# Repositories in Science & Technology: Preserving Access to the Record of Science

## CASE STUDY 1: THE DRYAD REPOSITORY

CENDI/NFAIS/FLICC

November 30, 2011

Jane Greenberg, Metadata Research Center
School of Information and Library Science
University of North Carolina at Chapel Hill

# Overview

- Acknowledgements
- Objectives
- Structure, standards, and workflow
  - *Responsibilities:  who does what?*
- Challenges
  - *Motivating challenges…*
- Conclusions

# 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

# Objectives

# DRYAD

Search Data

**Submit Data Now!**

See how to submit

## My Account

Login or Register

## Browse

Authors

Journal Title

## Information

Depositing Data

Using Data

Dryad Partners

Archiving Policy

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 Nov 22, 2011, Dryad contains **1099 data packages** and **2605 data files**, associated with articles in **95 journals**.

## Recent Posts from the Dryad Blog

○ 1E+3

○ Why does Dryad use CC0?

○ UK Parliament report supports Dryad and da access

## Recently Published Data

Beldade P, Jeronimo MA, Saenko SV (2011) Data from: Genetic basis of stage-specific melanism: a putative new role for a cysteine sulfinic acid decarboxylase in insect pigmentation. *Heredity*
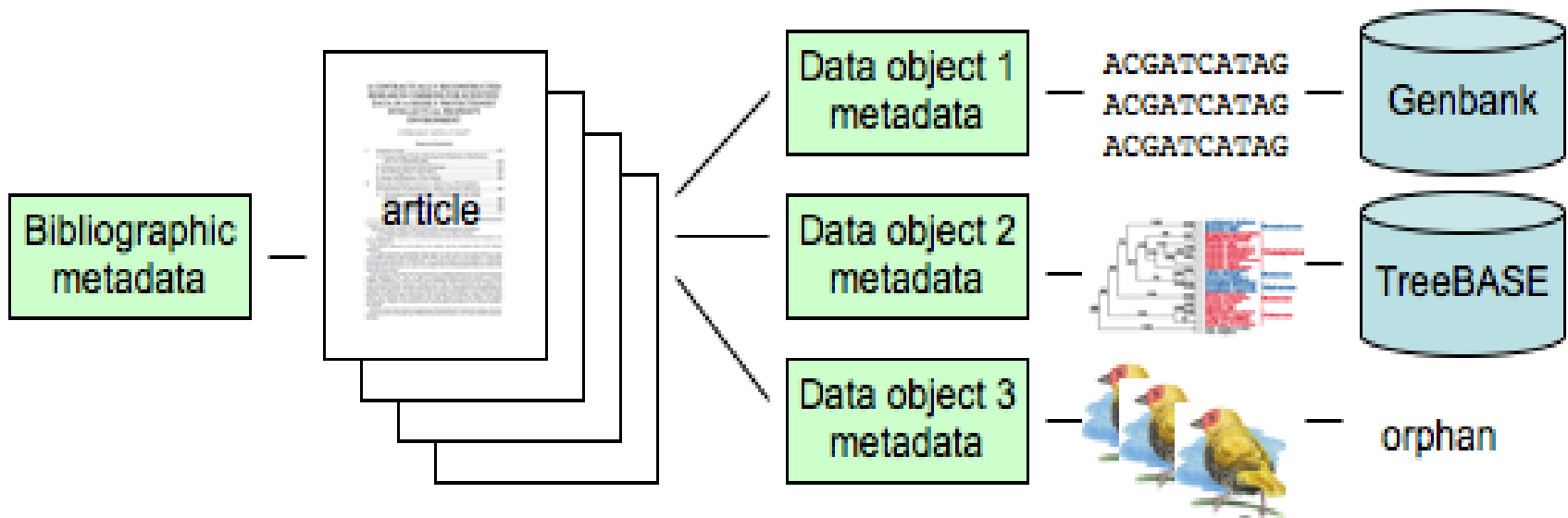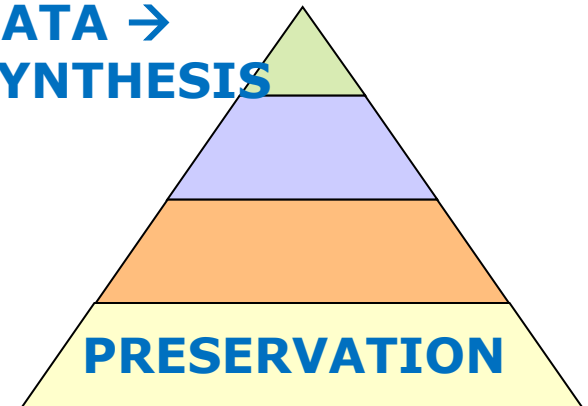
# Dryad's goals

