Scaling up the coupling STM articles with research data

Todd Vision
University of North Carolina at Chapel Hill
@tjvision

Academic Publishing in Europe, Berlin, Germany, 25 January 2012
Failure of peer-to-peer data sharing

Wicherts and colleagues requested data from 141 articles in American Psychological Association journals.

“6 months later, after … 400 emails, [sending] detailed descriptions of our study aims, approvals of our ethical committee, signed assurances not to share data with others, and even our full resumes…” only 27% of authors complied

The fallibility of published results

- More than half of published papers contain statistical errors.
- 5-10% of papers contain errors that change the conclusions.

Citations to 2007 GEO data

One or more author(s) in common

No author(s) in common

H. Piwowar
Do journal articles publish data?

<table>
<thead>
<tr>
<th>Function</th>
<th>Role of journal</th>
</tr>
</thead>
<tbody>
<tr>
<td>Registration</td>
<td>✓</td>
</tr>
<tr>
<td>Certification</td>
<td>✓ for methods</td>
</tr>
<tr>
<td>Dissemination</td>
<td>✓ limited</td>
</tr>
<tr>
<td>Archiving</td>
<td>✓ limited</td>
</tr>
</tbody>
</table>
Joint Data Archiving Policy (JDAP)

Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future.

As a condition for publication, data supporting the results in the article should be deposited in an appropriate public archive.

Authors may elect to embargo access to the data for a period up to a year after publication.

Exceptions may be granted at the discretion of the editor, especially for sensitive information.

Small science $\rightarrow$ long tail $\rightarrow$ orphan data

Volume

Specialized repositories
(e.g., EMBL)

Orphan data

Rank frequency of datatype

after B. Heidorn
Small science ➔ long tail ➔ orphan data

Specialized repositories (e.g., EMBL)

Orphan data

Dryad is an international repository of data underlying peer-reviewed articles in the basic and applied biosciences. Dryad enables scientists to make their research data available in a usable and interoperable format, to develop analysis methodologies, repurpose data for research questions unanticipated by the original authors, and perform synthetic studies. Dryad collaboratively promote data archiving and ensure the sustainability of the repository.

As of Nov 13, 2011, Dryad contains 1070 data packages and 2539 data files, associated with articles in 93 journals.

Recently Published Data


Delplancke M (2011) Data from: Gene flow among wild and domesticated almond species, insights from chloroplast and nuclear markers. *Genetics* doi:10.5061/dryad.5f41f1q8


Luijckx P, Fienberg H, Duneau D, Ebert D (2011) Data from: Resistance to a bacterial parasite in the Crustacean *Daphnia magna* shows a major QTL near Dup-C. *Genetics* doi:10.5061/dryad.8j1q1m3


When using this data, please cite the original article:

Additionally, please cite the Dryad data package:
**Materials and Methods**

The analyzed alignment file was formatted for and deposited in TreeBASE (http://www.treebase.org/; study accession URL: http://purl.org/phylo/treebase/phylows/study/TB2:S12029; submission 12029) and analysis files were made available through the Dryad data repository (http://datadryad.org/; data package URL: http://dx.doi.org/10.5061/dryad.j1g5dh23; data package DOI: 10.5061/dryad.j1g5dh23; Hodkinson et al. 2011).

**References**


Mitochondrial Capture by a Transmissible Cancer

Cancer metastatic-reassembled virus (CTVT) is a highly adapted cancer, transmitted between cell lines, using the matrix of host mitochondria as a transport system. This unique mechanism of intracellular transport suggests that the cellular mitochondria might be harvested by cancer cells. Indeed, experiments have shown that cancer cells can capture mitochondrial DNA from host cells, leading to the formation of chimeric tumor cells with a mix of mitochondrial genomes. This phenomenon may have significant implications for cancer biology and therapy, as it suggests a new strategy for cancer cell dissemination and immune evasion. Further research is needed to fully understand the mechanisms underlying mitochondrial capture by cancer cells and to explore potential therapeutic targets.

