Focusing on Scalable Citations to Improve Data Usability and FAIRness

Deb Agarwal (daagarwal@lbl.gov), Martina Stockhause (stockhause@dkrz.de),
Lesley Wyborn (lesley.wyborn@anu.edu.au), Shelley Stall (SStall@agu.org)

Limits on the number of references in a publication often lead to references being included in the text or in supplementary material. This means references to supporting individual datasets, software, samples are hard for authors and funders of that object to track usage and measure impact of their research.

A method is needed for handling citation of large numbers of objects, particularly datasets, subsets of datasets, software, and physical samples, in scholarly work. These objects are usually stored in several different repositories. The method should allow the citation of any individual object to be counted as a primary citation (as if they were cited individually in the reference section). The solution should also enable aggregators of large numbers of objects to publish '*object collections*'/'*reliquaries*' that can be included in the reference section of a scholarly work without losing the primary citation credit and the ability to track credit for the underlying object contributions.

The need for this method arises from a common use case, in which an author collects large numbers (e.g., 50 to millions+) of existing objects (e.g., datasets, software, or physical samples), or subsets of objects. Each object has its own creators, repositories, funders and persistent identifiers, but collectively the objects are used together in a single analysis. The ability to trace references to datasets used is critical to the reuse part of FAIR and enhances the transparency of research outputs.

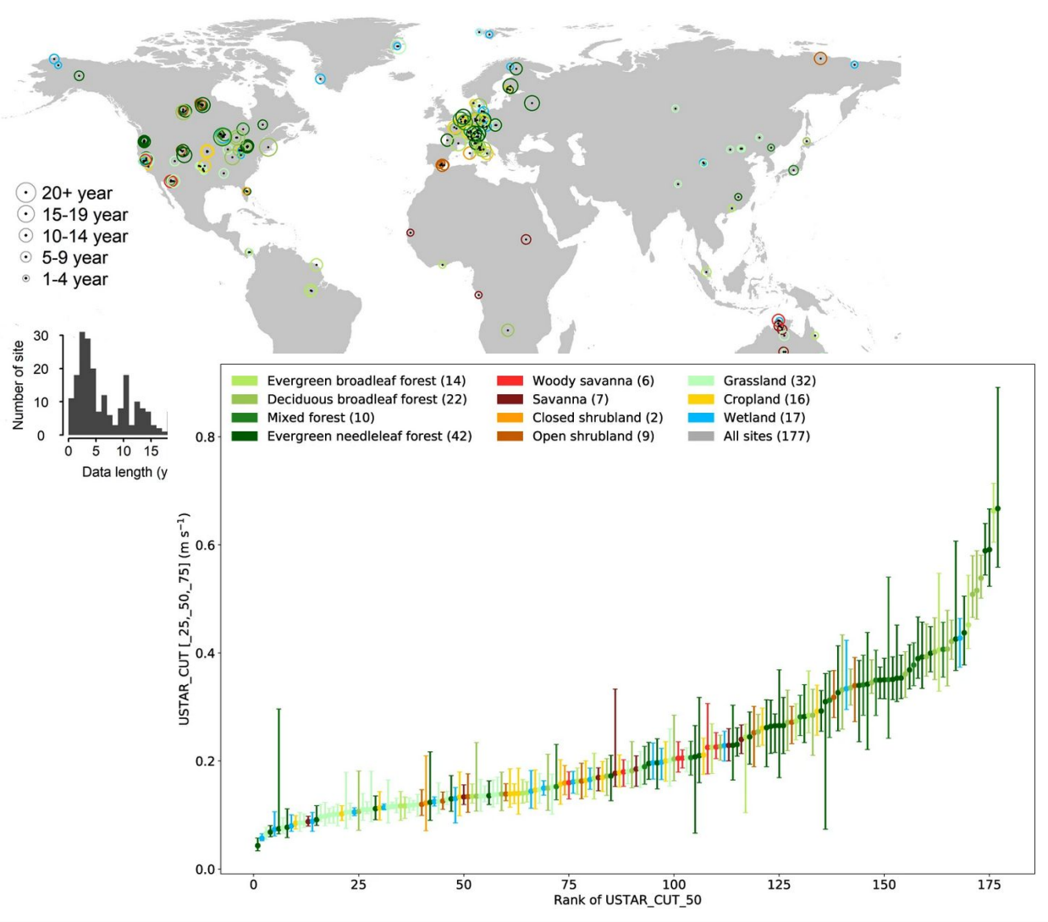
A community of practice on data citations which has been meeting for two years to address this pressing problem. The community is a world-wide group that includes scientists, repositories, publishers, and indexers. The group has identified citation use cases that exemplify the properties needed from across earth sciences and developed a 'cocktail napkin' concept for the new object collection and an idea for how it will work. Within the RDA Complex Citation WG, we reach out to other domains for additional use cases. We are also working with technology providers to investigate whether the concept can be realized using expanded versions of existing mechanisms. This poster describes the motivation, use cases, and current concept for solving the challenge.