## Dryad's Goals

- One-stop deposition/access for data objects supporting published research…
- Acquisition, preservation, discovery, and reuse of heterogeneous digital datasets
- Allow journals and societies to pool their resources

**DATA → SYNTHESIS**

**PRESERVATION**

# Structure and standards

# Dryad development and governance

- Dryad development - a joint project of <u>NESCent</u>, the <u>UNC Metadata Research Center</u>, and a growing number of <u>partner organizations</u>.

- Dryad Consortium Board
  - Sets policy and long-term strategic goals
  - <u>Interim Governance Agreement</u> (Wiki @: <u>Dryad_Consortium_Board</u>)
    - Permanent governance plan underway
  - One representative from each partner journals
  - Project director serves an *ex officio* member

## Journals and Societies

1. **The American Naturalist** (American Society of Naturalists)
2. **The Biological Journal of the Linnean Society** (Linnean Society of London)
3. **Ecological Monographs** (Ecological Society of America)
4. **Evolution** (Society for the Study of Evolution)
5. **Evolutionary Applications**
6. **Heredity** (The Genetics Society)
7. **Journal of Evolutionary Biology** (European Society for Evolutionary Biology)
8. **Journal of Heredity** (The American Genetic Assoc.)
9. **Molecular Biology and Evolution** (Society for MBE)
10. **Molecular Ecology**
11. **Molecular Ecology Resources**
12. **Molecular Phylogenetics and Evolution**
13. **Systematic Biology** (Soc. for Systematic Bio.)

Pensoft

1. **BioRisk**
2. **Comparative Cytogenetics**
3. **International Jrnl. of Myriopodology**
4. **MycoKeys**
5. **Nature Conservation**
6. **NeoBiota**
7. **PhytoKeys**
8. **ZooKeys**

Additional journals, societies, etc.: http://datadryad.org/partners

# Joint Data Archiving Policy
()

**<< 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.

- Whitlock, M. C., M. A. McPeek, M. D. Rausher, L. Rieseberg, and A. J. Moore. 2010. Data Archiving. American Naturalist. 175(2):145-146. DOI:10.1086/650340
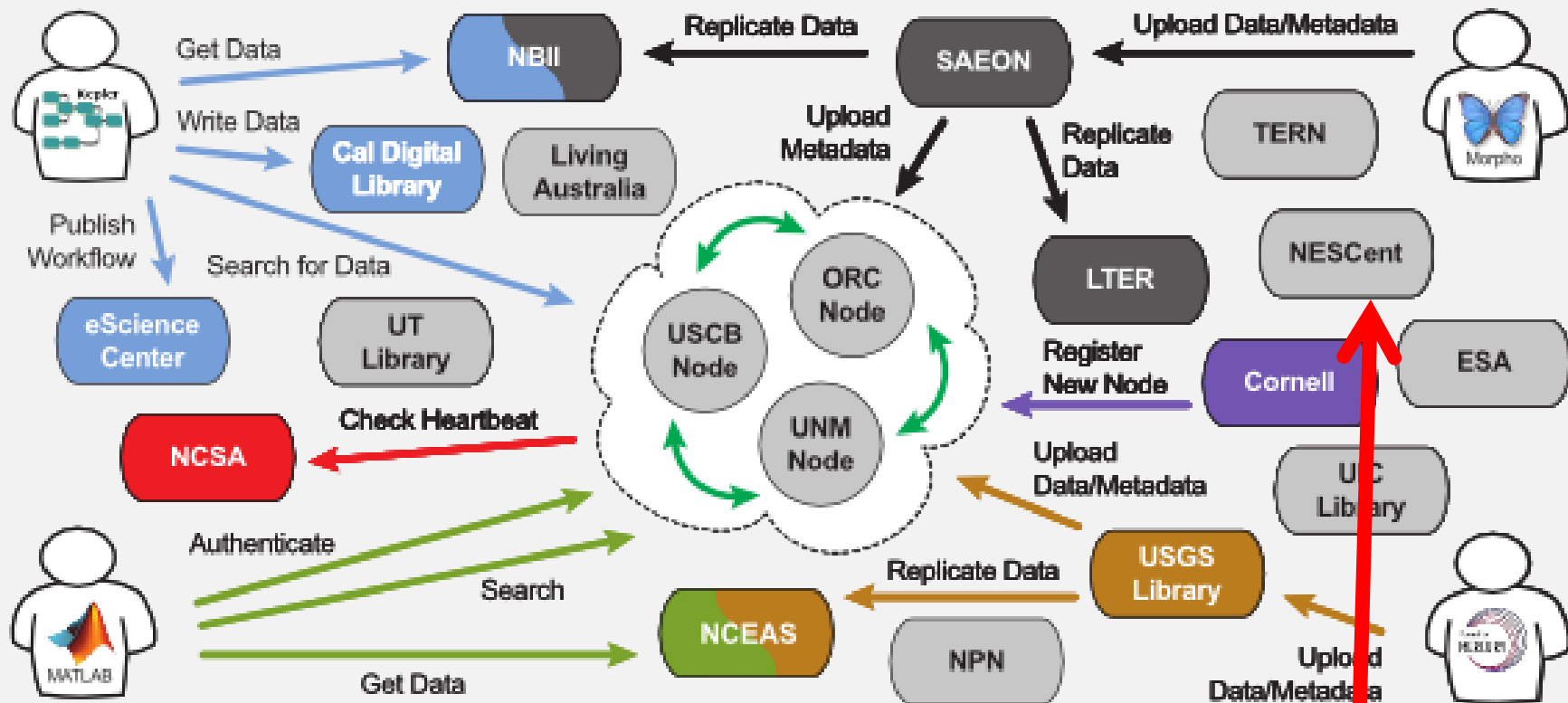
# Dryad Technology

