Build it and they will come? Strategies for populating a self serve data repository
Build it and they will come?*

Who are “they”? Introduction to CSIRO researchers

What are we building? About the CSIRO Data Access Portal (DAP)

What might get them to come?
- Strategies employed so far to populate the DAP

What worked? Outcomes of the strategies

What are we going to do next?
62% of our people hold university degrees

Top 1% of global research institutions in 14 of 22 research fields

Top 0.1% in 4 research fields

With our university partners, we develop 650 postgraduate research students

2000 doctorates

500 masters
What we built: Data Access Portal (DAP)

• Developed by CSIRO Information Management & Technology (IM&T) and research partners

• **Secure** repository

• Enables **external publication** but options for other access levels

• **Self serve** and enterprise wide

• Options for **searching, retrieving** and **downloading**
How we built it: DAP development and implementation

First steps: Project partnering
Dec 09 - Jan 11
Phase 1- MDBSY

Testing the waters: Early adopters/pilots
May 10 - Feb 11
Phase 2- ATNF

Taking it to the masses
Feb 11 - Nov 11
Phase 3

Jan 12 - Jul 12
Phase 4

Apr 12
General release announced
Strategies: Develop a user friendly system

- Respond to feedback from early adopters & reference group (Agile methodology)
- UI consultant – four easy steps to deposit

Deposit Data Collection

Four steps to deposit data:

- Description
- Citation
- Data
- Restrictions

- Context sensitive help
- Support for domain specific metadata schema
- Scheduled data deposits
- XML metadata import
Strategies: Incentives in the infrastructure

• Persistent identifiers for every collection
• DOIs for public collections
• Enable linking to “related materials”, including records in our publications repository
• Flexible access and permissions framework
• Approval system to mitigate risk
• Image gallery view and image metadata
• Discovery- public records also findable in Research Data Australia and Google
Strategies: add personal support

• **Research Data Service (RDS)** Support team practical, hands-on support for data deposits

• **Advanced Scientific Computing (ASC)** technical advice and assistance with preparing data for publication

• **DAP Development team** trouble-shooting system issues
Strategies: Train and promote

Online

• Help pages and trouble shooting guide
• Sandbox environment
• Intranet pages
  Quick Reference Cards
  Short video tutorials
• Newsletter articles
• Internal social media
Strategies: Train and promote

Personal

- We talk and demo to anyone anytime!
- Face to face or video-conference – individuals, project and business groups, internal conferences, tea room chats ...

Training sessions through Library Services- reached 181+ (as of end Sept 2012)
  Approver training – reached 80
  Face to face demo sessions on sites- group (reached 57+)

All delivering consistent key messages
Would you like to publish your data?

Did you know that having citable data in a publication will increase your citation rates?
Find out how to cite your data collections.
Strategies: Targeting our approaches

Targeting researchers:

- Already sharing
  just get data redirected to DAP
  audit of what was already on our web sites
- At appropriate point in time in research process
- With journal or funder mandates
  getting together lists of journals and funders
  getting reports from project system on who funders are
Targeting researchers at appropriate point in research process

• Started 15 September 2012

• Phone calls to authors who had submitted articles to peer reviewed journals in previous 2 weeks - report from publications approval system

• Little awareness of DAP and Research Data Service but most were interested in DAP and wanted extra information. Good awareness raising exercise
The numbers: Collections

At general release date end April 2012

• 356 (all stages and levels)
• Published 310
• Public 304

At end September 2012

• 564 (all stages and levels) increase of 208
• Published 419 increase of 109
• Public 364 increase of 60
The numbers: Depositors