OBJECTIVE

- Enable primary citation of a large numbers of papers, software, and datasets and subsets of data (research objects) in a paper by providing a means to collapse them into a small number of references.
- Allow direct citation of the group of research objects and/or the constituents within the group
- Empower an individual to directly cite large numbers of research objects that span multiple repositories
- Enable appropriate tracking of these direct citations

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

MOTIVATION



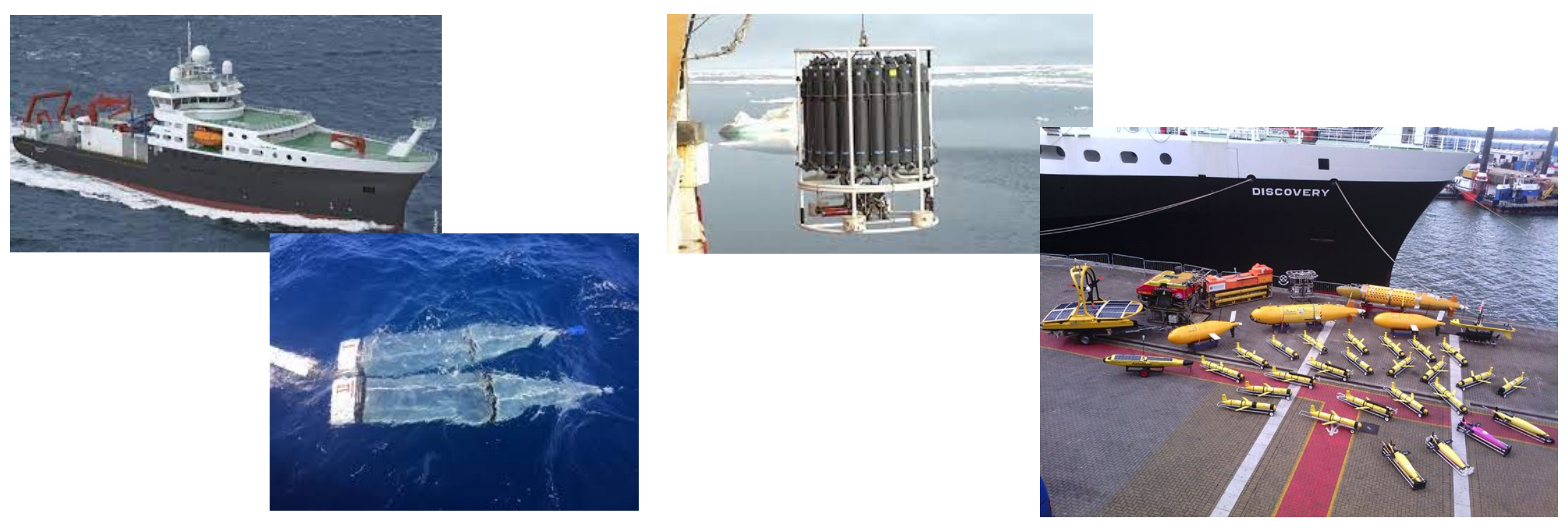
Example – Global FLUXNET2015

- The graphs below combine (left 212 and to the right 177) separate citable data sources
  The paper where this graph appeared had 259 references without referencing the data
- A highly cited data paper (>= 500 citations) on FLUXNET was published, having only the senior contributors to each dataset as authors (Pastorello, G., Trotta, C., Canfora, E. et al., 2020. https://doi.org/10.1038/s41597-020-0534-3).

