The Long Tail of Data
Wagging the Institutional Repository

Open Repositories 2013

Chuck Humphrey
University of Alberta
Is everything that is digital also data?

There is digital content that has research potential but is not research data.

Research data are the products that provide evidence in the research process.
Various stages of development

- We are all at various stages of dealing with research data and our repositories
  - Some may not yet deal with research data in their repository but are now investigating how to go about incorporating research data into their digital collections
  - Some may have started ingesting research data and are now looking at their next steps in this area
  - Some may have well established research data collections and are looking at ways to collaborate with other repositories or at how to fit into the emerging global research data ecosystem
The challenges of research data

- The heterogeneous nature of research data brings challenges to repositories in the following areas:
  - Policy foundation
  - Extent of processing for ingest and the resulting workflow
  - Formats
  - The design of the Archival Information Package
  - Identity of the repository
  - Skilled professionals
Two major environmental drivers

- Two significant sources of motivation for increased organizational interest in better managing research data
  - e-Science movement, now generalized to e-Research and expanded by Jim Gray’s Fourth Paradigm argument
  - Academic integrity and interests in the replication of research findings

- These drivers carry different expectations
  - Collection versus product
  - Interoperable versus reproducible
Data and a policy foundation

- Assumption: a set of policies exists based on the TRAC checklist or on an adaptation of the OAIS Reference Model
- Policies will likely need to be modified to support research data, especially if just getting into data
  - This is illustrated in the next slide, which shows the policy documents framework from the report by the Canadian Polar Data Network to the Canadian High Arctic Research Station on scientific and technological research data management infrastructure
Data policy document framework
Data policy, procedures & guidelines

1. Research Policy
2. Research Records Stewardship Guidance Procedure
3. Privacy Breach Procedure
4. Research Records Risk Classification Procedure
5. Research Data Management and Preservation Guidelines
Data policy document framework

CPDN

Governance
- Charter
- Memorandum of Understanding
- Business Plan

Operations
- Collection Development
- Data License Agreement
- Canadian IPY Data Policy, 2007-2008
- Technical Specifications
- Archiving & Access
- Succession Plan

Orange and Yellow: Organizational structure
Green: Completed documents
Blue: Documents to be crafted
Extent of processing prior to ingest

- Best practices for preparing data for ingest exist through some well-established domain archives, such as the ICPSR and UK Data Archive
  - Policy should guide whether research data get additional processing prior to ingest
    - Decision to accept “as is” or to do additional processing
  - Processing are steps needed to ensure completeness of documentation and data files, to screen for sensitive information, to conduct quality evaluations, to assign administrative content, to prepare generalized formats, etc.
- Obstacles
  - Not being able to provide guidance soon enough in lifecycle
  - Not having an adequate data curation toolkit
Research data formats

- This is complicated by the wide range of analytic software formats that researchers and data producers use and by the unusual or inappropriate naming conventions that are employed.

- The next slide shows the long tail of file formats in the repository for the Statistics Canada Data Liberation Initiative (DLI).
  - Files are received “as is” from STC author divisions, except for some of the production of SPSS syntax files to read microdata files.
Research data formats: an example

