Creating a sustainable business model for a digital repository: the Dryad experience

"Cherish old knowledge so that you may acquire new"
- The Analects of Confucius

Peggy Schaeffer
Research Data Access and Preservation (RDAP) summit
March 22, 2012
• **The End**
  – To make data archiving and reuse standard within scientific communication.

• **The Means**
  – Enable low-burden data archiving at the time of manuscript submission.
  – Promote researcher benefits from data archiving.
  – Promote responsible data reuse.
  – Empower journals, societies & publishers in shared governance.
  – Ensure sustainability and long-term preservation.

• **The Scope**
  – Research data in the basic and applied biosciences, broadly defined
  – Primarily data underlying findings in peer-reviewed articles
  – Also data from some non-peer reviewed publications (e.g. dissertations)
  – And some non-data content (e.g. software scripts, figures)
The value proposition

- For researchers, Dryad...
  - increases the impact of, and citations to, published research
  - preserves and makes available others’ data
  - frees researchers from the burden of data preservation and access
- For journals, publishers and societies, Dryad...
  - frees journals from the burden of maintaining supplemental data
- For libraries and institutions, Dryad...
  - makes data available at no cost, under clear terms of use
  - helps fulfill their research data management mandates
- For funders, Dryad...
  - provides a cost-effective mechanism to make research more accessible
<table>
<thead>
<tr>
<th>Year</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>2007</td>
<td>NSF/ESA Data Sharing and NESCent Small Science workshops&lt;br&gt;Beginning negotiation of Joint Data Archiving Policy (JDAP)</td>
</tr>
<tr>
<td>2008</td>
<td>Journals/societies join NESCent &amp; others to fund Dryad through NSF&lt;br&gt;Initial NSF funding for Dryad begins (through 2012)</td>
</tr>
<tr>
<td>2009</td>
<td>Repository went online&lt;br&gt;First consortium board meetings&lt;br&gt;Initial development of sustainability and revenue plans&lt;br&gt;Debut of integrated data submission</td>
</tr>
<tr>
<td>2010</td>
<td>Announcement of Joint Data Archiving Plan&lt;br&gt;JISC funding begins&lt;br&gt;Discussions with potential charter partners</td>
</tr>
<tr>
<td>2011</td>
<td>JDAP (and NSF DMP mandate) takes effect</td>
</tr>
<tr>
<td>2012</td>
<td>New funding from NSF begins (through 2016)&lt;br&gt;Approval of cost-recovery plan and governance structure&lt;br&gt;Cost-recovery begins&lt;br&gt;Transitional funding begins</td>
</tr>
</tbody>
</table>
Dryad integrates article and data submission

• Why?
  – Ensures permanent link to data within each article (and vice versa).
  – Makes data deposition fast and easy for authors (once files are prepared)

• Options are customized to meet the requirements of individual journals
  – Submission of data prior to manuscript review OR upon acceptance
  – Whether authors have the option of a 1-yr no-questions asked embargo OR not*

*By default, data is released upon article publication, and editors can permit longer embargoes for special cases
Over 20 integrated partner journals

.. and more being added regularly

The American Naturalist
BMJ Open
Biological Journal of the Linnean Society
Ecological Monographs
Evolutionary Applications
Evolution
Heredity
Journal of Evolutionary Biology
Journal of Fish and Wildlife Management
Journal of Heredity
Journal of Paleontology
Molecular Ecology and M.E. Resources
Paleobiology
PLoS Biology
Systematic Biology
ZooKeys & 7 other Pensoft journals
Researchers are using Dryad for archiving

Dryad is an international repository of data underlying peer-reviewed articles in the basic and applied biosciences. Dryad enables scientists to validate published findings, explore new analysis methodologies, repurpose data for research questions unanticipated by the original authors, and perform synthetic studies. Dryad is governed by a consortium of journals that collaboratively promote data archiving and ensure the sustainability of the repository.

As of Mar 21, 2012, Dryad contains 1,456 data packages and 3,800 data files, associated with articles in 117 journals.
And using the data for research and education
Dryad principles & priorities

• Enable data archiving as an extension of traditional publication
• Assert the value-added benefits of
  – Citable data (for depositors)
  – Economies of scale (for journals and publishers)
  – Professional curation (for users of data)
  – Long-term preservation of data (for all)
• Align incentives with the business model
  – Lower costs for partner journals
• Ensure international participation
• Understand the impact of data citation
• Understand the true costs of hosting supplementary data
Sustainability planning

• Long-term preservation requires an organization with a viable business model
  – Not one dependent on the success of future grant proposals.
• Goal: a business model based on the added value of repository to stakeholders:
  – Depositors of data
  – Users of archived data
  – Journals, publishers, societies
  – Universities, research institutions, and libraries
  – Funding organizations
Assumptions (2009)

• Institutional support: host provides efficiencies (accounting, contracts & grants, legal, shared staff, IT network, facilities)

• Hardware and storage costs decline faster than repository growth

• Curation effort
  – is primary staff expense
  – scales with level of curation tasks and volume
Potential sources of revenue (2009)

• an archiving charge (similar to a page charge)
• pay-per-use, or individual subscriptions, for access to repository contents *(never seriously considered)*
• institutional subscriptions (possibly for higher service levels?)
• subscriptions from societies and journals (possibly in return for full partnership benefits?)
• fees from publishers
• recovery of cost from archiving of large data packages
• grants from government funding agencies across the globe as well as private foundations
• angel donors
Two consultancies (2009)