At general release date end April 2012

21 across 6 Divisions

At end September 2012

45 across 12 Divisions
<table>
<thead>
<tr>
<th>Division</th>
<th>Draft</th>
<th>Submitted</th>
<th>Published</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSIRO Astronomy and Space Science (ATNF)</td>
<td>326</td>
<td></td>
<td></td>
<td>326</td>
</tr>
<tr>
<td>Animal, Food and Health Sciences (AAHL)</td>
<td>48</td>
<td>52</td>
<td></td>
<td>100</td>
</tr>
<tr>
<td>CSIRO Information Management and Technology (IM&amp;T)</td>
<td>64</td>
<td>3</td>
<td>12</td>
<td>79</td>
</tr>
<tr>
<td>Land and Water</td>
<td>11</td>
<td>7</td>
<td></td>
<td>18</td>
</tr>
<tr>
<td>ICT Centre</td>
<td>5</td>
<td>1</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>Ecosystem Sciences</td>
<td>3</td>
<td>5</td>
<td></td>
<td>8</td>
</tr>
<tr>
<td>Marine and Atmospheric Research</td>
<td>1</td>
<td>3</td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>Mathematics Informatics and Statistics</td>
<td>2</td>
<td>1</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>CSIRO Materials Science and Engineering</td>
<td>2</td>
<td>1</td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>Plant Industry</td>
<td>2</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>(blank)</td>
<td>2</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>CSIRO Earth Science and Resource Engineering</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>Food and Nutritional Sciences</td>
<td>1</td>
<td></td>
<td></td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>139</strong></td>
<td><strong>6</strong></td>
<td><strong>419</strong></td>
<td><strong>564</strong></td>
</tr>
</tbody>
</table>
## Published Collections by Access (end Sept 2012)

<table>
<thead>
<tr>
<th>Division</th>
<th>CSIRO_ONLY</th>
<th>SPECIFIC_USERS</th>
<th>PUBLIC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CSIRO Astronomy and Space Science</td>
<td></td>
<td></td>
<td>326</td>
<td>326</td>
</tr>
<tr>
<td>Animal, Food and Health Sciences</td>
<td></td>
<td></td>
<td>45</td>
<td>7</td>
</tr>
<tr>
<td>CSIRO Information Management and Technology</td>
<td>2</td>
<td>1</td>
<td>9</td>
<td>12</td>
</tr>
<tr>
<td>ICT Centre</td>
<td></td>
<td></td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>Land and Water</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>7</td>
</tr>
<tr>
<td>Ecosystem Sciences</td>
<td>1</td>
<td>1</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>Marine and Atmospheric Research</td>
<td>2</td>
<td></td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>(NA)</td>
<td></td>
<td></td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>CSIRO Materials Science and Engineering</td>
<td></td>
<td></td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>6</strong></td>
<td><strong>49</strong></td>
<td><strong>364</strong></td>
<td><strong>419</strong></td>
</tr>
</tbody>
</table>
Cumulative Collections added by Date

General release date
What worked?: Assessing the strategies

• DOIs opened ears and eyes. They resonate!
• DAP was built for self service but we needed to incorporate help with metadata creation and data preparation into our support services
• Catching people at the right point in the data/research lifecycle matters
• Talking to top levels doesn’t necessarily reach those with the datasets to share
• Responding to feedback and implementing wanted features in the DAP. Having a (flexible) roadmap for future development
• Automated deposit and batch uploads work best if numbers are the success criteria
What worked?: Assessing the strategies

Best uptake is where:

• There is recognition of a data management issue the DAP could solve.
• There are motivating factors eg. funder requirements, journal policies.
• There is a pre-existing data sharing culture or high demand for the data
• The project is current or very nearly
  people quickly move on to the next project
  no time or resources to revisit
  IP/ownership issues become more difficult to unravel
  data format/structure issues harder to resolve
What’s next?: Further development work

Ongoing system development taking on board user feedback

• Google analytics – carrot for seeing reuse
• Sensor network auto uploads
• Interoperability. Eg. TERN
• Combine/integrate e-Publish (publications approval) and DAP
• Search by location – collections within a geographic area
• Linking to other data repositories within CSIRO, using the Cloud infrastructure currently being developed by IM&T
• Machine to machine functionality: web services, APIs
What’s next?: Strategies?

Incentives:
• Tracking citation of data.

Promotion:
• Increase visibility on Intranet
• Increase awareness through CSIRO Learning and Development courses

Barriers:
• Address “do not have time and resources” issues
• Develop organisational guidelines/procedures incorporating the Data management plan template
• Continue to address the required cultural change
Acknowledgements

The rest of CSIRO Research Data Support Team past and present:

- Janet Applegate
- Anna Borg
- Anne Stevenson
- Dominic Hogan
- Kathryn Holt
- Tricia Kelly
- Meredith Hepburn
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 Industry
Innovation, Science, Research and Tertiary Education
Image credits:


Thank you
What we are building: DAP technical

- Java web application running under Java 6 on Tomcat 6.
- makes use of an Oracle database but uses the Hibernate library to avoid any specific dependency on the database being Oracle.
- uses a Fedora Commons repository for XML storage.
- SOLR is used to power the searches.
- successfully adapted for use as our publications repository
- should be adaptable to any SOLR indexed collection source.
- currently preparing DAP software for release to open source.
- public records auto harvested to RDA by OAI-PMH