

# RUBRIC Toolkit: Populating the Repository

## Identifying Content and Developing Policy

It is advisable to translate all key repository decisions into Policy and Procedure documents as soon as possible. This approach assists in managing project scope and resourcing once the system is in production.

A policy helps to define the intended scope of the collection: i.e. subject matter, object type and quality. It can also contain submission workflows (including intellectual property rights clearance), access and sustainability issues. These types of statements can be referred to as collection management or collection development policies or collection and access guidelines.

Such a policy or statement document may include a variety of decisions, some of which will need broader consultation with stakeholders and external policy makers. This checklist will assist in determining the policy areas to be covered, stakeholders to be consulted and the review processes required.

## Policy Development Checklist

### **Purpose of the policy:**

- to guide the management of the repository and its collection
- to ensure quality service provision
- to outline the principles of decision making for development, maintenance and use of objects in the repository
- to outline the repository's objectives, for example:
  - to make publicly funded research available
  - to showcase the research and intellectual output of the university
  - to disseminate more widely the research of the university to the global research community

### **Responsibility for the policy and for collection management:**

- identify responsibility and boundaries for governance, decision making and management on all aspects of policy and day to day operations (library staff, committees and other stakeholders)
- outline how the repository will be managed
- outline the process for endorsing the policy: committee channels, chain of command
- outline who will manage quality assurance

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>)

**Criteria for inclusion and exclusion:**

- identify authorised contributors and their affiliations:
  - staff and students of the university
  - conjoins
  - affiliates and others
- any conditions for submission
- criteria for inclusion and exclusion

A comparison of 10 Australian policy documents provides an overview of the range of material listed for inclusion.

The University of Newcastle Deposit Guidelines also compares policies from a range of organisations

**Copyright and contractual compliance:**

- author submission agreements
- management of copyright issues in theses
- any other deposit conditions

**Removal and disclaimers:**

- criteria for removal
- circumstances under which objects would be removed (for example, if it were found to infringe copyright or confidentiality)
- legal requirements
- identify the decision makers
- editorial rights

**Archiving, access and discovery:**

- what to archive:
  - citation
  - metadata (type)
  - object (what sort)
- what repository users can expect to be able to access, e.g. citation, link to library subscription, online full-text copy of the author's version
- under what circumstances could they expect to see a facsimile version
- relationship with other systems if appropriate: library system, copyright management
- registration with the appropriate agencies to harvest metadata and ongoing management of this process
- levels of access (internal/external) under what conditions (if any)

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>)

- statement on monitoring and applying (where applicable) best practice for the long-term preservation of metadata and objects.

The development of this type of policy can be a long process occurring over the length of the implementation period and will be refined as decisions are made and considered. It is advisable to establish an initial draft version which will develop with input from stakeholders and committees. Consider frequency and a process for policy review.

#### Sample Collection Policies:

- University of Southern Queensland Collection Development Policy
- University of the Sunshine Coast Collection Development Statement
- A Discussion Paper on Collection Development Policies was created by Flinders University for their Steering Committee to raise awareness of issues.

#### Other useful policy guides:

- [University College London](#)<sup>1</sup> has a guide for depositors on the eligibility of papers, copyright clearance, eligible formats, metadata creation and upload procedures
- [Framework of Guidance for Building Good Digital Collections](#)<sup>2</sup> NISO's principles for building good digital collections
- [Model Collection Policy](#)<sup>3</sup> guidelines from AUSEAccess wiki
- [LEADIRS workbook](#)<sup>4</sup>, particularly chapter 4: "Legal and regulatory environment and policy development" which contains another Policy checklist
- [Tool for generating policies](#)<sup>5</sup> a timesaving checklist from the OpenDOAR Project. It covers:
  - Metadata Policy- for information describing items in the repository
  - Data Policy - for full-text and other full data items
  - Content Policy - for types of document and dataset held
  - Submission Policy - concerning depositors, quality and copyright
  - Preservation Policy – for retention, preservation, withdrawal, version control and closure

## Policy at the Organisational Level

In the short term, repository managers will be developing policy relating to the management of the IR. In the long term, policy issues should be addressed at an organisational level where they have the most significant impact on scholarly communication processes.

