Update on data citation

Heather Piwowar
DataONE postdoc with NESCent and Dryad
@researchremix
There haven’t been many clear policies on how to cite data.
Journal policies rarely included useful best practices.

Nicholas Weber, *Data citation in the wild* IDCC 2010 poster.


<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Repositories</td>
<td>8 of 26, 31%</td>
<td></td>
</tr>
<tr>
<td>Journals</td>
<td>16 of 307, 6%</td>
<td></td>
</tr>
<tr>
<td>Funders</td>
<td>1 of 52, 2%</td>
<td></td>
</tr>
</tbody>
</table>
No consistent practice
The references used to build this database are available in the appendix.  GenBank accession numbers for those species can be found in the original publications.

We reanalyzed the Barro Colorado Island (BCI) permanent forest plot data in order to learn about the...

<table>
<thead>
<tr>
<th>Species</th>
<th>12S rDNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pezophaps solitaria</td>
<td>AF483300</td>
</tr>
<tr>
<td>Raphus cucullatus</td>
<td>AF483301</td>
</tr>
<tr>
<td>Electroenas madagascariensis</td>
<td>AF483307</td>
</tr>
<tr>
<td>Caloenas nicobarica</td>
<td>EF373289</td>
</tr>
</tbody>
</table>

(GenBank Accession nos EU873092–EU873150) were downloaded into BioEdit and compared to those in Genbank, accession number L19324 (Greenhalgh et al., 1993). Specimens sequenced are listed above.

Meteorological data were obtained from the North Carolina State Climate Service (North Carolina State Climate Office 2008) for the six weather stations surrounding Durm...
We reviewed 500 articles in six major evolution and ecology journals for evidence of data citation:

<table>
<thead>
<tr>
<th>Cited data article</th>
<th>Mentioned repository</th>
<th>Mentioned dataset ID</th>
<th>Pattern frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td></td>
<td></td>
<td>17%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td></td>
<td>22%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>36%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>12%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>7%</td>
</tr>
<tr>
<td></td>
<td></td>
<td>✓</td>
<td>5%</td>
</tr>
<tr>
<td>53%</td>
<td>47%</td>
<td>13%</td>
<td></td>
</tr>
</tbody>
</table>

Sarah Judson, *Data citation in the wild* IDCC 2010 poster.
We reviewed 500 articles in six major evolution and ecology journals for evidence of data citation:

<table>
<thead>
<tr>
<th>Cited data article</th>
<th>Mentioned repository</th>
<th>Mentioned dataset ID</th>
<th>Pattern frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>17%</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td>✓</td>
<td>22%</td>
</tr>
<tr>
<td>✓</td>
<td></td>
<td>✓</td>
<td>36%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>12%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>7%</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>5%</td>
</tr>
</tbody>
</table>

53% 47% 13%

Sarah Judson, *Data citation in the wild* IDCC 2010 poster.
As a result, very time consuming to track
Citation:
Randerson, J. T., G. R. van der Werf, L. Giglio, G. J. Coloma, and Oak Ridge National Lab.

Publications Using This Data Set

GLOBAL FIRE EMISSIONS DATABASE, VERSION 2.1

The following 15 publications cited the product above:


Fox, A., Williams, M., Richardson, A. D., Cameron, D., Covey, I. D., Quaife, T., Ricciuto, D., Reichstein, M., To...
In 2009, 116 articles cited ORNL DAAC data.

Finding these articles took **70-80 hours**

across at least **12 resources**

all chosen from a **deep understanding**

of this specific research domain

then the full text of all the hits were manually reviewed

Valerie Enriquez interview with James Kidder

More consistent standards and tool support should make this easier.
Active conversations:

- Cross-disciplinary meeting at Harvard in June on Data Citation Principles

- DataCite is refining its standards

- NISO Data Citation

- As Web of Science and Scopus begin to support, will become clear what they need...
Dryad’s policy

When using this data, please cite the original article:


Additionally, please cite the Dryad data package:

Very similar to the consensus that is emerging from community discussions.
• Dryad is an recognized leader in data citation best practice.
Consensus building around:

- as much as possible like an article
- in the references section
- with a unique identifier
- inertia behind DOIs
Benefits to this approach

- social benefits thru similarity to articles
- piggyback on infrastructure: tool support, clean data
Of course, a few issues remain
Some we can do things about,

others just worth knowing
Lack of tool support for tracking citations through DOIs
Citation:

Publications Using This Model

U.S. doi:10.3334/ORNLDAAC/567
Advanced Search. Use 2-character tags, Boolean operators, parentheses, and set references to create your query. Results appear in the Search History at the bottom of the page.

