Across Canada, across disciplines: Research data management practices and needs in the social sciences and humanities

Leanne Trimble¹, Dylanne Dearborn¹, Tatiana Zaraiskaya², Jane Burpee³, Eugene Barsky⁴, Catie Sahadath⁵, Melissa Cheung⁵, Marjorie Mitchell⁴, Matthew Gertler¹

¹University of Toronto, ²Queen's University, ³McGill University, ⁴University of British Columbia, ⁵University of Ottawa

IASSIST 2017
Canadian RDM Survey Consortium

Background

Several Canadian universities have committed to working together, using a common survey instrument to gather information about their respective researcher communities and generate a richer understanding of their users' RDM practices and attitudes.

Faced with looming changes in Canadian funding requirements around data sharing, data preservation and the submission of data management plans, librarians at these institutions are collaborating to better understand both national and local needs, as well as to generate a richer understanding of disciplinary practices by generating comparative data for cross analysis.
Who is involved?
Methods – data collection and analysis
Demographics
Respondents by institution

- Queens University: 38.4%
- University of Toronto: 23.3%
- University of Ottawa: 23.1%
- University of British Columbia: 11.7%
- University of Waterloo: 3.4%

n=437
Breakdown by discipline

- Social Sciences: 41.6%
- Arts/Humanities: 34.1%
- Business/Management: 11.9%
- Education: 7.8%
- Law: 3.4%
- Other: 1.1%

n=437
Respondents by rank

- Assistant/Associate/Full Professor: 73.5%
- Adjunct Professor: 11.7%
- Professor Emeritus: 5.3%
- Postdoctoral Fellow: 3.7%
- Lecturer: 3.7%
- Other: 2.3%

n=437
Working with Research Data
Storage volume, by number of research projects

- >4TB: 3 2
- 1TB to 4TB: 3 9 4
- 500GB to <1000GB: 10 5 3
- 50GB to <500GB: 19 11 2
- 10GB to <50GB: 30 11 5 3
- 1GB to <10GB: 47 33 9 13
- <1GB: 44 21 6 5
- Not sure: 41 20 5 24
- Not applicable: 2 26

Number of Respondents
Types of research data generated

- Text: 79.6%
- Numerical: 42.8%
- Multimedia: 40.5%
- Models: 14.9%
- Software: 14.0%
- Other: 8.2%
- Geospatial: 6.4%
- Instrument-specific: 3.0%
Storage media used

- Computer hard drive: 61.8%
- Laptop hard drive: 60.0%
- External hard drive: 58.1%
- Flash drive/USB: 50.3%
- Cloud/web-based solution: 47.6%
- Physical copy retained: 31.8%
- Hard drive of instrument/sensor: 25.4%
- CD/DVD: 11.7%
- External data repository: 5.5%
- Other: 3.7%
- Grid/high performance computing: 1.4%
- Not sure: 1.4%
Documentation & description of data

Is there sufficient documentation and description (e.g., variable and field definitions, codebooks, data dictionaries, metadata, scripts to run) for another person outside your research team to:

- **to understand and use the research data**
  - Yes: 30.0%
  - No: 46.7%
  - Not sure: 23.2%
  - n=413

- **to replicate the methodologies that produced the data**
  - Yes: 33.3%
  - No: 42.7%
  - Not sure: 24.0%
  - n=412
How long data is kept?

- Source/raw data (n=410)
- Intermediate/working data (n=403)
- Processed data (n=402)

<table>
<thead>
<tr>
<th>Duration</th>
<th>Source/raw data</th>
<th>Intermediate/working data</th>
<th>Processed data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Length of project</td>
<td>5.1%</td>
<td>3.0%</td>
<td>5.2%</td>
</tr>
<tr>
<td>Less than 3 years</td>
<td>2.0%</td>
<td>1.0%</td>
<td>2.0%</td>
</tr>
<tr>
<td>3-5 years</td>
<td>10.0%</td>
<td>9.0%</td>
<td>11.4%</td>
</tr>
<tr>
<td>5-10 years</td>
<td>22.2%</td>
<td>23.6%</td>
<td>19.9%</td>
</tr>
<tr>
<td>More than 10 years</td>
<td>16.8%</td>
<td>14.1%</td>
<td>18.9%</td>
</tr>
<tr>
<td>Until the data has become inaccessible or lost</td>
<td>43.9%</td>
<td>48.3%</td>
<td>38.2%</td>
</tr>
</tbody>
</table>
Sharing Research Data
Current sharing methods vs Future sharing methods
Sharing restrictions

- “Privacy or ethics restrictions” (37.8%) was the most identified restriction
  - Highest in Education (82.4%) and the Social Sciences (SS) (45.1%)

- “Needing to apply for a patent” was only selected by one researcher (SS)

- Low number of responses identifying “commercial concerns”
  - Only Business/Management (15.4%) and Law (6.7%) had higher response rates

- Low across board for “public safety/sensitive data” (all <=6%)
Perceived benefits & Reasons for not sharing