• Cost model (Lorraine Richards)
  – Examined current literature & environment
  – Developed list of potential exemplar repositories
  – Interviewed Dryad staff
  – Identified relevant cost categories & assumptions
  – Made best estimates

• Broad sustainability plan (Charles Beagrie, Ltd.)
  – Strategy, performance indicators and measures
  – Comparators and understanding of the costs
  – Advantages, benefits and revenue options
  – Drafted a proposal for sustainability
Development of cost model

• Based on JISC *Keeping Research Data Safe*2
• Total and per paper costs estimated
• Per paper cost estimates, by volume
  – 5,000 papers per yr = $40 (approximate)
  – 10,000 papers per yr = $32 (approximate)

Cost categories:
  – Repository management
  – **Curation**
  – Storage and hardware
  – Outsourcing
  – Infrastructure, facilities, & administration
  – R&D
  – Maintenance
  – Outreach and promotion
  – Documentation
<table>
<thead>
<tr>
<th>Curation level</th>
<th>Minutes</th>
<th>Tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (Low)</td>
<td>5</td>
<td>Verify that the DOI points to the correct article&lt;br&gt;Spell check&lt;br&gt;Verify that article metadata is correct&lt;br&gt;Verify that data files have expected kind of data</td>
</tr>
<tr>
<td>2 (Medium)</td>
<td>20</td>
<td>Expand keywords based on submitted metadata&lt;br(Convert data files to preservable formats)&lt;br(Deposit additional supplemental data at publisher site)&lt;br(Create/approve relationships to content in partner repositories)&lt;br(Approve updates submitted by the author)&lt;br(View the contents of metadata fields across the repository, and enforce consistency)</td>
</tr>
<tr>
<td>3 (High)</td>
<td>140</td>
<td>• Enter/verify authors in name authority file (LCNAF)&lt;br• Expand keywords based on text of the article&lt;br• Within-file annotations (spreadsheet columns, taxon names in trees)&lt;br• Evaluate comments from end users and relay to the author</td>
</tr>
</tbody>
</table>
50-100 integrated journals
Growing Dryad

• Enlarging repository scope
  – Biomedical data
  – Dissertation data
  – Software & other supplementary materials

• Building journal & publisher connections
  – DryadUK at the British Library, funded by JISC
    • Wiley Blackwell
    • Oxford Univ. Press
    • Nature Publishing Group
    • Elsevier
    • PLoS

• Expanding Consortium membership
Dryad as an organization

• Dryad Consortium, soon to be a 501(c)3 nonprofit
• Membership composed of journal & organization representatives
  – Open to the full spectrum of stakeholder organizations, including scientific societies, publishers, funding agencies, universities & institutes
  – Nominal annual fee - no more than $1000 USD
• Governed by a Board of Directors (12 members)
  – Nominated and elected by the Membership
• Next board meeting July in North Carolina
  – Transition to 501(c)3 status, hosted at Duke Univ.
  – Adopt governance model
  – Adopt cost-recovery model
Dryad’s sustainability model

• **Deposit fees** are the primary source of revenue, for several reasons:
  – The time of deposit is when the majority of costs are incurred
  – Revenue scales with costs (i.e. volume of deposits)
  – The costs are distributed both fairly and widely
  – This enables Dryad to make access to the data free *in perpetuity*

• **Membership fees** will cover costs of annual Membership meetings

• **Additional revenue**
  – Project grants will supplement the operational budget for R&D activities
## Payment plans (proposed)

<table>
<thead>
<tr>
<th>Plan</th>
<th>Contract?</th>
<th>Paid by</th>
<th>Cost&lt;sup&gt;2&lt;/sup&gt; (approximate)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal subscription</td>
<td>yes</td>
<td>Journal&lt;sup&gt;1&lt;/sup&gt;, in advance</td>
<td>Based on annual volume of research articles ($25-30/article)</td>
</tr>
<tr>
<td>Pre-paid per-deposit</td>
<td>yes</td>
<td>Journal&lt;sup&gt;1&lt;/sup&gt;, in advance</td>
<td>$50-60/data package</td>
</tr>
<tr>
<td>Pay-as-you-go per-deposit</td>
<td>yes</td>
<td>Journal&lt;sup&gt;1&lt;/sup&gt;, invoiced periodically for prior deposits</td>
<td>$60-70/data package</td>
</tr>
<tr>
<td>Individual deposit</td>
<td>no</td>
<td>Author, at time of deposit</td>
<td>$70-80/data package, with a process for granting waivers under development</td>
</tr>
</tbody>
</table>

<sup>1</sup> Or other sponsoring organization  
<sup>2</sup> Up to a fixed deposit size (currently 10GB). Additional charges for larger deposits.
Projections and issues

- Rate of deposit
- High volume journals and publishers
- How long before sustainability achieved?
- Potential for growth
- Enlarged scope?
To learn more

- Repository home: http://datadryad.org
- News: http://blog.datadryad.org
- Project documentation: http://wiki.datadryad.org
- Announce and User mailing lists: http://datadryad.org/about
- Twitter: @datadryad
- Code: http://code.google.com/p/dryad

or contact me: Peggy Schaeffer, pschaeffer@nescent.org
References


The complete list of Dryad publications and presentations is at  
http://wiki.datadryad.org/Publications