Publishing data in Dryad: infrastructure for the long tail

Todd Vision
@tjvision
NESCent & UNC Chapel Hill

http://orcid.org/0000-0002-6133-2581

You may reuse any of the original content in these slides as you wish, provided you attribute the source.
Goals

- Survey the landscape of journal policies & repositories for long tail data
- Understand what is needed to prepare a submission to Dryad and how to choose among submission options
- Understand what happens to data in Dryad after publication and how to track & get credit for data publication
- Some other data-related resources
Open data citation advantage

Long tail data: Bumpus’ sparrows

Table IIIa

Measurements of Twenty-eight Adult and Young Females which Perished.

<table>
<thead>
<tr>
<th>Trial</th>
<th>Length</th>
<th>Age</th>
<th>Nest</th>
<th>Sex</th>
<th>Body</th>
<th>Beak</th>
<th>Femur</th>
<th>Beak</th>
<th>Femur</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>154</td>
<td>26.8</td>
<td>31.4</td>
<td>.109</td>
<td>.110</td>
<td>1.155</td>
<td>.911</td>
<td>.612</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>156</td>
<td>26.8</td>
<td>31.3</td>
<td>.110</td>
<td>.110</td>
<td>1.156</td>
<td>.912</td>
<td>.613</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>160</td>
<td>26.8</td>
<td>31.4</td>
<td>.109</td>
<td>.110</td>
<td>1.155</td>
<td>.911</td>
<td>.612</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>154</td>
<td>26.8</td>
<td>31.4</td>
<td>.109</td>
<td>.110</td>
<td>1.155</td>
<td>.911</td>
<td>.612</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>156</td>
<td>26.8</td>
<td>31.4</td>
<td>.109</td>
<td>.110</td>
<td>1.155</td>
<td>.911</td>
<td>.612</td>
<td></td>
</tr>
</tbody>
</table>

Long tail data: includes data behind figures, tables, statistical analyses; files where context matters

DNA barcodes, sampling localities, specimen info

Long tail data (e.g. input to statistical analyses)

Intellectual complexity

After Heidorn (2008) http://hdl.handle.net/2142/9127

12-Mar-2015 iDigBio Field2DB workshop
Example: what and where are the data?

  http://doi.org/10.1371/journal.pone.0058978
Why data publishing is a useful metaphor

Vines TH et al. (2013) Current Biology DOI:10.1016/j.cub.2013.11.014
Data papers and data journals
Review Criteria: Biodiversity Data Journal

- Quality of the data
  - Are the data completely and consistently recorded within the dataset(s)?
  - Does the data resource cover scientifically important and sufficiently large region(s), time period(s) and/or group(s) of taxa to be worthy of publication?
  - Are the data consistent internally and described using applicable standards (e.g. in terms of file formats, file names, file size, units and metadata)?
  - Are the methods used to process and analyses the raw data, thereby creating processed data or analytical results, sufficiently well documented that they could be repeated by third parties?
  - Are the data correct, given the protocols? Authors are encouraged to report any tests undertaken to address this point.
  - Is the repository to which the data are submitted appropriate for the nature of the data?

- Consistency between manuscript and data
  - Does the manuscript provide an accurate description of the data?
  - Does the manuscript properly describe how to access the data?
  - Are the methods used to generate the data (including calibration, code and suitable controls) described in sufficient detail?
  - Is the dataset sufficiently novel to merit publication?
  - Have possible sources of error been appropriately addressed in the protocols and/or the paper?
  - Is anything missing in the manuscript or the data resource itself that would prevent replication of the measurements, or reproduction of the figures or other representations of the data?
  - Are all claims made in the manuscript substantiated by the underlying data?
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.

http://datadryad.org/pages/jdap
Effects of JDAP since 2011

Figure 5. Availability of archived phylogenetic data as a function of age. We estimated the effect of publication age on our ability to procure partial (top panels) and complete (bottom panels) phylogenetic datasets from online archives. Overall, the probability of recovering archived phylogenetic data increases toward the present, with a conspicuous recent increase for partial datasets (left panels). The recent surge of archived phylogenetic data likely reflects recent policy changes (middle panels): studies with NSF funding are more likely to archive alignment (but not tree) files (c.f., Table S.14); whereas studies published in journals with JDAP membership are dramatically more likely to archive both partial and complete phylogenetic datasets. The effects of these policy initiatives are not strictly additive (right panels): the correlation of these predictor variables suggests that studies published in JDAP journals are likely to have NSF funding. Shaded areas reflect the 95% credible intervals.

