Dryad’s Evolving Proof of Concept and the Metadata Hook

Wolfram Data Summit

September 6, 2012

Jane Greenberg
Professor, School of Info.& Lib.Sci /UNC-CH
Director, Metadata Research Center
Overview

- **PART 1: Dryad**
  - Goals, governance, and workflow
  - Size, growth, and use
- **PART 2: Dryad metadata R&D**
  - Principles and objectives
  - Questions, methods, and findings
- Conclusions
- Q&A
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 Sep 5, 2012, Dryad contains 1969 data packages and 5189 data files, associated with articles in 150 journals.

Today: Dryad contains 1971 data packages and 5193 data files, associated with articles in 150 journals.

Recently Published Data


Joint Data Archiving Policy
(http://datadryad.org/jdap)

<< Journal >> requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive, such as << list of approved archives here >>. Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. Authors may elect to have the data publicly available at time of publication, or, if the technology of the archive allows, may opt 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.

Dryad’s goals

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.”

(http://datadryad.org/)
Dryad development and governance

- Dryad development - a joint project of [NESCent](https://nescent.org), the [UNC Metadata Research Center](https://metadata.unc.edu), and a growing number of [partner organizations](https://datadryad.org/partners).
- Stakeholders: journals, publishers and scientific societies, and researchers
- Governance
  - 2009 to 2012 Dryad Interim Board
  - May 2012 members of the Dryad Interim Board approved the [Bylaws](https://datadryad.org/governance) of the organization, establishing Dryad as an **“independent organization, applying for non-profit status, with a 12 member Board of Directors”**
    - Reps from science, journals, societies, OCLC, MS, etc.
  - Board: Sets policy and long-term strategic goals
  - [http://wiki.datadryad.org/Governance](http://wiki.datadryad.org/Governance)
Dryad’s workflow
Multiple Benefits Drive Helping Behavior in a Cooperatively Breeding Bird: An Integrated Analysis

Sjouke A. Kingma,¹,* Michelle L. Hall,¹,²,³ and Anne Peters¹,⁴

1. Max Planck Institute for Ornithology, Vogelwarte Radolfzell, Schlossallee 2, 78315 Radolfzell, Germany; 2. Mornington Wildlife Sanctuary, Australian Wildlife Conservancy, PMB 925, Derby, Western Australia 6728, Australia; 3. Research School of Biology, Australian National University, Canberra, Australian Capital Territory 0200, Australia; 4. School of Biological Sciences, Monash University, Clayton, Victoria 3800, Australia

Submitted July 23, 2010; Accepted January 3, 2011; Electronically published March 10, 2011

Dryad data: http://dx.doi.org/10.5061/dryad.8210.
Data from: Patterns of morphological and plastid DNA variation in the Corallorhiza striata species complex (Orchidaceae)

When using this data, please cite the original article:


Additionally, please cite the Dryad data package:


Corallorhiza striata is a wide-ranging, morphologically variable, mycoheterotrophic species complex distributed across North America. Objectives of this study were to assess relationships and test validity of previously delimited varieties of C. striata, including the recently described C. bentleyi. Two plastid DNA regions were sequenced for individuals from 45 populations across North America, identifying four major clades. The large-flowered C. striata var. striata (northern U.S.A., southern Canada) was sister to the smaller-flowered var. vreelandii (southwestern U.S.A., Mexico), and C. vreelandii was sister to a Californian clade with relatively intermediate-sized flowers. C. bentleyi (eastern U.S.A.) shared a close relationship, sister to the remaining two clades, and Nonparametric Multivariate Analysis of Variance on nine quantitative morphological traits and plastid DNA sequence divergence grouped C. bentleyi with the remaining two clades.
<table>
<thead>
<tr>
<th></th>
<th>Abbreviation</th>
<th>Full name</th>
<th>Review Workflow?</th>
<th>Blackout?</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>amNat</td>
<td>The American Naturalist</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>2</td>
<td>BJLS</td>
<td>Biological Journal of the Linnean Society</td>
<td>N</td>
<td>N</td>
</tr>
<tr>
<td>3</td>
<td>biorisk</td>
<td>BioRisk</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>4</td>
<td>bmjOpen</td>
<td>BMJ Open</td>
<td>Y</td>
<td>N</td>
</tr>
<tr>
<td>21</td>
<td>....</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Increasing submission rate of data packages through June 2011

Today: Dryad contains 1971 data packages and 5193 data files, associated with articles in 150 journals.

74,466 download, mid-July 2012
Data reuse...

- (1) Mascaro et al (2011) combine the Zanne et al (2009) dataset that is in Dryad with new data to perform their own - similar but different - analysis.
- (2) They deposited the new data that they collected into Dryad.
- (3) Both the data and article are cited correctly in the references.

Dryad DCAP (Dublin Core Application Profile), ver. 3.0
- bibo (The Bibliographic Ontology)
- dcterms (Dublin Core terms)
- dryad (Dryad) (property: Dryadstatus)
- DwC (Darwin Core)

1. **Simple**: automatic metadata gen; heterogeneous datasets

2. **Interoperable**: harvesting, cross-system searching

3. **Semantic Web compatible**: sustainable; supporting machine processing

**Data-package centric**
Dryad Technology

- DSpace repository software (open source)
- DOIs via California Digital Library/DataCite
- CCZero (CC0) (Metadata and data)
- Integration with specialized repositories and databases
  - Federated searching with TreeBASE and KNB LTER
  - TreeBASE submission (using BagIt and OAI-PMH)
  - GenBank (currently in development)
Describe publication

Submitting data to Dryad consists of three simple steps:

1. **Describe your publication**
2. Upload and describe your data files
3. Approve data for publication

Please describe your publication in as much detail as possible. Providing a detailed description will make it easier for others to find and use your data in Dryad. Please describe the **publication only**. Do not enter information specific to your data files on this page.

Fields marked with an asterisk (*) are required. For more information on expected contents for a field, hold your mouse over the question.

### Publication metadata

- **Title**: Adaptive responses and disruptive effects: how major wildfires
- **Authors**
  - Last name, e.g. Smith
  - First name + initial, e.g. Donald F.
  - Banks, Sam
  - Blyton, Michaela
  - Blair, David
  - McBurney, Lachlan
  - Lindenmayer, David
- **Journal name**: Molecular Ecology
- **Abstract**: Environmental disturbance is predicted to play a key role in the evolution of animal social behaviour. This is because disturbance affects key factors underlying...
<table>
<thead>
<tr>
<th>Data file description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Title</strong>:</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td>Description:</td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>

Please upload your data file or provide the identifier of a file located in another repository.

- **Data file**: *
  
  Please upload your data file or provide the identifier of a file located in another repository.

- **Browse...**

- **External file identifier**
  
  Select a repository from the list:
  - TreeBASE
  - GenBank
  - KNB
  - OTHER REPOSITORY

- **Description**: 
  
  Enter a description for your data file here.
No controlled subject indexing, yet!!

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 Sep 6, 2012, Dryad contains 1971 data packages and in 150 journals.

Recently Published Data 📰

<table>
<thead>
<tr>
<th>Subject</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>454</td>
<td>1</td>
</tr>
<tr>
<td>Abscission</td>
<td>1</td>
</tr>
<tr>
<td>abundance</td>
<td>26</td>
</tr>
<tr>
<td>accretion</td>
<td>1</td>
</tr>
<tr>
<td>accumulation</td>
<td>2</td>
</tr>
<tr>
<td>acorn woodpecker</td>
<td>2</td>
</tr>
<tr>
<td>AFLP</td>
<td>1</td>
</tr>
<tr>
<td>Agriculture</td>
<td>1</td>
</tr>
<tr>
<td>algae</td>
<td>12</td>
</tr>
<tr>
<td>alkalinity</td>
<td>3</td>
</tr>
<tr>
<td>Name</td>
<td>Count</td>
</tr>
<tr>
<td>-------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Tango, Jazzlyn M.</td>
<td>1</td>
</tr>
<tr>
<td>Taniguchi, Fumiya</td>
<td>2</td>
</tr>
<tr>
<td>Taper, Mark, L.</td>
<td>1</td>
</tr>
<tr>
<td>Tarailo, David</td>
<td>1</td>
</tr>
<tr>
<td>Taraschewski, Horst</td>
<td>1</td>
</tr>
<tr>
<td>Tarone, Aaron</td>
<td>1</td>
</tr>
<tr>
<td>Tartarini, Stefano</td>
<td>1</td>
</tr>
<tr>
<td>Tarver, James E.</td>
<td>1</td>
</tr>
<tr>
<td>Tarver, James</td>
<td>1</td>
</tr>
<tr>
<td>Tattersall, Glenn J.</td>
<td>1</td>
</tr>
<tr>
<td>Author</td>
<td>Count</td>
</tr>
<tr>
<td>------------------------</td>
<td>-------</td>
</tr>
<tr>
<td>Smith, Alan R.</td>
<td>1</td>
</tr>
<tr>
<td>Smith, Andrew B.</td>
<td>1</td>
</tr>
<tr>
<td>Smith, Andrew B</td>
<td>1</td>
</tr>
<tr>
<td>Smith, Carl</td>
<td>2</td>
</tr>
<tr>
<td>Smith, Christian T.</td>
<td>4</td>
</tr>
<tr>
<td>Smith, Damian</td>
<td>1</td>
</tr>
<tr>
<td>Smith, Douglas W.</td>
<td>1</td>
</tr>
<tr>
<td>Smith, Eric N.</td>
<td>1</td>
</tr>
<tr>
<td>Smith, Felisa A.</td>
<td>1</td>
</tr>
<tr>
<td>Smith, Gilbert</td>
<td>1</td>
</tr>
</tbody>
</table>

