NSF Arctic Data Center

Overview and NSF Standards and Policies

Matthew B. Jones

https://arcticdata.io

NSF Award #: 1546024

NCEAS

DataONE

Photo credit: duncan c (http://flic.kr/p/bBtZhy)


Photographer: Jeremy Potter NOAA/OAR/OER https://flic.kr/p/8F3Q6M

Troms Fylke

Rama

Detroit Publishing Co.

Key Deliverables

- **Data Archive**
- **Portal** for data discovery
- **Tools** for data and metadata submission
- **Repository** cyberinfrastructure
  - *Provenance features*
  - *Replication features*
  - *Metadata quality check*
- **Support** services
- **Training**
- **Outreach**

- **Overview: http://bit.ly/arctica-summary**

Key Deliverables

- **Data Archive**
- **Portal** for data discovery
- **Tools** for data and metadata submission
- **Repository** cyberinfrastructure
  - *Provenance features*
  - *Replication features*

DataONE NCEAS knb

Team


M.Jones


Arzayus


Baker-Yeboah


Budden


Casey


Dozier


Schildhauer


Walker*


C.Jones*


Mecum*


Clark


Goldstein*


Lortie

- Student Interns
  - J.Graybiel, A.Prescott, I.Su, J.S.Raquel, K.McGill, S.Halperin, S.Lee, A.Gordee, D.Mullen, H.Kim

Milestones: Year 1

Cyberinfrastructure	✓	***CI-1: Launch initial system***
	✓	***CI-2: Prototype content replication***
	✓	***CI-3: Improve data portal***
	✓	***CI-4: Customize web submission***
	✓	***CI-5: Deploy client tools***
Support Systems	✓	***SS-1: Deploy support infrastructure***
	✓	***SS-2: Repository operation***
	✓	***SS-3: Support Data Management Planning***
Community Engagement	✓	***CE-1: Build public website***
	↻	***CE-2: Create Data Science fellowship***
	↻	***CE-3: Conduct training***
	↻	***CE-4: Conduct synthesis project***