<table>
<thead>
<tr>
<th>ExtType</th>
<th>N_Rels</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>pdf</td>
<td>29899</td>
<td>46.51%</td>
</tr>
<tr>
<td>zip</td>
<td>7922</td>
<td>12.32%</td>
</tr>
<tr>
<td>exe</td>
<td>6381</td>
<td>9.93%</td>
</tr>
<tr>
<td>ivd</td>
<td>3363</td>
<td>5.23%</td>
</tr>
<tr>
<td>txt</td>
<td>2048</td>
<td>4.59%</td>
</tr>
<tr>
<td>doc</td>
<td>2620</td>
<td>4.08%</td>
</tr>
<tr>
<td>irt</td>
<td>2276</td>
<td>3.54%</td>
</tr>
<tr>
<td>sps</td>
<td>2239</td>
<td>3.48%</td>
</tr>
<tr>
<td>sas</td>
<td>1508</td>
<td>2.35%</td>
</tr>
<tr>
<td>xis</td>
<td>945</td>
<td>1.47%</td>
</tr>
<tr>
<td>tgt</td>
<td>851</td>
<td>1.32%</td>
</tr>
<tr>
<td>wpd</td>
<td>839</td>
<td>1.31%</td>
</tr>
<tr>
<td>xnl</td>
<td>388</td>
<td>0.60%</td>
</tr>
<tr>
<td>ppt</td>
<td>350</td>
<td>0.54%</td>
</tr>
<tr>
<td>htm</td>
<td>222</td>
<td>0.35%</td>
</tr>
<tr>
<td>gz</td>
<td>188</td>
<td>0.31%</td>
</tr>
<tr>
<td>rtf</td>
<td>165</td>
<td>0.26%</td>
</tr>
<tr>
<td>wp</td>
<td>148</td>
<td>0.23%</td>
</tr>
<tr>
<td>cso</td>
<td>93</td>
<td>0.14%</td>
</tr>
<tr>
<td>can</td>
<td>58</td>
<td>0.09%</td>
</tr>
<tr>
<td>dl</td>
<td>56</td>
<td>0.09%</td>
</tr>
<tr>
<td>rl</td>
<td>51</td>
<td>0.08%</td>
</tr>
<tr>
<td>dl</td>
<td>50</td>
<td>0.08%</td>
</tr>
<tr>
<td>gxf</td>
<td>46</td>
<td>0.07%</td>
</tr>
<tr>
<td>dat</td>
<td>42</td>
<td>0.07%</td>
</tr>
<tr>
<td>dbf</td>
<td>41</td>
<td>0.06%</td>
</tr>
<tr>
<td>jpg</td>
<td>34</td>
<td>0.05%</td>
</tr>
<tr>
<td>dr</td>
<td>31</td>
<td>0.05%</td>
</tr>
<tr>
<td>dwt</td>
<td>28</td>
<td>0.04%</td>
</tr>
<tr>
<td>tpi</td>
<td>26</td>
<td>0.04%</td>
</tr>
<tr>
<td>cvo</td>
<td>26</td>
<td>0.04%</td>
</tr>
<tr>
<td>cat</td>
<td>23</td>
<td>0.04%</td>
</tr>
<tr>
<td>fra</td>
<td>22</td>
<td>0.03%</td>
</tr>
<tr>
<td>tpc</td>
<td>18</td>
<td>0.03%</td>
</tr>
<tr>
<td>spc</td>
<td>18</td>
<td>0.03%</td>
</tr>
<tr>
<td>hip</td>
<td>18</td>
<td>0.03%</td>
</tr>
<tr>
<td>xss</td>
<td>17</td>
<td>0.03%</td>
</tr>
<tr>
<td>png</td>
<td>16</td>
<td>0.02%</td>
</tr>
</tbody>
</table>
The AIP for research data

- Research data require thought as to the design of the Archival Information Package (AIP), that is, to the digital object produced from the Submission Information Package (SIP) and that is placed in archival storage.
- Because the context in which research data are produced is vitally important for others to understand the data, efforts are made to document as much context as possible.
  - Integrating contextual information with the research data needs to be considered in the AIP design.
  - Multiple data files that are related also need thought.
AIP design for research data: example
AIP data workflow: example
Repository: as a brand

- Do you think of your repository as a digital collection or as a platform?
- Quaecumque Vera: whatsoever things are true
- ERA: Educational and Research Archive
  - ERA-data
  - ERA-theses
  - ERA-text
- Collection distinctions should help direct decisions around repository infrastructure and services
- The next slide is an example of a mixed infrastructure model to support the data repository for the Canadian Polar Data Network
Mixed infrastructure model for data
Research data curation expertise

- Build a team environment for data curation
  - While not a sustainable solution, strategically select research projects to serve as an embedded data curator
- Within the Library, develop a team of experts
  - U of Alberta example: Digital initiatives coordinator, Preservation officer, Digital initiatives technology librarian, Digital initiatives applications librarian, Institutional repository librarian, Metadata librarians, GIS librarian, Data library coordinator, Data curator intern
  - Develop liaison librarian roles (mainstreaming research data)
  - Capitalize on Co-op, Intern, and Post-doc data curators
Summary

- The special requirements of research data need to be rooted in a repository’s policy foundation.
- Preparing research data for submission often requires additional processing, degrees of intervention, mediation, best practices in data management, policy support, and a data curation toolkit.
- The variety of formats for research data requires a community effort to manage.
- The design of the AIP is important in creating sound digital objects for research data (don’t defer to technology on this point!)
- Research data should have its own identity in a repository of mixed digital content.
- Build data curation teams with complementary expertise.