Moving research datasets to the cloud

CAUL Research Repository Community Days
28-29 October 2013

Elke Dawson
Manager, Library Services
CQUniversity Library
ANDS ‘Seeding the Commons’

• Identify, catalogue and provide permanent access to 30 research datasets
• Majority under 50MB (original or zipped) and stored in our repository, ACQUIRE
• About 6 gb-size datasets looking for a permanent home.
Research Data Storage Infrastructure (RDSI) Project

- Initiative of the National Collaborative Research Infrastructure Strategy (NCRIS)
- Funded from the Education Investment Fund under the Super Science (Future Industries) initiative with University of Queensland (UQ) the lead agent.
- Nodes to provide significantly increased storage capacity for research data holdings;
- Two staff at each node to support technical infrastructure and ingest of data.
8 nodes in total
7 service providers

- INTERSECT
- NCI
- qcif
- iVEC
- eRESEARCH SA
- VicNode
- TPAC
- University of QUT Australia
Queensland Cyber Infrastructure Foundation (QCIF)

• Not-for-profit public established in 2000 under the name QPSF (Queensland Parallel Supercomputing Foundation) becoming QCIF in 2006.

• Its mission is to accelerate and strengthen innovation throughout Queensland research and business through broad-based and effective application of high performance computing, large scale data management and regional, national and international collaboration.

• Six members – University of Queensland, Queensland University of Technology, Griffith University, James Cook University, CQUniversity, and the University of Southern Queensland.

• University of the Sunshine Coast is an associate member.

• Selected in 2012 to operate 2 of the nodes in Brisbane and Townsville.
1. Researcher fills in application form and the collections are assessed against 3 RDSI ReDS criteria -

Is the collection under review:
• of importance as an input to future research;
• relevant to research questions of interest to a variety of researchers;
• accessible, for the purposes of research, subject to appropriate access rules;
• discoverable through registration in discipline/domain registries, or generic registries such as Research Data Australia; and
• an input to research supported by at least one participant organisation in the node.
Is the collection under review:
• observational data that is collected, curated and made available for general researcher use;
• generated data that is developed as an information product to underpin the research of third party researchers;
• aggregated data that is assembled by individuals or research communities in order to make research questions more addressable; and/or
• a mirror of international collections that are held locally, for purposes of improved performance and reduced cost.
Is the collection under review fit for its intended purpose, which may be:

- relevant to a broad user base within or across disciplines and domains;
- readily accessible, with respect to connectivity and integration, recognising any constraints such as legal, licensing or ethical requirements;
- in reusable forms, well described and documented, to support long-term access and management;
- citable through a mechanism such as a digital object identifier; and
- of sufficiently high quality and fit for purpose, as indicated by domain-relevant measures such as completeness, coverage, and resolution.
2. Applications are sent to the Resource Allocation Committee (RAC) for approval as a merit based collection or for collection development.

3. RAC approves, rejects or refers application back for more detail (eResearch Analyst will be in touch if application is rejected or referred back)

4. QCIF aims to have requests approved and allocated within three to four weeks.

5. ReDS storage is provisioned for ReDS-approved data collections.
CQUniversity approved datasets

- Coral community data of the Wide Bay - Burnett coastal fringing reefs main study - https://widebaycoral.qcloud.qcif.edu.au

- Historical coastlines (community perspectives) : manuscript and images archive - https://historcoast.qcloud.qcif.edu.au

- Coral community data of coastal fringing reefs : pilot study - https://coastalreef.qcloud.qcif.edu.au

- Plants of Central Queensland images archive - https://plantscq.qcloud.qcif.edu.au
### BMRG Coral Mapping Dataset

#### Index of /cqu6918q0018/BMRG coral Mapping data

<table>
<thead>
<tr>
<th>Name</th>
<th>Last modified</th>
<th>Size</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Parent Directory</td>
<td></td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Aerial photography/</td>
<td>09-Aug-2013 13:19</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>BMRG Coral Mapping Area.kml</td>
<td>23-Jun-2011 15:49</td>
<td>1.6K</td>
<td></td>
</tr>
<tr>
<td>Coral Mapping report.pdf</td>
<td>28-Jun-2011 17:20</td>
<td>1.4M</td>
<td></td>
</tr>
<tr>
<td>Maps/</td>
<td>09-Aug-2013 13:19</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Photos/</td>
<td>09-Aug-2013 13:25</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Quickbird natural colour images/</td>
<td>09-Aug-2013 13:36</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Spatial Data/</td>
<td>09-Aug-2013 13:36</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Thumbs.db</td>
<td>19-Aug-2011 12:35</td>
<td>15K</td>
<td></td>
</tr>
</tbody>
</table>

Apache/2.2.15 (CentOS) Server at widebaycoral.qcloud.qcif.edu.au Port 443
Lessons learnt

- There were a few “teething issues”, but things have improved since we first started and RDSI and QCloud have learnt a lot from the “initial” tests.
- Used “rsync - http://en.wikipedia.org/wiki/Rsync” which allows users to “restart a copy” if it stops, so that it doesn’t have to start the copying all over again.
- Tested new “file transfer” tool, aspera for very large datasets which performed amazingly fast.
- RDSI is looking to license this software, so that it will greatly improve the process of transferring data much easier and faster (over 100 times performance increase).
The Future

- Looking for even larger data sets.
- Web front end to the datasets.
- Add links to datasets from our IR.
- QCloud to be upgraded, with a significant increase in storage and CPU’s to occur at the end of this year.
Thank you

e.dawson@cqu.edu.au
07 4930 6838
www.library.cqu.edu.au