[Building the Infrastructure for Data Access and Reuse in Collaborative Research: An Analysis of the Legal Context](#)<sup>6</sup> produced by the OAK Law project in 2007, draws on the 2006 consultation draft of the Joint NHMRC/AVCC Statement and Guidelines on Research Practice ([Australian code for the responsible conduct of research](#))<sup>7</sup> to say that organisations

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>)

should be addressing the following policy issues:

- ownership of research records and data
- security and confidentiality of research data and records
- protection of confidentiality
- management of intellectual property

The purpose of these policy and strategic initiatives is to support the researcher in making decisions about data management.

## Deposit Licences

The repository manager, in consultation with the Steering Committee, needs to decide whether a deposit licence will be applied to all material coming into the IR.

The [Sherpa Report on a deposit licence for e-Prints](#)<sup>8</sup> is a useful place to start considering the issues. The main reasons cited for having a deposit licence include:

- establishing a formal contract between author and IR
- reassuring the author about the IR's claims to rights on their work
- providing permission for the IR to manage the work
- reducing the IR's liability if a work is found to infringe copyright

The [EThOS Toolkit](#)<sup>9</sup> says that a deposit licence should cover:

- a depositor's declaration:
  - to determine whether the depositor is the copyright owner, or
  - has permission by proxy to deposit on behalf of the copyright owner
  - to determine whether permission has been sought from third party copyright owners where necessary
- the IR's rights and responsibilities of an IR:
  - permission needed for future acts of digital preservation
  - access and distribution rights
  - removal and ownership rights of the metadata record based on the work
  - protocols for removal of work from the IR
  - declaration of IR liability regarding mistakes, omissions or infringements
- re-use terms and conditions:
  - outlining the rights of end-users to access, download or reproduce works or parts of works

Some organisations may also wish to develop a Repository Distribution (End User)

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>)

Agreement which covers this last point in more depth and can be used to cover any other access issues.

An [OAK-Law presentation at the 2006 RUBRIC Reports conference](#)<sup>10</sup> explains the distinction between a Deposit Licence and Distribution Agreement.

[How to Protect your rights with a licence agreement](#)<sup>11</sup> explains the stages of development of the policy, with short, medium and long term consideration.

The IR Manager will need to decide whether it will be more effective to have a short and simple licence or to cover more complex matters. This should be discussed with the institution's lawyer and signed off by the Steering Committee.

#### **Examples of Australian Deposit Licences:**

- [Flinders Academic Commons Contributor's Licence](#)<sup>12</sup>
- [Adelaide Research & Scholarship](#)<sup>13</sup> deposit licence
- [Australian Social Science Data Archive](#)<sup>14</sup> deposit licence
- [QUT ePrints](#)<sup>15</sup> deposit agreement

#### **Examples of overseas Deposit Licences:**

- [EthOS Deposit Agreement](#)<sup>16</sup> deposit licence
- [Bristol Repository of Scholarly ePrints](#)<sup>17</sup> deposit licence

## Data Entry

The Data Management section provides more detailed information on data sources and data migration and a number of different methods for loading data into the IR which are briefly outlined below.

Data can be added to an IR using any of the following methods:

- batch ingest
- managed or mediated submission
- self submission

Data entry can be driven by either:

- voluntary submission, or
- mandatory submission

## Batch Ingest

The Data Management section guides you through locating and evaluating data sources that may be useful for batch loading into the IR. Batch ingest dramatically reduces the time taken to populate an IR if the content is available from another digital source.

A [Migration Toolkit](#)<sup>18</sup> was developed by RUBRIC Central technical staff. The migration toolkit can be used as a basis for migrating data from another digital system to a repository. It is based on Python and XSL transformations to alter the format in which the data is exported and ingested. The components of the kit can be used and altered according to the migration required.

## Managed or Mediated Submission

Managed or mediated submission involves centrally managed data entry by trained staff.

