Getting Research Data ‘Out There’

Collaborative solutions to identifying, describing and making research data more visible

Craig Milne – Research Data Librarian
Ellen Thompson – Research Data Librarian
Lance De Vine – High Performance Computing and Research Support Specialist

Information Online 2011
ALIA 15th Conference and Exhibition
1-3rd February 2011, Sydney
What is Research Data?

“....information that is generated or collected to be used as primary sources in the production of original research results and would be required to validate or replicate research findings.”

Guidelines for the Management of Research Data at QUT

QUT Policy - D/2.8 Management of research data
http://www.mopp.qut.edu.au/D/D_02_08.jsp
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>iso 3166 country</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>AU Australia</td>
<td>25.4</td>
<td>27.3</td>
<td>28.4</td>
<td>27.3</td>
</tr>
<tr>
<td>3</td>
<td>US United States</td>
<td>30.6</td>
<td>33.3</td>
<td>23.6</td>
<td>14.8</td>
</tr>
<tr>
<td>4</td>
<td>GB United Kingdom</td>
<td>7.3</td>
<td>6.3</td>
<td>7.2</td>
<td>7.2</td>
</tr>
<tr>
<td>5</td>
<td>KR Korea, Republic of</td>
<td>0.8</td>
<td>0.7</td>
<td>1</td>
<td>4.6</td>
</tr>
<tr>
<td>6</td>
<td>IN India</td>
<td>2.5</td>
<td>2.6</td>
<td>3.8</td>
<td>4.5</td>
</tr>
<tr>
<td>7</td>
<td>CN China</td>
<td>4.1</td>
<td>2.8</td>
<td>2.5</td>
<td>2.6</td>
</tr>
<tr>
<td>8</td>
<td>CA Canada</td>
<td>2.7</td>
<td>2.4</td>
<td>2.7</td>
<td>2.5</td>
</tr>
<tr>
<td>9</td>
<td>JP Japan</td>
<td>0.8</td>
<td>0.5</td>
<td>0.7</td>
<td>2.1</td>
</tr>
<tr>
<td>10</td>
<td>MY Malaysia</td>
<td>1.2</td>
<td>1.2</td>
<td>1.7</td>
<td>2</td>
</tr>
<tr>
<td>11</td>
<td>FR France</td>
<td>0.9</td>
<td>0.8</td>
<td>1.1</td>
<td>2</td>
</tr>
<tr>
<td>12</td>
<td>DE Germany</td>
<td>1.8</td>
<td>1.6</td>
<td>1.8</td>
<td>1.8</td>
</tr>
<tr>
<td>13</td>
<td>Blanks Blanks</td>
<td>0.4</td>
<td>0.2</td>
<td>0</td>
<td>1.5</td>
</tr>
<tr>
<td>14</td>
<td>ID Indonesia</td>
<td>0.9</td>
<td>1.1</td>
<td>1.5</td>
<td>1.3</td>
</tr>
<tr>
<td>15</td>
<td>IR Iran, Islamic Republic</td>
<td>0.8</td>
<td>0.9</td>
<td>1.1</td>
<td>1.3</td>
</tr>
<tr>
<td>16</td>
<td>NL Netherlands</td>
<td>1</td>
<td>0.9</td>
<td>1.1</td>
<td>1</td>
</tr>
<tr>
<td>17</td>
<td>NZ New Zealand</td>
<td>1</td>
<td>0.9</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>18</td>
<td>ZA South Africa</td>
<td>0.7</td>
<td>0.7</td>
<td>0.9</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>IT Italy</td>
<td>0.9</td>
<td>0.8</td>
<td>0.9</td>
<td>0.9</td>
</tr>
<tr>
<td>20</td>
<td>SG Singapore</td>
<td>0.8</td>
<td>0.7</td>
<td>0.9</td>
<td>0.9</td>
</tr>
</tbody>
</table>
live audience feedback
SMS messaging (time-stamped)
feedback written on tablecloths
**Why Do We Want to Get Research Data ‘Out There’**

- Australian Code for the Responsible Conduct of Research
- Publically funded research
- Reuse and repurposing
  - Save time and money
- Sharing
  - Speed up scientific discovery
- Multi-disciplinary research projects

- Store it
- Register it
- Describe it
- Expose it
- Find it
- Access it
- Reuse it
What’s already ‘out there’

• Genbank – Gene sequence data
• TARDIS – x-ray diffraction images
• IMOS – Integrated Marine Operating System
• PANGAEA – geoscientific / environmental
• Survey Question Bank (UK Data Archive)

• ABS
• data.australia.gov.au
• Research Data Australia
Research Data Australia

A Window on the Australian Research Data Commons

Research Data Australia is a discovery service for Australian research data.

Collection Providers
- Australian Institute of Marine Science (371)
- Australian Nuclear Science and Technology Organisation (4)
- Breast Cancer Tissue Bank (18)
- Commonwealth Scientific and Industrial Research Organisation (27)
- Griffith University (11)
- IVFC MEST - Western Australian Marine Data and Projects (674)
- Monash University (30)
- PARADISEC (145)
- Polar Information Commons (26)
- Publish My Data (45)
- Queensland University of Technology (12)
- University of South Australia (4)
- University of Wollongong (4)

http://services.ands.org.au
**QUT Seeding the Commons**

- Managed by two Research Data Librarians with technical support from HPC Specialists

**Main Objectives**
- Identify QUT’s Category 1 research data
  - ARC and NHMRC funded research
- Describe the research data
  - Create metadata records for the research data
- Provide a feed to Research Data Australia
  - RIF-CS compliant records to be fed to the ARDC

**Broader objectives**
- Raise awareness of research data management
- Develop a continuing culture, capability and capacity to manage research data
2 projects being undertaken simultaneously with the seeding the commons