EXAMPLE USE CASES

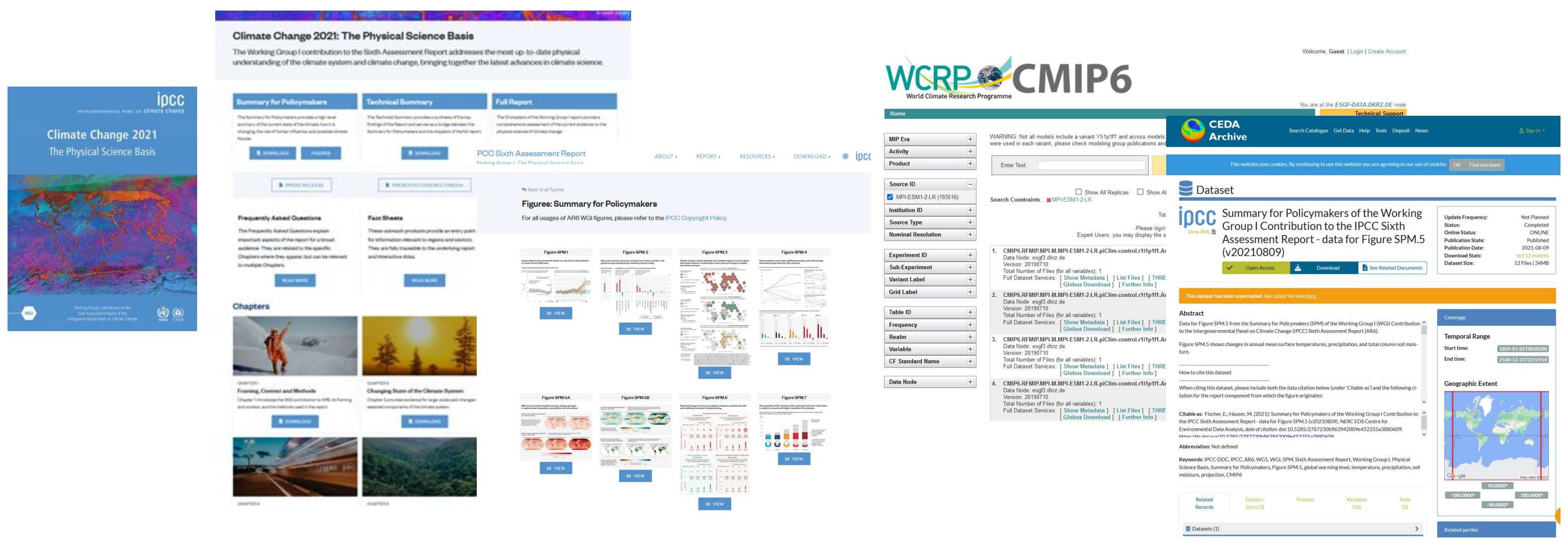
British Oceanographic Data Center: Citing subsets of a larger dataset

- Data and samples from many research cruises and other activities
- Datasets are organized in data collections
- Citation of the dataset subsets across several data collections should be enabled



