

# ESS-DIVE Data Citations Webinar

Joan Damerow and Deb Agarwal

Please fill out survey at [PollEv.com/essdive](https://PollEv.com/essdive)



February Community Webinar

# Webinar Overview

- **Data Contributors:** Sharing/Publishing Data and Citations
- Dataset citation purpose
- Recommendations and examples of data citations
- **Data Users:** Challenges and options for citing large numbers of datasets

*Data citation promotes data sharing, is often a legal requirement, and arguably essential to fully understand and judge scientific conclusions*

# ESS-DIVE Data Citations Pre-Survey

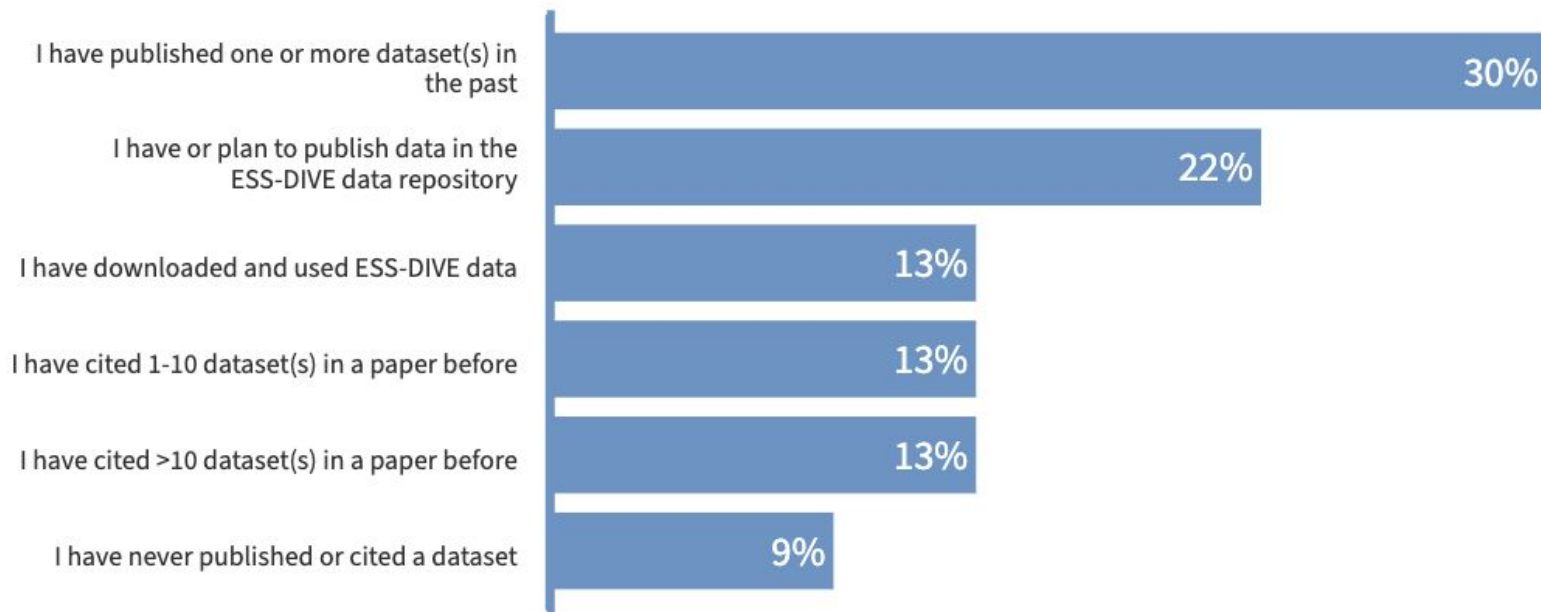
When survey is active, respond at [Pollev.com/essdive](https://pollev.com/essdive)

