Advancing Open Science and Rapid Data Access Using Preprint and Experimental Datasets

Bruce E. Wilson[1], Chris Lindsley[1], Debjani Singh[1], Michele M Thornton[1], Tammy Walker[1], Yaxing Wei[1], and Daine Wright[1]

[1]Affiliation not available

January 15, 2024

IN41B-0583

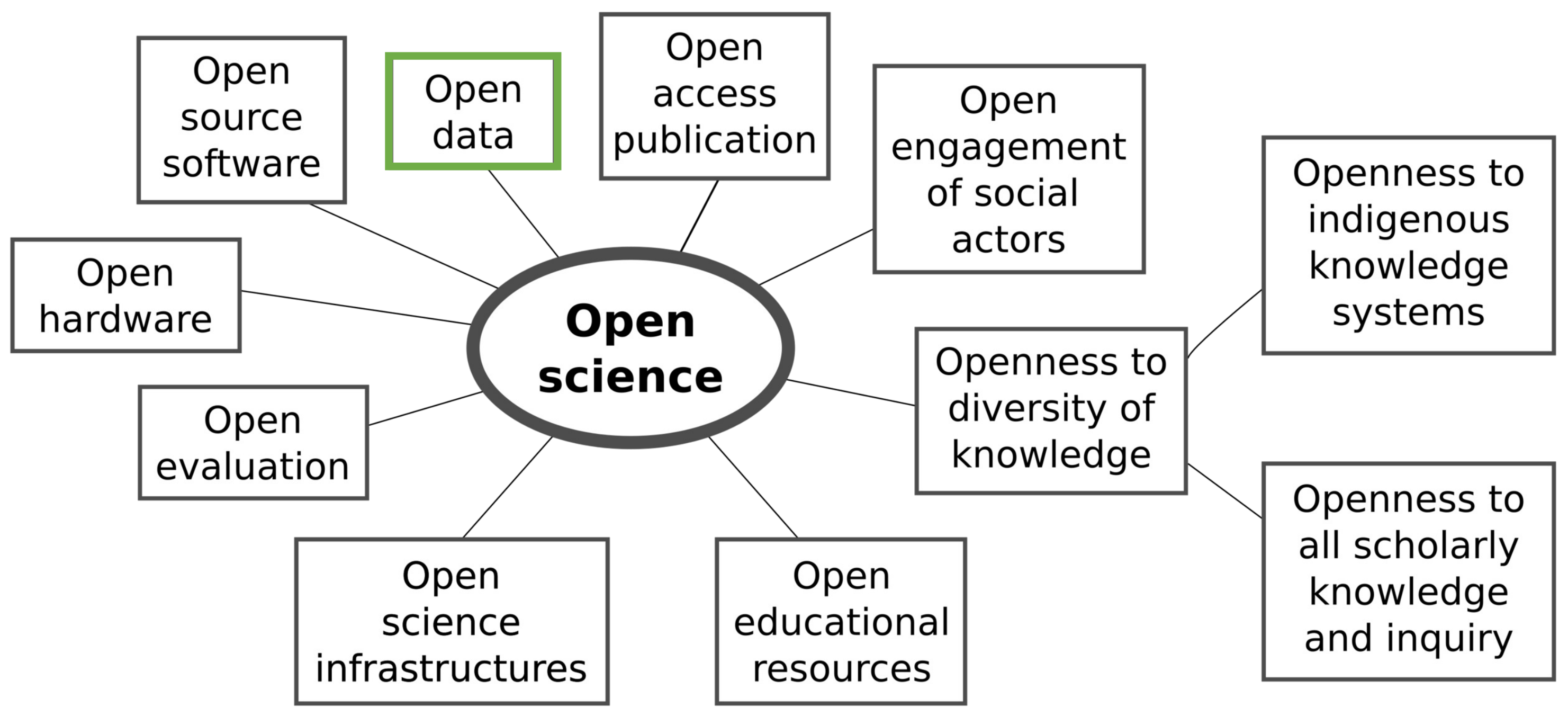
Advancing Open Science and Rapid Data Access Using Preprint and Experimental Datasets

Bruce E. Wilson (wilsonbe@ornl.gov), Chris Lindsley, Debjani Singh, Michele M. Thornton, Tammy Walker, Yaxing Wei, Daine Wright
Oak Ridge National Laboratory Distributed Active Archive Center (ORNL DAAC; https://daac.ornl.gov)

Summary

Preprint and "experimental" datasets help advance open science by providing rapid access to data – when done with clear communication of science quality.