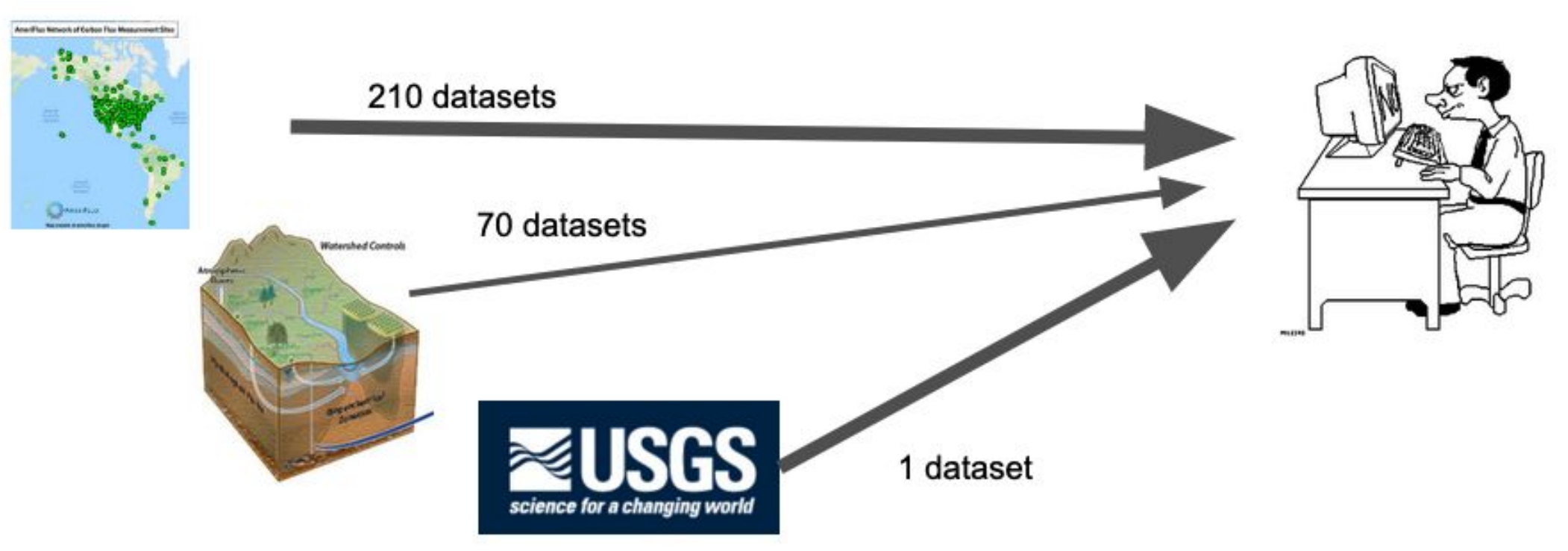
Intergovernmental Panel on Climate Change: Tracing results/figures back to their origins

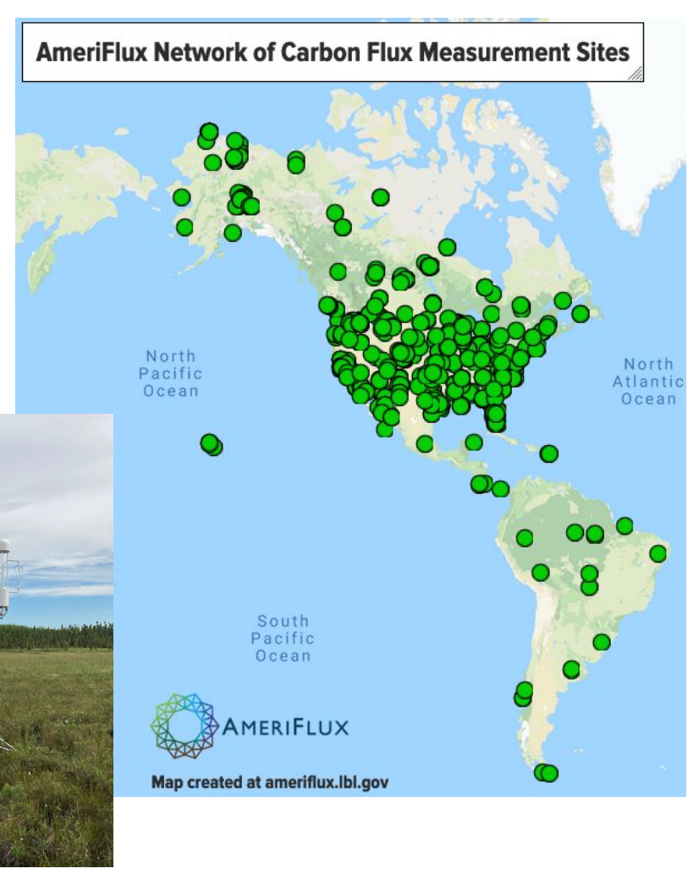
- IPCC figures and other outcomes combine input data (subsets) from many repositories applying scripts to create figure datasets
- Citation of input data (subsets) and code should be enabled through a single reference in the caption
- Figure reproducibility requires information on inputs, code and provenance.