**0 done**

 **1 underway**

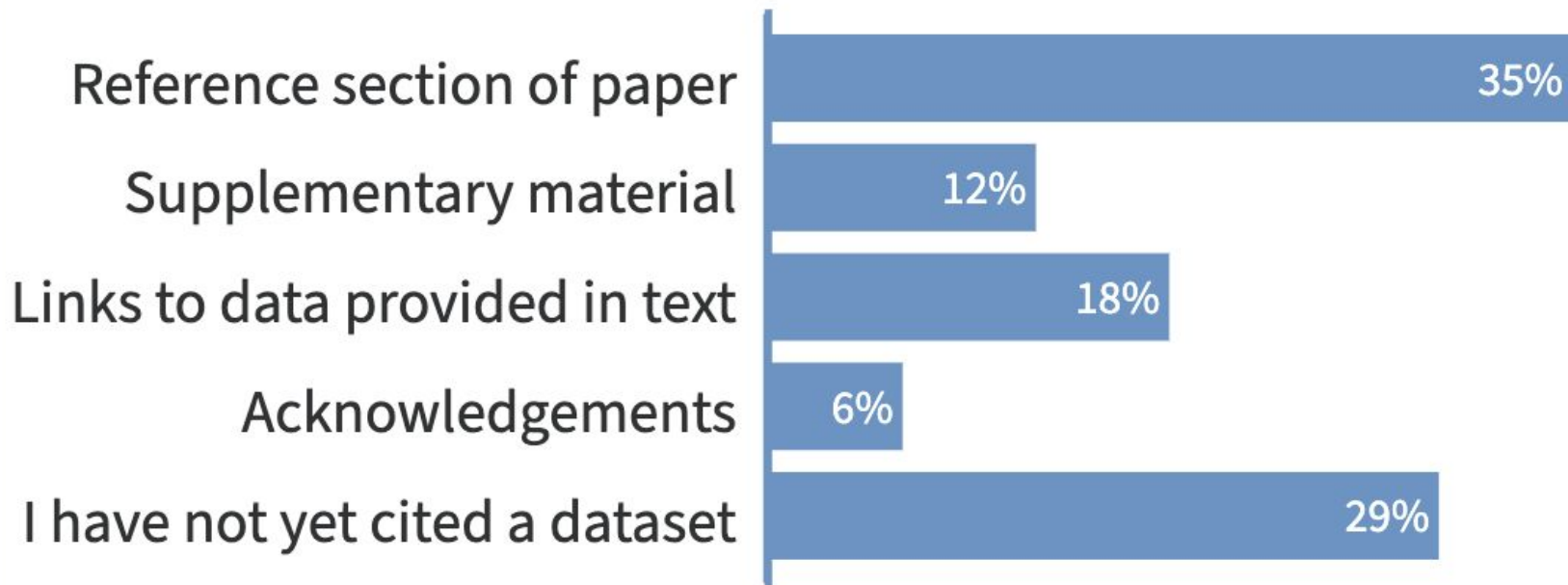


# What is your experience with data publication and citation?

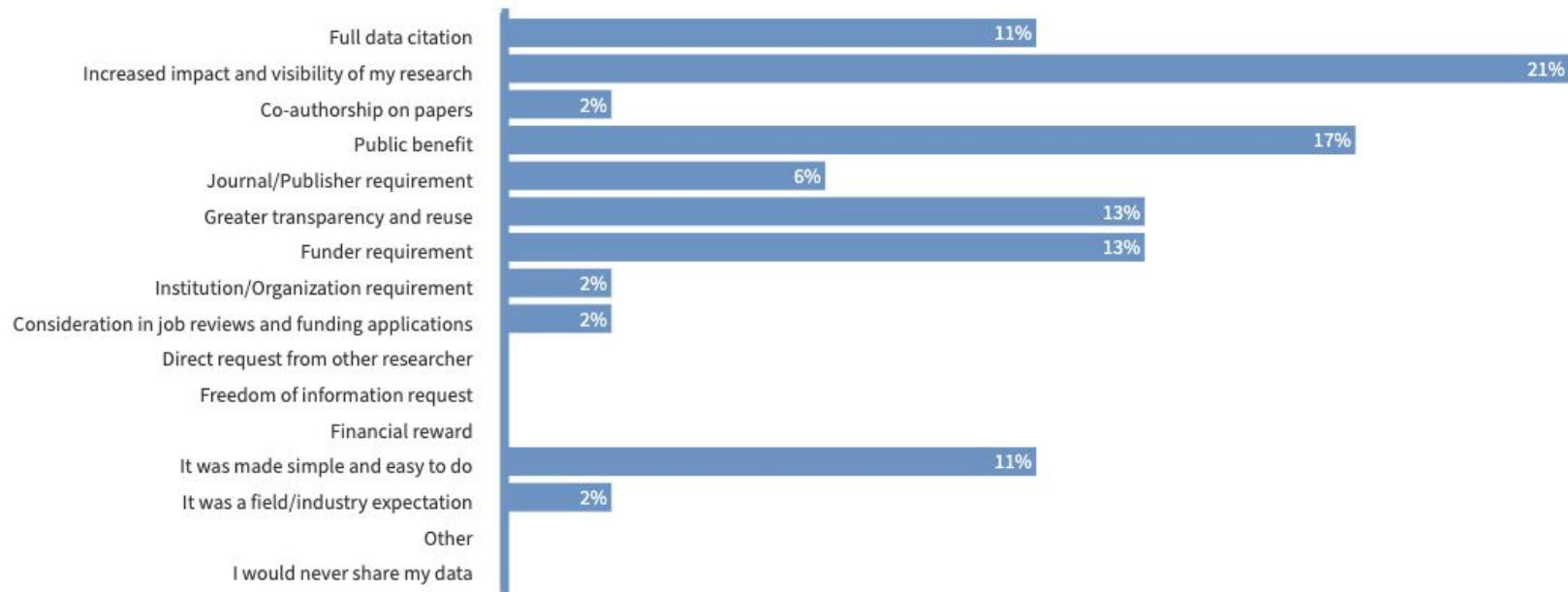