How does Dryad work?

1. DEPOSIT DATA
2. GET PERMANENT IDENTIFIER
3. WATCH YOUR CITATIONS GROW!
4. RELAX, YOUR DATA ARE DISCOVERABLE AND SECURE

Submit data now
How and why?

Search for data
Enter keyword, author, title, DOI, etc.
Go

Advanced search

Browse for data
Recently published Popular By Author By Journal

Recently Published Data


Sremba AL, Martin AR, Baker CS (2014) Data from: Species Identification and likely catch time period of whale bones from South Georgia. Marine Mammal Science
Submission integration of publications & data
Metadata exchange via submission systems

Abstract: Genomic scans often reveal marker loci with greatly increased differentiation between populations. Often this differentiation coincides in space with contrasts in ecological factors, forming a genetic-environment association (GEA). GEAs imply a role for local adaptation, and so it is tempting to conclude that the strongly differentiated markers are themselves under ecologically-based divergent selection, or are closely linked to loci under such selection. Here, we highlight an alternative and neglected explanation: intrinsic genetic incompatibilities rather than local adaptation can be responsible for increased differentiation. Intrinsic genetic incompatibilities create tension zones whose location can shift over time. However, tension zones have a tendency to become trapped by exogenous barriers due to ecological selection. This coupling of endogenous and exogenous barriers can occur easily, even if the loci involved are unlinked. We show that (i) the coupling of endogenous and exogenous barriers can easily occur even when ecological selection is weak; (ii) when environmental heterogeneity is fine-grained, GEAs can emerge at incompatibility loci, but only locally, in places where habitats and gene pools are sufficiently intermingled to maintain linkage disequilibria between genetic incompatibilities, local-adaptation genes and neutral loci. Furthermore, the sign of linkage disequilibrium between the locally adapted and intrinsically incompatible alleles is arbitrary. Reviewing results from the literature, we find that many predictions of our model are supported, including endogenous genetic barriers that coincide with environmental boundaries, local GEA in mosaic hybrid zones, and inverted or modified GEAs at distant locations.
Files in this package

Content in the Dryad Digital Repository is offered "as is." By downloading files, you agree to the Dryad Terms of Service. To the extent possible under law, the authors have waived all copyright and related or neighboring rights to this data.

<table>
<thead>
<tr>
<th>Title</th>
<th>Global Wood Density Database</th>
</tr>
</thead>
<tbody>
<tr>
<td>Downloaded</td>
<td>7318 times</td>
</tr>
<tr>
<td>Description</td>
<td>Please direct all correspondence to G. Lopez-Gonzalez <a href="mailto:G.Lopez-Gonzalez@leeds.ac.uk">G.Lopez-Gonzalez@leeds.ac.uk</a></td>
</tr>
<tr>
<td>Download</td>
<td>GlobalWoodDensityDatabase.xls (2.047Mb)</td>
</tr>
<tr>
<td>Details</td>
<td>View File Details</td>
</tr>
</tbody>
</table>

When using this data, please cite the original article:


Additionally, please cite the Dryad data package:

Bidirectional linkage with publications

Molecular Phylogenetics and Evolution

Molecular systematics of armadillos (Xenarthra, Dasypodidae): contribution of maximum likelihood and Bayesian analyses of mitochondrial and nuclear genes

Frédéric Delage, Michael J Stanhope, Emmanuel J.P Douzery

Laboratoire de Paléontologie, Paléobiologie et Paléogéographie, Institut de Sciences de l'Évolution, Université Montpellier II, Montpellier, France

Queen's University of Belfast, Biology and Biochemistry, 97 Lisburn Road, Belfast BT9 7BL, UK

http://dx.doi.org/10.1016/S1055-7903(03)00111-8, How to Cite or Link Using DOI

Data in DRYAD

Purchase $39.95

More information on this application

Data for this article is available at the following data repositories:

12-Mar-2015

iDigBio Field2DB workshop
Linkouts from PubMed, Genbank
When to submit data?