Managing the submission process in the early phases of establishing a repository is a good idea in order to gain experience with data and to monitor quality issues that may arise. This is a valuable staff training exercise, providing the skills and experience to assist users if the organisation later moves to a self submission model.

A managed submission method enables staff to:

- examine workflows
- determine any data issues
- discover aspects of the system which might cause problems for users

## Self Submission

Self submission is a user-driven method of populating an IR, with the obvious benefits of leverage (given the potential rapid growth of deposits) which is not tied to the staffing resources involved in maintaining the IR.

An [FAQ on self submission](#)<sup>19</sup> maintained by Stevan Harnad contains a detailed overview of this method of building a digital collection, as well as many useful links to documents and websites debating the benefits of mandatory submission and managed submission. There are powerpoint slides available from this site which can be used (with acknowledgement) to support the cause for Open Access IRs.

## Submission Workflows

Flinders University adopts a range of workflows to suit various requirements:

- **Library submission:** Documents are received initiating the following procedures: file conversion, copyright checking, metadata creation and deposit into the repository.

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>)

- **Academic self-submission:** It is anticipated that training early career academics in self submission will be standard by the end of 2007. It is essential to have academics within each of the faculties who are prepared to support new submitters.
- **Delegated academic submission:** Post graduate students can be employed as archivists to deposit materials on behalf of academic staff. This works extremely well as library resources are not stretched and academics are able to dedicate their time to the work.
- **The future:** Processes will improve as further areas of the University become involved. It is expected that the library will do the initial work and then discuss ways to maintain things submissions in the long term.

The University of Southern Queensland's submission charts may be useful:

- Self Submission
- Editorial Submission

At USQ, users (authors) are registered by ePrints staff, but can then login using the university's LDAP authentication system. Initial training is provided to enable authors to self submit. The benefits of this process are:

- personal contact gives a good impression; academics feel a personal commitment to the process
- important issues can be discussed immediately
- improvements or further information can be relayed
- as first submissions are monitored, instructions can be amended and catalogued for use by others
- an introductory or follow up email with an attachment of a Quick Guide to Submitting can act as a prompt

## Mandated Deposit

Mandatory deposit policies require the deposit of all digitally produced research output in line with the IR's Collection Development Policy.

[The Impact of Mandatory Policies on ETD Acquisition](#)<sup>20</sup> (ETD: Electronic Thesis and Dissertation) suggests that mandatory submission policies for electronic theses causes submission rates to rise to 50 - 80% where voluntary submission may only yield 5 - 15% . The paper includes a table of the state of mandatory policies in Australian universities.

The [Queensland University of Technology policy document](#)<sup>21</sup> is an example of a mandatory submission policy.

In summary, this article proposes that:

- universities that establish an ETD repository seem to be wasting their money if they maintain a voluntary deposit policy. Deposits are poor, running at most at 12% to 20%

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>)

- mandatory policies pay off handsomely in capturing all or most theses
- mandatory policies established from date of submission are 5-6 years faster in achieving 80% compliance than policies dated from enrolment.

[The Alliance for Taxpayer Access](#)<sup>22</sup> (a USA site) Media Release in June 2007, endorsed the effectiveness of mandatory policies, pointing out that since the voluntary NIH Public Access Policy was put into effect “in May 2005, investigators have deposited less than five percent of eligible manuscripts”. Congress has expressed concern about the voluntary policy’s failure to meet its goals” and has proposed legislative action in the form of a bill to go to the Senate in July 2007 requiring all NIH funded research to be made publicly available on the internet.

[The Patchwork Mandate](#)<sup>23</sup> was written for “repository managers who are at a loss as to what policies they (or their universities or research institutions) ought to deploy in order to ensure that most, if not all, of the institution's scholarly output is deposited in the institution's repository”. The paper suggests that institutional mandatory deposit policies can be built up by securing mandatory policies at a departmental level. In other words, gradual compliance.