Answers to this poll are anonymous

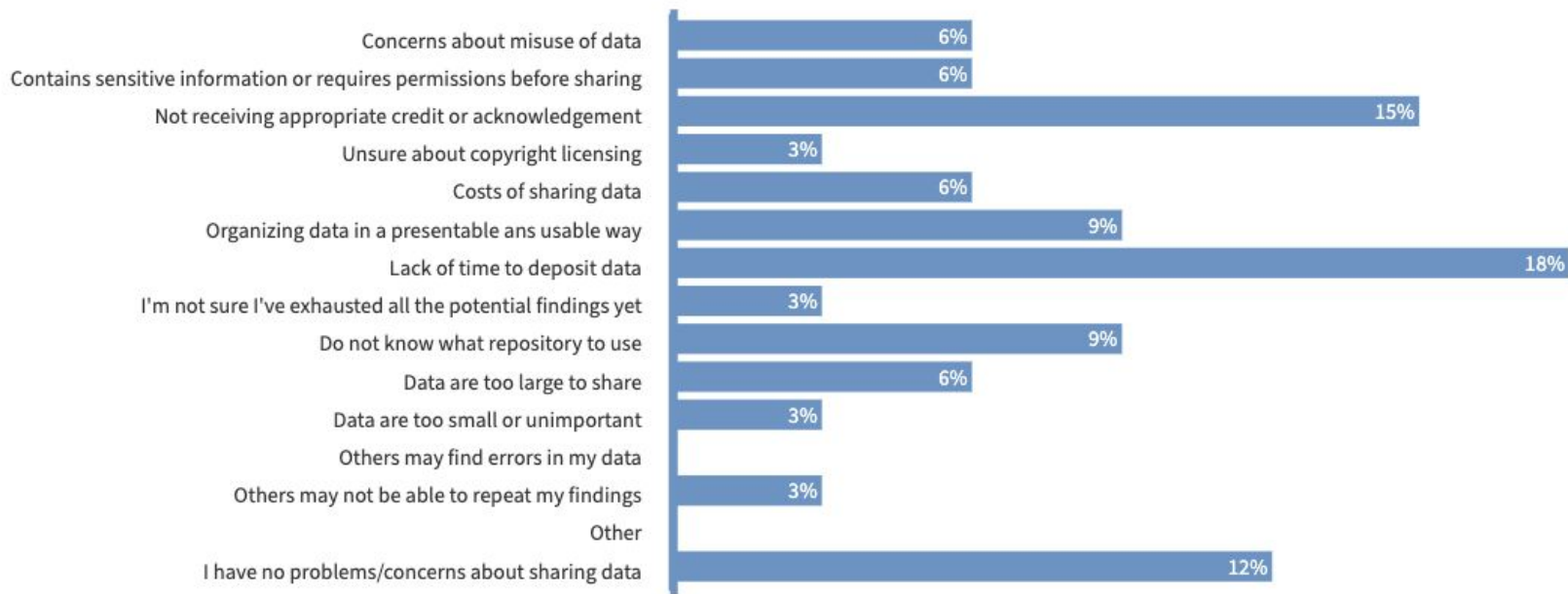
# Where have you cited data thus far in your publications?



