The Data Citation Community Of Practice (CoP): an Introduction and Summary

Shelley Stall
Senior Director for the American Geophysical Union's Data Leadership Program

Caroline Coward
Library Group Supervisor at NASA Jet Propulsion Laboratory

Chris Erdmann
Assistant Director of Data Stewardship
American Geophysical Union's Data Leadership Program

Deb Agarwal
Data Science & Technology Department
Lawrence Berkeley National Laboratory
Inria, Rennes
Berkeley Institute for Data Science

Motivation – Tackle Remaining Data Citation Challenges

- Extensive work on citations exists
- Still difficult in some cases to have proper data citation
- AGU 2020 – First meeting – Citation of large numbers of datasets in papers
- Began meeting in April 2020
- Process
    - Define problem
    - Get buy-in from broad community (publishers, indexers, authors, technologists, digital library community, data repositories,…)
    - Articulate use cases
    - Review existing mechanisms that could help
    - Prototype potential solutions

First CoP Challenge:

Enabling Large Numbers of Dataset Citations

The data citation community of practice

Academic publisher

- AGU
- JATS4R
- Citation styles
- Outreach/guidance/training

Infrastructure

- DataCite
- CrossRef
- Scholix / OpenAire
- RO-Crate
- Zenodo
- Schema.org
- Web of Science

