Research Data Repositories
review, gaps, & minimum requirements

Presented by: Amber Leahey, Scholars Portal, Ontario Council of University Libraries

Co-authors: Research Data Canada, Standards and Interoperability Committee (RDC-SINC) http://www.rdc-drc.ca/our-work/standards-interoperability/
Research Data Canada, Standards & Interoperability Committee (RDC-SINC)

- Formed in 2013;
- Brings together data people from government, research, and libraries/archives;
- Identify and promote the use of standards to support the preservation and reuse of research data.

http://www.rdc-drc.ca/our-work/standards-interoperability/
1. Research Data Repositories Review

What is the state of the research data sharing and repositories landscape?
“the potential of data-intensive research is progressively and rapidly outstripping our ability to manage and to grow the digital ecosystem to meet 21st century needs” (GoC 2013).
Research questions we asked

◎ Is such an infrastructure(s) currently in place?
◎ What standards exist?
◎ How can we develop best practice criteria to assist with repository selection by researchers, managers, librarians, etc.?
Criteria for assessment

Developed criteria using digital preservation assessment tools:

- Data Seal of Approval (DSA)
- Trustworthy Repositories Audit & Certification (TRAC)
- Others (see references)

### TABLE 2. Features checklist used to compare 32 online data platforms

<table>
<thead>
<tr>
<th>Category</th>
<th>Sub-category</th>
<th>Detailed features</th>
</tr>
</thead>
</table>
| Hardware & Infrastructure         | Server (server resources, platforms etc.)         | ○ Cloud (i.e. Amazon S3)  
○ Dspace  
○ Local - IBM server and storage, VMware ESXi virtualized redundant server farm  
○ Other |
| Cost                              |                                                   | ○ Free to access, download, and deposit data  
○ Free to access, but contribution suggested or required for deposit, i.e. funding structure for access/deposit beyond the threshold  
○ Publishing charge  
○ Formal agreement with research/monitoring program for funding to support archiving and serving the datasets |
| Size                              |                                                   | ○ Size of repository (number of files, datasets) |
| Description                       | Domain                                           | ○ Multidisciplinary  
○ Earth & environmental science  
○ Medical & life sciences  
○ Social Sciences (Economics, Sociology, Political Science, etc.)  
○ Physics  
○ Biological and Life Sciences |
| Redundancy                        |                                                   | ○ Multiple redundant copies  
○ CLOCKSS - Geographically and geopolitically distributed network of redundant archive nodes |
| Preservation                      | Persistent identifiers                            | ○ DOI (specify where possible)  
○ DSpace Handle (HDL)  
○ Other persistent IDs  
○ Other unique resource identifiers (i.e. URIs) (not persistent)  
○ EZID registration management or other persistent identifier registration |
|                                   | Persistent data deposit                           | ○ Long Term preservation of data |
|                                   | Curation                                          | ○ Data curation (specify where possible) |
| Privacy & Security                | Security                                          | ○ Authentication mechanisms  
○ Distinction between public and private data |
|                                   | Author identifier                                 | ○ ORCID ID  
○ SCOPUS ID  
○ Digital Author Identifier |
| Archiving                         | Timestamping and version control                 | ○ Timestamped upon upload  
○ Data can be edited following upload  
○ Version statement  
○ Universal Numeric Fingerprint (UNF) |
|                                   | Citation and references                           | ○ Citation provided (specify format) |
| Submission                        | Data types accepted [list exceptions per]        | ○ Datasets  
○ Metadata (supported upload of exchange formats (XML))  
○ Computer code |
Results

What did we find out?
Repository Models ($ vs free)

Percentage of repositories that are "free" vs. "restricted use ($)" (n=30)

- Restricted access ($) 25%
- Free up to an initial point 68.8%
- Requires affiliation 6.3%

Percentage of repositories that charge for publishing data ($) (n=30)

- 25%
- 75%
Percentage of repositories that are using persistent identifiers (n=30)

- Unique IDs/local system: 40.6%
- Persistent Identifiers (URNs, DSpace Handle, DOI): 59.4%

Percentage of repositories that are using the Digital Object Identifier (DOI) for study or data set level persistent identification (n=30)

- Yes: 53.1%
- No/NA: 46.9%
Preservation options

Percentage of repositories that offer users long-term preservation of their data (n=30)
- Long term deposit: 40%
- No mentions: 56.7%
- Soon: 3.1%

Percentage of repositories that are storing multiple redundant copies (n=30)
- Yes: 43.8%
- No/NA: 53.1%
- Soon: 3.1%
Percentage of repositories that use distributed archiving such as LOCKSS/CLOCKSS (n=14)

- LOCKSS/CLOCKSS: 22.2%
- Other/NA: 77.8%

Percentage of repositories that timestamp data sets/studies at the point of submission (n=30)

- Timestamped: 71.9%
- None/NA: 28.1%
Percentage of repositories that version datasets (n=30)

- No versioning/metadata available/NA: 59.4%
- Versioning: 40.6%
Other findings - Metadata

- Prevalence of local or custom metadata for describing datasets
- Supported metadata standards included:
  - Dublin Core
  - DataCite
  - ISO 19115
  - DDI
- Very little support for file-level, data-level description
Major gaps

**Metadata**
While there does appear to be support for standard metadata, this varies across platforms.

Granularity is a concern, especially as it impacts reuse.

**Versioning**
Few repositories support standard version control and persistent identification of versions.

**Curation**
Few repositories provided detail about curation activities.

**Preservation & succession planning**
Preservation and succession planning documentation was limited. Very repositories made these activities known and openly available. Only 2 repositories had stated succession plans.
What does this mean for RDM and services?

Without existing standards / compliance for infrastructure and repository platforms…

- Build understanding in light of the complexities and heterogeneity;
  - Example - re3data.org etc.
  - National tools and guidance

- Develop criteria and guidance specifically for data repository selection
Certification

Percentage of repositories that are certified (DSA, ICSU WDS etc.) (n=32)

- 93.8% no
- 6.3% yes
Datasets - Whoa! That’s a big number, aren’t you proud?
Next Steps
In progress

- Guidelines for the deposit of research data in Canada, 2015
- List of Standards for research data and repository interoperability
- Guidelines and criteria for the selection of repositories for data deposit and sharing
Thanks!

Any questions?

You can find me at:

amber.leahey@utoronto.ca