

Data Citation in the Wild



Valerie Enriquez^{1,2}, Sarah Walker Judson^{1,3}, Nicholas M. Weber^{1,4,10}

and mentors Suzie Allard^{1,5}, Robert B. Cook^{1,6}, Heather A. Piwowar^{1,7}, Robert J. Sandusky^{1,8}, Todd J. Vision^{1,7,9}, Bruce Wilson^{1,5,6}

Data attribution is important

- Rewards those who contribute
- Allows measurement of reuse to demonstrate value and evaluate policy

How is data reuse currently attributed?

We investigated the **policies**, **practices**, and **implications** of data attribution practices in the **environmental sciences**.

Policies

We reviewed policies of various stakeholders within the environmental sciences, looking for **evidence of data citation policies and attribution suggestions**.

Policies in our sample that

Repositories	8 of 26,	31%
Journals	16 of 307,	6%
Funders	1 of 52,	2%

Repository instructions varied. Several patterns:

How do I cite data from Dryad?
When using data from Dryad, please cite the original article.

Sidauskas, B. 2007. Testing for unequal rates of morphological diversification in the absence of a detailed phylogeny: a case study from characiform fishes. *Evolution* 61: 299-316.

For example:
Sidauskas, B. 2007. Data from: Testing for unequal rates of morphological diversification in the absence of a detailed phylogeny: a case study from characiform fishes. Dryad Digital Repository. doi:10.5061/dryad.20

If you are using a large number of data sources, it may be appropriate to provide a list of referenced data packages, rather than citing each individually in the references section. This list of data packages can then

Data Citation Policy

In the event that data distributed from the Land Processes DAAC are incorporated into your research, please supply the following acknowledgment within your published work: "These data are distributed by the Land Processes Distributed Active Archive Center (LP DAAC), located at the U.S. Geological Survey (USGS) Earth Resources Observation and Science (EROS) Center (lpdaac.usgs.gov)." If possible, please e-mail or send us reprints/citations of papers or oral

Policy links can be found at: http://openwetware.org/wiki/User:Valerie_Enriquez/Notebook/DataONE_Web_resources

GenBank Sequences

If GenBank sequence entries are to be cited, the citation should include the sequence name and GenBank accession number. For example:

The GenBank accession number for the sequence of the left arm of Chromosome 1 reported in this paper is U12980.

Data Product Citation Policy

To acknowledge the scientists who have provided products, we request that you include a bibliographic citation to all ORNL DAAC data products or services, please contact the ORNL DAAC User Services Office (USO).

Citation information is provided in the documentation that accompanies all our data products. If you have questions about how to cite ORNL DAAC data products or services, please contact the ORNL DAAC User Services Office (USO).

An editorial "Citations to Published Data Sets" describes the rationale and advantages for data set citations.

Citation Style

- On-Line Data Set

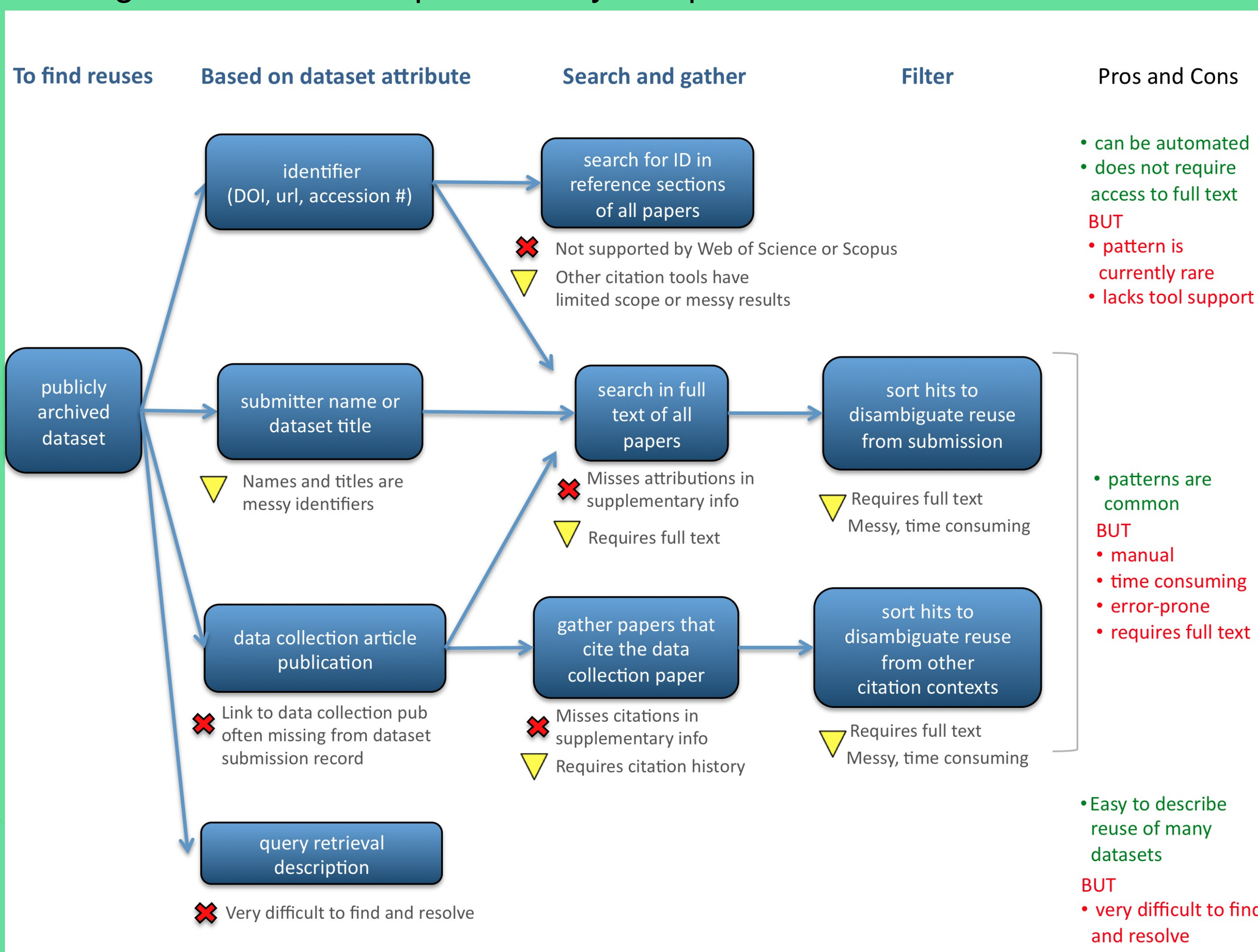
Turner, D.P., W.D.Ritts, and M. Gregory. 2006. BigFoot NPP Surfaces for North and South American Sites, 2002-2004. Data set. Available online [<http://dasac.ornl.gov/>] from Oak Ridge National Laboratory Distributed Active Archive Center, Oak Ridge, Tennessee, U.S.A. doi:10.3334/ORNLDAAC/750.