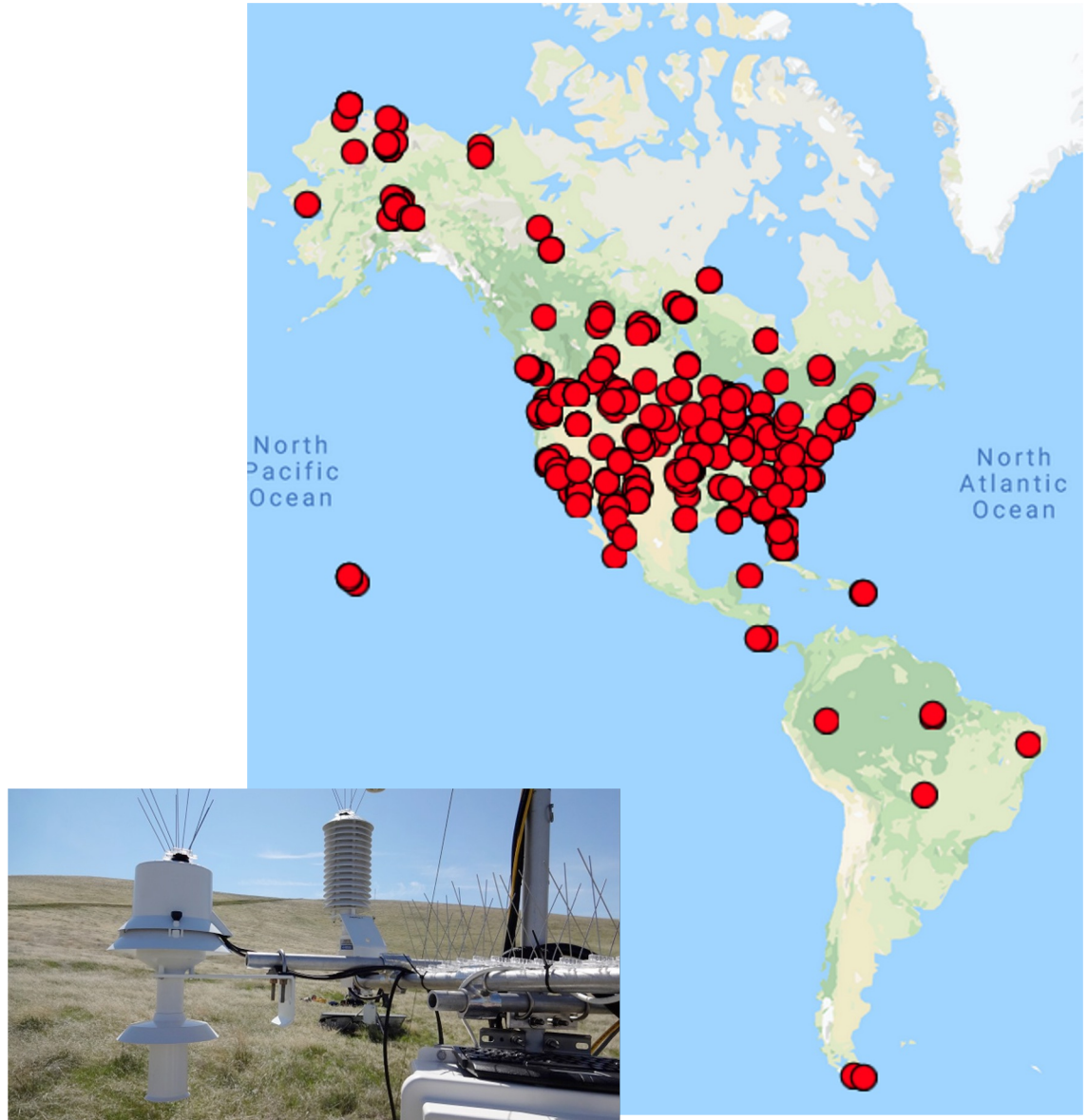
Community use cases

- BioStudies
- Global Biodiversity Information Facility (GBIF)
- PANGAEA
- Intergovernmental panel on climate change (IPCC)
- British Oceanographic Data Centre

Use Case: AmeriFlux Datasets

AmeriFlux (Americas)

- Downloadable data from 426 sites
- Total unique downloads = 26,675 (since November 2015)
- Data updated quarterly/yearly
- Data citation policy = one per site used



Use Cases – Oceanographic Campaign and Large-Scale Data Analysis

British Oceanographic Data Center

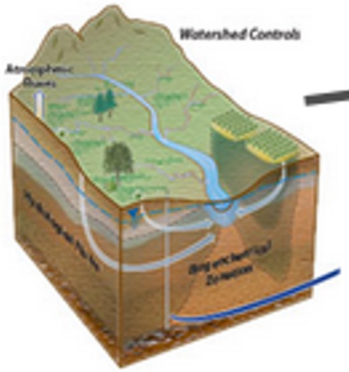
Justin Buck & James Ayliffe

Intergovernmental Panel on Climate Change

Martina Stockhause

An Example Problem

210 datasets



70 datasets

USGS
science for a changing world

1 dataset

Objective

Enable citation of a large numbers of papers, software, and datasets (research objects) in a paper by providing a means to collapse them into a small number of references.

Allow citation of the group of research objects and/or the constituents within the group

Empower an individual to create a group of research objects that might span repositories

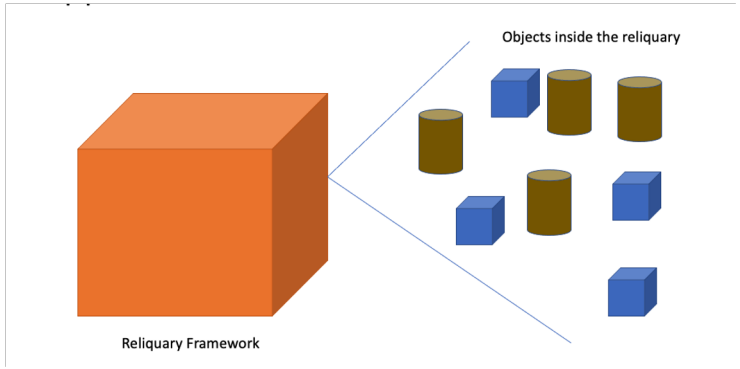
Ideally enable both credit tracking and reuse

*For the purposes of this discussion, we will refer to a group of research objects as a **reliquary** - an object collection where precious objects (relics) are held for posterity.*

Starting Assumptions

- For the near term, the limits on references in the main paper will remain
- Credit is important to research object publishers
- It is important to enable citation of individual research objects
- Citations of research objects and groups of research objects need to be countable without requiring extraordinary effort on the part of an individual or repository
- The solution needs to be easy to use

Proposed Properties of a Reliquary



- Has a unique identifier and well-defined citation components.
- Each object in a reliquary has a unique identifier and descriptive metadata in the reliquary.
- The type and role of objects in the reliquary can be designated from well defined vocabularies.
- The reliquary appears in the reference section of a paper and each citation of the reliquary indicates what subset of the reliquary is being cited if it is not the whole reliquary.
- A reliquary can contain items from many sources and repositories.
- Tracking of primary citation credit for objects in a reliquary is possible without additional help from the entity that created the reliquary.
- Access to an object in the reliquary possible from the description of the object.