Climate Change 2021: The Physical Science Basis
ipcc
WCRP CMIP6
CEDA Archive
Summary for Policymakers of the Working Group I Contribution to the IPCC Sixth Assessment Report - data for Figure SPM.5

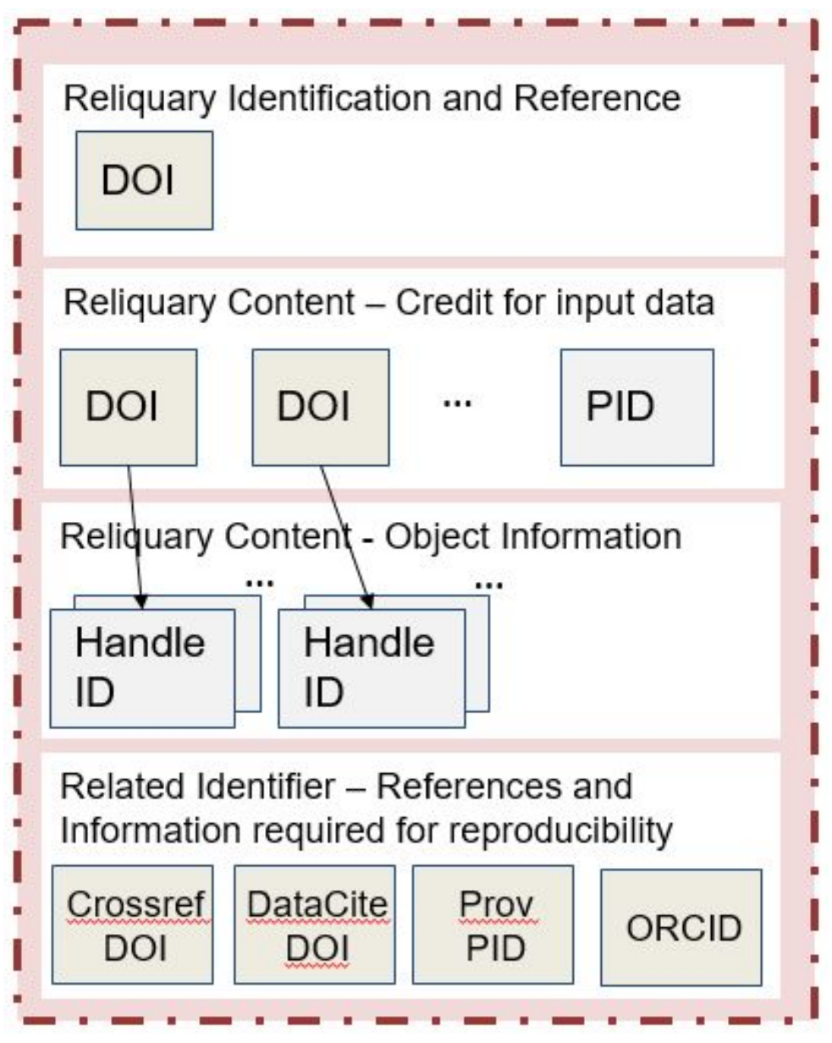
Large Collaborative Earth Science Projects: Citing many datasets across repositories