Operation Metrics

4,700
DATA SETS

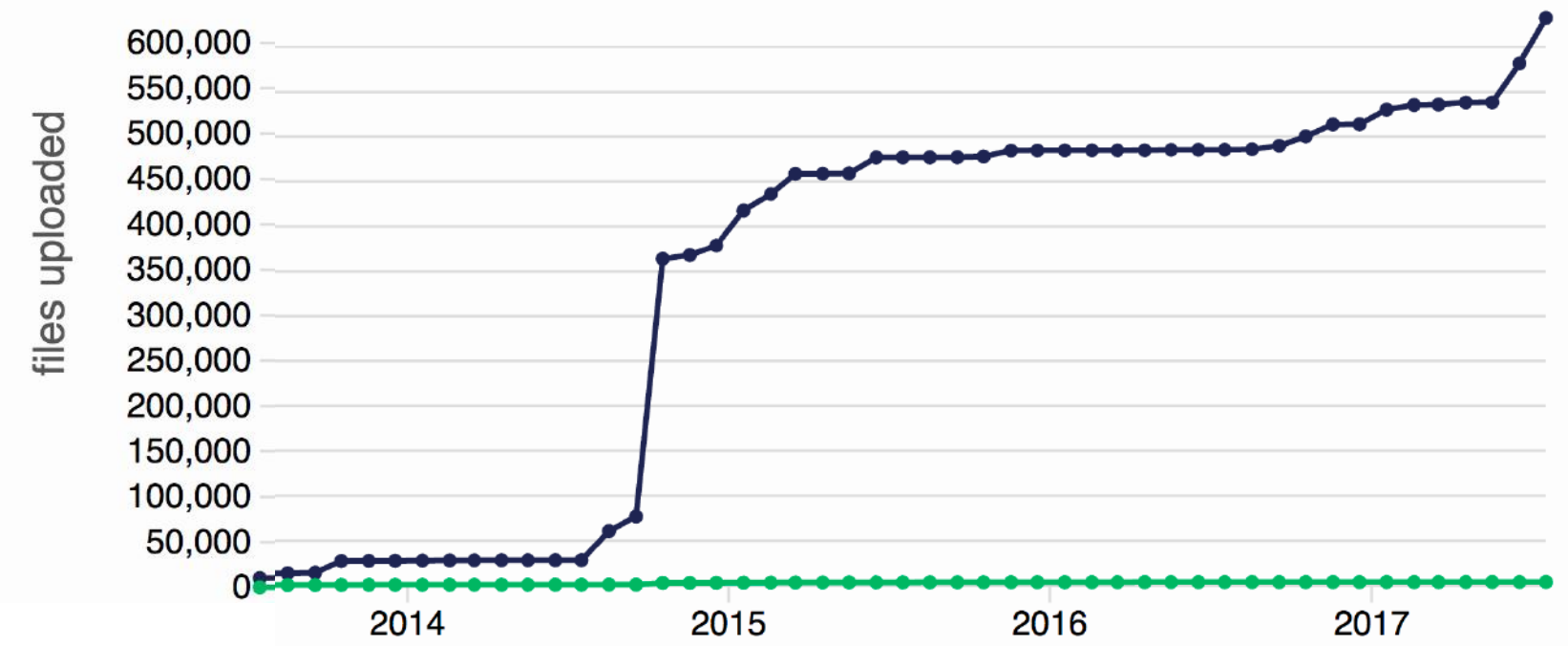
631K
DATA FILES

56K
FILE DOWNLOADS

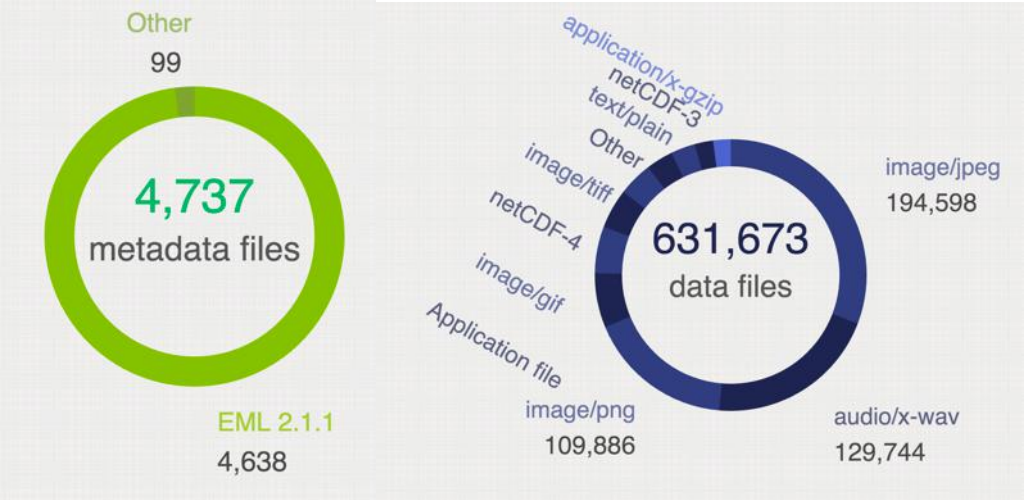
1,900
CREATORS

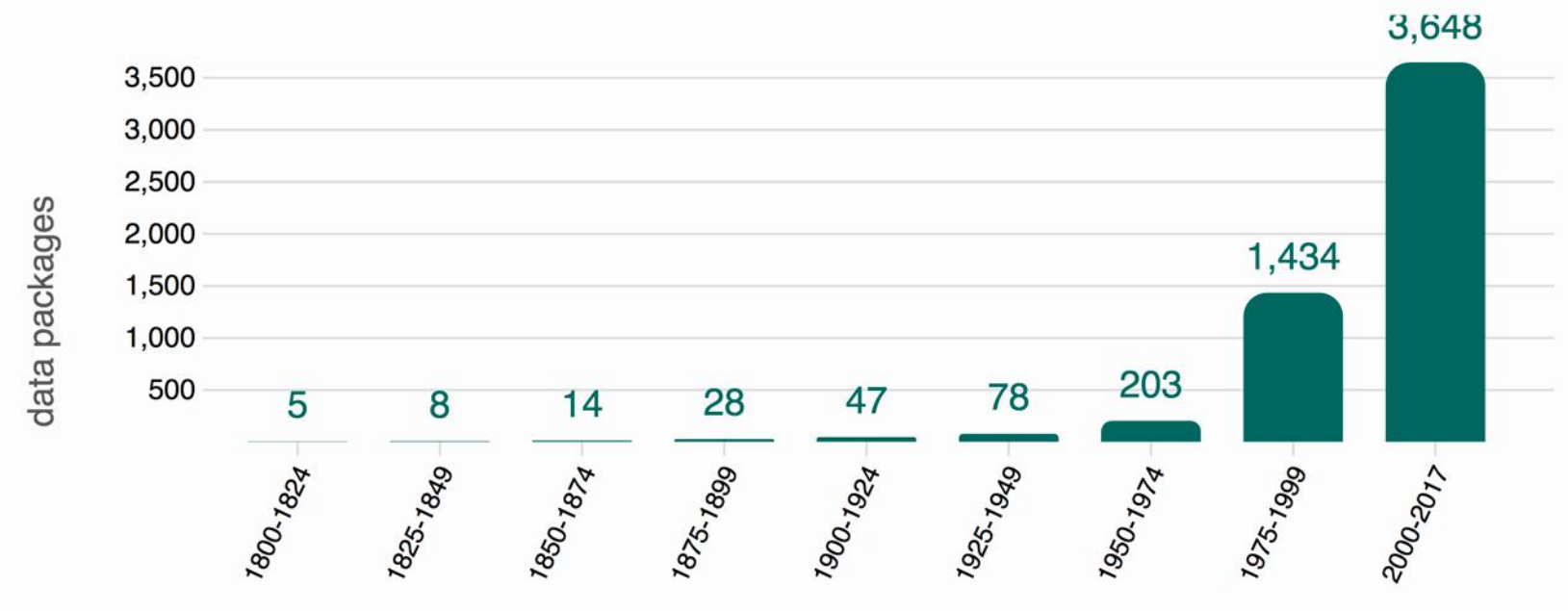
3,800
USERS

An Arctic Research Data Archive

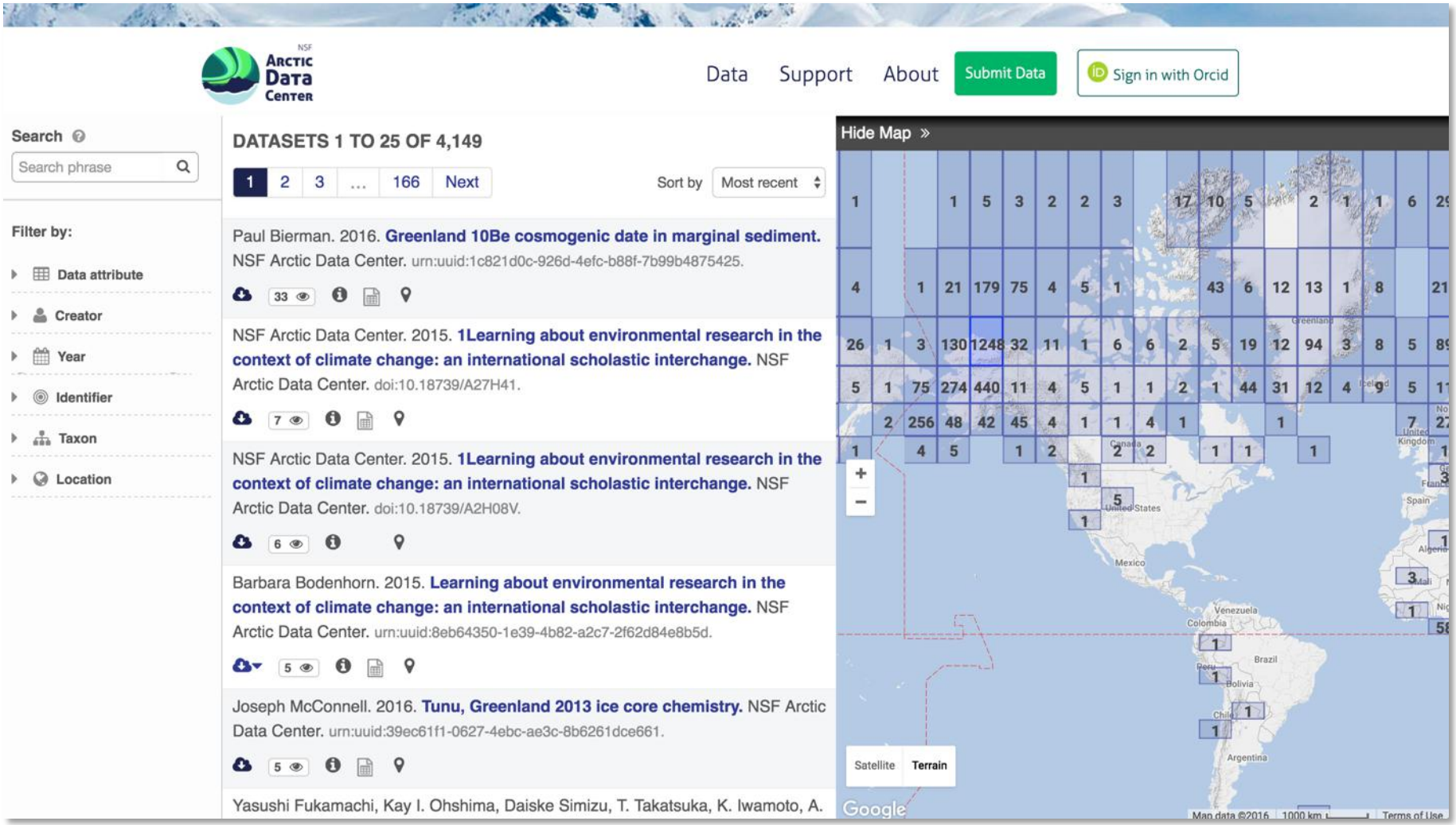


Data Holdings

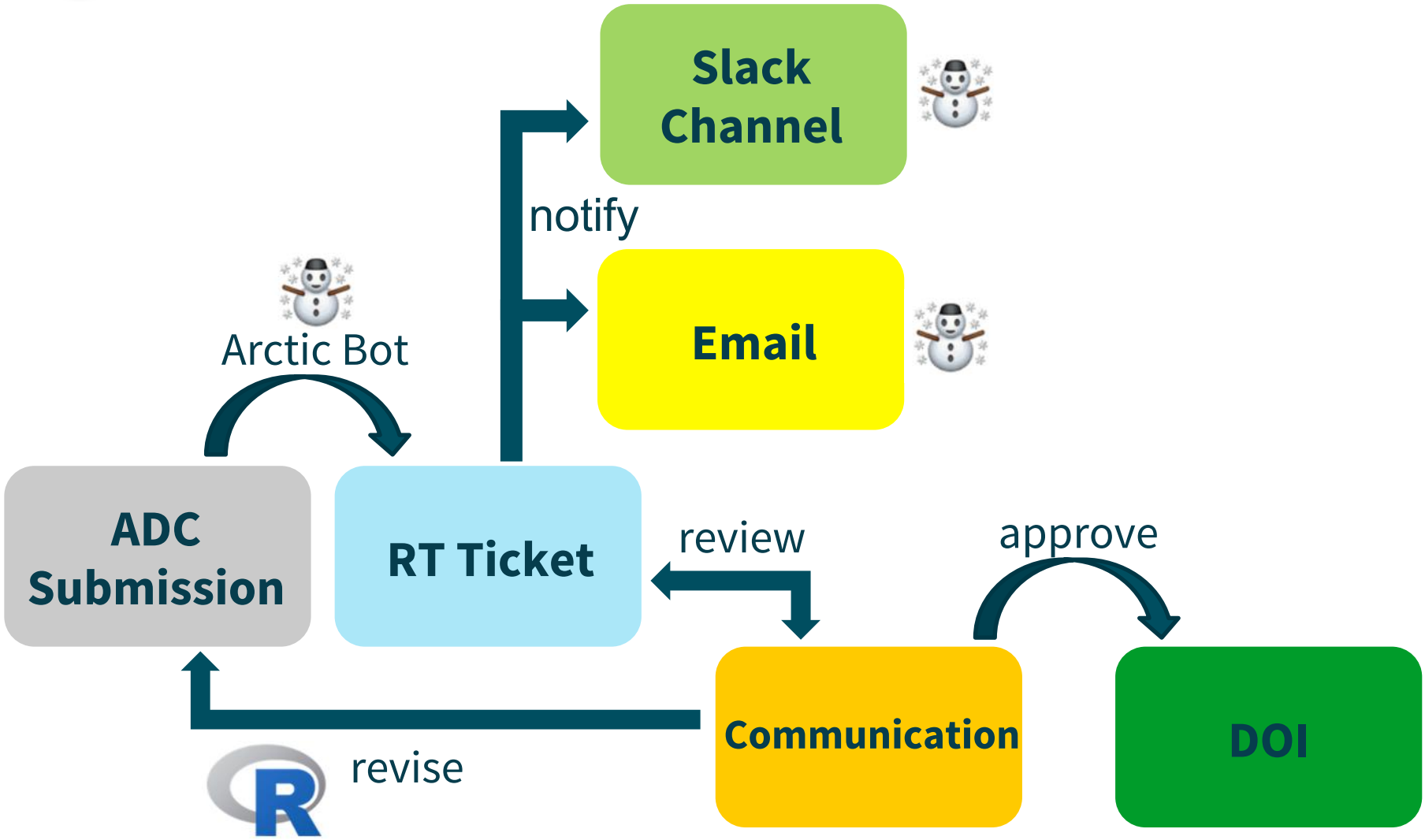




Data Discovery Portal



Curation process



Policies

- **Data producers**
  - Who, what, and how
- **Data consumers**

Who Must Submit

https://arcticdata.io/submit/#who-must-submit

- **Arctic Research Opportunities (ARC):**
  - Complete metadata and all appropriate data and derived products
  - Within 2 years of collection or before end of award, whichever comes first
- **ARC Arctic Observing Network:**
  - Complete metadata and all data
  - Real-time data made public immediately
  - Within 6 months of collection

Who Must Submit: Social Sciences

https://arcticdata.io/submit/#who-must-submit

- **Arctic Social Sciences Program (ASSP):**
  - NSF policies include special exceptions for ASSP and other awards that contain sensitive data
  - Human subjects, governed by an Institutional Review Board, ethically or legally sensitive, at risk of decontextualization
  - Metadata record that documents non-sensitive aspects of the project and data
    - Title
    - Contact information
    - Abstract
    - Methods

Terms of Use: Licensing and Distribution

https://arcticdata.io/submit/#license

- **All metadata and (non-sensitive) data will be released under either:**
  - CC-0 Public Domain Dedication:

"… can copy, modify, distribute and perform the work, even for commercial purposes, all without asking permission."

  - Creative Commons Attribution 4.0 International License:

"… free to… copy,… redistribute,… remix, transform, and build upon the material for any purpose, even commercially,… [but] ***must give appropriate credit****, provide a link to the license, and indicate if changes were made."*

Data Citation

- We assign a DOI to each published data set
- Researchers should cite data they use

Nina J. Karnovsky and Ann M. A. Harding. 2016. At-sea density of foraging little auks (Alle alle) near Hornsund Fjord. Arctic Data Center. doi:10.5065/D6MK6B17.

