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

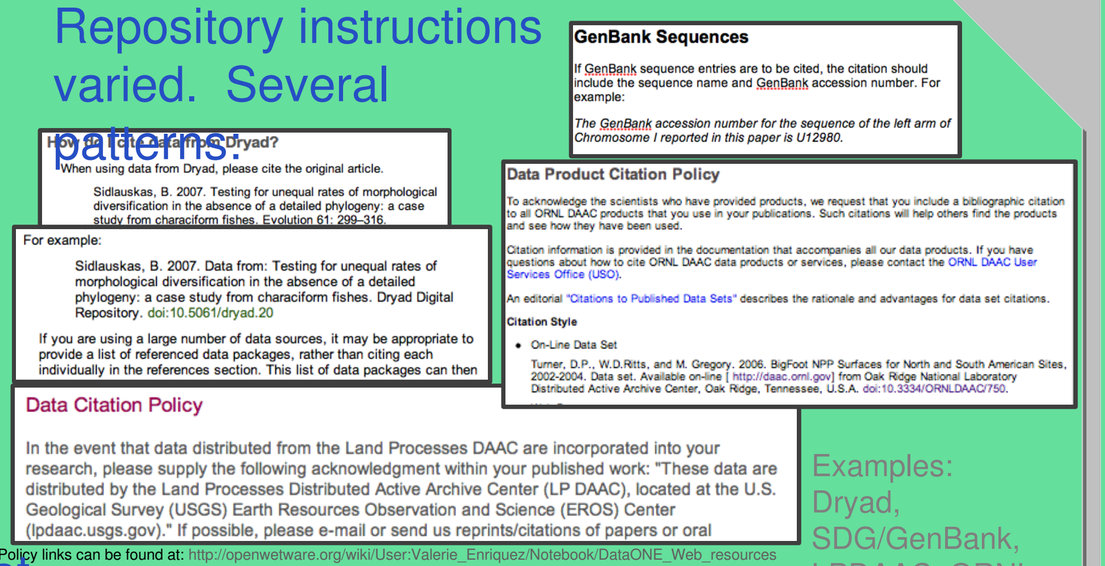
How is data reuse currently attributed?

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***.

Repository instructions varied. Several patterns:



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

Examples: Dryad, SDG/GenBank, LPDAAC, ORNL DAAC

Policies in our sample that mention data citation:

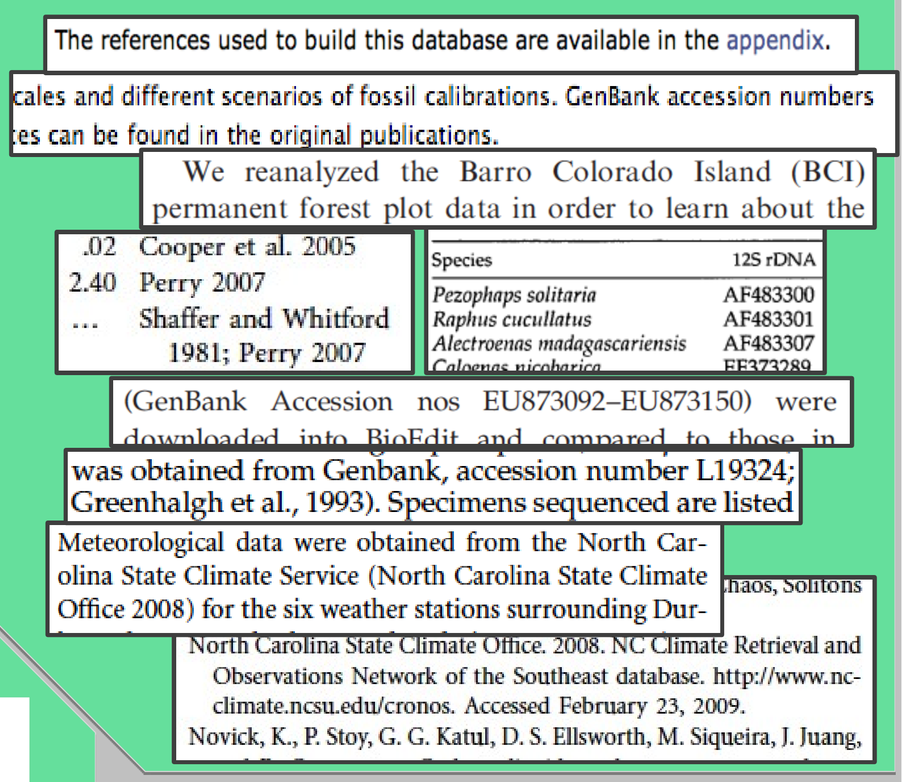
Repositories	8 of 26,	**31%**
Journals	16 of 307,	**6%**
Funders	1 of 52,	**2%**

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.

