The Dryad Digital Repository: Published data as part of the greater data ecosystem

Todd Vision, Hilmar Lapp
National Evolutionary Synthesis Center (NESCent)

NESCent
Kevin Clarke
Heather Piwowar
Peggy Schaeffer
Ryan Scherle

UNC-CH <MRC>
Sarah Carrier
Elena Feinstein
Jane Greenberg
Hollie White

Kristin Antelman (NCSU)
Bill Michener (UNM / DataONE)
Bill Piel (Yale / TreeBASE)

Funding: NSF, IMLS
PHILOSOPHICAL TRANSACTIONS: GIVING SOME ACCOUNT OF THE PRESENT Undertakings, Studies, and Labours OF THE INGENIOUS IN MANY CONSIDERABLE PARTS OF THE WORLD

Vol. I.
For Anno 1665, and 1666.

In the Savoy,
Printed by T. N. for John Martyn at the Bell, a little without Temple-Bar, and James Allestry in Duck-Lane; Printers to the Royal Society.
Use and reuse of archived data in evolutionary biology

- \( n = 27 \) articles from 5 journals

- Use previously published data: 48%
- Provide supplementary data: 7%
- Provide supplementary materials: 41%
- GenBank submission honored: 100%
Sharing data on request is not effective

• Wicherts et al (2006) requested data from 141 articles in the field of psychology.
  ▪ “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

• In a survey among geneticists by Campbell et al. (2002) the most frequent reason for withholding data was the effort required to share it (80%).
  ▪ 28% were unable confirm others published research because of data withholding.
Archiving at the time of publication is effective

- The point in time when authors are most prepared to archive their data.
  - No opportunity for loss, corruption, etc., of data files
- Publication can be both carrot and stick.
- The “GenBank model” is uniquely successful.
Further incentives to authors

- Increases impact of one’s own work
- A *quid pro quo* for access to others’ data
- Relief from the burden of *ad hoc* data sharing
View articles in this issue online

Data sets and supplementary material for articles this issue can be downloaded here. Files are (typically) in NEXUS, Word, or HTML formats. Note that the authors may also have deposited their data in GenBank and TreeBASE, or have additional data on their own web sites.

Phylogeny of Eunicida (Annelida) and Exploring Data Congruence Using a Partition Addition Bootstrap Alteration (PABA) Approach
Torsten H. Struck, Günter Purschke, Kenneth M. Halanych
doi:10.1080/10635150500354910
“Do you think the data underlying published scientific results should be made publicly accessible?”
- Yes: 395 (95.4%)
- No: 19 (4.6%)

“If yes, do you think journals should require data sharing of their authors, or should it be voluntary?”
- Required: 220 (55.6%)
- Voluntary: 176 (44.4%)
Joint Data Archiving Policy

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

This journal requires, as a condition for publication, that 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 such as human subject data or the location of endangered species.

DOI:10.1086/650340
So where is this “appropriate public archive”? 
Potential archiving solutions

Author-managed websites
   Avoids some of the hazards of informal sharing, but is fragile.

Supplementary materials online
   Publisher provides basic infrastructure, but with low level of service.

Specialized databases (e.g. GenBank, TreeBase)
   Will cover some datatypes well, some not at all; High quality data, but with greater submission burden; May have issues with sustainability.

Or ...
Dryad - A shared public archive

• Functional goals
  ▪ To publish and preserve the data reported in the biological literature.
  ▪ To promote reuse of the data.

• Organizational goals
  ▪ Governance is shared by a consortium of journals.
  ▪ Responsible long-term stewardship.

Dryad - A shared public archive

- Permanent identifiers (DOIs), trackable data citations
- Explicit terms (CCZero) for reuse
- No paywall to access
- Searchable across publishers & repositories
- Metadata enhanced for discoverability
- Support for standard APIs
- Commitment to preservation in perpetuity
- Migration of formats, files updatable
- Support for embargoes

Data from: Angiosperm wood structure: global patterns in vessel anatomy and their relationship to wood density and potential conductivity

When using this data, please cite the original article:


Additionally, please cite the Dryad data package:


Dryad Package Identifier

doi:10.5061/dryad.1138

Dryad Data Files

http://hdl.handle.net/10255/dryad.1139

Abstract

Woody stems comprise a large biological carbon fraction and determine water transport between roots and leaves; their structure and function can influence both carbon and hydrological cycles. While angiosperm wood anatomy and density determine hydraulic conductivity and mechanical strength, little is known about interrelations across many species. We compiled a global dataset comprising two anatomical traits for 3005 woody angiosperms: mean vessel lumen area (A) and number per unit area (N). From these, we calculated vessel lumen fraction (F = N) and size/number ratio (S = /N), a new vessel composition index. We examined extent to which F and S influenced potential sapwood specific stem conductivity (KS) and wood density (D; dry mass/fresh volume). F and S varied essentially independently across angiosperms. Variation in KS was driven primarily by S, and variation in D was virtually unrelated to F and S. Tissue density outside vessel lumens (DN) must predominantly influence D. High S should confer faster KSs but incur greater freeze-thaw embolism risk. F should also affect KS, and both F and DN should influence mechanical strength, capacitance, and construction costs. Improved theory and quantification are needed to better understand ecological costs and benefits of these three distinct dimensions.
Dryad is a digital library

not a traditional bioinformatics database
Repository priorities

1. Discovery
2. Sharing
3. Integration
4. Preservation
Repository priorities

Integration
Sharing
Discovery
Preservation

Dryad’s scope
Low-burden for deposition

1. Author prepares manuscript and related data files.
2. Manuscript is submitted to the journal.
3. Manuscript is reviewed by the editor.
4. If accepted, the manuscript is published, and the author receives the DOI for the published article.
5. Data is uploaded to DRYAD.
6. Data description is sent to DRYAD.
7. Data is packaged.
8. Data is curated.
9. Published data (with article citation) is made available.
DataONE: An Interoperating Consortium

Providing *universal access to data about life on earth and the environment that sustains it*

- engaging the scientist in the data curation process
- supporting the full data life cycle
- encouraging data stewardship and sharing
- promoting best practices
- engaging citizens
- developing domain-agnostic solutions

1. **Build on existing cyberinfrastructure**

2. **Create new cyberinfrastructure**

3. **Support new communities of practice**
Distributed framework

Flexible, scalable, sustainable network of Member Nodes and Coordinating Nodes

Investigator 1..N Toolkit
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 think the Arabidopsis community is a little out of whack," said an NSF official.
Lessons from Dryad (so far)

• The importance of journals in data publication.
• The value of a shared public repository to promotion of data reuse.
• The delicate balance of benefit and burden to data authors.
• The need to break down data silos.
• Achieving long-term data preservation by achieving long-term organizational sustainability.
To learn more:

Blog: http://blog.datadryad.org
Wiki: http://datadryad.org/wiki
Users list: dryad-users@nescent.org
Developers: dryad-dev@googlegroups.com
Code: http://code.google.com/p/dryad

Follow us on Facebook & Twitter