- We are working with DataCite to track the citations to data
  - To be displayed next year

Data Support About Submit Data Sign in with Orcid

NOTE: A newer version of this dataset exists

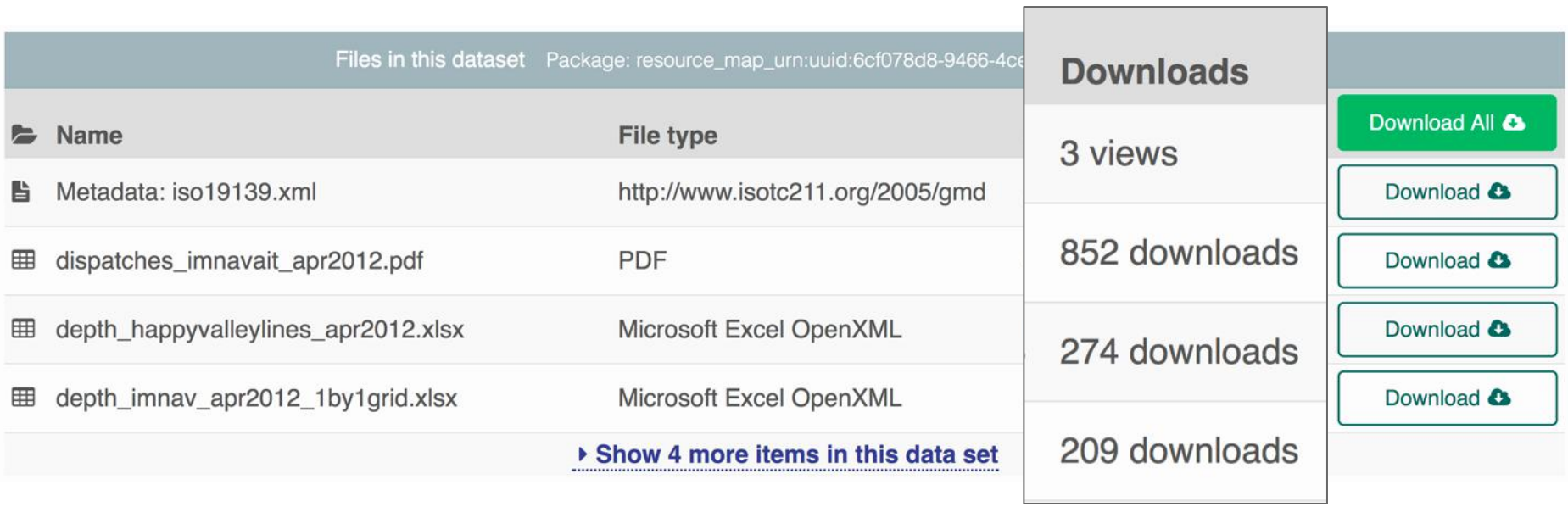
Home / Search / Metadata

Nina J. Karnovsky, Pomona College, Ann M. A. Harding, Environmental Science Department, Alaska Pacific University, and UCAR/NCAR - Earth Observing Laboratory. 2016. **At-sea density of foraging little auks (Alle alle) near Hornsund Fjord.** Arctic Data Center. urn:uuid:849a7036-8dc4-400e-a584-9d1aafacca63.

- Each update has a unique identifier
- Cite the exact version used

- Newer versions are clearly indicated

Data usage counts



Files in this dataset Package: resource_map_urn:uuid:6cf078d8-9466-4ce

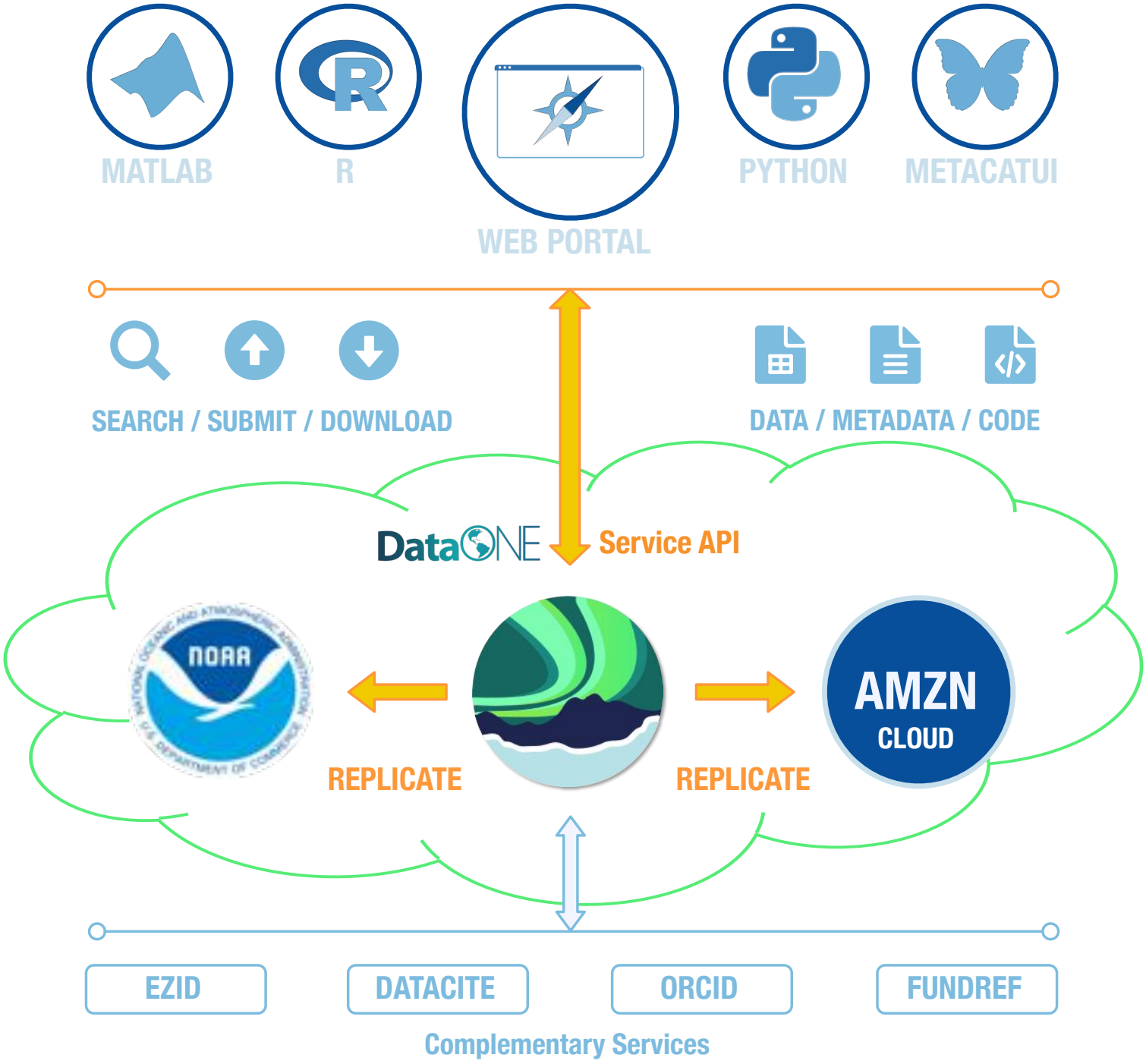
Name	File type	Downloads	
			Download All
Metadata: iso19139.xml	http://www.isotc211.org/2005/gmd	3 views	Download
dispatches_imnavait_apr2012.pdf	PDF	852 downloads	Download
depth_happyvalleylines_apr2012.xlsx	Microsoft Excel OpenXML	274 downloads	Download
depth_imnav_apr2012_1by1grid.xlsx	Microsoft Excel OpenXML	209 downloads	Download

Show 4 more items in this data set

Preparing Data and Metadata

https://arcticdata.io/submit/#preparing-data

- **ADC supports upload of any data file format**
- **Sharing greatly enhanced if using open-source formats**
  - Microsoft Excel files are commonplace, but better to export to CSV text files (do not require Microsoft products)
  - Export GIS data to ESRI shapefiles
  - Export data created in MATLAB or other matrix-based programs to NetCDF (an open binary format)
- **Describe data with complete metadata**
  - What, who, when, where, how, and why


MATLAB
R
WEB PORTAL
PYTHON
METACATUI
SEARCH / SUBMIT / DOWNLOAD
DATA / METADATA / CODE
DataONE
Service API
NOAA
REPLICATE
REPLICATE
AMZN
CLOUD
EZID
DATACITE
ORCID
FUNDREF
Complementary Services

Key Features

- Archival formats
  - Data, Code, Figures/Images
- Identification
  - DOIs assigned on publication
- Versioning
- Replication and audit

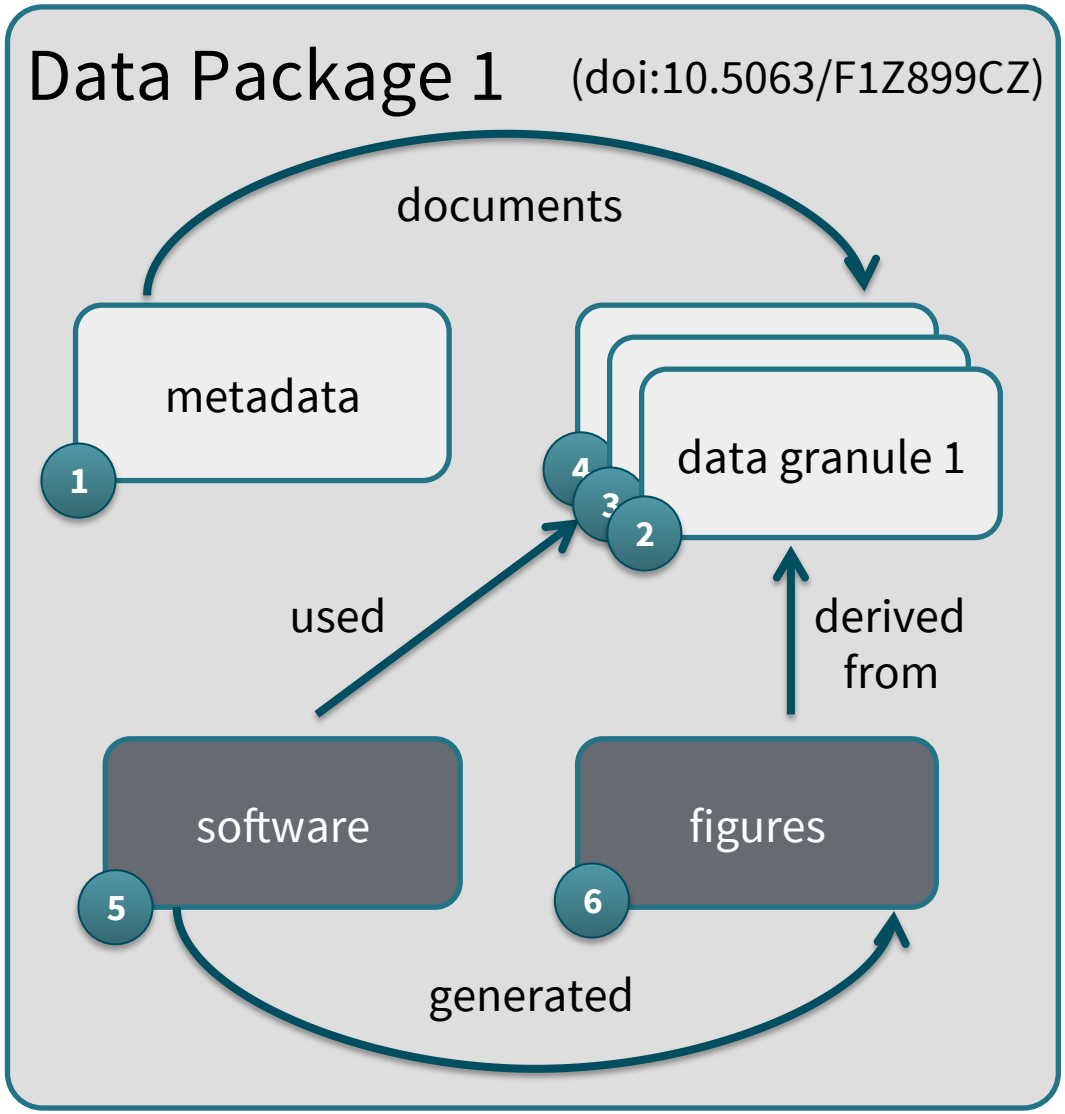
- Usage tracking
- Provenance
- Quality engine
- R and Matlab packages