Practices

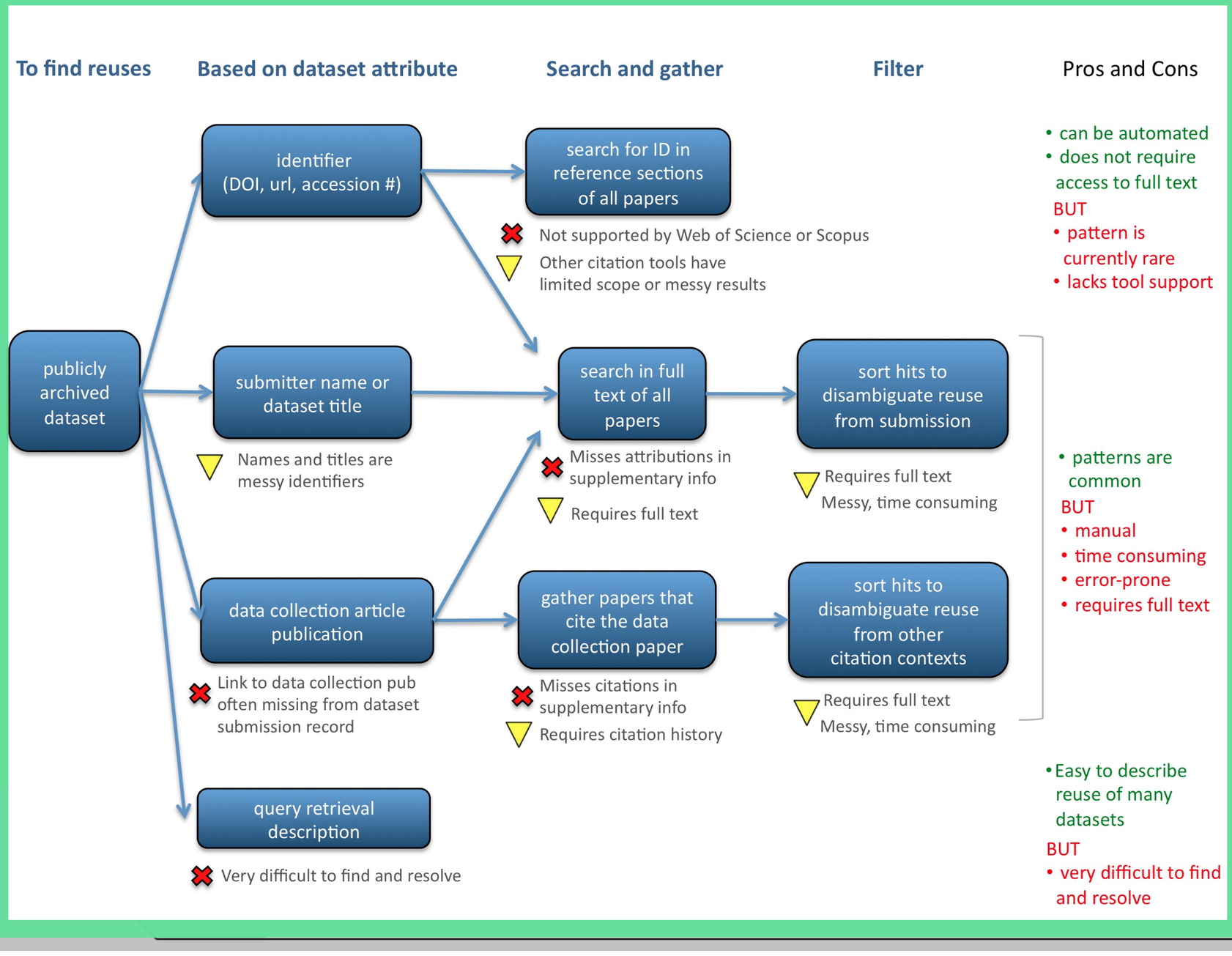
We reviewed 500 articles in six leading environmental science journals, 2000-2010. **Data attribution patterns** varied widely in the body of the articles that reused data (*n*=221):

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



Implications

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



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.

This research was done as open science: ask us about it!

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

For more information: hpiwowar@nescent.org