Open Data is a Key Part of Open Science


Open science
Open source software
Open data
Open access publication
Open engagement of social actors
Openness to indigenous knowledge systems
Open hardware
Openness to diversity of knowledge
Open evaluation
Openness to all scholarly knowledge and inquiry
Open science infrastructures
Open educational resources

Open science elements based on UNESCO presentation by Ana Persic (February 17, 2021).

Image CC Attribution International 4.0 by Robbie Morrison on Wikipedia Open Science page. Redrawn by Morrison from presentation by Ana Persic, Division of Science Policy and Capacity-Building (SC/PCB), UNESCO (France) presentation to Open Science Conference 2021.

What Are the Problems to Be Solved?

- As an Author of a manuscript being submitted for publication, I need to make my data available to reviewers.
- As an Open Science Researcher, I want to make some of my data available for community review and comment, even before it is ready for publication.
- As a Data Publisher supporting data producers, I need to make some datasets publicly available within days of receiving them, even though I have not had time to complete quality assurance and documentation.
- As a Data User, I need to clearly understand the quality of data I am considering, so that I can determine its fitness to my use.
- As a Workshop Organizer, I need to make data available to participants and I want them to correctly cite the data (by DOI) in publications and other products resulting out of that workshop.
- As a Science Funder, I need to see data made publicly available in data repositories as quickly as practical, with clear provenance and clear version history, and at the lowest practical cost.

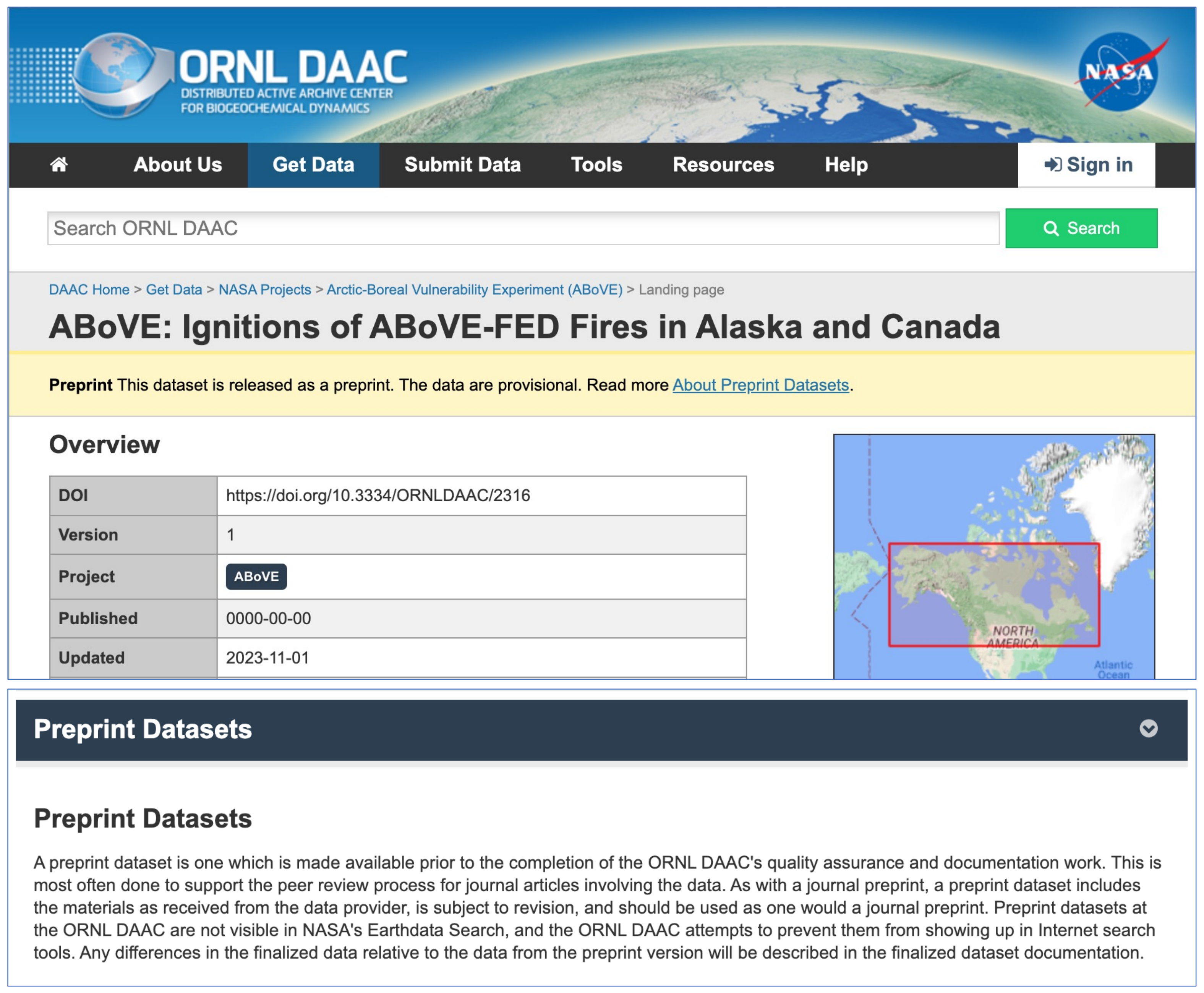
Oak Ridge National Laboratory
Oak Ridge National Laboratory is managed by UT-Battelle for the US Department of Energy (DOE)