Archival Packages: Provenance inside

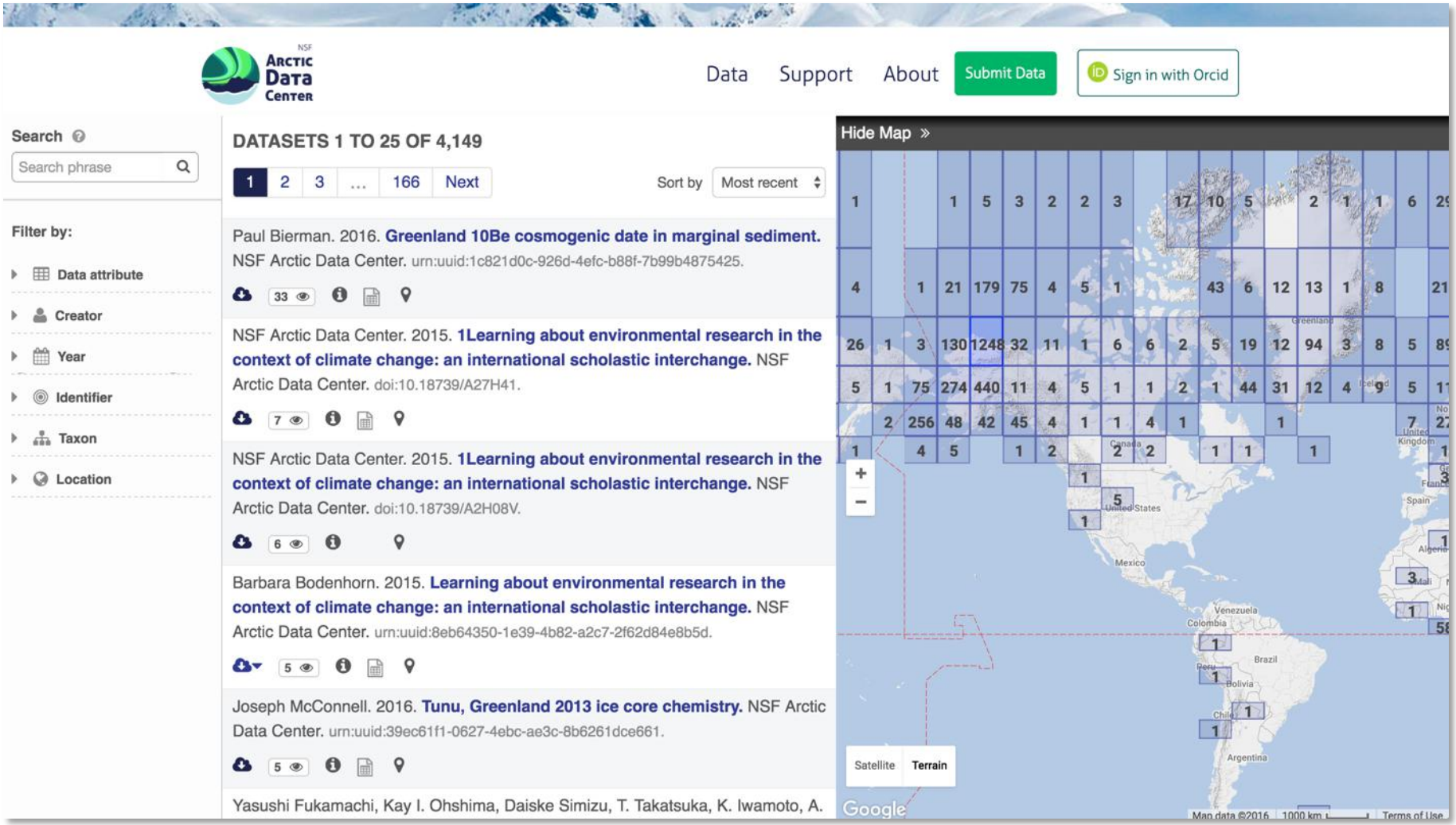
- Globally unique identifiers
- Strong versioning