Software development work being undertaken by HPC

Data Repository
  - based on Mediaflux, an XML object orientated database
  - Used to store the “seeding the commons metadata records
  - Good for handling multiple schemas

Metadata Exchange Hub
  - a joint project with Griffith University
  - Metadata aggregator
Identifying existing metadata

- Identify existing metadata with QUT systems
  - Reduce duplication of effort and ensure data integrity
- Research Master
  - Office of Research System
  - 563 potential Research Projects and 271 primary contacts to interview within the project scope
  - Information exported to very large spreadsheets
- Other University systems
  - QUT Staff Profile (up to date contact details, title)
  - QUT ePrints (linking publications and datasets)
Creating the metadata records

- QUT data mapped to RIF-CS
  - To identify what was missing from existing data
- RIF-CS (Registry Interchange Format – Collection Services)
  - is an object orientated, relational model
  - Party, Collection, Activity Objects, Service
- QUT created own schema
  - More granular version of RIF-CS
  - Added in an Interview object
  - Future proofing
  - Map to RIF-CS as required
Pilot interviews in 2009

195 interviews undertaken in relation to 424 research activities over approx. 4 months

15 minutes – 1 hour

Semi-structured interviews using a data interview template

Liaison Librarians attended where possible

### Obtaining additional metadata

**The Data Interview**

<table>
<thead>
<tr>
<th>Interview Topic</th>
<th>Typical Metadata collected</th>
</tr>
</thead>
</table>
| **Description and coverage of dataset** | • Title of dataset, description of dataset  
• Keywords (author and controlled vocabularies)  
• Data source (instruments, surveys, existing data, simulations, variables, cohort size, demographics)  
• Temporal and spatial coverage  
• Investigators, Primary Data Contact, Who collected the data |
| **Documentation and file information** | • What documentation exists about the data (codebooks, data management plans)  
• Type of data (digital or analogue, quantitative or qualitative, documents, video, image, audio, GIS)  
• Format and file sizes  
• Specialist software required  
• Publications associated with data (journal articles, thesis) |
| **Access and location of data** | • Data retention & Disposal  
• Location of data  
• Can data be shared (ethics, contracts, confidentiality, IP)  
• Embargoes, copyright, ownership |
Workflow and Systems
# The Data Librarian Interface

## Metadata (Collection)

<table>
<thead>
<tr>
<th>RDA Key *</th>
<th>Chemical concentrations of pollutants from outboard engines</th>
</tr>
</thead>
<tbody>
<tr>
<td>co0020000292</td>
<td>excel</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Experimental</th>
<th>Shared</th>
<th>yes</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Sharing Issue</th>
<th>ed for two years. Data needs some 'cleaning' before data can be placed in data repository.</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Collection Type</th>
<th>dataset</th>
<th>RDA Status</th>
<th>Approved</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Data Collected</th>
<th>2004</th>
<th>Collected From</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Collect to</th>
<th>Disposal Date</th>
<th></th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Retention Start</th>
<th>Retention Period</th>
<th>indefinite</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Subject</th>
<th>Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>00702</td>
<td>991399</td>
<td>FOR-08</td>
</tr>
</tbody>
</table>

| Description (Full) | lubricants and different outboard motors at varying engine speeds. Data was obtained in laboratory and field environments for both fresh water and sea water. 20 pages of tables in Excel format. |


* Compulsory fields

* Related Info URL:

<table>
<thead>
<tr>
<th>Spatial Coverage</th>
<th>Name</th>
<th>Type</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Relationship</th>
<th>Party</th>
<th>Activity</th>
<th>Collection</th>
<th>Interview</th>
<th>Publication</th>
</tr>
</thead>
</table>

[Save] [Update] [Save to file] [Delete] [Close]
Working Collaboratively

• Sourcing metadata
  – Researchers (data interviews and follow ups)
  – Office of Research (Research Master Data)
  – QUT ePrints (linking data and publications)
  – Future work with Ethics Office

• Development work
  – HPC and Research Support
  – ANDS Staff

• Party Identifiers
  – National Library of Australia, Party Infrastructure Project
Project Challenges

• Short project timeline
• Skilling up (xml, schemas and data sharing)
• Simultaneous data interview and data repository interface development
• Working in an evolving environment
  – Starting from scratch
  – Requirements for Research Data Australia changed as the project was progressing
Barriers to getting Research Data ‘Out There’

• Ethics
  – In some cases overly restrictive consent forms have been used

• Contracts
  – Despite public funding there are also commercial partners and therefore IP issues

• Data Condition
  – Poor data management practices – i.e. little or no metadata, poor storage procedures, where is it?

• Researcher
  – often overcome once it is made clear it is only the metadata and they still control the research data

• Data Management / Research Culture
  – Being addressed by data management guidelines and data management workshops
  – Advocacy by Liaison Librarians
Project Outcomes

• QUT metadata records were fed to Research Data Australia
  – 12 collections, several more with datasets waiting approval
• Rich anecdotal data collected in interviews
  – Important feedback for IT areas and developing guidelines
• Enhanced collaborations between areas internally at QUT
  – HPC and Research Support and the Library (Technical / Advocacy)
• Starting to see a change in data management culture
  – Requests for presentations / workshops
• Liaison Librarians advocating for good data management practices
  – Referral to HPC / ITS
  – Raising awareness of data management guidelines
• Lessons learned and systems developed are being used and fed into ongoing projects looking at more automated methods of data capture

• QUT is now well placed to enable sharing and discovery of QUT Research Data
This project is supported by the Australian National Data Service (ANDS)

ANDS is supported by the Australian Government through the National Collaborative Research Infrastructure Strategy Program and the Education Investment Fund (EIF) Super Science Initiative

Australian Government
Department of Innovation, Industry, Science and Research