Supporting Online Material
www.sciencemag.org/cgi/content/full/331/6015/303/DC1
Materials and Methods
Table S1
References

13 September 2010; accepted 20 December 2010
10.1126/science.1197696

When using this data, please cite the original article:


Additionally, please cite the Dryad data package:

submit manuscript

manuscript review

accepted?

no

accepted?

yes

send article description

Dryad data package

curation

published data (with article citation)

upload data

send data identifier (DOI)

published article (with data citation)

JOURNAL

prepare manuscript and related data files

author

editor

DRYAD

Peer review

data curator
Integrated manuscript and data submission

- 23 journals integrated to date
- 21 allow a data embargo (91%)
  - Authors choose an embargo for 33% of files
- 10 require submission prior to peer review (43%)
- 6 prohibit pre-publication metadata release (26%)
Integrated manuscript and data submission

• 23 journals integrated to date
• 21 allow a data embargo (91%)
  ▪ Authors choose an embargo for 33% of files
• 10 require submission prior to peer review (43%)
• 6 prohibit pre-publication metadata release (26%)

• 27% of submissions are from non-integrated journals
• Dryad hosts content from 108 journals in total
20 papers from Delsuc and Douzery going back to 2002
What do the data look like?

- Currently, 3,103 data files
- Size
  - Mean number of files: 2.4
  - Mean storage requirement: 9.7 Mb
What do the data look like?

• Currently, 3,103 data files
• Size
  ▪ Mean number of files: 2.4
  ▪ Mean storage requirement: 9.7 Mb
• Format types
  ▪ 60% plain text (unrecognized content standard)
  ▪ 11% Excel
  ▪ 5% PDF
  ▪ 3% Word
## Contrast of Dryad with (typical) SOM

<table>
<thead>
<tr>
<th>Feature</th>
<th>Dryad</th>
<th>SOM</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Discoverable</strong>: indexed and exposed to both web and bibliographic search engines</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td><strong>Identifiable</strong>: DataCite DOIs within articles serve as permanent, resolvable identifiers</td>
<td>✓</td>
<td>x*</td>
</tr>
<tr>
<td><strong>Permanent</strong>: processes in place to promote preservation (incl. format migration)</td>
<td>✓</td>
<td>✓/x**</td>
</tr>
<tr>
<td><strong>Curated</strong>: quality control by both automated processes and human inspection</td>
<td>✓</td>
<td>x*</td>
</tr>
<tr>
<td><strong>Ease of deposit</strong>: streamlined deposit, allowance for large and complex datasets</td>
<td>✓</td>
<td>✓/x**</td>
</tr>
<tr>
<td><strong>Formatted for reuse</strong>: support for non-PDF file formats</td>
<td>✓</td>
<td>✓/x**</td>
</tr>
<tr>
<td><strong>Updatable</strong>: new versions of data files can be added, metadata can be enhanced</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td><strong>Support for embargoes</strong>: can delay release of data in accordance with journal policy</td>
<td>✓</td>
<td>x</td>
</tr>
<tr>
<td><strong>Free reuse</strong>: no paywall, clear terms of reuse (all data released under CC Zero)</td>
<td>✓</td>
<td>✓/x**</td>
</tr>
<tr>
<td><strong>Economy of scale</strong>: cost efficiency from shared infrastructure</td>
<td>✓</td>
<td>✓/x**</td>
</tr>
</tbody>
</table>

* A few publisher SOM sites are exceptions to the general rule
The value proposition

- **For researchers**
  - Dryad increases the impact of, and citations to, published research. It preserves and makes available others’ data to verify their published results, to refine methodologies, and to repurpose. It frees researchers from the burden of data preservation and access.

- **For journals, publishers and societies**
  - Dryad frees journals from the burden of managing supplemental data, allows them to increase the discoverability and impact of their articles, and to increase their value to the community they serve.

- **For funders**
  - Dryad provides a cost-effective mechanism to make research more accessible, and to leverage existing investments in order to enable new science.
Governance: the Dryad Consortium

- Independent nonprofit with a board elected by its membership
  - Members may include any stakeholder organization: journal, society, publisher, funder, research institution
  - Board need not be drawn from membership organizations
- The membership provides an invaluable forum
  - Coordinating data policies
  - Raising awareness of research, best practices, etc.
  - Helping shape Dryad’s business model
  - Ensuring that Dryad’s features satisfy stakeholder’s evolving needs
Revenue model

- **Deposit fees**
  - Pay upfront for both ingest and long-term preservation
  - At 10K data packages/yr, ~ €40 per deposit
  - Non-operational costs covered by membership fees, grants for R&D, etc

- **Payment schemes**
  - Journal subscription
  - Per-deposit fees, covered by journal (pre or post-submission)
  - Author-pays (last resort!)