- **Perceived benefits**
  - Retains data integrity: 131
  - Re-analysis helps verify results: 165
  - Supports open access knowledge: 224
  - Reduces redundant data collection: 188
  - Moves field of research forward: 214
  - Interdisciplinary research: 188
  - Collaborative scholarship: 147
  - Increase research impact: 192
  - Training new researchers: 228
  - Safeguards against falsification: 145
  - Other: 24
  - None: 52

- **Reasons for not sharing**
  - I am willing to share them: 250
  - Other: 100
    - nproper citation or acknowledgement: 95
    - Privacy, legal or security issues: 116
    - Not useful to others: 72
    - No place to put them: 72
    - Lack of funding: 52
    - Lack of standards: 47
    - Insufficient time: 37
    - Did not know I could: 12
    - Believe they should not be shared: 30
    - Not required by funding body: 12
    - Do not hold the rights: 79
    - Lacking technical skills or knowledge: 29
    - Still wish to derive value: 138
    - Data are incomplete: 124
Funding Mandates & RDM Services
Drafting a data management plan

- 51.5%: I would be able to draft a DMP without assistance
- 36.1%: I would be able to draft a DMP, but would prefer to have assistance/guided documentation
- 12.4%: I would need assistance and/or guided documentation to appropriately address some or all of the sections

n=396
Level of interest in services from libraries

- Interest in assistance with DMP preparation (n=384): 77.1%
- Interest in faculty workshops (n=391): 76.0%
- Interest in personalized consultation (n=388): 72.4%
- Interest in communication/info about funding/journal requirements (n=384): 71.6%
- Interest in data storage during active projects (n=381): 70.6%
- Interest in institutional repository (n=384): 69.6%
- Interest in assistance w/ preservation/sharing (n=382): 68.6%
- Interest in finding/accessing data sources (n=374): 65.0%
- Interest in external repositories (n=382): 62.8%
- Interest in assistance w/ metadata creation (n=383): 58.7%
- Interest in digitization of physical records (n=378): 55.6%
- Interest in grad student workshops (n=380): 54.5%
- Interest in permanent identifiers/DOIs (n=381): 54.1%
Digital Humanities/
Digital Scholarship
*Digital Humanities or Digital Scholarship (DS)*, can be defined as the collection and use of digital research data (either through digitization of print resources, or using born-digital resources) combined with methodologies from traditional Humanities and Social Science scholarship.

**Do you feel your research falls under this definition?**

- Yes: 41% (n=176)
- No: 39% (n=165)
- Not Sure: 20% (n=85)
Nature of Digital Scholars’ Data

Type of data generated in a typical research project

- Instrument Specific: 3%
- Other: 6%
- Geospatial: 10%
- Software: 17%
- Models: 20%
- Numerical: 45%
- Multimedia: 59%
- Text: 88%

Data storage required for average research project

- Not applicable: 2%
- Not sure: 2%
- >4TB: 4%
- 1TB to 4TB: 7%
- 500GB to <1000GB: 8%
- 50GB to <500GB: 9%
- 10GB to <50GB: 17%
- 1GB to <10GB: 23%
- <1GB: 20%
Data topics taught in digital scholarship

- Data Version Control: 7%
- Data Sharing: 11%
- Data Retention: 11%
- Data Archiving: 12%
- Data Security: 12%
- Data Backup: 14%
- Data Documentation: 16%
- Data Privacy: 18%
- Data Ethics: 24%
- None: 70%
Summary: Key findings

- Data being produced:
  - Most commonly text; small storage sizes (i.e. requirements are not complex)

- Knowledge gaps:
  - Data storage: range of non-optimal storage options in use
  - Over half were not confident about the quality of their documentation

- Sharing:
  - Plan to share more in the future than they do now
  - Privacy/ethics most common reason for not sharing
  - Strong awareness of the benefits of sharing

- Interest in support from libraries:
  - 87% would like/need some support in drafting DMPs
  - Large proportion of respondents expressed interest in services from the library in general

- Some variance in disciplinary responses - useful for targeting services
Lessons learned

● Conducting a survey with both local and consortial goals can be a challenge
  ○ Survey design must consider local and group needs
  ○ Group data management planning
  ○ Wrangling data from different institutions
  ○ Communication with partners is key

● Portage clearinghouse aims to improve this process for additional institutions joining the project
Future steps...

Engineering & Science

Humanities & Social Sciences

Health & Medical Sciences

Survey templates, data and reports will be made available on the Portage website
How the data can be used

- Informing institutional decisions, e.g. shaping services
- Informing national initiatives, e.g. Portage, funding agencies

Data, templates and reports available: 
https://goo.gl/VD5hTY

Hosted by Portage, https://portagenetwork.ca/
Acknowledgements – data contributors

- Queen’s University
  - Alexandra Cooper
- University of British Columbia
  - Sheryl Adam, Megan Meredith-Lobay
- University of Ottawa
  - Jessica McEwan, Patrick Labelle
- University of Toronto
  - Leslie Barnes, Nadia Muhe
- University of Waterloo
  - Kathy Szigeti, Sandra Keys
Questions?