- OAI-ORE manifest
- BagIt serialization

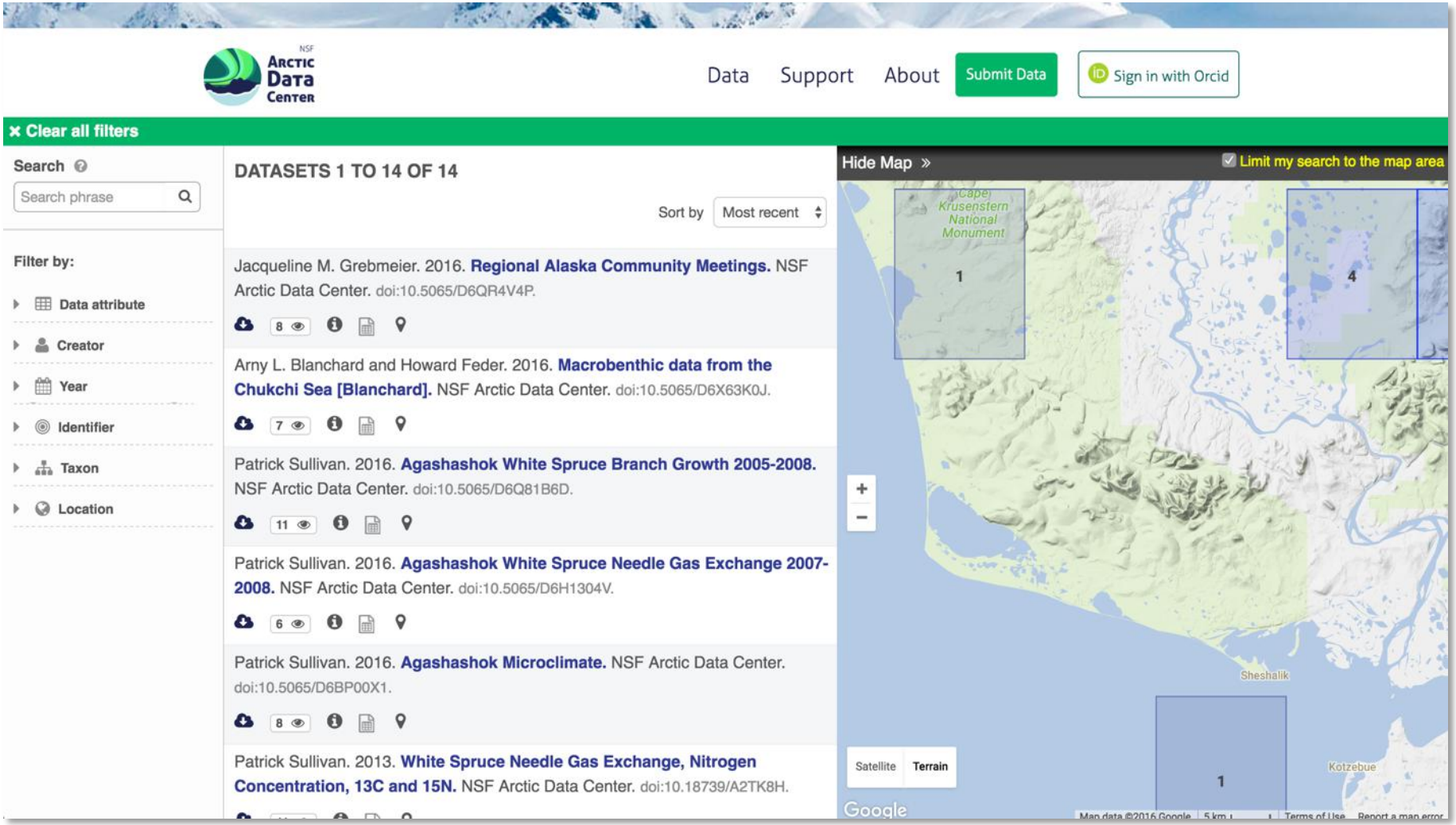
- Hierarchical
- Provenance



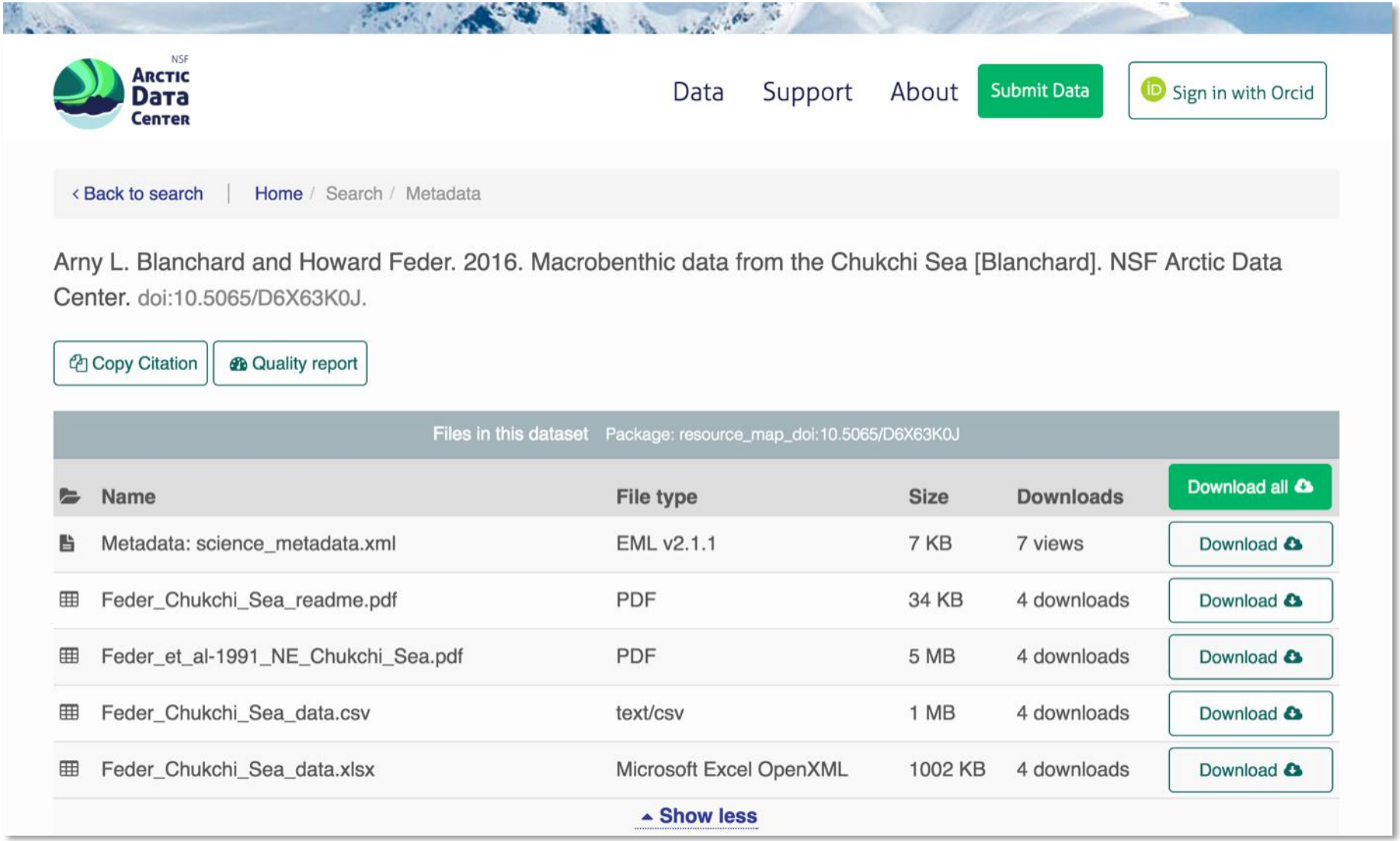
Data Discovery Portal



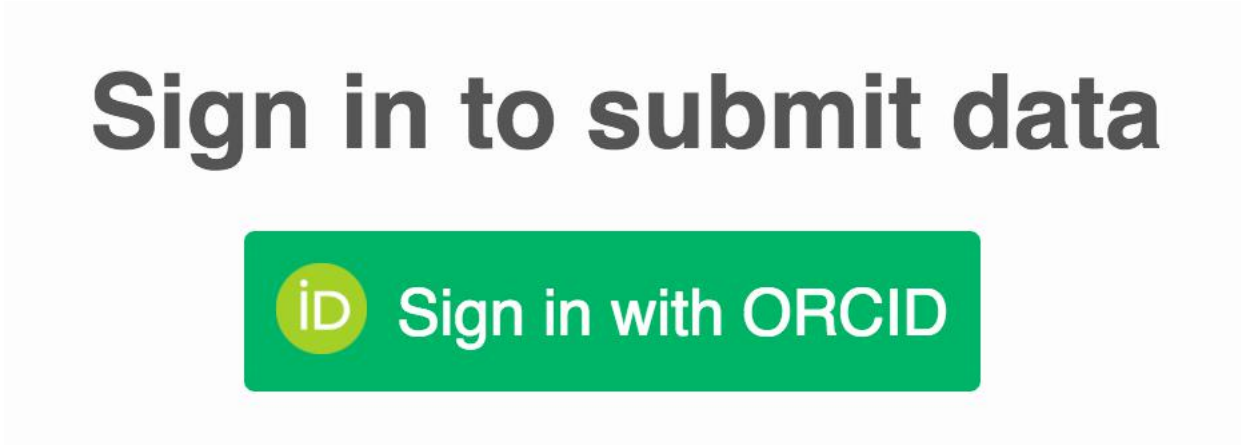
Precise Drill Down



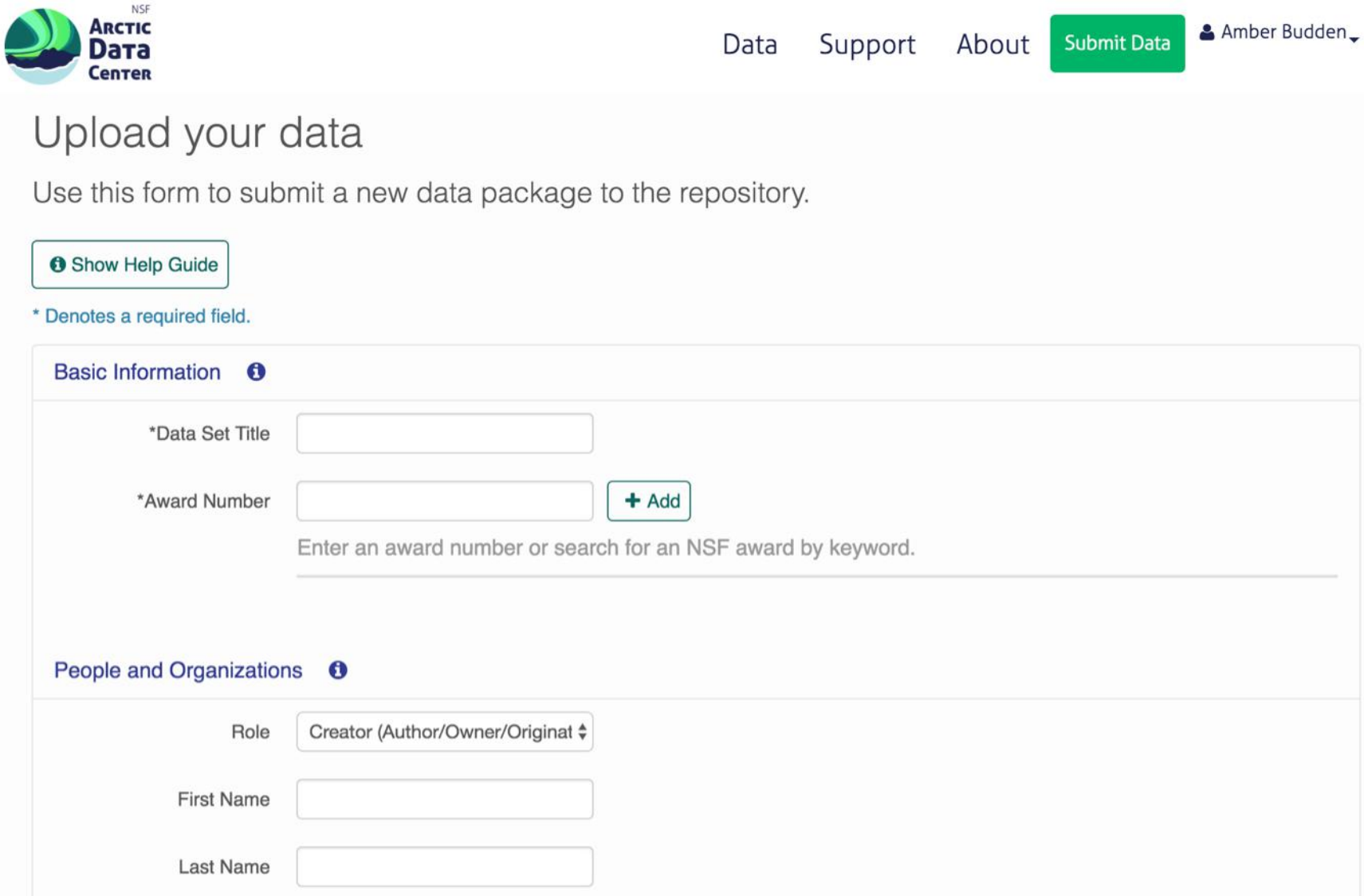
Metadata Record



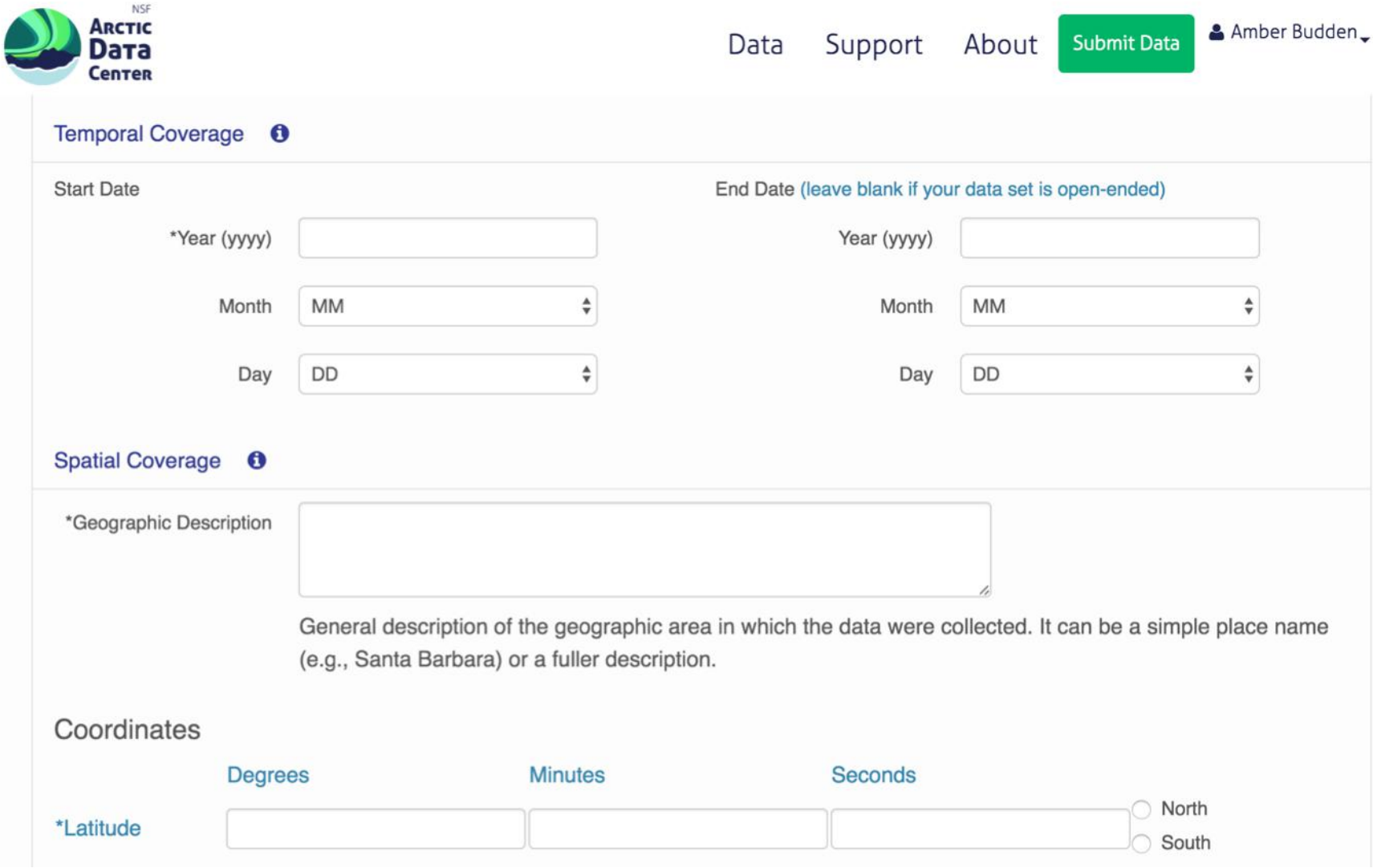
Data Submission



Data Submission



Data Submission



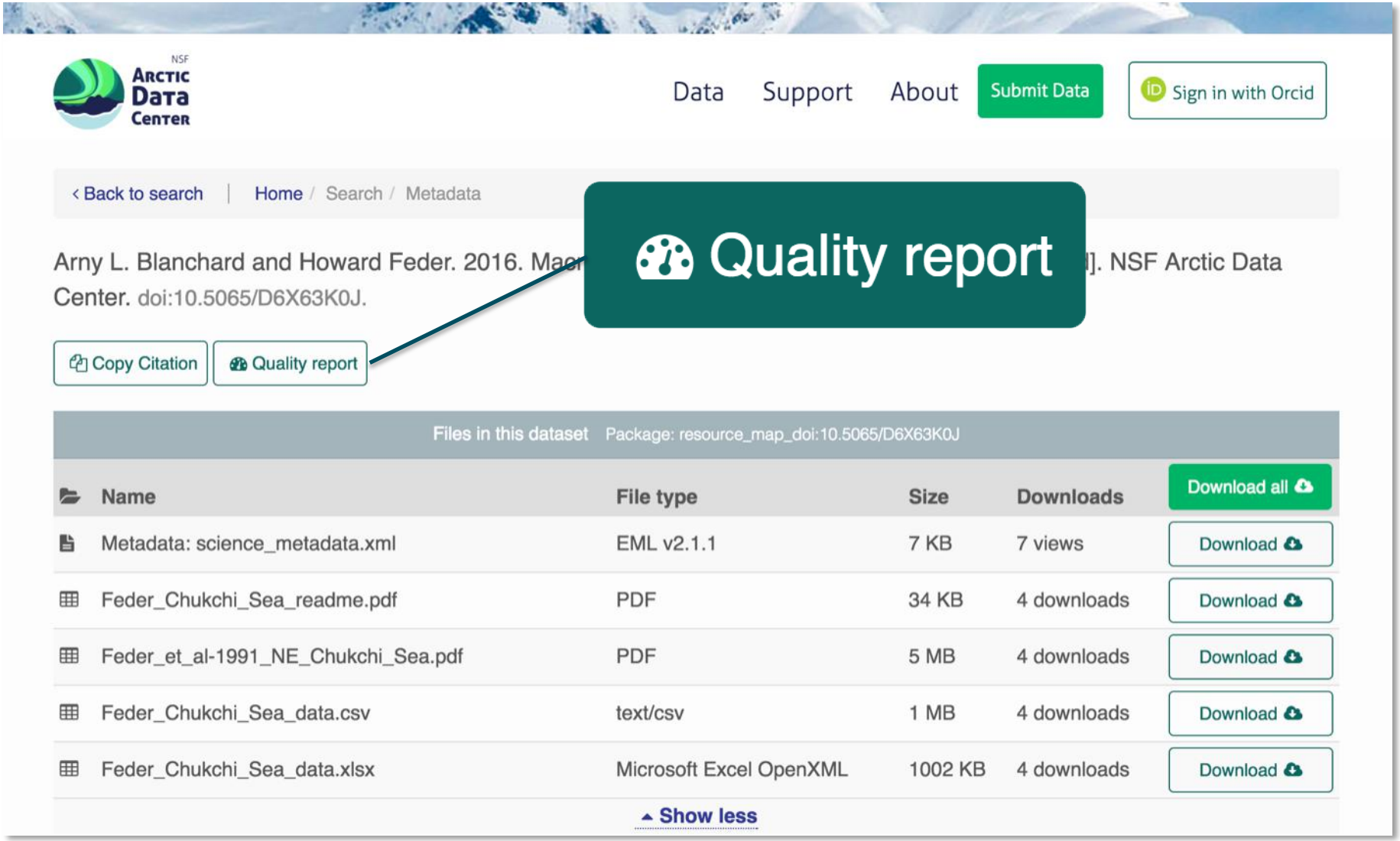
Other Submission Methods

DataONE R and Matlab Tools

Scientists can submit data files and code programmatically through R or Matlab.

But...EML Metadata stills needs to be created separately or by using the limited EML R library.

Metadata Quality Reports: MetaDIG



Metadata Quality Report