Examples: Dryad, SDG/GenBank, LPDAAC, ORNL

Journal policies rarely included useful best practices. For example, some policies only discussed hyperlinks to datasets, others emphasized that only certain databases were permitted in their reference lists.

Implications

Because of **tool limitations** and these **diverse data attribution practices**, tracking dataset reuse is prohibitively complicated:



Practices

We reviewed 500 articles in six leading environmental science journals, 2000-2010. **Data attribution patterns** varied widely in the body of the articles that reused data ($n=221$):

Cited data article	Mentioned repository	Mentioned dataset ID	Pattern frequency
			17%
	✓		22%
✓			36%
✓	✓		12%
	✓	✓	7%
✓	✓	✓	5%
53%	47%	13%	

The references used to build this database are available in the appendix.

cales and different scenarios of fossil calibrations. GenBank accession numbers can be found in the original publications.

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

.02 Cooper et al. 2005	Species	12S rDNA
2.40 Perry 2007	<i>Pezophaps solitaria</i>	AF483300
...	<i>Raphus cucullatus</i>	AF483301
Shaffer and Whitford 1981; Perry 2007	<i>Alectroenas madagascariensis</i>	AF483307
	<i>Calocypus nicobarica</i>	EF373289

(GenBank Accession nos EU873092-EU873150) were downloaded into BioEdit and compared to those in

was obtained from Genbank, accession number L19324; Greenhalgh et al., 1993). Specimens sequenced are listed

Meteorological data were obtained from the North Carolina State Climate Service (North Carolina State Climate Office 2008) for the six weather stations surrounding Dur-

North Carolina State Climate Office. 2008. NC Climate Retrieval and Observations Network of the Southeast database. <http://www.nc-climate.ncsu.edu/cronos>. Accessed February 23, 2009. Novick, K., P. Stoy, G. G. Katul, D. S. Ellsworth, M. Siqueira, J. Juang,

We found that data citation policies are rarely articulated, lack standardization, and — even where commonly followed — may fail to support attribution discovery.

As a result of diverse practices and tool limitations, data citations are currently very difficult to track.

Call to action

For scholars to receive credit for reuse of their data:

1. Publishers and repositories need to standardize data citation policies
2. Data citations need to facilitate attribution inference (e.g. through unique dataset identifiers)
3. Automated tools for tracking data citations need to be developed

Until there is a widely adopted standard for data citations that enables automated tracking of data reuse, scholars will lack a key incentive for sharing their data publicly.

Acknowledgements

This work was conducted as part of the **DataONE Summer 2010 internship program**. The DataONE summer 2010 internship program was funded by INTEROP: Creation of an International Virtual Data Center for the Biodiversity, Ecological and Environmental Sciences, NSF grant #0753138, and the Data Observation Network for Earth (DataONE) NSF cooperative agreement #0830944.

Affiliations

- 1: DataONE; 2: Simmons College; 3: Brigham Young University;
- 4: University of Illinois Champaign-Urbana; 5: University of Tennessee, Knoxville;
- 6: Oak Ridge National Laboratory; 7: National Evolutionary Synthesis Center;
- 8: University of Illinois at Chicago; 9: University of North Carolina at Chapel Hill
- 10: Data Conservancy