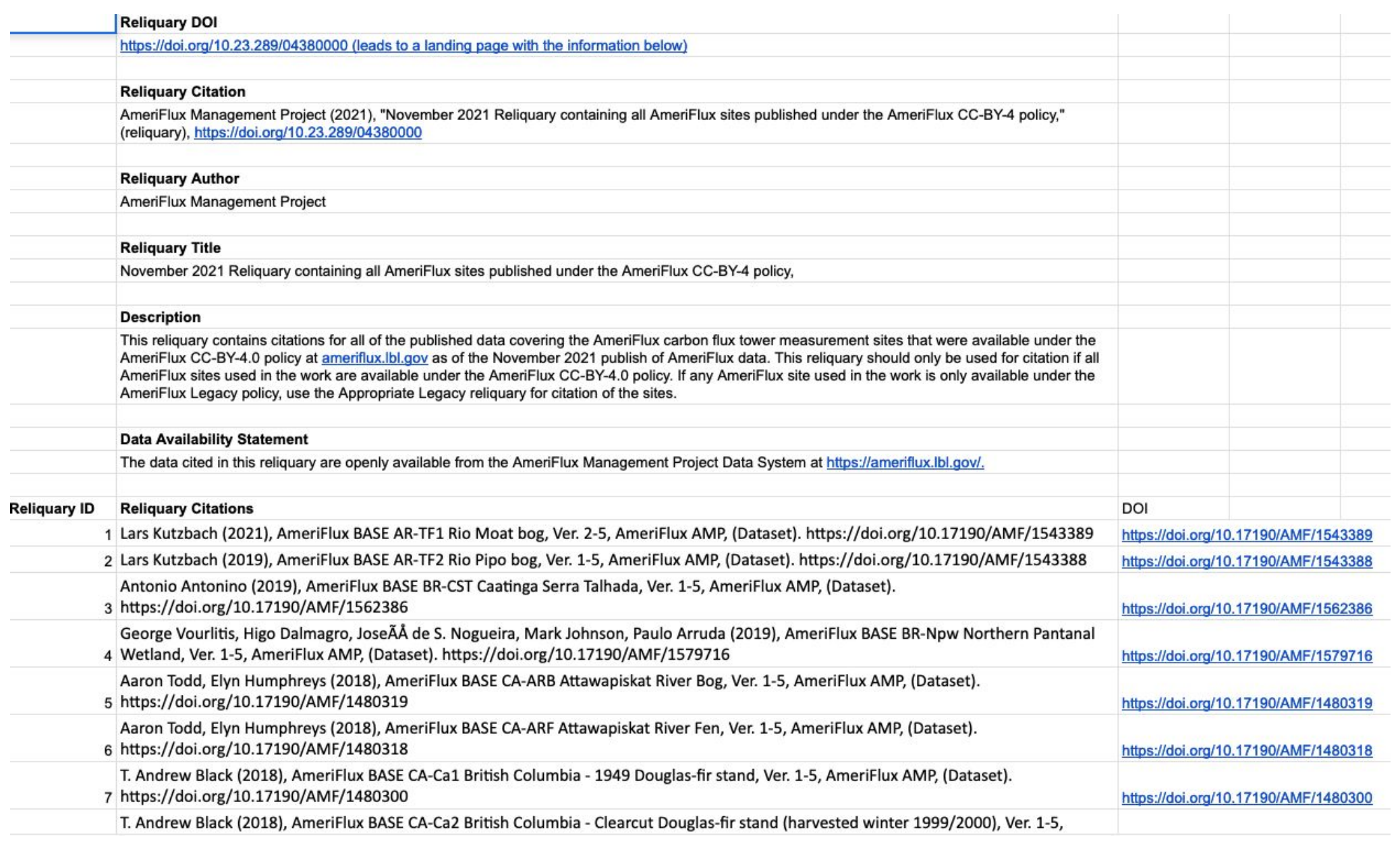
- 100s of datasets are used together
- Citations should enable to give credit to every dataset
- Citations should enable data reusability




AmeriFlux Network of Carbon Flux Measurement Sites
North Pacific Ocean
North Atlantic Ocean
South Pacific Ocean
AmeriFlux
Map created at ameriflux.lbl.gov