Mackenzie Grieman. 2016. Ice core vanillic acid and para-hydroxybenzoic acid concentrations, Akademii Nauk, Severnaya Zemlya archipelago, Russia, 2013-2015. Arctic Data Center. doi:10.18739/A23Q23.

After running your metadata against our standard set of metadata, data, and congruency checks, we have found the following potential issues. Please assist us in improving the discoverability and reusability of your research data by addressing the issues below.

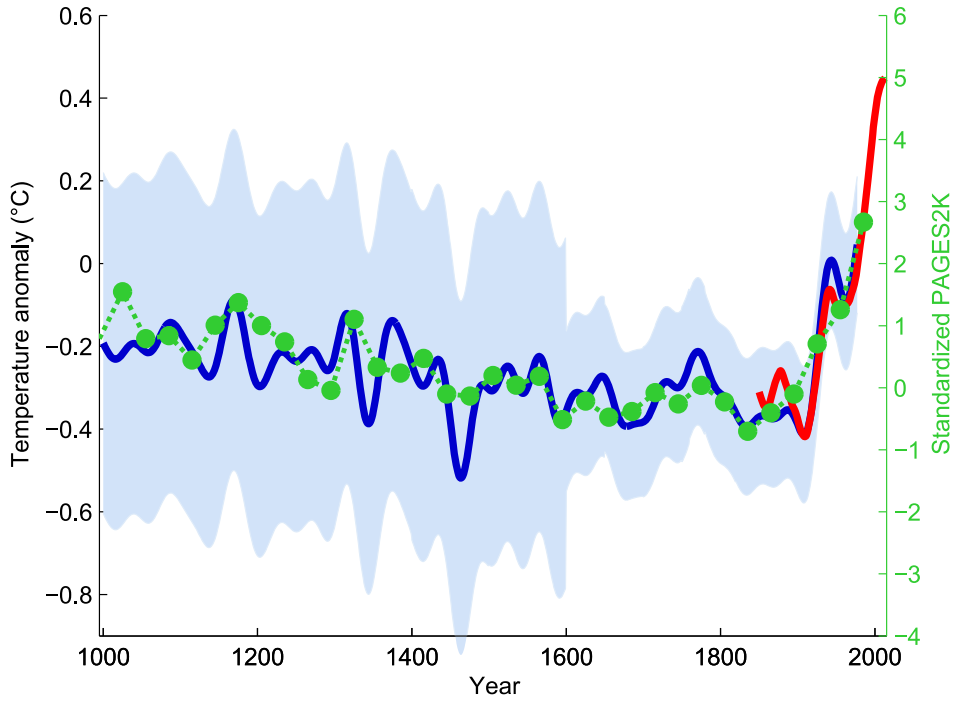
Identification: 100% complete

Discovery: 100% complete

Interpretation: 100% complete

- Passed 20 checks out of 22. Good job!
- Warning for 2 checks. Please review these warnings.
- Failed 0 checks.
- 4 informational checks. These may include skips, errors and failures.

Provenance



What ***input data*** went into this study?

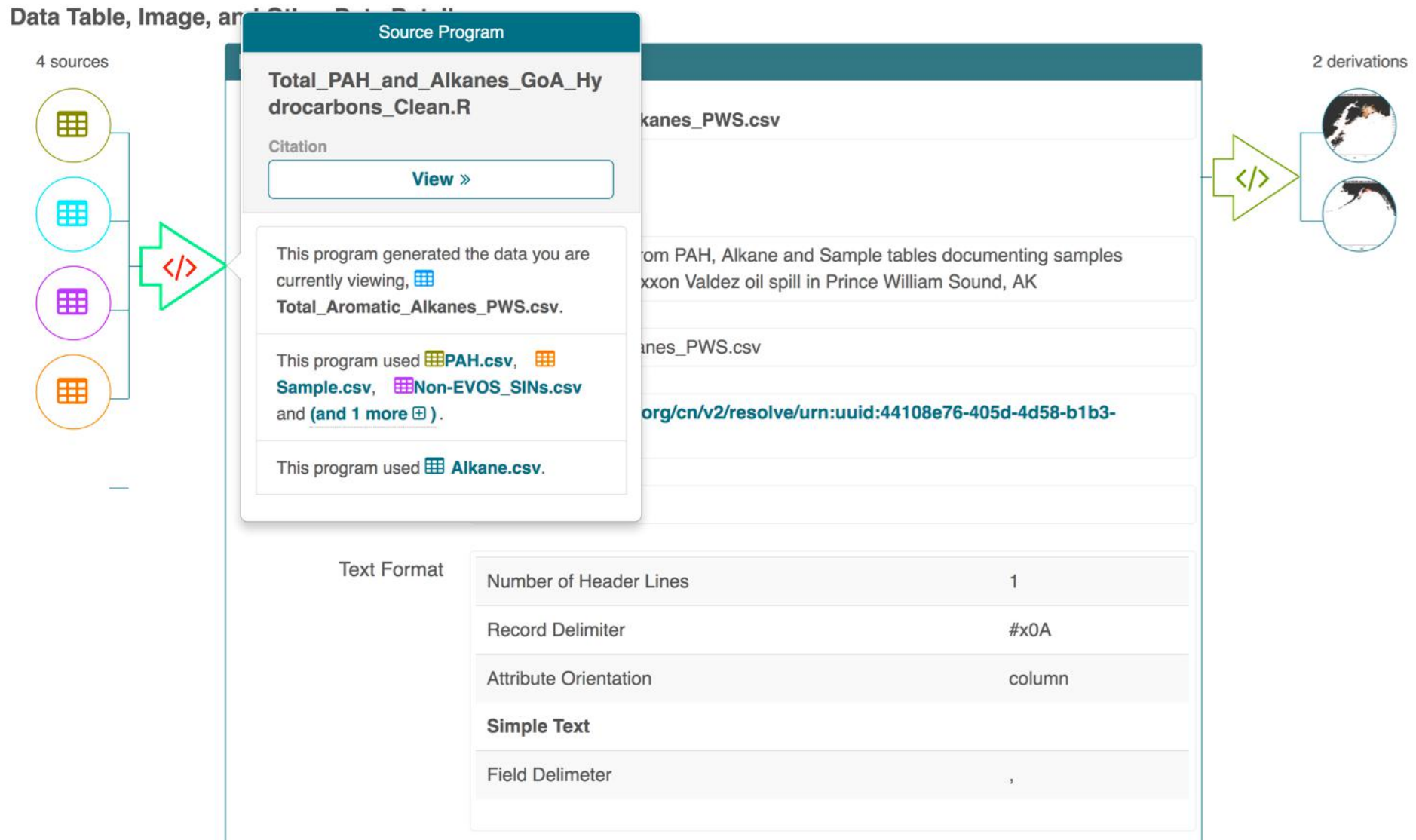
What ***methods*** were used?