# What circumstances would motivate you to share data?



# What problems/concerns, if any, do you have with sharing datasets?



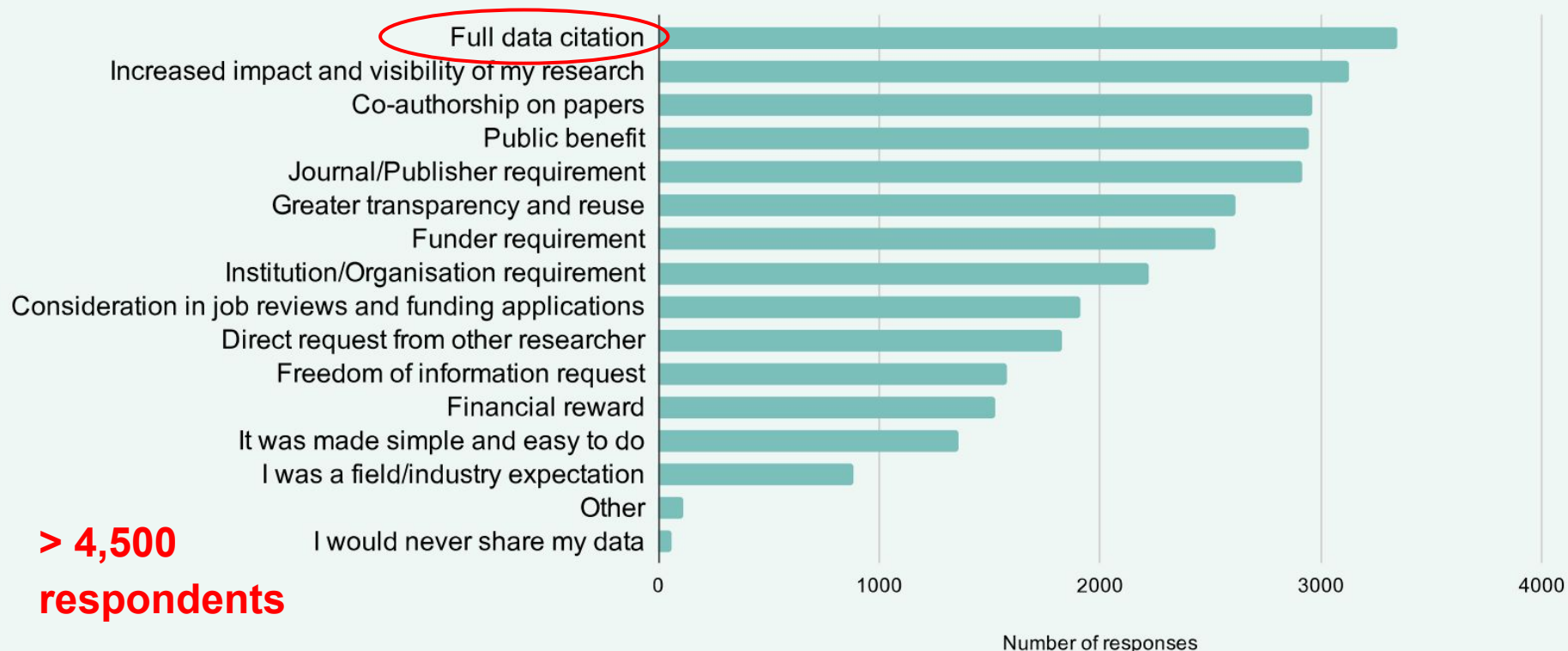
Answers to this poll are anonymous

# Data Contributors/Authors: Sharing Data and Citations



# Longitudinal Survey on Sharing Data 2020: Motivation

### What circumstances would motivate you to share your data?



# Longitudinal Survey on Sharing Data 2020: Concerns

What problems/concerns, if any, do you have with sharing datasets?



> 4,500  
respondents

# Sharing Data: ESS-DIVE Data Packages



Data Packages / Datasets have two primary components

- **Data Files**
- **Metadata:** a collection of information that describes the **content** and **scope** of the data files

**DOI** - permanent identifier and link to the data package

A screenshot of the ESS-DIVE website showing a data package page. The page header includes the ESS-DIVE logo and navigation links for DATA, SUPPORT, ABOUT, and a 'Submit Data' button. Below the header, the package title is 'Boden T ; Marland G ; Andres R J (1999): Global, Regional, and National Fossil-Fuel CO2 Emissions (1751 - 2014) (V. 2017). Carbon Dioxide Information Analysis Center (CDIAC), Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States), doi:10.3334/CDIAC/00001\_V2017'. Below the title, there are statistics for Citations (53), Downloads (4.9K), and Views (21.2K). There are also buttons for 'Copy Citation', 'Assessment report', and 'Edit'. The main content area is titled 'General' and contains a table with the following information:

Identifier	doi:10.3334/CDIAC/00001_V2017
Alternate Identifier	osti:1389331
Alternate Identifier	doi:10.3334/CDIAC/00001_V2017
Alternate Identifier	cdiac:doi 10.3334/CDIAC/00001_V2017
Abstract	Publications containing historical energy statistics make it possible to estimate fossil fuel CO2 emissions back to 1751. Etemad et al. (1991) published a summary compilation that tabulates coal, brown coal, peat, and crude oil production by nation and year. Footnotes in the Etemad et al. (1991) publication extend the energy statistics time series back to 1751. Summary compilations of fossil fuel trade were published by Mitchell (1983, 1992, 1993, 1995). Mitchell's work tabulates solid and liquid fuel imports and exports by nation and year. These pre-1950 production and trade data were digitized and CO2 emission calculations were made following the procedures discussed in Marland and Rotty (1984) and Boden et al. (1995). Further details on the contents and processing of the historical energy statistics are provided in Andres et al. (1999). The 1950 to present CO2 emission estimates are derived primarily from energy statistics published by the United Nations (2017), using the methods of Marland and Rotty (1984). The energy statistics were compiled primarily from annual questionnaires distributed by the U.N. Statistical Office and supplemented by official national statistical publications. As stated in the introduction of the Statistical Yearbook, "in a few cases, official sources are supplemented by other sources and estimates, where these have been subjected to professional scrutiny and debate and are consistent with other independent sources." Data from the U.S. Department of Interior's Geological Survey (USGS 2017) were used to estimate CO2 emitted during cement production. Values for emissions from gas flaring were derived primarily from U.N. data but were supplemented with data from the U.S. Department of Energy's Energy Information Administration (1994), Rotty (1974), and data provided by G. Marland. Greater details about these methods are provided in Marland and Rotty (1984), Boden et al. (1995), and Andres et al. (1999).

*Boden et al. (1999)<sup>2</sup>*

# Considering citation when deciding what to include in a data package

## Author contributions

Based level of contributor effort for portions of data - **author order**



## Data in a publication

All data (raw or processed) that went into the publication



## Field Campaign or Time Period

Data from a field campaign or season that need to be viewed together



## Data type

Particular data type from a project - e.g. continuously generated sensor data, sample data, data synthesis product



***Related references:*** Use the DOI/citation to link related datasets!

# Establishing the Citation Requirements for Data



All public datasets in ESS-DIVE are shared openly under one of two data usage licenses:

[Creative Commons Attribution \(CC BY 4.0\)](#) requires that the data package be cited by anyone using the data.

[Creative Commons Public Domain Dedication \(CC0 1.0\)](#) dedicates the data to the public domain without restriction.



