Research data in Dryad: Balancing preservation and re-use

Elizabeth Hull
Operations Manager
Dryad Digital Repository

PASIG
March 12, 2015
Dryad mission & vision

- The Dryad Digital Repository is a **curated** resource that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad provides a **general-purpose** home for a wide diversity of datatypes.

- Dryad’s **vision** is to promote a world where research data is openly available, integrated with the scholarly literature, and **routinely re-used to create knowledge**.

- Our **mission** is to provide the infrastructure for, and promote the re-use of, data underlying the scholarly literature.

- All data submitted to Dryad **must be associated with a publication**.


Dryad’s preservation challenges

“The rate at which the research world has become digital — in fact much earlier and faster than many other sectors — poses a threat to the longevity and retrievability of scientific information.

Science depends entirely on the knowledge gained in the past to further advance.”

  http://dx.doi.org/10.1087/20110107
Dryad’s preservation challenges

- Focus on stable formats that are easy to migrate clashes with need for data to remain useful

- General-purpose by design: no file format restrictions, wide diversity of file types - “long tail” data sets

- Common digital preservation tools like PRONOM and JHOVE don’t recognize many of our formats

- Specialized user base with *really* proprietary formats
Dryad’s preservation challenges

Data from: Baby on board: olfactory cues indicate pregnancy and fetal sex in a non-human primate

Files in this package

Content in the Dryad Digital Repository is offered "as is." By downloading files, you agree to the Dryad Terms of Service. To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data.

<table>
<thead>
<tr>
<th>Title</th>
<th>GCMS Files</th>
</tr>
</thead>
<tbody>
<tr>
<td>Downloaded</td>
<td>3 times</td>
</tr>
<tr>
<td>Description</td>
<td>GCMS files generated using Shimadzu QP2010 GC/MS. Files are compressed in RAR format. 'PC' corresponds to preconceptive; 'Preg' corresponds to pregnant.</td>
</tr>
<tr>
<td>Download</td>
<td>GCMSfiles.rar (72.97Mb)</td>
</tr>
<tr>
<td>Details</td>
<td>View File Details</td>
</tr>
</tbody>
</table>

http://dx.doi.org/10.5061/dryad.ps70b
Dryad’s preservation challenges

Example of a GC-MS instrument

Source: By Polimerek (own work), via Wikimedia Commons
Where we are now

“Forever – or 5 years, whichever comes first”
Where we are now

Step 1: admitting you have a problem.
Figure 7. Where do researchers store data?, n = 1,202
Source: PARSE.insight survey report.
Where we are now

Keeping the submission barriers low

- Permissive definition of “data”
- Proprietary = okay
- Good documentation = encouraged, not required
- Metadata schema = general
Where we are now

File types in Dryad
Where we are now

Quality control measures in place

- Encourage non-proprietary & open formats
- Link to the publication provides context
- Built-in checks: data review option, community policing, versioning
- Preservation Policy
The future: where we’d like to be

Moving towards a more holistic preservation strategy

- Use copies and preservation copies?
- Automatic checks, periodic analyses, improved metadata quality
- More hands-on curation/collaboration with researchers
- Crowdsourcing/user feedback mechanisms
- Taking a more active voice in digital preservation community
- Partnerships with DataONE and others
Organizational preservation

“Ensuring long-term preservation of research data requires financial sustainability...Support must come from the stakeholders that already support the business of scholarly publication: authors, scientific societies, research libraries, and academic institutions.”

“In short, if we are serious about seeing research data preserved and reused, we must be willing to support the enterprise.”

  http://dx.doi.org/10.1525/bio.2010.60.5.2

*and publishers, and funders
Learn more

- Repository home:  http://datadryad.org
- Documentation:  http://wiki.datadryad.org
- News:  http://blog.datadryad.org
- @datadryad
- /DataDryad

- Elizabeth Hull, Operations Manager -  ehull@datadryad.org
- Erin Clary, Senior Curator -  eclary@datadryad.org
- Christine Mayo, Curator -  kevral@live.unc.edu