… with what ***parameter*** settings, ***calibrations***, …?

*Can we **trust** the data and methods?*

- **Provenance** (*lineage*): track **origin** and **processing history** of data ➔ trust, data quality ~ audit trail for attribution, credit

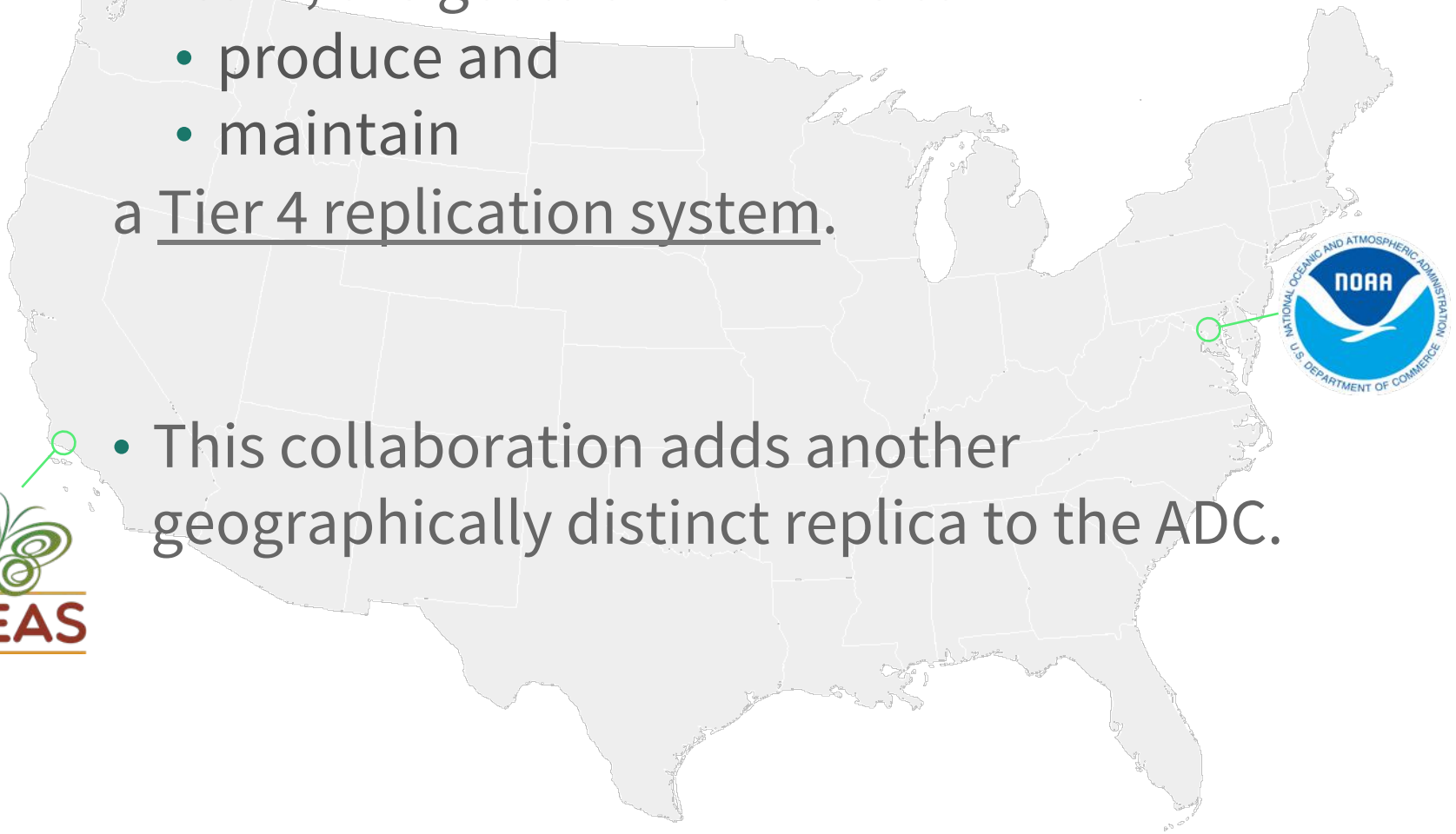
Provenance Web UI



Replication Goals

- In collaboration with the Arctic Data Center Team, the goals of NCEI are to
  - produce and
  - maintain

a Tier 4 replication system.

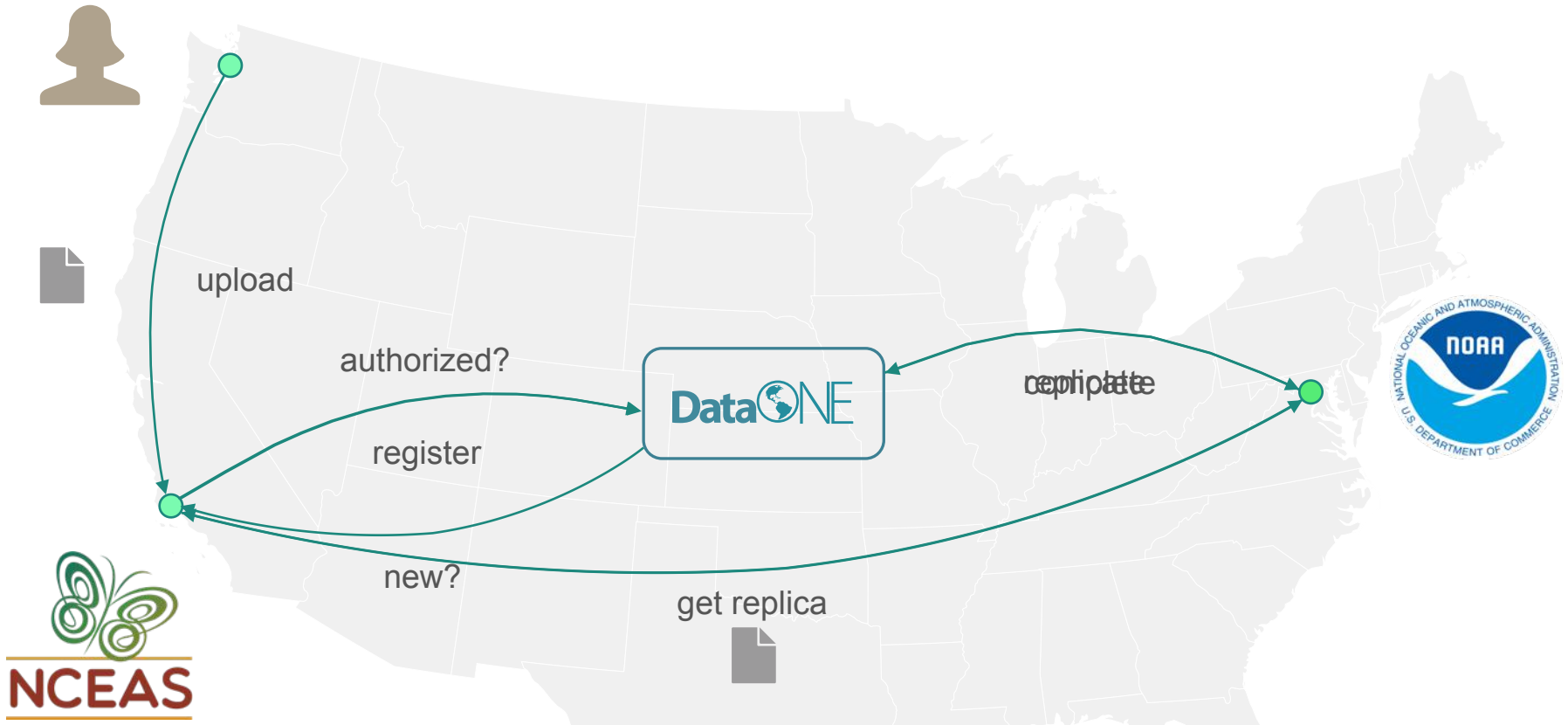


- This collaboration adds another geographically distinct replica to the ADC.