The ORNL DAAC is supported by an Interagency Agreement between NASA and the US DOE: NASA Agreement 80GSFC23TA025; DOE Project: 2052-Z565-23

Preprint Datasets

Preprint datasets are similar to journal article preprints

- Publicly available within days, with data as submitted
- Resolvable DOI, with versioning
- Not indexed for searching (Internet or Earthdata)
- Essential for manuscript review process
- Starting to be used even within project teams



DAAC Home > Get Data > NASA Projects > Arctic-Boreal Vulnerability Experiment (ABoVE) > Landing page

ABoVE: Ignitions of ABoVE-FED Fires in Alaska and Canada

Preprint This dataset is released as a preprint. The data are provisional. Read more About Preprint Datasets.

Overview

DOI	https://doi.org/10.3334/ORNLDAAC/2316
Version	1
Project	ABoVE
Published	0000-00-00
Updated	2023-11-01

Preprint Datasets

Preprint Datasets

A preprint dataset is one which is made available prior to the completion of the ORNL DAAC's quality assurance and documentation work. This is most often done to support the peer review process for journal articles involving the data. As with a journal preprint, a preprint dataset includes the materials as received from the data provider, is subject to revision, and should be used as one would a journal preprint. Preprint datasets at the ORNL DAAC are not visible in NASA's Earthdata Search, and the ORNL DAAC attempts to prevent them from showing up in Internet search tools. Any differences in the finalized data relative to the data from the preprint version will be described in the finalized dataset documentation.

Preprint datasets are a new name (2021) for something the ORNL DAAC has been doing for several years