Now showing items 1-10   Next Page >>
Dryad: Metadata R&D
1. Curation workflow - cognitive walkthroughs
3. Metadata reuse - content analysis (Greenberg, IDCC Research Summit, 2010)
5. Name-authority control - exploratory study (Haven, 2009, INLS 720)
6. KO/metadata community practices - Concurrent triangulation mixed methods (survey + simulation experiment) (White, 2010, ASIST, 2010 JLM)
8. Vocabulary needs (HIVE) – mapping study (Greenberg, 2009, CCQ; Scherle, 2010, Code4Lib)
9. Metadata theory – deductive analysis (Greenberg, 2009)
Helping Interdisciplinary Vocabulary Engineering (HIVE)

- <AMG> approach for integrating discipline CVs
- Model addressing **CV cost, interoperability, and usability constraints** (interdisciplinary environment)

*Building, Sharing, Evaluation* the HIVE....
Towards a worldwide wood economics spectrum

Abstract
Wood performs several essential functions in plants, including mechanically supporting aboveground tissue, storing water and other resources, and transporting sap. Woody tissues are likely to face physiological, structural and defensive trade-offs. How a plant optimizes among these competing functions can have major ecological implications, which have been under-appreciated by ecologists compared to the focus they have given to leaf function. To draw together our current understanding of wood function, we identify and collate data on the major wood functional traits, including the largest wood density database to date (8412 taxa), mechanical strength measures and anatomical...
Towards a worldwide wood economics spectrum

Abstract
Wood performs several essential functions in plants, including mechanically supporting aboveground tissue, storing water and other resources, and transporting sap. Woody tissues are likely to face physiological, structural and defensive trade-offs. How a plant optimizes among these competing functions can have major ecological implications, which have been under-appreciated by ecologists compared to the focus they have given to leaf function. To draw together our current understanding of wood function, we identify and collate data on the major wood functional traits, including the largest wood density database to date (8412 taxa), mechanical strength measures and anatomical features.
Package metadata harvested from email

DCContributor

Contr. 101 (gr. 99%, bl. 1%)

DDICorresp

DCDescription

DCTitle

DCSubject

Subj. 177 (gr. 97%, rd. 2%, bl. 1%)

DCSpatial

DCTemporal

Spat. 35
Temp. 2
DwCSci. 26

DwCSci.Name

Pkg metadata (exact harvest)
Pkg metadata (some editing)
Pkg metadata (not from email)
Email metadata (not used)
File metadata harvested from package metadata

Contr. 100 (gr. 93%, bl. 7%)

DCContributor

Subj. 185 (gr. 83 %, or. 1%, red 4 %, bl. 12 %)

DCSubject

DCSpatial

DCTemporal

DwCSci.Name

Subj. 177 (gr. 97%, rd. 2%, bl. 1%)

- File metadata (inherit exactly)
- File metadata (some editing)
- File metadata (created, not inherited)
- Pkg metadata not used for file
Automatic Metadata Generation R&D (SILS Metadata class)

This page was created in fall 2009 to document work done by students in Jane Greenberg's Metadata class in UNC SILS.

Final projects and independent study research

Dryad and HIVE: exploring automatic metadata generation and metadata quality.

1. SNIP-IT project (Semantic Nuggets in Pieces of Indexed Text): Article content and data object metadata generation—an evaluation of the suitability of published article content, found in pre and post data presentation sub-genres, for automatic metadata generation. (J. Alonzo and L. Skalla)

2. NACS project (Name Authority Control for Scientists): Name-authority control—an exploratory study on the need for name authority control for scientists depositing in Dryad, and the suitability of the LCNAF (LC Authorities: http://authorities.loc.gov/). (D. Haven)

3. Comparing HIVE/HIVE and the NCBO Biportal: Comparing concept retrieval indexing and basic term matching techniques. (J. Sherman)

4. HIVE Usability study: Pilot testing HIVE interface and system in Hollie's 520 class, and preparing for more involved HIVE testing for masters' paper research. (L. Huang)

5. MEOW[excel] project (MEdata Object Wrapper): Data object wrapper metadata—an exploratory/very small scale assessment of the suitability of represented resources.
Functional aspects/properties

1. Core set
2. Data lifecycle
3. Data portability
4. Scheme simplicity
5. Data comparability
6. Scheme stability
7. Provenance
8. Element refinement
9. Scheme harmonization
10. Intra-scheme Modularity
11. Comprehensiveness
12. Data retrieval
13. Data documentation
14. Scheme extensibility