- DSpace repository software (open source)

- DOIs via California Digital Library/DataCite

- CCZero (CC0)

- Integration with specialized repositories and databases
  - Federated searching with TreeBASE and KNB LTER
  - TreeBASE submission (using BagIt and OAI-PMH)
  - GenBank (currently in development)

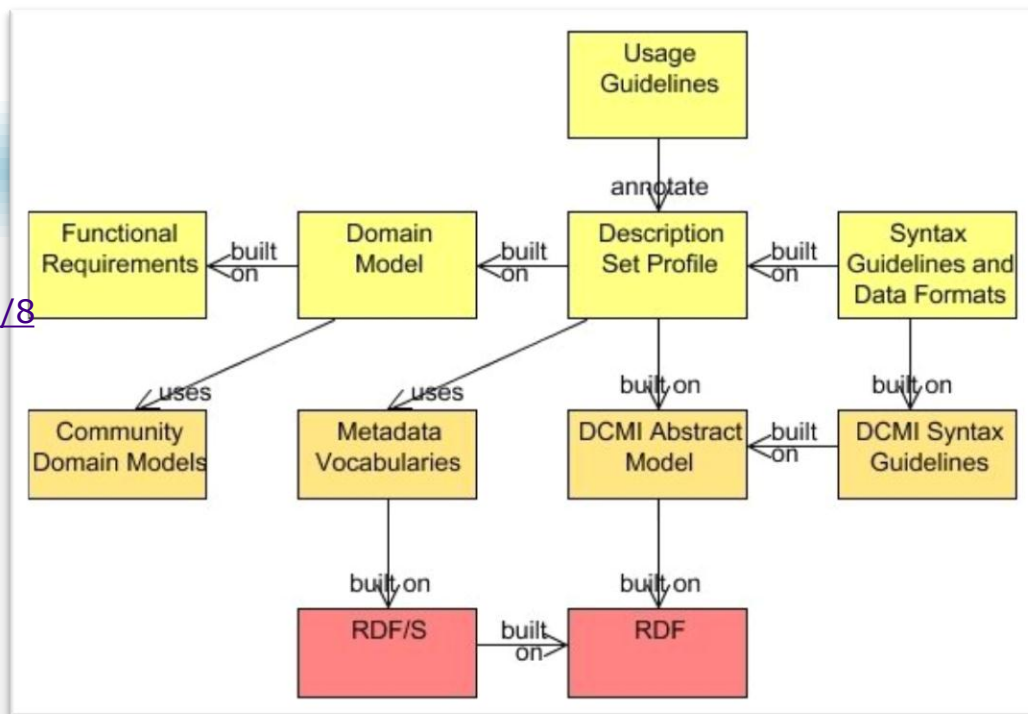# Dryad DCAP (Dublin Core Application Profile), ver. 3.0