Example: TS=(nanotub* SAME carbon) NOT AU=Smalley RE
#1 NOT #2 more examples | view the tutorial
<table>
<thead>
<tr>
<th>Document (sort by relevance)</th>
<th>Author(s)</th>
<th>Date</th>
<th>Source Title</th>
</tr>
</thead>
</table>
Impacts of experimentally imposed drought on leaf respiration and morphology in an Amazon rain forest


Functional Ecology 24 (3), pp. 524-533


“Tracking Dataset Citations Using Common Citation Tracking Tools Doesn’t Work”


Our best practice doesn’t scale to mega-reuse
<table>
<thead>
<tr>
<th>Reference</th>
<th>Platform</th>
<th>Disease</th>
<th>No. of samples</th>
<th>No. of differentially expressed features</th>
<th>No. of differentially expressed mapped features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Golub et al., 1999</td>
<td>Affymetrix HU6000</td>
<td>AML/ALL</td>
<td>72</td>
<td>100</td>
<td>78</td>
</tr>
<tr>
<td>Okutsu et al., 2002</td>
<td>Custom cDNA, 23,040 clones</td>
<td>AML</td>
<td>76</td>
<td>491</td>
<td>355</td>
</tr>
<tr>
<td>Schoch et al., 2002</td>
<td>Affymetrix U95Av2</td>
<td>AML</td>
<td>37</td>
<td>150</td>
<td>140</td>
</tr>
<tr>
<td>Debernardi et al., 2003</td>
<td>Affymetrix U95Av2</td>
<td>AML</td>
<td>28</td>
<td>77</td>
<td>75</td>
</tr>
<tr>
<td>Kohlmann et al., 2003</td>
<td>Affymetrix U95Av2 Affymetrix U133A</td>
<td>AML/ALL</td>
<td>90</td>
<td>156</td>
<td>147</td>
</tr>
<tr>
<td>Yagi et al., 2003</td>
<td>Affymetrix U95Av2</td>
<td>AML</td>
<td>54</td>
<td>1,910</td>
<td>1,753</td>
</tr>
<tr>
<td>Bullinger et al., 2004</td>
<td>Custom cDNA, 39,711 clones</td>
<td>AML</td>
<td>116</td>
<td>1,040</td>
<td>855</td>
</tr>
<tr>
<td>Lacayo et al., 2004</td>
<td>Custom cDNA, 42,749 clones</td>
<td>AML</td>
<td>100</td>
<td>436</td>
<td>329</td>
</tr>
<tr>
<td>Neben et al., 2004</td>
<td>Custom cDNA, 4211 clones</td>
<td>AML</td>
<td>29</td>
<td>170</td>
<td>162</td>
</tr>
<tr>
<td>Ross et al., 2004</td>
<td>Affymetrix U133A</td>
<td>AML</td>
<td>150</td>
<td>713</td>
<td>682</td>
</tr>
<tr>
<td>Valk et al., 2004</td>
<td>Affymetrix U133A</td>
<td>AML</td>
<td>293</td>
<td>779</td>
<td>745</td>
</tr>
<tr>
<td>Vey et al., 2004</td>
<td>DiscoveryChip, cDNA, 9,039 clones</td>
<td>AML</td>
<td>55</td>
<td>197</td>
<td>119</td>
</tr>
<tr>
<td>Alcalay et al., 2005</td>
<td>Affymetrix U133A</td>
<td>AML</td>
<td>78</td>
<td>554</td>
<td>541</td>
</tr>
<tr>
<td>Gutierrez et al., 2005</td>
<td>Affymetrix U133A</td>
<td>AML</td>
<td>43</td>
<td>181</td>
<td>167</td>
</tr>
<tr>
<td>Heuser et al., 2005</td>
<td>Custom cDNA, 41,424 clones</td>
<td>AML</td>
<td>137</td>
<td>178</td>
<td>146</td>
</tr>
<tr>
<td>Haferlach et al., 2005</td>
<td>Affymetrix U133A</td>
<td>AML</td>
<td>35</td>
<td>20</td>
<td>19</td>
</tr>
<tr>
<td>Neben et al., 2005</td>
<td>Custom cDNA, 4,211 clones</td>
<td>AML</td>
<td>110</td>
<td>250</td>
<td>250</td>
</tr>
<tr>
<td>Verhaak et al., 2005</td>
<td>Affymetrix U133A</td>
<td>AML</td>
<td>275</td>
<td>568</td>
<td>555</td>
</tr>
<tr>
<td>Radmacher et al., 2006</td>
<td>Affymetrix U133plus2.0</td>
<td>AML</td>
<td>64</td>
<td>314</td>
<td>304</td>
</tr>
<tr>
<td>Wilson et al., 2006</td>
<td>Affymetrix U95Av2</td>
<td>AML</td>
<td>170</td>
<td>705</td>
<td>674</td>
</tr>
<tr>
<td>Gal et al., 2006</td>
<td>Affymetrix U133A</td>
<td>AML</td>
<td>5</td>
<td>822</td>
<td>724</td>
</tr>
<tr>
<td>Bullinger et al., 2007</td>
<td>Custom cDNA, 39,711 clones</td>
<td>AML</td>
<td>93</td>
<td>4,120</td>
<td>3,828</td>
</tr>
<tr>
<td>Eisele et al., 2007</td>
<td>Affymetrix U133A</td>
<td>AML</td>
<td>11</td>
<td>82</td>
<td>82</td>
</tr>
<tr>
<td>Mullighan et al., 2007</td>
<td>Affymetrix U133A</td>
<td>AML</td>
<td>93</td>
<td>1,188</td>
<td>1,095</td>
</tr>
<tr>
<td>Wouters et al., 2007</td>
<td>Affymetrix U133plus2.0</td>
<td>AML</td>
<td>530</td>
<td>608</td>
<td>560</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>10 platforms</strong></td>
<td></td>
<td><strong>2,744</strong></td>
<td><strong>15,809</strong></td>
<td><strong>14,385</strong></td>
</tr>
</tbody>
</table>

* Included further analysis of data by Verhaak 2005\(^2\) and Wilson 2006\(^2\).

doi:10.1371/journal.pone.0009466.t001
Number of datasets referenced by articles that reused GEO data, cumulative
Number of datasets referenced by articles that reused GEO data, cumulative
FOR AUTHORS

MANUSCRIPT FORMATTING GUIDE

The maximum number of references, strictly enforced, is 50 for Articles and 30 for Letters. Acknowledgements, a conflict of interest statement, and references. They do not exceed 1200 words. The maximum number of references is 10.

The number of references is limited to 40.

http://www.nature.com/nature/authors/gta/

Citations in supplementary information are invisible.

Seeber F.
#3
Adoption of best practices erode incentives in short term (though not problem for Dryad given its policy to cite article+dataset)
Sharing Detailed Research Data Is Associated with Increased Citation Rate

Heather A. Piwowar*, Roger S. Day, Douglas B. Fridsma

~70% in multivariate analysis
Dryad’s policy

When using this data, please cite the original article:


Additionally, please cite the Dryad data package:

Lacking best practice for attributing data deposit
When using this data, please cite the original article:


Additionally, please cite the Dryad data package:

If your article has not yet been published, we recommend you add text specifying the location of the data, for example:

Data deposited in the Dryad repository: [doi:10.5061/dryad.#####]
male inbred lines from Sengwa and Victoria Falls respectively) were used for the analyses, where a comparison to an outgroup was required (Table S1). The raw data for the microsatellites are available in Dryad (doi: 10.5061/dryad.1731).

**Figure S3** Dominance estimates are not sensitive to changing v, keeping the haploid mutation rate equal to the diploid mutation rate.

To assess the sensitivity of dominance estimates to decreasing the haploid mutation rate.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be reorganized for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Data deposited at Dryad: doi: 10.5061/dryad.8048

**Supporting information**

Additional supporting information may be found in the online version of this article.

Data S1 Binary allele calls for MSAP data can be obtained from the Dryad repository using the following URL: http://dx.doi.org/10.5061/dryad.7851.

**Data and protocol availability**

We’ve made the following data and protocols available for the public: (1) GOS and reference sequences for RecA and RpoB; (2) Subfamilies of RecA and RpoB (Table 1.3); (3) Alignments and Newick format phylogenetic trees of RecA and RpoB (Figure 1.3); (4) Sequences of the genes that share assemblies with the novel recAs. (Table 2); (5) GOS ss-rRNA sequence reads; (6) the Lek clustering program. The data and protocols are available at http://bacterial.speciescenter.ucdavis.edu/GOSRecA_DATA/index.html. The data have also been submitted to the Dryad repository http://datadryad.org/doi:10.5061/dryad.38384.
Some recommendations for consideration
a) Dryad should recommend attributing data deposit the same way as data reuse (with a citation to the dataset in the references section)
Pros:

- more exposure,
  -> fewer people think it is weird to cite data
- educates and gives models
- data archives archives can train depositing investigators
- every dataset gets at least one citation ;)
- and most importantly, it creates explicit, unambiguous, best practice links between datasets and papers.

Better and consistently linked datasets and papers = wins for everybody
Here are some disadvantages:

- Different than how people usually reference Genbank, PDB, etc data
- Few journals have standardized on this approach so far
- More complex for authors to add a citation after data archiving than just a sentence in full text? Especially for papers that have a maximum-number-of-citations rule?
- It makes citation context around data citations more ambiguous, since a reference could be for sharing or reuse.
b) Dryad should encourage its partner journals to also follow this convention (with a citation to the dataset in the references section)
c) Dryad should remain flexible, adopting new aspects of best practices as they emerge.
In summary: Dryad is doing the right things.

Recommendations:
1. Suggest that authors cite data deposition using references section
2. Encourage partner journals adopt this policy too
3. Stay flexible to adopt community norms as they emerge
In summary: Dryad is doing the right things.

Recommendations:
1. Suggest that authors cite data deposition using references section
2. Encourage partner journals adopt this policy too
3. Stay flexible to adopt community norms as they emerge
thanks

Dryad and DataONE colleagues, in particular:
   Todd Vision
   Nicholas Weber, Sarah Judson, Valerie Enriquez

The open science online community
and those who release their articles, datasets and photos openly.