<table>
<thead>
<tr>
<th>Criterion</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Core set</td>
<td>The scheme is intended to provide a common set of elements used to describe the most common situations.</td>
</tr>
<tr>
<td>Data lifecycle</td>
<td>The scheme is intended to support documentation of the data lifecycle.</td>
</tr>
<tr>
<td>Data portability</td>
<td>Data created intended to be &quot;portable“...independent.</td>
</tr>
</tbody>
</table>

(Greenberg, 2005, MODAL (Metadata Objectives and principles, Domains, and Architectural Layout) Framework, CCQ; Willis, Greenberg, & White, CODATA, 2010)
<table>
<thead>
<tr>
<th>Scheme</th>
<th>Vers.</th>
<th>Initial Rel.</th>
<th>Maint. Body</th>
<th>Repository</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. DDI</td>
<td>3.1</td>
<td>2000</td>
<td>DDI Alliance</td>
<td>ICPSR (and others)</td>
</tr>
<tr>
<td>2. CIF</td>
<td>2.4.1</td>
<td>1991</td>
<td>IUCr</td>
<td>Cambridge Structural Database (CSD)</td>
</tr>
<tr>
<td>3. DwC</td>
<td>App.P</td>
<td>2001</td>
<td>TDWG</td>
<td>GBIF</td>
</tr>
<tr>
<td>4. EML</td>
<td>2.1.0</td>
<td>1997</td>
<td>KNB</td>
<td>Ecological Archives</td>
</tr>
<tr>
<td>5. mmCIF</td>
<td>2.0.09</td>
<td>2005</td>
<td>wwPDB</td>
<td>Protein Data Bank (PDB)</td>
</tr>
<tr>
<td>6. MINiML</td>
<td>1.16</td>
<td>2007?</td>
<td>NCBI</td>
<td>Gene Expression Omnibus (GEO)</td>
</tr>
<tr>
<td>7. MAGE</td>
<td>1.0</td>
<td>2002</td>
<td>FGED</td>
<td>ArrayExpress</td>
</tr>
<tr>
<td>8. NEXML</td>
<td>1.0</td>
<td>2009</td>
<td>NESCent</td>
<td>TreeBase</td>
</tr>
<tr>
<td>9. ThermoML</td>
<td>3</td>
<td>2002</td>
<td>IUPAC</td>
<td>ThermoML Archives</td>
</tr>
</tbody>
</table>
Roadmap
February 2007

Metadata research nodes

Metadata generation and quality evaluation
- Process model
- Statistical rating confidence score

Dynamic vocabulary Integration and maintenance
- Dynamic vocabulary server
- IR/QE answers

Instantiation
- Determine to what extent we might Dryad track instantiations

Outcomes/deliverables
Sustainability continued...

- **Revenue model under development**
  
  **Guiding principles:**
  1. Depositors assured that Dryad continues to have resources
  2. Protect integrity and accessibility of the content
  3. Dryad seeks to minimize costs
  4. Spreading the revenue burden

- **Possible payment plans**
  
  1. *Journal-based*: the journal (or group from a society or publisher) prepays, annual fee
  2. *Voucher*: pay in advance for a minimum number
  3. *Pay-as-you-go*: pay retrospectively for deposits during a certain time period
  4. *Author-pays*: individual pays for integrated or nonintegrated

---

Acknowledgments

- Dryad Consortium Board, journal partners, and data authors
- NESCent: Kevin Clarke, Hilmar Lapp, Heather Piwowar, Peggy Schaeffer, Ryan Scherle, Todd Vision (PI)
- UNC-CH <Metadata Research Center>: Jose R. Pérez-Agüera, Sarah Carrier, Elena Feinstein, Lina Huang, Robert Losee, Hollie White, Craig Willis
- U British Columbia: Michael Whitlock
- NCSU Digital Libraries: Kristin Antelman
- HIVE: Library of Congress, USGS, and The Getty Research Institute; and workshop hosts
- Yale/TreeBASE: Youjun Guo, Bill Piel
- DataONE: Rebecca Koskela, Bill Michener, Dave Veiglais, and many others
- British Library: Lee-Ann Coleman, Adam Farquhar, Brian Hole
- Oxford University: David Shotton
Concluding comments

- A contribution, have to start somewhere...
  - Good timing, the right discipline
- Confirmed use
- Machine capabilities, eScience/data synthesis
- An educative commons, intellectually engaging
http://datadryad.org
http://blog.datadryad.org
http://datadryad.org/wiki
http://code.google.com/p/dryad
dryad-users@nescent.org
Facebook: Dryad
Twitter: @datadryad
http://ils.unc.edu/mrc/hive/
http://code.google.com/p/hive-mrc/