- **Neutral w.r.t. the business model of journals/publishers**

What is the return on investment?

- A rigorous framework is lacking, but...
- Marginal cost of data archiving is small
  - €40 is ~2% of the publication cost per article
  - €40 is ~0.2% of the costs per article for a research funder (~€20K)

What is the return on investment?

• A rigorous framework is lacking, but…
• Marginal cost of data archiving is small
  ▪ €40 is ~2% of the publication cost per article
  ▪ €40 is ~0.2% of the costs per article for a research funder (~€20K)
• Is the data worth at least 2% of the investment?
  ▪ Tracked all papers ‘citing’ DNA array data submitted to NCBI in 2007
  ▪ >30% of citations actually reuse the data
  ▪ Only counting substantive reuses by different researchers
  ▪ Too early to have statistics for Dryad
  ▪ But reuse frequency is similar for Pangaea

Roles for the researchers
journals
publishers
funders
scientific societies
Roles for the researchers
journals
publishers
funders
scientific societies
• http://datadryad.org
• http://blog.datadryad.org
• http://datadryad.org/wiki
• http://code.google.com/p/dryad
• dryad-users@nescent.org
• @datadryad
• Dryad
Rapid growth in supplemental data

% of published research articles with supplemental data

<table>
<thead>
<tr>
<th>Year</th>
<th>%SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>1999</td>
<td>0</td>
</tr>
<tr>
<td>2000</td>
<td>2</td>
</tr>
<tr>
<td>2001</td>
<td>3</td>
</tr>
<tr>
<td>2002</td>
<td>5</td>
</tr>
<tr>
<td>2003</td>
<td>7</td>
</tr>
<tr>
<td>2004</td>
<td>22</td>
</tr>
<tr>
<td>2005</td>
<td>40</td>
</tr>
<tr>
<td>2006</td>
<td>55</td>
</tr>
<tr>
<td>2007</td>
<td>65</td>
</tr>
<tr>
<td>2008</td>
<td>75</td>
</tr>
<tr>
<td>2009</td>
<td>87</td>
</tr>
</tbody>
</table>
Does your organisation validate the data submitted to you?

If you allow for the submission of digital research data with publications, is it possible for users of that digital research data to link to it?

Do you have preservation arrangements for underlying digital data?

Sustainability

The word "sustainable" is unsustainable.

http://xkcd.com/1007/
Government funding not assured

Published online 18 November 2009 | Nature 462, 258-259 (2009) |
doi:10.1038/462258b

News

Plant genetics database at risk as funds run dry

National Science Foundation to cut support for Arabidopsis resource.

Alison Abbott

The world’s most valued plant database faces extinction because its funding is being phased out by the US National Science Foundation (NSF), and no alternative source is on the horizon.

"This is the wrong way to go," says genomics researcher Ernest Retzel of the National Center for Genome Resources in Santa Fe, New Mexico. "I believe it will set the field back."

The NSF says that it does not have a policy to support long-term, established research-infrastructure projects such as the Arabidopsis Information Resource (TAIR), which maintains a free, open-access database of genetic and molecular-biology data for Arabidopsis thaliana, or thale cress, the widely used model plant. "We..."
# Taxonomy of data archiving benefits

Modified from Beagrie et al. (2009) *Keeping Research Data Safe 2*

## Direct
- Verification of published research
- Preserving accessibility to data
- Allowing reuse and repurposing of data
- Discoverability of data

## Indirect (costs avoided)
- Redundant data collection
- Inefficient legacy data curation
- Burden of sharing-upon-request
- Opportunity cost of science not done

## Near term
- Protection against personnel turnover
- Availability for review and validation

## Long term
- Secure long-term stewardship
- Increased impact per publication

## Private
- Increased citations
- New collaborations
- New research opportunities
- Fulfilling funding mandates

## Public
- More efficient use of research dollars
- Public trust in science
- Educational opportunities
- Improved methodologies
- More informed policy
Brussels Declaration on STM Publishing

• “Raw research data should be made freely available to all researchers. Publishers encourage the public posting of the raw data outputs of research. Sets or sub-sets of data that are submitted with a paper to a journal should wherever possible be made freely accessible to other scholars”