

20





Regional Universities Building Research Infrastructure Collaboratively

<http://www.rubric.edu.au/>

Copyright<sup>21</sup> 2007 RUBRIC<sup>22</sup>

RUBRIC is supported by the Systemic Infrastructure Initiative as part of the Commonwealth Government's Backing Australia's Ability - An Innovative Action Plan for the Future (<http://backingaus.innovation.gov.au>)

- 1 [http://www.nhmrc.gov.au/funding/policy/\\_files/acrcr.pdf](http://www.nhmrc.gov.au/funding/policy/_files/acrcr.pdf)
- 2 <http://www.rin.ac.uk/data-principles>
- 3 <http://www.rubric.edu.au/techreports/index.htm>
- 4 <http://www.rubric.edu.au/>
- 5 [http://airway.lib.hokudai.ac.jp/index\\_en.html](http://airway.lib.hokudai.ac.jp/index_en.html)
- 6 <http://en.wikipedia.org/wiki/OAI-PMH>
- 7 [http://www.arl.org/sparc/bm~doc/IR\\_Guide\\_&\\_Checklist\\_v1.pdf](http://www.arl.org/sparc/bm~doc/IR_Guide_&_Checklist_v1.pdf)
- 8 <http://search.arrow.edu.au/>
- 9 <http://www.pictureaustralia.org/>
- 10 <http://adt.caul.edu.au/>
- 11 <http://www.google.com.au/>
- 12 <http://scholar.google.com/schhp?ie=UTF-8&oe=UTF-8&hl=en&tab=ws&q=>
- 13 <http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html>
- 14 <http://eprints.rclis.org/archive/00005512/>
- 15 <http://www.rubric.edu.au/techreports/index.htm>
- 16 [https://rubric-central.usq.edu.au/packages/RUBRIC\\_Toolkit/docs/System\\_Options.htm](https://rubric-central.usq.edu.au/packages/RUBRIC_Toolkit/docs/System_Options.htm)
- 17 <http://ro.uow.edu.au/asdpapers/44/>
- 18 [http://en.wikipedia.org/wiki/RSS\\_feeds](http://en.wikipedia.org/wiki/RSS_feeds)
- 19 [http://en.wikipedia.org/wiki/Web\\_service](http://en.wikipedia.org/wiki/Web_service)
- 20 <http://creativecommons.org/licenses/by-sa/2.5/au/>
- 21 <http://creativecommons.org/licenses/by-sa/2.5/au/>
- 22 <http://www.rubric.edu.au/>