(https://www.nescent.org/wg/dryad/images/8/8b/Dryad3.0.pdf)

- bibo (The Bibliographic Ontology)
- dcterms (Dublin Core terms)
- dryad (Dryad) (property: Dryadstatus)
- DwC (Darwin Core)



Baker, T. (2007), Singapore Framework

**Simple:** automatic metadata gen; heterogeneous datasets
**Interoperable:** harvesting, cross-system searching
**Semantic Web compatible:** sustainable; supporting machine processing

**Data-package centric

| page | discussion | view source | history |

# Main Page

This wiki is primarily for the use of the Dryad development team, but most of the content is open for read-only access. To access the Dryad repository, go to DataDryad.org ⧉. Dryad-specific source code and documentation is available at Google Code ⧉. For the latest news on the project, and the occasional opinion piece, check out Dryad News and Views ⧉.

## DRYAD

### search

Go    Search

### popular pages

- Development Plan
- Funding and People
- Publications
- Submission Integration

### categories

- Browse by Category
- Conferences
- Funding
- Help
- Metadata

Dryad ⧉ is a repository for data underlying scientific publications, with an initial focus on evolution, ecology, and related fields. Dryad allows investigators to validate published findings, explore new analysis methodologies, repurpose the data for research questions unanticipated by the original authors, and perform synthetic studies such as formal meta-analyses. All data files in Dryad are available for download and reuse, except those that are under a temporary embargo period, as permitted by editors of the relevant journals.

Dryad aims to provide one-stop data deposition upon publication by interfacing with specialized repositories which are already required for publication (such as GenBank and TreeBASE). However for many publications, these repositories do not capture the whole data package. As a result, many

# Workflow

**Author** submits manuscript to journal

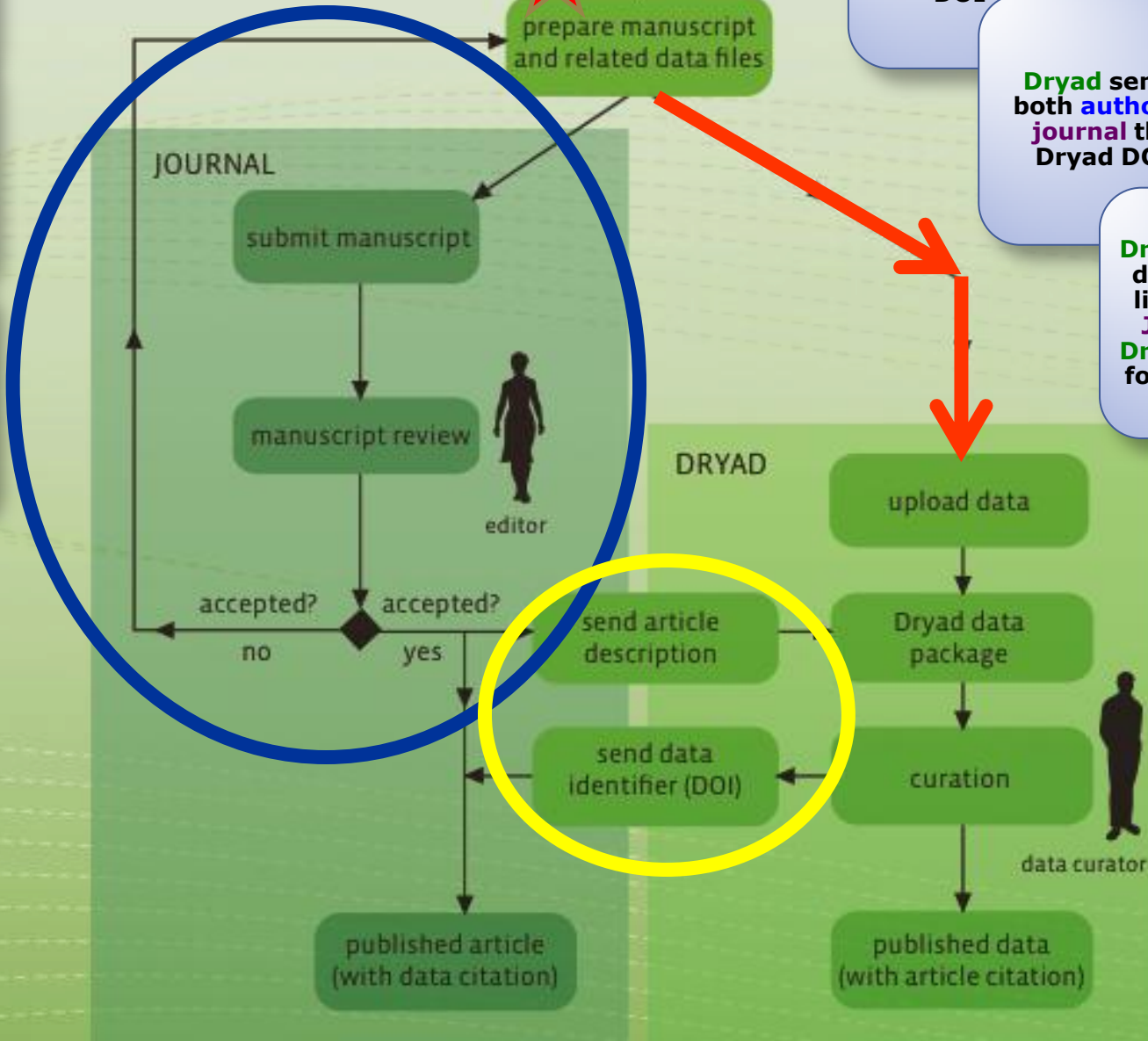**Journal** reports **accepted manuscript** to **Dryad; Dryad** creates provisional record