## File Naming for Repositories

It is important to decide why and how to apply consistent schema to file names in the planning stage. Prior planning of file naming conventions to be used can reduce risks of:

- having to change file names later on
- avoiding broken links or bookmarks
- managing editorial and submission workflows

File naming guidelines might include information about:

- acceptable file formats (will there be restrictions imposed at an editorial level or a software level?)
- conversion into acceptable file formats (user access to software to create pdf versions)
- file sizes (will large files be broken into smaller components?)
- whether compressed files can be managed (or will this pose usability issues?)

There is not a single convention for filenames. For example, multimedia or image files, unpublished preprints or institutional working papers may be better handled with alternative naming methods.

[Preparing files for Contribution](#)<sup>24</sup> at the Flinders Academic Commons provides guidelines about file naming, file types and formats along with other useful information for submitters, which demonstrates that IR Managers can present this information succinctly to users online.

[File Format Guidelines](#)<sup>25</sup> outlines guidelines on file formats, naming, sizes and compression.

“Remember that you are not only storing information but also providing a collaborative environment for others.” ([Alexport](#)<sup>26</sup> 2007).

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>)

## Preliminary considerations

### Relevance to repository clients:

- Filenames are not always visible in the main portal display of all repositories but they are visible in the address box of the browser and as the name of the file that is downloaded by the user
- If the filename assigned by the depositor is visible to users when the file is downloaded, the filename will become the reference for the user to locate, store and cite the file. For these reasons RUBRIC considers it best practice to apply a meaningful filename convention

### Relevance to administrative staff:

- Consistent and well named files can make it easier to locate files for maintenance and updating. Filenames can assist in the efficient retrieval of file copies e.g. if checking the integrity of uploaded files. A useful convention in assigning filenames is considered best practice.

## Filenaming Conventions: Do's, Dont's and Considerations

### DO:

- **factor in self-submission.** If encouraging academics to submit their own works in the repository, then consider how a file naming convention may impact on them and their willingness to adapt. Alternatively the library may change the names of submitted files as part of the editorial workflow. Some faculties or centres may wish to adopt their own file naming schemes
- **keep it relatively short.** Adding authors' names and initials, detailed dates (20060810), item types (e.g. prp for preprint, ja for journal article, etc.) may be informative but consider the usefulness of such information once the file is uploaded and archived. Too much information makes filenames unwieldy. Theoretically a filename can have 128 or more characters, however, many software systems break long filenames and it's more likely there will be errors in both creating and retrieving it
- **use a brief yet meaningful filename.** This increases its potential for impact and recall by users, both public and administrative. Make it clear and mnemonic and there will be less likelihood of making mistakes in assigning names. Consider a filename consisting of just a few title keywords
- **use only one "." in the entire filename.** It should be right before the file extension which is usually 3 characters such as .pdf, .doc, .htm etc

### DON'T:

- **use a repository's PID** (e.g. rubric156, vital1043) as this will lose its meaning and possibly cause confusion in the event of a future migration to another repository
- **use punctuation characters.** Some characters such as / : \* ? " < are reserved by the operating system's shell. Spaces should be avoided or replaced with hyphens or

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>)

underscores particularly with images (Underscores, however, may sometimes be used for spaces)

- **use upper case characters.** Best practice is not to use upper case characters since upper and lower case characters are not treated uniformly across all operating systems. Consistency makes them easier to work with, too
- **use spaces.** Some web browsers will discard anything after a space.

## CONSIDERATIONS

- **use of abbreviations.** Before deciding on institutional or other abbreviations consider the longevity of their potential relevance and obvious meaning. Also avoid anything difficult to spell
- **use of an author's name,** especially as the first characters of the filename, is a logical choice for files authored by the one person. However, many documents have multiple authors within the institution. Does one want two different conventions: one for documents with a sole author and another for documents with multiple authors? Or is it best to avoid author names altogether and focus on document type, title, date?
- **use of numbers.** Avoid the use of numbers unless they clearly represent, say, a date or a publication issue number. If using a date then ensure that it will always be entered in a consistent format
- The most short and meaningful file naming scheme may require a minimum of record keeping offline. Hence if a filing scheme is to have a unique number for each year, this may require a simple and easy to manage offline list of used numbers for each year. Or it may be even simpler and useful for users if the filename consisted of a few keywords from the title of the document

## Examples of Filenaming Conventions

Some specific filenaming conventions are described at the following websites. Note that their purposes may not coincide with university repositories and are supplied as a guide to the principles of filenaming.

