ESS-DIVE Data Citations Webinar

Joan Damerow and Deb Agarwal

Please fill out survey at PollEv.com/essdive
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
What is your experience with data publication and citation?

- I have published one or more dataset(s) in the past: 30%
- I have or plan to publish data in the ESS-DIVE data repository: 22%
- I have downloaded and used ESS-DIVE data: 13%
- I have cited 1-10 dataset(s) in a paper before: 13%
- I have cited >10 dataset(s) in a paper before: 13%
- I have never published or cited a dataset: 9%

Answers to this poll are anonymous.
Where have you cited data thus far in your publications?

- Reference section of paper: 35%
- Supplementary material: 12%
- Links to data provided in text: 18%
- Acknowledgements: 6%
- I have not yet cited a dataset: 29%

답변은 익명입니다.
What circumstances would motivate you to share data?

- Increased impact and visibility of my research: 21%
- Full data citation: 11%
- Public benefit: 17%
- Journal/Publisher requirement: 6%
- Greater transparency and reuse: 13%
- Funder requirement: 13%
- Institution/Organization requirement: 2%
- Consideration in job reviews and funding applications: 2%
- Direct request from other researcher: 2%
- Freedom of information request: 2%
- Financial reward: 2%
- It was simple and easy to do: 11%
- It was a field/industry expectation: 2%
- Other: 2%
- I would never share my data: 2%

Answers to this poll are anonymous.
What problems/concerns, if any, do you have with sharing datasets?

- Concerns about misuse of data: 6%
- Contains sensitive information or requires permissions before sharing: 6%
- Not receiving appropriate credit or acknowledgement: 15%
- Unsure about copyright licensing: 3%
- Costs of sharing data: 6%
- Organizing data in a presentable and usable way: 9%
- Lack of time to deposit data: 18%
- I’m not sure I’ve exhausted all the potential findings yet: 3%
- Do not know what repository to use: 9%
- Data are too large to share: 3%
- Data are too small or unimportant: 6%
- Others may find errors in my data: 3%
- Others may not be able to repeat my findings: 3%
- Other: 3%
- I have no problems/concerns about sharing data: 12%

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?

- Full data citation
- Increased impact and visibility of my research
- Co-authorship on papers
- Public benefit
- Journal/Publisher requirement
- Greater transparency and reuse
- Funder requirement
- Institution/Organisation requirement
- Consideration in job reviews and funding applications
- Direct request from other researcher
- Freedom of information request
- Financial reward
- It was made simple and easy to do
- I was a field/industry expectation
- Other

> 4,500 respondents

Number of responses
Longitudinal Survey on Sharing Data 2020: Concerns

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

- Concerns about misuse of data
- Contain sensitive information or require permissions before sharing
- Not receiving appropriate credit or acknowledgment
- Unsure about copyright licensing
- Unsure I have permission to share
- Costs of sharing data
- Organising data in a presentable and usable way
- Lack of time to deposit data
- I'm not sure I've exhausted all the potential findings yet
- Not know what repository to use
- Another lab may make a different interpretation
- Data are too large to share
- Data are too small or unimportant
- Others may find errors in my data
- Others may not be able to repeat my findings
- Other
- I have no problems/concerns about sharing data
- I have no desire to share my data

> 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

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

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

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

Boden et al. (1999)²
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]


- 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 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**: Findable, Accessible, Interoperable, Reusable

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.

https://makedatacount.org/
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

- **November 1999**: Crossref founded
- **February 1996**: Bermuda Principles
- **October 2010**: Altmetrics Manifesto
- **December 2010**: DataCite founded
- **1996**:
- **1999**:
- **2010**:
- **2013**:
- **2014**:
- **2016**:
- **2017**:
- **2018**:
- **March 2013**: Research Data Alliance (RDA) first plenary
- **December 2016**: FAIR Guiding Principles
- **July 2018**: COUNTER Code of Practice for Research Data

Credit: [https://makedatacount.org/](https://makedatacount.org/)
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
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

Date Published
24

Title
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

Certain metadata fields determine how the citation will look

Boden et al. (1999)²
## Potential Changes to ESS-DIVE Citations

<table>
<thead>
<tr>
<th>Authors</th>
<th>Date Published</th>
<th>Title</th>
<th>Publisher/Repository:</th>
<th>Resource Type</th>
<th>DOI</th>
<th>Accessed via persistent link, date</th>
</tr>
</thead>
</table>
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
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

Join our mailing list ESS-DIVE Community mailing list

Follow us on Twitter! 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].


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 - 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.