**Journal** invites **author** to submit data to **Dryad** & provides link to provisional record

**Author** submits data to **Dryad; curator** reviews, issues DOI

**Dryad** sends both **author** & **journal** the Dryad DOI

**Dryad** publishes data files with link to article; **Journal** adds **Dryad** DOI to all forms of article

author

prepare manuscript and related data files

JOURNAL

submit manuscript

manuscript review

editor

accepted?  accepted?

no  yes

published article (with data citation)

DRYAD

upload data

Dryad data package

curation

data curator

send article description

send data identifier (DOI)

published data (with article citation)

From: managing.editor@molecol.com
Date: April 19, 2011 3:09:22 PM EDT
To: Author
Cc: journal-submit@datadryad.org
Subject: Dryad entry for MEC-11-0140.R1

Dear Author

Many thanks for agreeing to participate in the Dryad project. To upload your data, please click the link below- it will take you directly to your entry in the Dryad database.

http://datadryad.org/submit?journalID=MolEcol&manu=223330

<deleted text>

Once you have uploaded your data please include the Dryad identifier in your manuscript. Please let me know if you have any questions about this process.

All the best,
Tim Vines,
Managing Editor, Molecular Ecology

## 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 othe 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 o question.

### Publication metadata

**Title*:** Adaptive responses and disruptive effects: how major wildfire

**Authors*:**

Last name, *e.g. Smith*          First name + initial, *e.g. Donald F.*          Add

☐ Banks, Sam
☐ Blyton, Michaela
☐ Blair, David
☐ McBurney, Lachlan
☐ Lindenmayer, David

Remove selected

**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

Pre-populated metadata field

**DOI:**

**Journal issue:**

Volume       Number       Year

**Primary contact for data associated with this article:**

Banks, Sam ▾

**Subject keywords:**                               Add

- ☐ Behavior/Social Evolution
- ☐ Evolutionary Theory
- ☐ Mammals
- ☐ Phenotypic Plasticity
- ☐ Population Dynamics
- ☐ Population Genetics - Empirical

Remove selected

**Taxonomic names:**              Add

**Geographic areas covered by this publication:**

Add

**Geologic timespans covered by this publication:**

Add

Save & Exit    Continue to describe data file

## Data file *

Please upload your data file or provide the identifier of a file located in another re[...]