# Check ESS-DIVE Data Usage Rights



**Data usage rights** determine citation requirement

- Check Data Usage Rights listed at the bottom of every Data Package
- Either Creative Commons **Attribution** (default) or **Public Use**

A screenshot of the ESS-DIVE website interface. At the top, there is a navigation bar with "DATA", "SUPPORT", "ABOUT", "Submit Data", and "Sign in with Orcid". The main content area shows a search result for a dataset by Boden T.; Marland G.; and Andres R. J. (1999). The dataset title is "Global, Regional, and National Fossil-Fuel CO2 Emissions (1751 - 2014) (V. 2017). Carbon Dioxide Information Analysis Center (CDIAC), Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States). doi:10.3334/CDIAC/0001\_V2017". Below the title, there are statistics for Citations (53), Downloads (4.9K), and Views (21.2K). A table lists files in the dataset, including metadata files and data files. The "General" section lists the authors: G. Marland (Principal Investigator) and R. J. Andres (Principal Investigator). The "Funding" section shows "DOE.NONE". The "Data Set Usage Rights" section is highlighted with an orange box and an arrow, showing "This work is licensed under the Creative Commons Attribution 4.0 International License. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/". Below this, the citation "Boden et al. (1999)<sup>2</sup>" is displayed.

# Citations in Data Package: Related References



Place to provide citations for any related references

- Associated paper(s)
- Other datasets used
- Related data
- Methods/protocols...

Change to related identifiers with specific relationships (e.g. associated paper, new version, datasets included...)

## Related References

Gu, L., Pallardy, S. G., Hosman, K. P., and Sun, Y.: Drought-influenced mortality of tree species with different predawn leaf water dynamics in a decade-long study of a central US forest, *Biogeosciences*, 12, 2831-2845, doi:10.5194/bg-12-2831-2015, 2015.

- **Supports multi-disciplinary sciences:** link related and diverse data
- Clear and consistent linking between **papers and datasets**

# Dataset Citation Purpose and Examples



# Purpose of a Data Citation



‘Data citation is a reference to data for the purpose of **credit attribution** and facilitation of **access** to the data’ (TGDCSP 2013: CIDCR6; Parsons et al. 2019)



- Recognize value of data
- In many disciplines the paper alone is not sufficient to understand and judge the strength of scientific conclusions
- Translate attributions into reward for individuals



# Joint Declaration Data Citation Principles

**Importance:** Data should be considered legitimate, citable products of research.

**Credit and Attribution:** facilitate giving scholarly credit and normative and legal attribution to all contributors to the data

**Evidence:** Whenever and wherever a claim relies upon data, the corresponding data should be cited.

**Unique Identification:** persistent method for identification that is machine actionable, globally unique, and widely used

**Access:** facilitate access to the data, metadata, documentation, code, and other materials, to make informed use of the the data.

**Persistence:** Unique identifiers, and metadata describing the data, and its disposition, should persist, even beyond data they describe

**Specificity and Verifiability:** Data citations should facilitate identification of, access to, and verification of the specific data that support a claim.

**Interoperability and Flexibility:** Data citation methods should be sufficiently flexible to accommodate the variant practices...

# Data Citations help make data FAIR



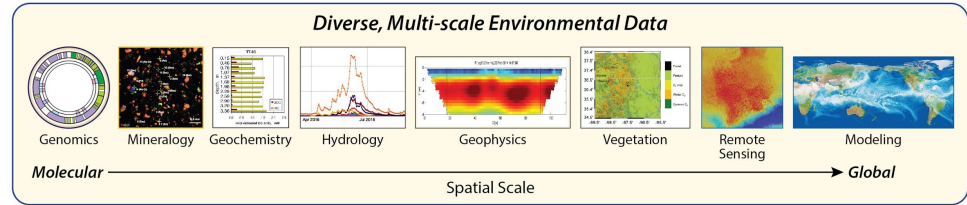
Making data **FAIR**:

**F**indable

**A**ccessible

**I**nteroperable

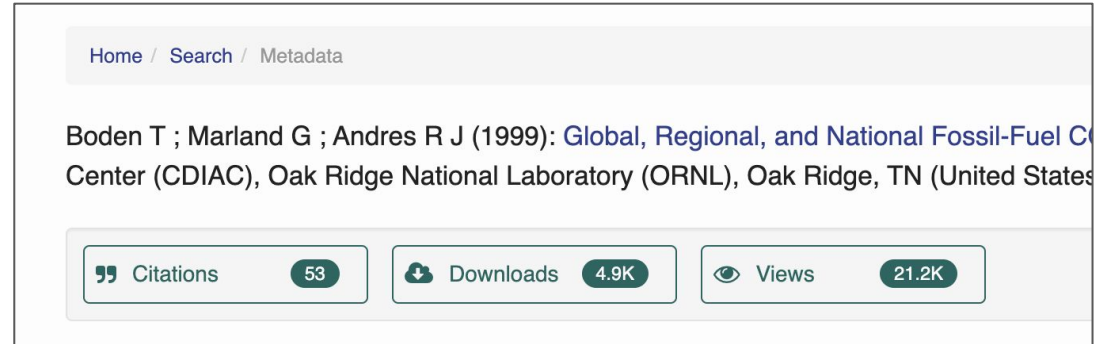
**R**eusable



**Citation Metadata** can make **finding** and **accessing** much easier

# Credit: Citation Metrics

- Every data package has metrics on number of data package views, downloads, and citations
- Citation counts are not fully accurate because we don't always get this information from journals

A screenshot of a data package metadata page. At the top, there is a breadcrumb trail: "Home / Search / Metadata". Below this, the title of the package is displayed: "Boden T ; Marland G ; Andres R J (1999): Global, Regional, and National Fossil-Fuel CO2 Emissions from the Global Energy System (GES) and Energy Infrastructure (EI) Center (CDIAC), Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States)". At the bottom of the screenshot, there are three metric boxes: "Citations" with a value of 53, "Downloads" with a value of 4.9K, and "Views" with a value of 21.2K.