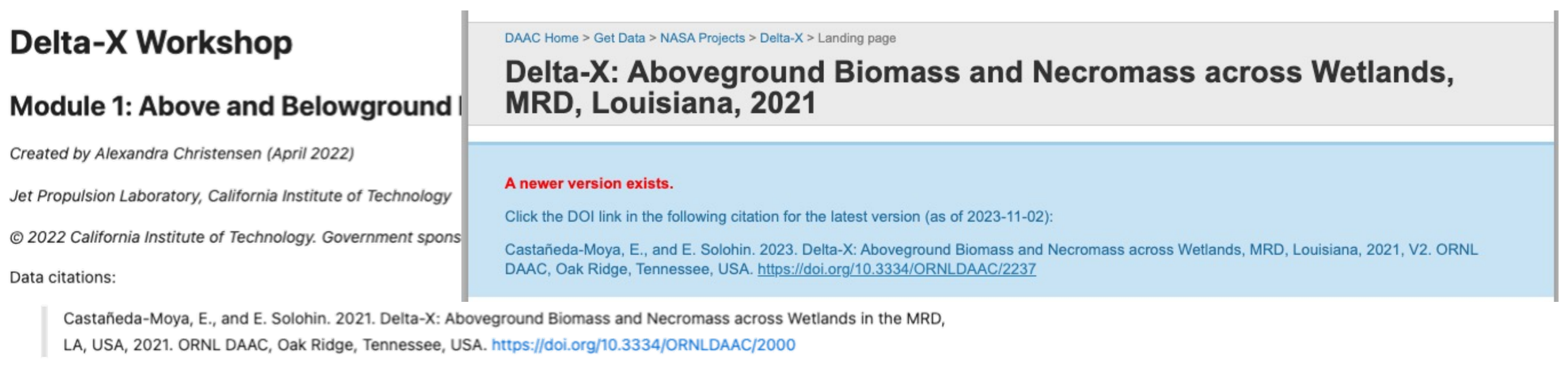
FY19	FY20	FY21	FY22	FY23
23	34	50	73	45

Number of preprint datasets/fiscal year published by the ORNL DAAC. Prior to 2021, these were called Manuscript Datasets.

Preprints and the 2022 Delta-X Science Workshop

Part of the larger preprint count in FY22 is the May 4-5, 2022 Delta-X Applications Workshop (Baton Rouge, LA and on-line).

- All data available through the ORNL DAAC (preprint or final)
- All preprints have since been finalized
- Workshop materials reference data by DOIs
- Links to newer versions on landing pages, where appropriate



Delta-X Workshop

Module 1: Above and Belowground

Created by Alexandra Christensen (April 2022)

Jet Propulsion Laboratory, California Institute of Technology

© 2022 California Institute of Technology. Government spons

Data citations:

> Castañeda-Moya, E., and E. Solohin. 2021. Delta-X: Aboveground Biomass and Necromass across Wetlands in the MRD, LA, USA, 2021. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/2000

DAAC Home > Get Data > NASA Projects > Delta-X > Landing page

Delta-X: Aboveground Biomass and Necromass across Wetlands, MRD, Louisiana, 2021

A newer version exists.

Click the DOI link in the following citation for the latest version (as of 2023-11-02):

Castañeda-Moya, E., and E. Solohin. 2023. Delta-X: Aboveground Biomass and Necromass across Wetlands, MRD, Louisiana, 2021, V2. ORNL DAAC, Oak Ridge, Tennessee, USA. https://doi.org/10.3334/ORNLDAAC/2237

Excerpt from a workshop tutorial pointing to a dataset that was released as a preprint in November 2021 and finalized just before the workshop. A newer version was released in 2023.

Exploring "Experimental" Data

Assertion: Getting feedback and suggestions from diverse perspectives improves science.

- The SBG Space-based Imaging Spectroscopy and Thermal Pathfinder (SISTER) science team is creating and testing "SBG-like" workflows and experimental data products. How can we share these so that a broad community can provide feedback to the project team, while ensuring users understand this isn't science-ready data?
- A university research team is creating a new, high-resolution Land Cover Classification. How can they easily share this large dataset to gain diverse feedback (including from under-resourced communities) while being very clear that this data is not ready for science or application use? How can/should they credit that feedback in the final data product?

Earthdata Data Maturity Levels page

Is "Experimental" different from existing Earthdata Data Maturity Levels?

If so, is there a better name than "Experimental?

- Beta: Products intended to enable users to gain familiarity with the parameters and the data formats.
- Provisional: Product was defined to facilitate data exploration and process studies that do not require rigorous validation. These data are partially validated, and improvements are continuing; quality may not be optimal since validation and quality assurance are ongoing.

What Would You Call This Type of Data?

Proposed definition: Products are based on algorithms and processes which are in development. The data products are typically distributed to support Open Science and to gain diverse input useful in improving them. Some validation work has been done, but the data products are not ready to be used in scientific or process applications.

Beta ____________________

Provisional ____________________

Experimental ____________________

Prototype ____________________

Something else: