Troubleshooting public data archiving

suggestions to increase participation

@dom_roche
dominique.roche@anu.edu.au
Important people

Rob Lanfear
Sandra Binning
Tonya Haff
Lisa Schwanz
Kristal Cain
Hanna Kokko
Mike Jennions
Loeske Kruuk

(Roche, Lanfear, Binning, Haff, Schwanz, Cain, Kokko, Jennions, Kruuk 2014 *PLOS Biology*: e1001779)
What is PDA (Public Data Archiving)?

andra thomer
@an_dre_a_

this HAS to be acronym trolling: "PDA has obvious group benefits for the scientific community" - plosbiology.org/article/info%3...
What is PDA?

this HA
obviously
community
…
What is Public Data Archiving?

The process of storing data and associated metadata in a repository that is open to the public and where data can be accessed and downloaded freely by a third party.

Data typically archived under CC0 license:

- no legal requirement to cite data when they are re-used

- recognition of PDA relies on scientific ethics and good practice
Why do it?

• avoids data loss from hardware malfunction/obsolescence or from researchers moving on to different projects or retiring
• encourages good metadata production to ensure that datasets are interpretable
• increases the ability to evaluate and reproduce studies
• increases opportunities for teaching and learning
• encourages a stronger sharing culture
• improves the return per research dollar

(Huang & Qiao 2011 TREE, Molloy 2011 PLOS Biol, Piwowar et al 2011 Nature, Poisot et al 2013 IEE, Reichman et al 2011 Science, Tenopir et al 2011 PLOS One, Whitlock 2011 TREE, Whitlock et al 2010 Am Nat)
The Joint Data Archiving Policy (JDAP) describes a requirement that supporting data be publicly available. This policy was adopted in a joint and coordinated fashion by many leading journals in the field of evolution in 2011, and JDAP has since been adopted by other journals across various disciplines. Additional journals are welcome to endorse and implement JDAP, or use it as a model.

Journals that adopt JDAP often recommend Dryad as an appropriate data repository, however the JDAP initiative is distinct from Dryad.

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

Representative journals that require data archiving

• The American Naturalist
• Biological Journal of the Linnean Society
• Biology Letters
• BMC Ecology
• BMC Evolutionary Biology
• BMJ
• BMJ Open
• Ecological Applications
• Ecological Monographs
• Ecology
• Ecosphere
• Evolution
• Evolutionary Applications
• Frontiers in Ecology and the Environment
• Functional Ecology
• Genetics
• Heredity

...
DataDryad.org is a curated general-purpose repository that makes the data underlying scientific publications discoverable, freely reusable, and citable. Dryad has integrated data submission for a growing list of journals; submission of data from other publications is also welcome.
How to do it?
Many researchers still harbour concerns about making their data publicly available. This is particularly true in fields such as ecology and evolutionary biology, where datasets are often complex, have a long shelf life, and can be used to test multiple hypotheses.
Lost branches of the tree of life

6,193 papers surveyed in more than 100 journals

17% had accessible trees and alignments

contacting lead authors to get data sets only 19% successful

(Drew et al 2013 PLOS Biol)
More lost data...

Drew BT, Gazis R, Cabezas P, Swithers KS, Deng J, Rodriguez R, Katz LA, Crandall KA, Hibbett DS, Soltis DE. 2013 Lost branches on the tree of life. *PLoS Biol.* 11: e1001636.

Savage C, Vickers A 2009 Empirical study of data sharing by authors publishing in PLoS journals. *PLoS ONE* 4: e7078.

Alsheikh-Ali A, Qureshi W, Al-Mallah M, Ioannidis J 2011 Public availability of published research data in high-impact journals. *PLoS ONE* 6: e24357.

Wicherts JM, Bakker M, Molenaar D. 2011 Willingness to Share Research Data Is Related to the Strength of the Evidence and the Quality of Reporting of Statistical Results. *PLoS ONE* 6: e26828.

Vines TH, Albert AY, Andrew RL, Débarre F, Bock DG, Franklin MT, Gilbert KJ, Moore J-S, Renaut S, Rennison DJ. 2014 The availability of research data declines rapidly with article age. *Curr. Biol.* 24: 94-97.

Gilbert KJ et al. 2012 Recommendations for utilizing and reporting population genetic analyses: the reproducibility of genetic clustering using the program structure. *Mol. Ecol.* 21: 4925-4930.
How do we increase participation?
How do we increase participation?

1. Encourage communication between data generators and re-users
How do we increase participation?

2. Disclose data re-use ethics

Data re-use: The data on embargo selections of Dryad data authors [42] were kindly compiled and archived by Todd J. Vision for this paper. We are grateful for the contribution. Co-authorship was not offered as the data are only a small part of the study.

(Roche et al 2014 PLOS Biol)
How do we increase participation?

3. Encourage increased recognition of publicly archived data

1. **REGISTER**
   Get your unique ORCID identifier [Register now!](#)
   Registration takes 30 seconds.

2. **ADD YOUR INFO**
   Enhance your ORCID record with your professional information and link to your other identifiers (such as Scopus or ResearcherID or LinkedIn).

3. **USE YOUR ORCID ID**
   Include your ORCID identifier on your Webpage, when you submit publications, apply for grants, and in any research workflow to ensure you get credit.
How do we increase participation?

4. Facilitate more flexible embargoes on archived data

Do I have the option to embargo release of my data?

![Bar chart showing number of embargoes granted vs. length of embargo (years)].
Thanks!

Many thanks to Ainsley Seago, Luke Holman, Scott Keogh, Pat Backwell, Andrew Cockburn, Todd Vision, Mark Hahnel and the Evolutionary Ecology Reading group at the Australian National University.

Image / illustration credits: A. Saego, Mr. Google