'COCKTAIL NAPKIN' RELIQUARY


Reliquary Identification and Reference
DOI
Reliquary Content – Credit for input data
DOI
DOI
...
PID
Reliquary Content - Object Information
...
...
Handle ID
Handle ID
Related Identifier – References and Information required for reproducibility
Crossref DOI
DataCite DOI
Prov PID
ORCID



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

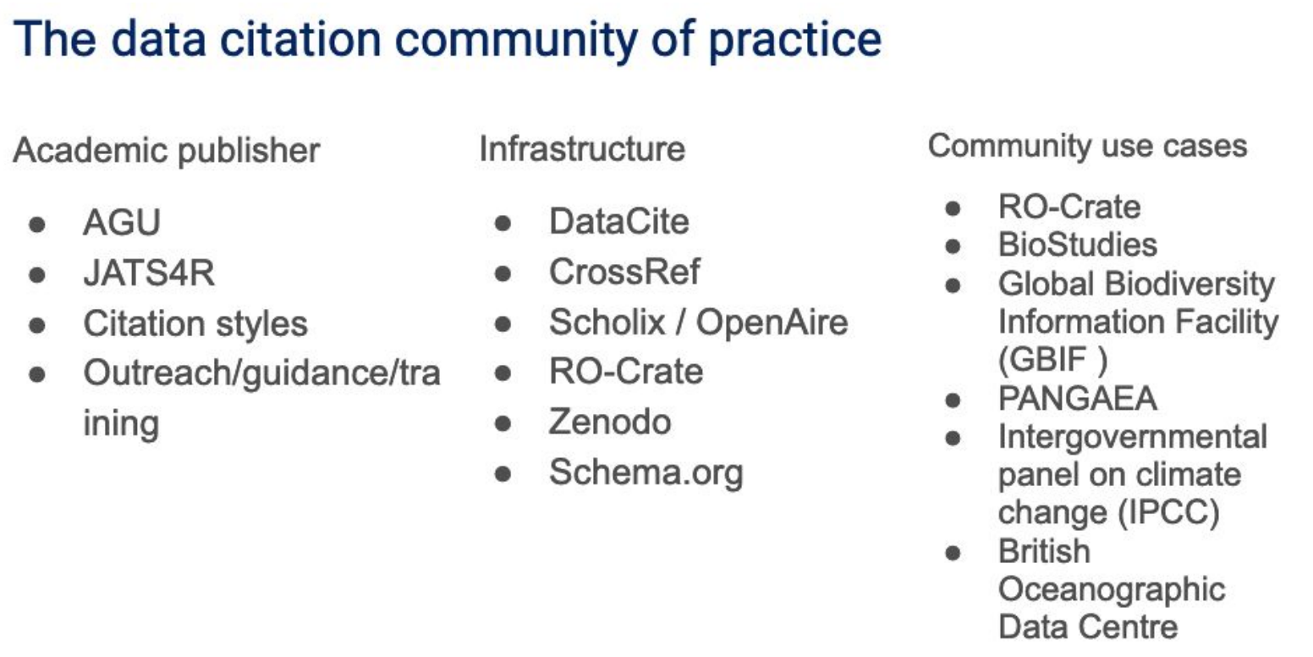
Reliquary Title
November 2021 Reliquary containing all AmeriFlux sites published under the AmeriFlux CC-BY-4 policy,

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,	

COMMUNITY OF PRACTICE PARTICIPANTS



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

Community use cases

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

Why now?

- Journals are pushing back on large (50+) numbers of data citations.
- Research reproducibility packages that include the data, software, physical samples and more need ways to be cited.
- Need for attribution and usage tracing for datasets, software, physical samples that are "children" of a "reliquary".

ENGAGE

Connect with us through:

1. The AGU Data Citation Community of Practice
   https://agu-data.github.io/DataCitationCoP/
2. Research Data Alliance (RDA) Complex Data Citation Working Group
   https://rd-alliance.org/groups/complex-citations-working-group (Email sign up)

U.S. DEPARTMENT OF ENERGY | Office of Science
DKRZ
AGU ADVANCING EARTH AND SPACE SCIENCE
NCI AUSTRALIA
BERKELEY LAB Lawrence Berkeley National Laboratory