Discussed Existing Mechanisms

- Collections - definition of a group of objects (DataCite)
- RO-Crate – definition of a research object
- Scalable dynamic data citations – query-based collection
- Data papers – adds methods and explanations +++
- Structured supplementary material for papers
- Zenodo
- ….

Building Rough Reliquary Examples

	Reliquary DOI	
	https://doi.org/10.23.289/04380000 (leads to a landing page with the information below)	
	Reliquary Citation	
	AmeriFlux Management Project (2021), "November 2021 Reliquary containing all AmeriFlux sites published under the AmeriFlux CC-BY-4 policy," (reliquary), https://doi.org/10.23.289/04380000	
	Reliquary Author	
	AmeriFlux Management Project	
	Reliquary Title	
	November 2021 Reliquary containing all AmeriFlux sites published under the AmeriFlux CC-BY-4 policy,	
	Description	
	This reliquary contains citations for all of the published data covering the AmeriFlux carbon flux tower measurement sites that were available under the AmeriFlux CC-BY-4.0 policy at ameriflux.lbl.gov as of the November 2021 publish of AmeriFlux data. This reliquary should only be used for citation if all AmeriFlux sites used in the work are available under the AmeriFlux CC-BY-4.0 policy. If any AmeriFlux site used in the work is only available under the AmeriFlux Legacy policy, use the Appropriate Legacy reliquary for citation of the sites.	
	Data Availability Statement	
	The data cited in this reliquary are openly available from the AmeriFlux Management Project Data System at https://ameriflux.lbl.gov/.	
Reliquary ID	**Reliquary Citations**	DOI
1	Lars Kutzbach (2021), AmeriFlux BASE AR-TF1 Rio Moat bog, Ver. 2-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1543389	https://doi.org/10.17190/AMF/1543389
2	Lars Kutzbach (2019), AmeriFlux BASE AR-TF2 Rio Pipo bog, Ver. 1-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1543388	https://doi.org/10.17190/AMF/1543388
3	Antonio Antonino (2019), AmeriFlux BASE BR-CST Caatinga Serra Talhada, Ver. 1-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1562386	https://doi.org/10.17190/AMF/1562386
4	George Vourlitis, Higo Dalmagro, JoseÃÅ de S. Nogueira, Mark Johnson, Paulo Arruda (2019), AmeriFlux BASE BR-Npw Northern Pantanal Wetland, Ver. 1-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1579716	https://doi.org/10.17190/AMF/1579716
5	Aaron Todd, Elyn Humphreys (2018), AmeriFlux BASE CA-ARB Attawapiskat River Bog, Ver. 1-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1480319	https://doi.org/10.17190/AMF/1480319
6	Aaron Todd, Elyn Humphreys (2018), AmeriFlux BASE CA-ARF Attawapiskat River Fen, Ver. 1-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1480318	https://doi.org/10.17190/AMF/1480318
7	T. Andrew Black (2018), AmeriFlux BASE CA-Ca1 British Columbia - 1949 Douglas-fir stand, Ver. 1-5, AmeriFlux AMP, (Dataset). https://doi.org/10.17190/AMF/1480300	https://doi.org/10.17190/AMF/1480300
	T. Andrew Black (2018), AmeriFlux BASE CA-Ca2 British Columbia - Clearcut Douglas-fir stand (harvested winter 1999/2000), Ver. 1-5,	

SubGroup Participants

Name	Institution
Deb Agarwal	Berkeley Lab
Stian Soiland-Reyes	University of Manchester
James Ayliffe	British Oceanographic Data Centre, NOC
Justin Buck	BODC, NOC
Chris Erdmann	AGU
Carole Goble	The University of Manchester, ELIXIR-UK, RO-Crate
Ugis Sarkans	EBI
Martina Stockhause	DKRZ / IPCC DDC
Shelley Stall	AGU
Martin Fenner	
Uwe Schindler	PANGAEA

CoP Website: https://data.agu.org/DataCitationCoP/

Contact info: sstall@agu.org, cerdmann@agu.org, daagarwal@lbl.gov, and caroline.m.coward@jpl.nasa.gov