Metric	Value
Citations	53
Downloads	4.9K
Views	21.2K



<https://makedatacount.org/>

When poll is active, respond at [Pollev.com/essdive](https://Pollev.com/essdive)

Text **ESSDIVE** to **22333** once to join

# Should the academic credit system incorporate data authorship and citations, along with papers?

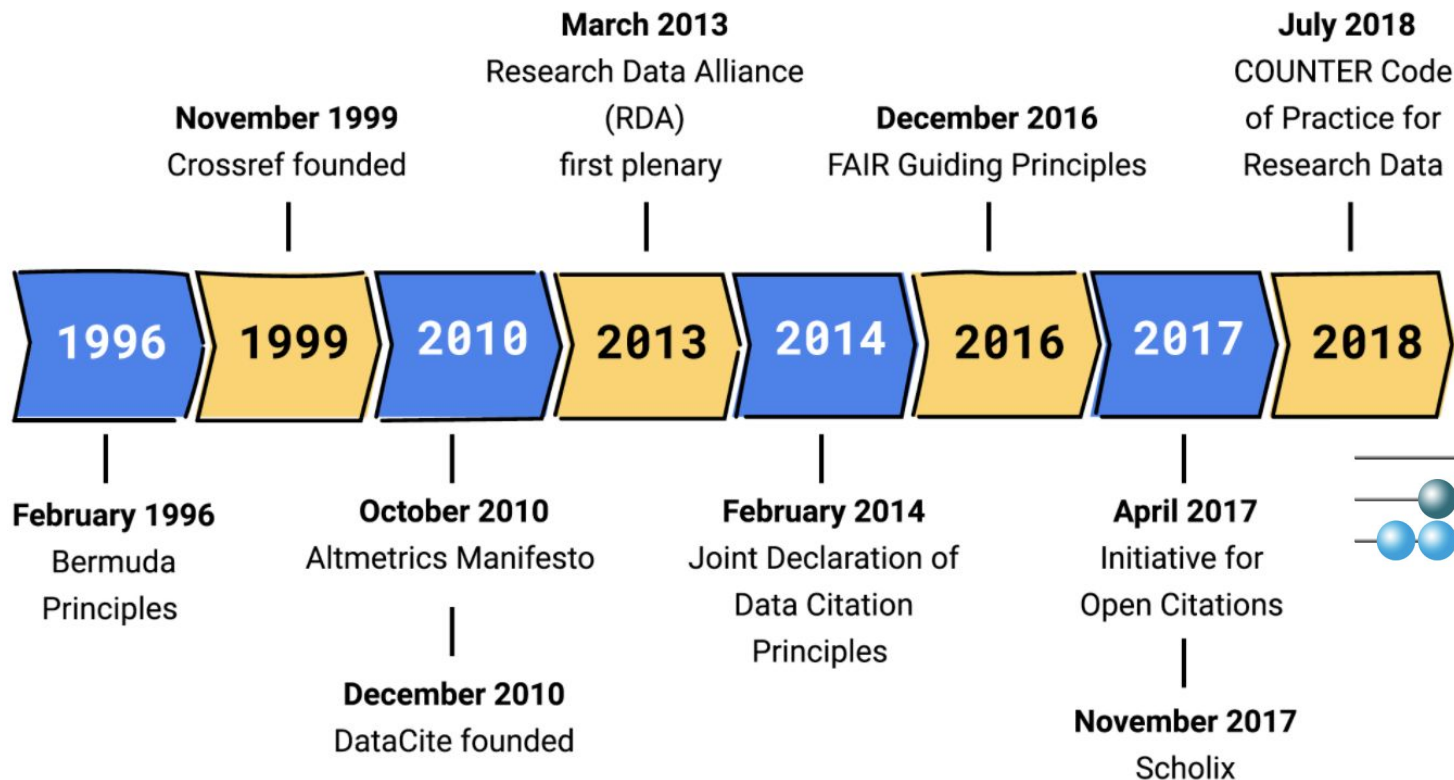
It is critical that dataset authorship and citations are equally counted in the academic credit system

Dataset authorship and citations should be incorporated into the academic credit system, but with less weight than papers

Dataset authorship and citations should not be considered in the academic credit system



# Data Citation and Tracking is Evolving



# Basic Components of a Data Citation

Authors

Date Published

Title

Publisher/Repository: Project

DOI: globally unique persistent ID - access and track impact of a particular dataset over time

Ely K ; Rogers A ; Crystal-Ornelas R  
(2020): ESS-DIVE reporting format for  
leaf-level gas exchange data and  
metadata. ESS-DIVE.  
doi:10.15485/1659484

# Basic Components of a Data Citation

Authors

Date Published

Title

Publisher/Repository

DOI - doesn't change

Rogers D B ; Newcomer M ; Raberg J ; Dwivedi D ; Steefel C ; Bouskill N ; Nico P ; Faybishenko B ; Fox P ; Conrad M ; Bill M ; Brodie E ; Arora B ; Dafflon B ; Williams K ; Hubbard S (2020): Modeling the impact of riparian hollows on river corridor nitrogen exports, *Frontiers in Water: Dataset*. **Watershed Function SFA**. doi:10.15485/1734795



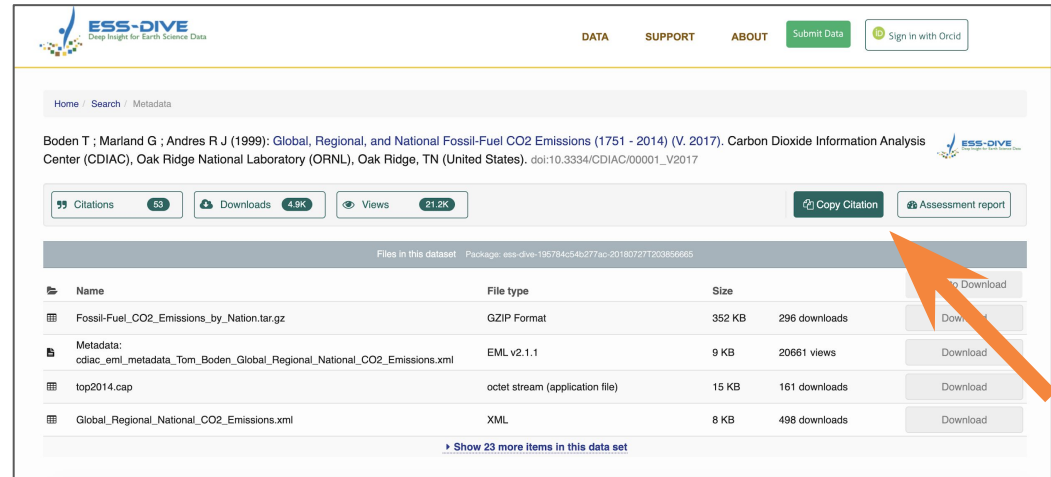
D. Brian Rogers, Michelle Newcomer, Jonathan Raberg, Dipankar Dwivedi, Carl Steefel, et al. 2020. Modeling the impact of riparian hollows on river corridor nitrogen exports, *Frontiers in Water: Dataset*. **ESS-DIVE**. doi:10.15485/1734795, **version: ess-dive-ee844e1c82243f9-20201209T083635575**.



# Citing an ESS-DIVE Data Package

Citing individual data packages is easy on ESS-DIVE

- Use **Copy Citation** button to store Data Package Citation to Clipboard



Home / Search / Metadata

Boden T ; Marland G ; Andres R J (1999): Global, Regional, and National Fossil-Fuel CO2 Emissions (1751 - 2014) (V. 2017). Carbon Dioxide Information Analysis Center (CDIAC), Oak Ridge National Laboratory (ORNL), Oak Ridge, TN (United States). doi:10.3334/CDIAC/00001\_V2017

[Citations](#) 53
 [Downloads](#) 4.9K
 [Views](#) 21.2K
 [Copy Citation](#)
[Assessment report](#)

Name	File type	Size	Downloads	Views
Fossil-Fuel_CO2_Emissions_by_Nation.tar.gz	GZIP Format	352 KB	296 downloads	
Metadata: odiaac_eml_metadata_Tom_Boden_Global_Regional_National_CO2_Emissions.xml	EML v2.1.1	9 KB	20661 views	
top2014.cap	octet stream (application file)	15 KB	161 downloads	
Global_Regional_National_CO2_Emissions.xml	XML	8 KB	498 downloads	

[Show 23 more items in this data set](#)

*Boden et al. (1999)<sup>2</sup>*

***Certain metadata fields determine how the citation will look***

# Potential Changes to ESS-DIVE Citations

Authors

Date Published

Title

Publisher/Repository: add ESS-DIVE

Resource Type

DOI

Accessed via persistent link, date

Iversen C ; Ontl T ; Brice D ; Childs J (2017).

SPRUCE S1 Bog Plant-Available Nutrients

Assessed with Ion-Exchange Resins from

2011-2012 in the Southern End of the S1 Bog.

Climate Change-Terrestrial Ecosystem Science

SFA, ESS-DIVE. Dataset.

doi:10.3334/CDIAC/SPRUCE.022. Accessed via

<https://data.ess-dive.lbl.gov/view/doi:10.3334/CDIA>

C/SPRUCE.022 on 2019-10-21.

# Data Users: Citation Complexities and Challenges

# Data Versions

Versions recorded by access date

- Allow editing - same DOI
- Looking at implementing formal versioning in citation (ESIP 2019)

Research on sample PIDs - sample tracking across facilities, linking related data, citation in the future

<https://github.com/ess-dive-community/essdive-sample-id-metadata>

# Large numbers of related datasets



Balance needs of data producers and users

- **Data producers:** want their data to be **reused** and **want credit**
- **Data users:** need practical guidelines for citing large numbers of datasets and subsets of data packages

DOE data collections from large interdisciplinary teams have **complex citation challenges**

- Options: Thematic data collections, Dynamic DOIs

# AmeriFlux DOIs

- 374 AmeriFlux sites with data
- 2482 Site years of data
- Team members change at each site over the years
- Whole or part of data record is reprocessed often to correct issues
- One DOI per site
- Many analyses use most or all of the data in AmeriFlux - citations are a challenge

*Similar challenge for large ESS projects e.g. NGEEs and SFAs*



# Citing Large Numbers of Datasets

Current practices when large number of datasets used for a paper

- Table of data in paper
- List of DOIs in acknowledgement
- DOI provided in-line
- List in supplementary material
- Author creates a data paper
- No mention

None of these options are indexed by citation trackers like crossref

# Collective Citation Methods Available

- Existing methods for creating collective citations
  - Data Collection - group together several datasets, papers, other items
  - Data Paper - paper describing a number of datasets and data
  - Dynamic Data Citation - a query that identifies a specific retrieval of data
- Properties
  - Single DOI citation to cite in a paper
  - Data citation is through a citation of a citation
  - Most require implementation by data repository
  - Difficult to track specific contributions of different datasets



When poll is active, respond at [Pollev.com/essdive](https://Pollev.com/essdive)

Text **ESSDIVE** to **22333** once to join

# If all the citation tracking systems are in place, how important is it for you to be cited directly, versus being part of a collection or data paper and that being cited?

I want a direct citation when my dataset is included in a collection or data paper (i.e. citation for a collection propagates to each dataset included)

A citation for the collection or data paper is sufficient, if I am an author of the collection or data paper

A citation for the collection or data paper is sufficient, even if I am not an author of the collection or data paper



# Recommendations for Authors and Reviewers of Scientific Papers



- Data used to develop analyses/conclusions in a paper should be cited in the reference section using citation text provided at ESS-DIVE
- Data should ideally be published (publicly available) before citing it in a paper
- Data publications should appropriately credit all the people substantially involved in the creation of the dataset (including processing, QA, analyses, etc)

# ESS-DIVE Next Steps for Citations

- Recommendations for citations
- Broad community discussion of method and tools needed for citing large numbers of datasets
- Citation analysis of ESS-DIVE datasets
  - How ESS-DIVE datasets are cited and reused
  - Characteristics of highly-cited datasets
  - Identify gaps in ability to cite and reuse datasets

# Questions?



Provide ideas/feedback for future webinars:

<https://github.com/ess-dive-community/essdive-webinars-and-events>

Contact us at [ess-dive-support@lbl.gov](mailto:ess-dive-support@lbl.gov)

Join our mailing list [ESS-DIVE Community mailing list](#)

Follow us on Twitter! [twitter.com/ESSDIVE](https://twitter.com/ESSDIVE)

# References and Resources

Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014

[<https://www.force11.org/group/joint-declaration-data-citation-principles-final>].

Agarwal, D., Damerow, J. Varadharajan, C., Christianson, D., Pastorello, G., Cheah, Y., Ramakrishnan. L. 2021. Balancing the needs of consumers and producers for scientific data collections. Ecological Informatics. <https://doi.org/10.1016/j.ecoinf.2021.101251>.

ESIP Data Preservation and Stewardship Committee, 2019. ESIP Data Preservation and Stewardship Committee Data Citation Guidelines for Earth Science Data, Version 2. Earth Science Information Partners (2019), <https://doi.org/10.6084/m9.figshare.8441816>.

T. Weigel, B. Almas, F. Baumgardt, T. Zastrow, U. Schwardmann, M. Hellström, J. Quinteros, D. Fleischer. Recommendation on research data collections. Res. Data Alliance (2017), <https://doi.org/10.15497/RDA00022>.

# ESS-DIVE Glossary



- **DataONE** - The Data Observation Network for Earth (DataONE) is a distributed framework and sustainable cyberinfrastructure that provides open and secure access to Earth observational data. ESS-DIVE is a DataONE member.
- **DOE** - The U.S. Department of Energy (DOE) is a Cabinet-level department of the United States whose mission is to ensure America's security and prosperity by addressing its energy, environmental and nuclear challenges through transformative science and technology solutions.

## ESS-DIVE Glossary (*cont.*)

- **DOI** - A Digital Object Identifier (DOI) is a unique alphanumeric string assigned by a registration agency (e.g., The Office of Scientific and Technical Information (OSTI)) to identify content and provide a persistent link to its location on the internet. ESS-DIVE assigns a DOI when your data package is published and made available electronically.
- **Metadata** - Descriptive information about data / data that provides information about other data.
- **Data Package / Dataset** - Data files with associated descriptive metadata and a DOI.

# ESS-DIVE Glossary (*cont.*)



- **ESS** - Environmental Systems Science (ESS) is a U.S. Department of Energy Office of Science program under the Biological and Environmental Research Program seeking to advance a robust predictive understanding of terrestrial surface and subsurface ecosystems.
- **ESS-DIVE** - Environmental System Science Data Infrastructure for a Virtual Ecosystem (ESS-DIVE) is a U.S. Department of Energy repository for earth and environmental science data, models and software generated from research on terrestrial and subsurface environments.