- [File Naming Conventions for Digitally-stored Images](#)<sup>27</sup>
- [TASI \(Technical Advisory Service for Images\)](#)<sup>28</sup>
- [Ontolog-forum](#)<sup>29</sup>
- [The Alexport](#)<sup>30</sup>

The Data Management section will provide further information if required

## 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: Populating the Repository” produced July 2007



31

32

Copyright<sup>33</sup> 2007 RUBRIC<sup>34</sup>

- 1 <http://eprints.ucl.ac.uk/Deposit.html>
- 2 <http://www.niso.org/framework/Framework2.html>
- 3 <http://leven.comp.utas.edu.au/AuseAccess/pmwiki.php?n=General.SampleQA>
- 4 <http://www.dspace.org/implement/leadirs.pdf>
- 5 <http://opendoar.org/tools/en/policies.php>
- 6 [http://www.oaklaw.qut.edu.au/files/Data\\_Report\\_final\\_web.pdf](http://www.oaklaw.qut.edu.au/files/Data_Report_final_web.pdf)
- 7 <http://www.nhmrc.gov.au/funding/policy/code.htm>
- 8 [http://www.sherpa.ac.uk/documents/D4-2\\_Report\\_on\\_a\\_deposit\\_licence\\_for\\_E-prints.pdf](http://www.sherpa.ac.uk/documents/D4-2_Report_on_a_deposit_licence_for_E-prints.pdf)
- 9 [http://ethostoolkit.rgu.ac.uk/?page\\_id=99](http://ethostoolkit.rgu.ac.uk/?page_id=99)
- 10 <http://www.slideshare.net/arrowcentral/dea2006creating-a-legal-framework-for-open-access/>
- 11 <http://www.ukoln.ac.uk/qa-focus/documents/briefings/briefing-53/briefing-53-A5.doc>
- 12 <http://www.lib.flinders.edu.au/~dspace/license.html>
- 13 <http://digital.library.adelaide.edu.au/dspace/faq/license.jsp>
- 14 [http://assda.anu.edu.au/forms/ASSDA\\_Deposit\\_Licence.pdf](http://assda.anu.edu.au/forms/ASSDA_Deposit_Licence.pdf)
- 15 <http://eprints.qut.edu.au/copyright.html#depositagree>
- 16 <http://ethostoolkit.rgu.ac.uk/wp-content/ethos-content/DEPOSIT%20AGREEMENT.htm>
- 17 <http://www.bristol.ac.uk/is/library/collections/rose/rose-licence.html>
- 18 <http://www.rubric.edu.au/techreports/index.htm>
- 19 <http://www.eprints.org/openaccess/self-faq/#self-archiving>
- 20 <http://www.dlib.org/dlib/april06/sale/04sale.html>
- 21 [http://www.mopp.qut.edu.au/F/F\\_01\\_03.jsp](http://www.mopp.qut.edu.au/F/F_01_03.jsp)
- 22 <http://www.taxpayeraccess.org/media/release07-0628.html>
- 23 <http://www.dlib.org/dlib/january07/sale/01sale.html>
- 24 <http://www.lib.flinders.edu.au/~dspace/contribution.html>
- 25 <http://www.usq.edu.au/eprints/policies/fileformatguidelines.htm>
- 26 <http://www.alexport.net/>
- 27 <http://www.northernjourney.com/photo/articles/filenaming.html>
- 28 <http://tasi.ac.uk/advice/creating/filenaming.html>
- 29 <http://ontolog.cim3.net/forum/ontolog-forum/2004-02/msg00029.html>
- 30 <http://www.alexport.net/>
- 31 <http://creativecommons.org/licenses/by-sa/2.5/au/>
- 32 <http://creativecommons.org/licenses/by-sa/2.5/au/>
- 33 <http://creativecommons.org/licenses/by-sa/2.5/au/>
- 34 <http://www.rubric.edu.au/>