- ( ) [ Browse… ]

- ( ) [ External file identifier ]  [ (please select a repository) ▲▼ ]

  - (please select a repository)
  - TreeBASE
  - GenBank
  - KNB

## Data file description

**Title*:** [ ]

**Description:** 

---

# Multiple Benefits Drive Helping Behavior
## Breeding Bird: An Integrated A

Sjouke A. Kingma,[1,*] Michelle L. Hall,[1,2,3] and Anne Peters[1,4]

1. Max Planck Institute for Ornithology, Vogelwarte Radolfzell, Schlossallee 2, 78315 Radolfze[...] Sanctuary, Australian Wildlife Conservancy, PMB 925, Derby, Western Australia 6728, Austral[...] Australian National University, Canberra, Australian Capital Territory 0200, Australia;   4. Sch[...] University, Clayton, Victoria 3800, Australia

# Data from: Patterns of morphological and plastid DNA variation in the Corallo... species complex (Orchidaceae)

**Dryad Package Identifier**    doi:10.5061/dryad.1013    **44 views**

**Individual Data Files**    Supplementary Figure 1    **37 views    9 downloads**

Supplementary Figure 2    **34 views    11 downloads**

**Abstract**    Corallorhiza striata is a wide-ranging, morphologically variable, mycoheterotrophic species complex distributed a North America. Objectives of this study were to assess relationships and test validity of previously delimited varie striata, including the recently described C. bentleyi. Two plastid DNA regions were sequenced for individuals fro populations across North America, identifying four major clades. The large-flowered C. striata var. striata (northe U.S.A., southern Canada) was sister to the smaller-flowered var. vreelandii (southwestern U.S.A., Mexico), and t sister to a Californian clade with relatively intermediate-sized flowers. C. _____ d the en C. bentleyi (eastern U.S.A.) shared a close relationship, sister to the rem_____nponents and Nonparametric Multivariate Analysis of Variance on nine quantitative_____g plastic clades as independent variables, demonstrated strong correlations betw_____ groupi

# Challenges

# Challenges

*motivating challenges...*

- Operational with ongoing development
  - R&D, metadata, and team logistics
- Different workflows
- Growth and sustainability

# Operational...with ongoing development

- Real customers, real needs, still building
  - Manual proof-of-concept → rapid automation

- Metadata research
  - Cognitive walkthroughs (Curation workflow)
  - Crosswalk analyses (Metadata scheme design)
  - Content analyses (Metadata reuse, Metadata functions, Vocabulary needs and name authority control)
  - Experiments + exploratory approaches (Instantiation, PIM)

- Synergy and collective knowledge *vs.* different roofs, goal alignment

Linked data

xml
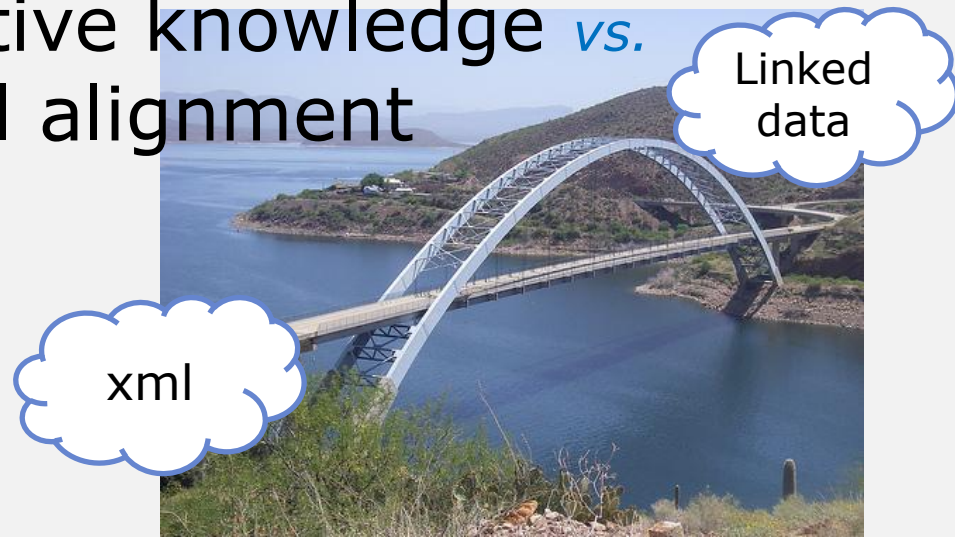
# Workflows

| | Abbreviation | Full name | Review Workflow? | Blackout? |
|---|---|---|---|---|
| 1 | amNat | The American Naturalist | N | N |
| 2 | BJLS | Biological Journal of the Linnean Society | N | N |
| 3 | biorisk | BioRisk | Y | N |
| 4 | bmjOpen | BMJ Open | Y | N |
| : | | | | |
| : | | | | Y |
| 21 | .... | | | |