6 journals

MS preparation → article submission → article review → article acceptance → article publication

30 journals

R (Red)

30 journals

A (Green)
Over 60 integrated journals, ~3/4 sponsored

<table>
<thead>
<tr>
<th>Journal</th>
<th>Submission</th>
<th>Payment</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Int.</td>
<td>Submit: R. Prior to Review</td>
</tr>
<tr>
<td>American Naturalist, The</td>
<td>Y</td>
<td>A</td>
</tr>
<tr>
<td>Biodiversity Data Journal</td>
<td>N</td>
<td>A</td>
</tr>
<tr>
<td>Biological Journal of the Linnean Society</td>
<td>Y</td>
<td>A</td>
</tr>
<tr>
<td>Biology Letters</td>
<td>Y</td>
<td>A</td>
</tr>
<tr>
<td>BioRisk</td>
<td>Y</td>
<td>R</td>
</tr>
<tr>
<td>BMJ Open</td>
<td>Y</td>
<td>R</td>
</tr>
<tr>
<td>Comparative Cytogenetics</td>
<td>Y</td>
<td>R</td>
</tr>
</tbody>
</table>
Also books, theses, other vetted publications

Odd couples in the animal kingdom, but not in a data repository
2013/04/30 by Peggy Schaeffer | Edit

We are celebrating the recent publication in Dryad of the first data to accompany a book [1, 2]. Odd Couples: Extraordinary Differences Between the Sexes in the Animal Kingdom, from Princeton University Press, examines the occasionally surprising gender differences in animals, and what it means to be male or female in the animal kingdom. It is intended for both general and scientific readers.

A dominant male northern elephant seal attempts to copulate with a female. Photo by Derek Roff, courtesy Princeton Univ. Press.


40 Years of Data
2014/04/03 by Peggy Schaeffer | Edit

Availability of the data underlying the book “40 Years of Data: Evolutionary Genetics of Darwin’s Finches on Daphne Major Island” by Peter R. Grant and B. Rosemary Grant. This book is the first comprehensive account of Darwin’s finches from the biological, ecological, and evolutionary perspectives, and is intended for evolutionary biologists, ornithologists, conservation biologists, and general readers interested in evolution in nature. The book is based on a 40-year-long study of Darwin’s finches on Daphne Major Island, located in the Galapagos Islands.

40 YEARS OF EVOLUTION
Darwin’s Finches on Daphne Major Island

Peter R. Grant
B. Rosemary Grant
Dryad is not just for tabular data
Dryad is not just for tabular data II
Embargoes are the exception, not the rule

A. Embargo selections of Dryad data authors for the 10,108 files in Dryad deposited from inception to September 20, 2013. Data include only datasets related to articles published in journals for which the authors had the option of selecting an embargo.

B. Long-term embargoes (>1 year) by journal that granted them.


ReadMe files

- One per data package or one per file
- Recommendations
  - Plain text
  - For each filename, a short description of contents
  - For tabular data: definitions of column headings and row labels; data codes (including missing data); and measurement units
  - Data processing steps not described elsewhere that may affect interpretation of results
  - What if any associated datasets are stored elsewhere
  - Whom to contact with questions
- Côrtes et al. [http://dx.doi.org/10.5061/dryad.b003f](http://dx.doi.org/10.5061/dryad.b003f)
Giving credit for data: citation

Novel forests maintain ecosystem processes after the decline of native tree species

Joseph Musicaro, R. Flint Hughes, and Stefan A. Schnitzer

Department of Biological Sciences, University of Wisconsin, Milwaukee, Wisconsin 53211 USA
Institute for Pacific Islands Forestry, USDA Forest Service, Hilo, Hawaii 96720 USA
Smithsonian Tropical Research Institute, Apartado 2072, Balboa, Republic of Panama

Abstract. The positive relationship between species diversity (richness and evenness) and critical ecosystem functions, such as productivity, carbon storage, and nutrient cycling, is often used to predict the consequences of extinction. At regional scales, however, plant species richness is mostly increasing rather than decreasing because successful plant species introductions far outnumber extinctions. If these regional increases in richness lead to local Dombrois, and P. A. Matson. 1987. Biological invasion by Myrica faya alters ecosystem development in Hawaii. Science 238:802–804.
Taking credit for data I

Holly M Bik
Postdoctoral Researcher

Publications
2012


Taking credit for data II

Genome-scale phylogenetics: Inferring the plant tree of life from 18,896 gene trees: Systematic Biology 2011
DOI: 10.1093/sysbio/syq072
http://www.scopus.com/inward/record.url?eid=2-s2.0-79951569533&partnerID=MN8TOARS

Data from: Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees 2010
DOI: 10.5061/DRYAD.788I
Burleigh, J. Gordon; Bansal, Mukul S.; Eulenstein, Oliver; Hartmann, Stefanie; Wehe, André; Vision, Todd J.;, (2010). "Data from: Genome-scale phylogenetics: inferring the plant tree of life from 18,896 gene trees"
Tracking data citation and reuse

Data from: Crossing the divide: gene flow produces intergeneric hybrid in feral transgenic creeping bentgrass population.