# Growth and sustainability
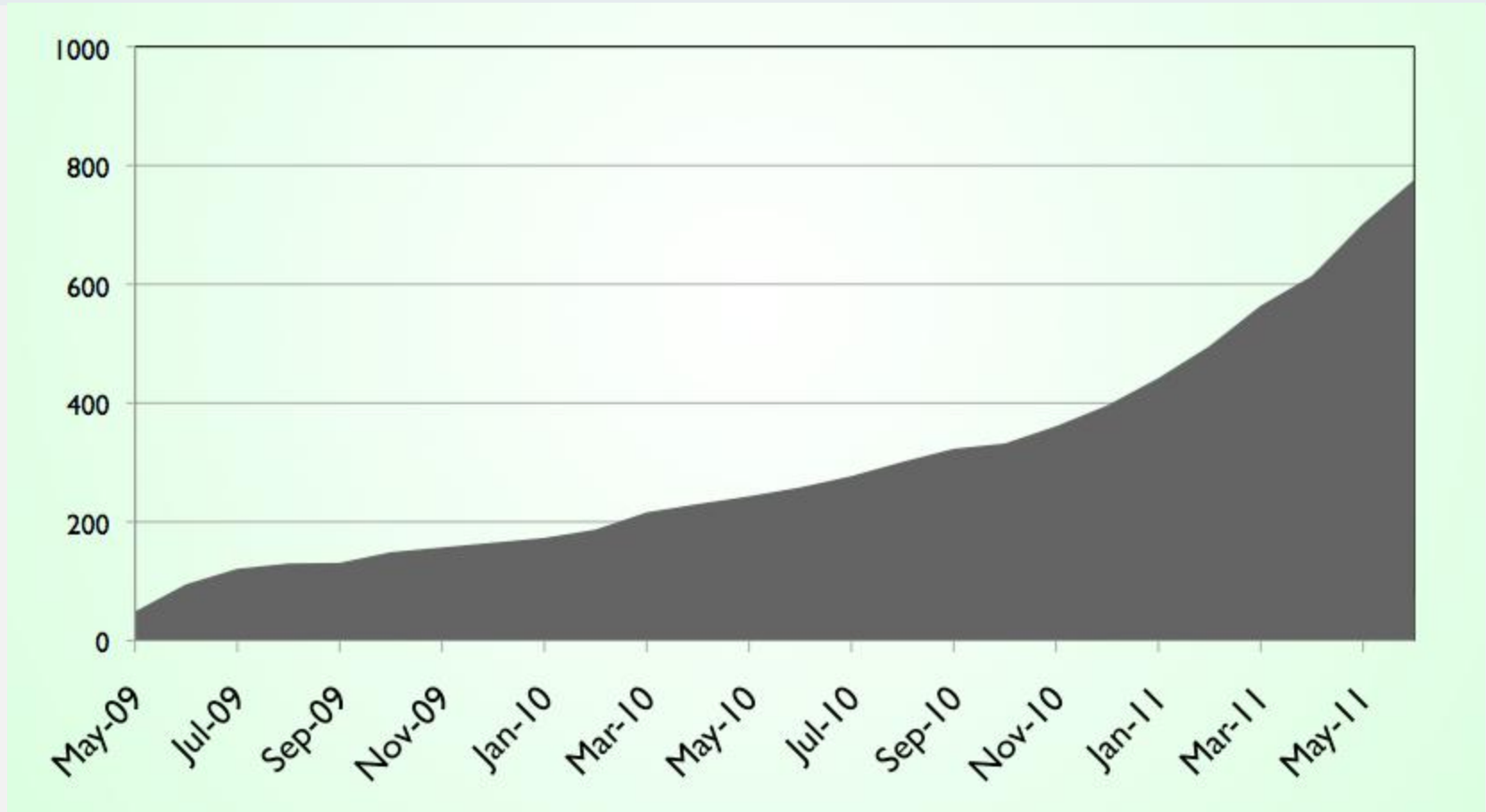


Increasing submission rate of data packages through June 2011

# 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

Beagrie N, Eakin-Richards L, Vision TJ (2010) Business Models and Cost Estimation: Dryad Repository Case Study, iPRES, Vienna: http://www.ifs.tuwien.ac.at/dp/ipres2010/papers/beagrie-37.pdf.

# 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

Facebook & Twitter (#datadryad)