From Repository: Dryad.
By: Zapiola, Maria L; Mallory-Smith, Carol A
Dryad
DOI: http://dx.doi.org.libproxy.lib.uno.edu/10.5061/dryad.kg521
Viewed Date: 24 Oct 2012
Published: 2012

Abstract
Gene flow is the most frequently expressed public concern related to the deregulation of transgenic events (Snow 2002; Elistrand 2003). However, assessing the potential for transgene escape is complex because it depends on the opportunities for unintended gene flow, and establishment and persistence of the transgene in the environment (Warwick et al. 2008). Creeping bentgrass (Agrostis stolonifera L.), a turfgrass species widely used on golf courses, has been genetically engineered to be resistant to glyphosate, a nonselective herbicide. Outcrossing species, such as creeping bentgrass (CB), which have several compatible species, have greater chances for gene escape and spontaneous hybridization (i.e. natural, unassisted sexual reproduction between taxa in the field), which challenges transgene containment. Several authors have emphasized the need for evidence of
Tracking data citation and reuse II
Members and sponsors other than journals

Welcome to the Canadian Healthy Oceans Network website!

The Canadian Healthy Oceans Network, or CHONE (pronounced Ko-nee), is an NSERC strategic network focused on biodiversity science for the sustainability of Canada’s three oceans. CHONE, a marine worm that occurs in all of Canada’s three oceans is the equivalent of the canary in the coal mine: it indicates a healthy benthic habitat.

The network includes about 150 researchers from 14 universities across Canada, the federal Department of Fisheries and Oceans, and seven other government laboratories, to carry out thirty-five collaborative research projects in three interconnected themes.

CHONE’s research in Canada’s three oceans: the Pacific, the Arctic and the Atlantic, showing Discovery Corridors in white. These locations are focal areas for biodiversity research.
Example: Dryad submission

Let’s resubmit this one:

Other resources: best practice docs


- See also DataONE resources: [http://www.dataone.org/best-practices](http://www.dataone.org/best-practices)
### Integrated Digitized Biocollections

<table>
<thead>
<tr>
<th>General</th>
<th>Institutions</th>
<th>Terms</th>
<th>Standards</th>
</tr>
</thead>
<tbody>
<tr>
<td>Name of repository</td>
<td>Integrated Digitized Biocollections</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Additional name(s)</td>
<td>iDigBio</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repository URL</td>
<td><a href="https://www.idigbio.org/">https://www.idigbio.org/</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Subject(s)</td>
<td>Life Sciences, Biology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>The National Resource for Advancing Digitization of Biodiversity Collections (ADBC) funded by the National Science Foundation. Through ADBC, data and images for millions of biological specimens are being made available in electronic format for the research community, government agencies, students, educators, and the general public</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Contact</td>
<td><a href="https://www.idigbio.org/contact">https://www.idigbio.org/contact</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Content type(s)</td>
<td>Audiovisual data, Images</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Keyword(s)</td>
<td>biodiversity, evolution, paleontology</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repository size</td>
<td>23,925,716 Specimen Records; 3,419,648 Media Records; 348 Recordsets</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Repository type(s)</td>
<td>disciplinary</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mission statement for designated community</td>
<td><a href="https://www.idigbio.org/about/mission">https://www.idigbio.org/about/mission</a></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Research data repository language(s)</td>
<td>eng</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Data and/or service provider</td>
<td>dataProvider</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Other resources: DMPTool

The DMP Tool allows you to:

- Meet funder requirements for data management plans.
- Get step by step instructions and guidance for your data management plan as you build it.
- In many cases, get institution specific advice and assistance.

Create ready to use data management plans for specific funding agencies.

Sign up and start building your data management plan now!

Recent DMP News

- Open Access and Climate Research Data
- Data, Data Everywhere... A Deluge of Data Management Articles
- University of Illinois at Urbana-Champaign joins DMPTool partners
- Funder X now available in DMPTool

more news >
A sample data management plan

- https://dmptool.org/plans/8332.pdf
More information

- http://datadryad.org
- http://blog.datadryad.org
- @datadryad

Dryad Community Meeting
Taking a closer look at data

27 May 2015. Washington, DC
8:30AM - 4:00PM
Questions?

CC-BY, Davidk79, wikipedia