Replication Year-2 Milestones

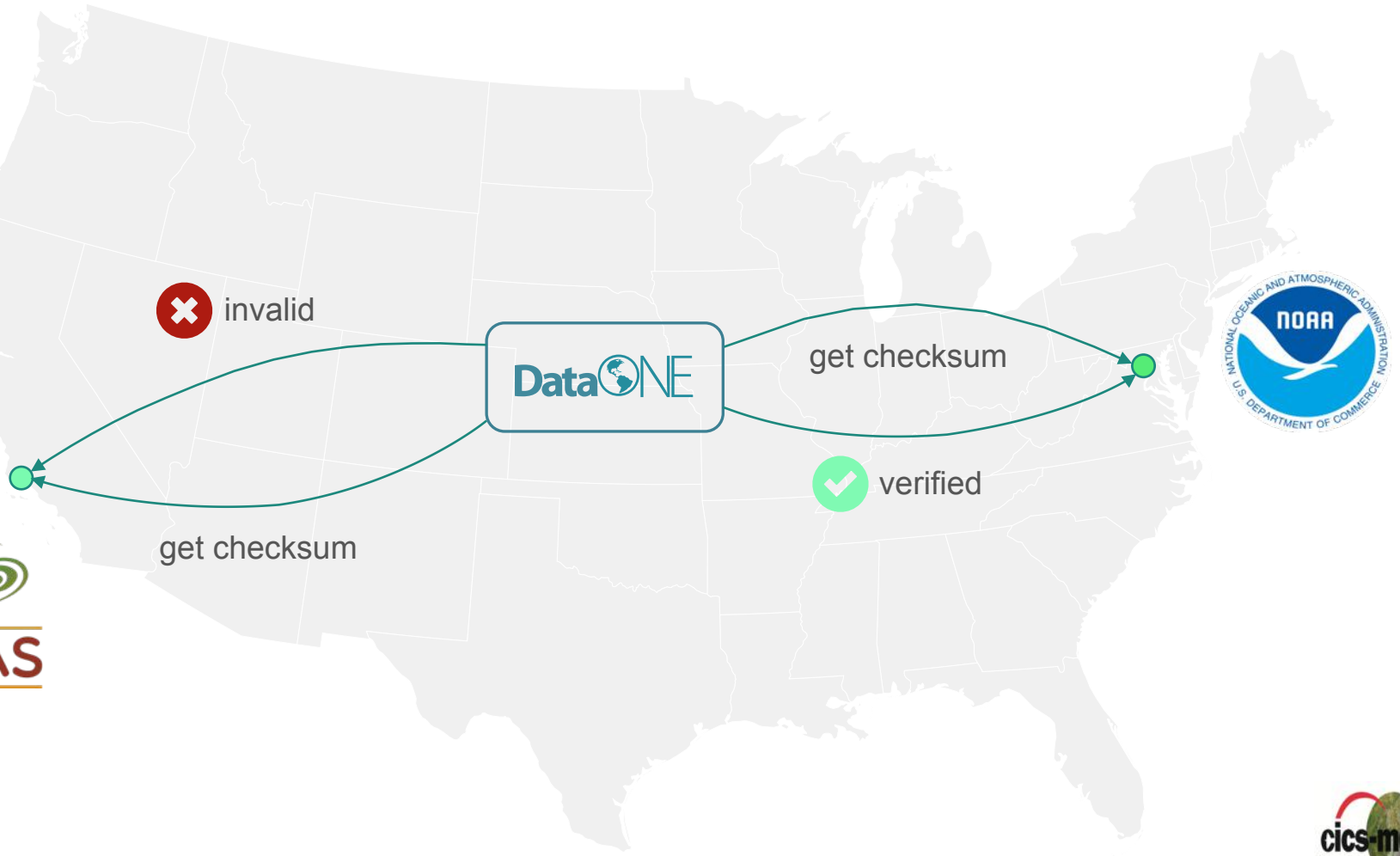
- In Year 2, NCEI will
  - complete the prototype replication system
  - move the prototype into production

Replication Steps



Replica Audit Steps

- All files are audited every 3 months

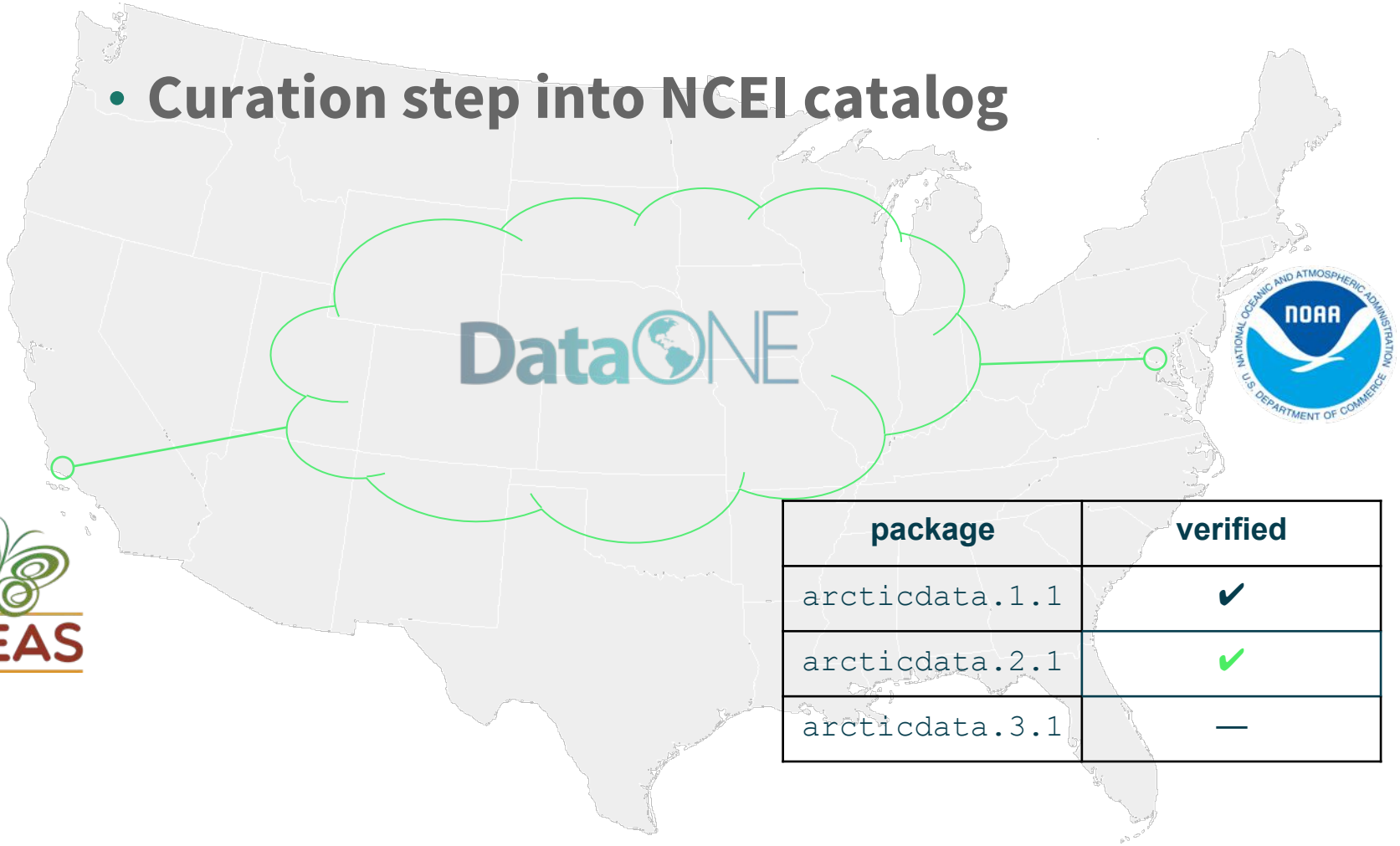


NCEAS

cics-md

Replication with NCEI Member Node

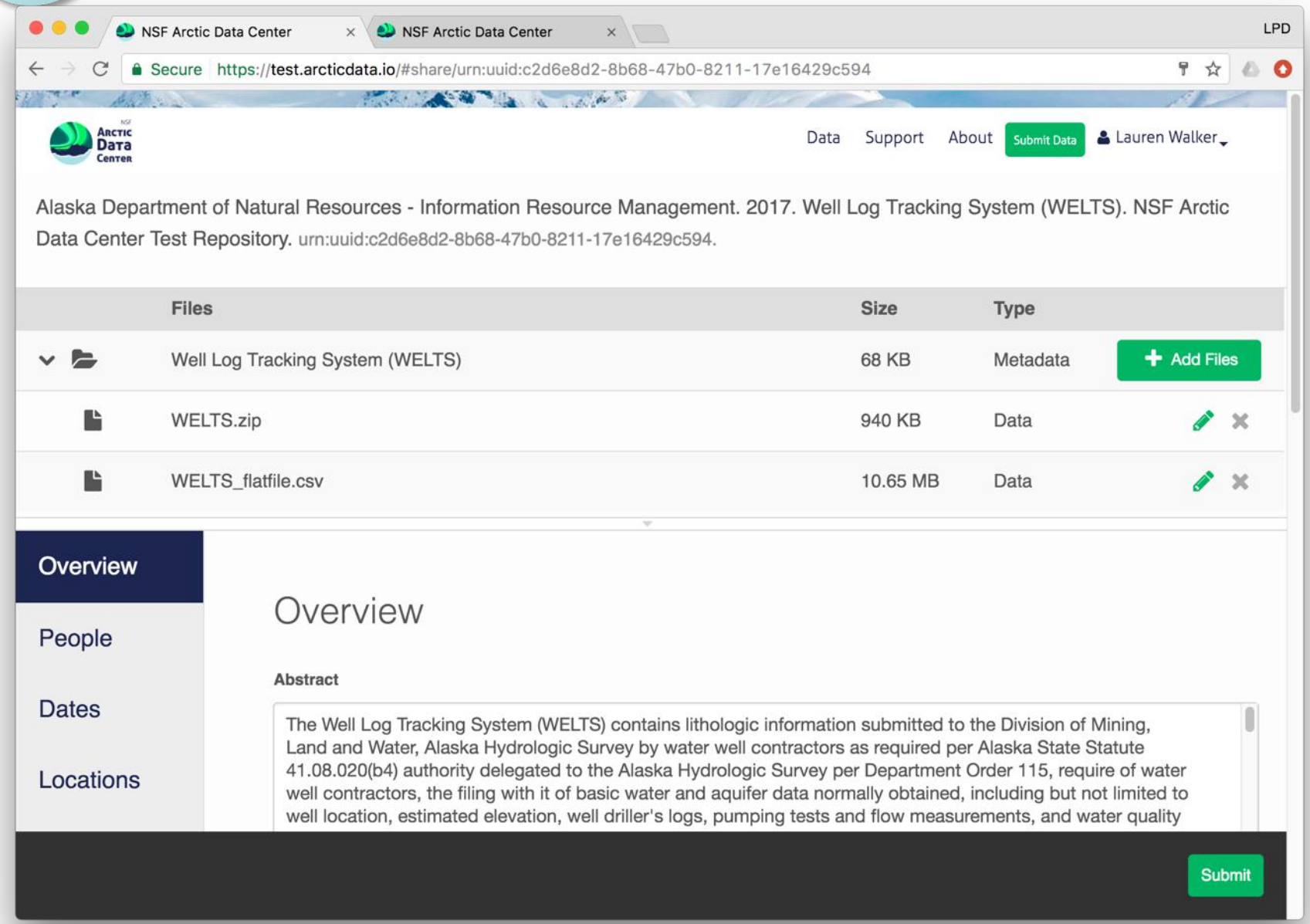
- Automated replication via DataONE
- Curation step into NCEI catalog



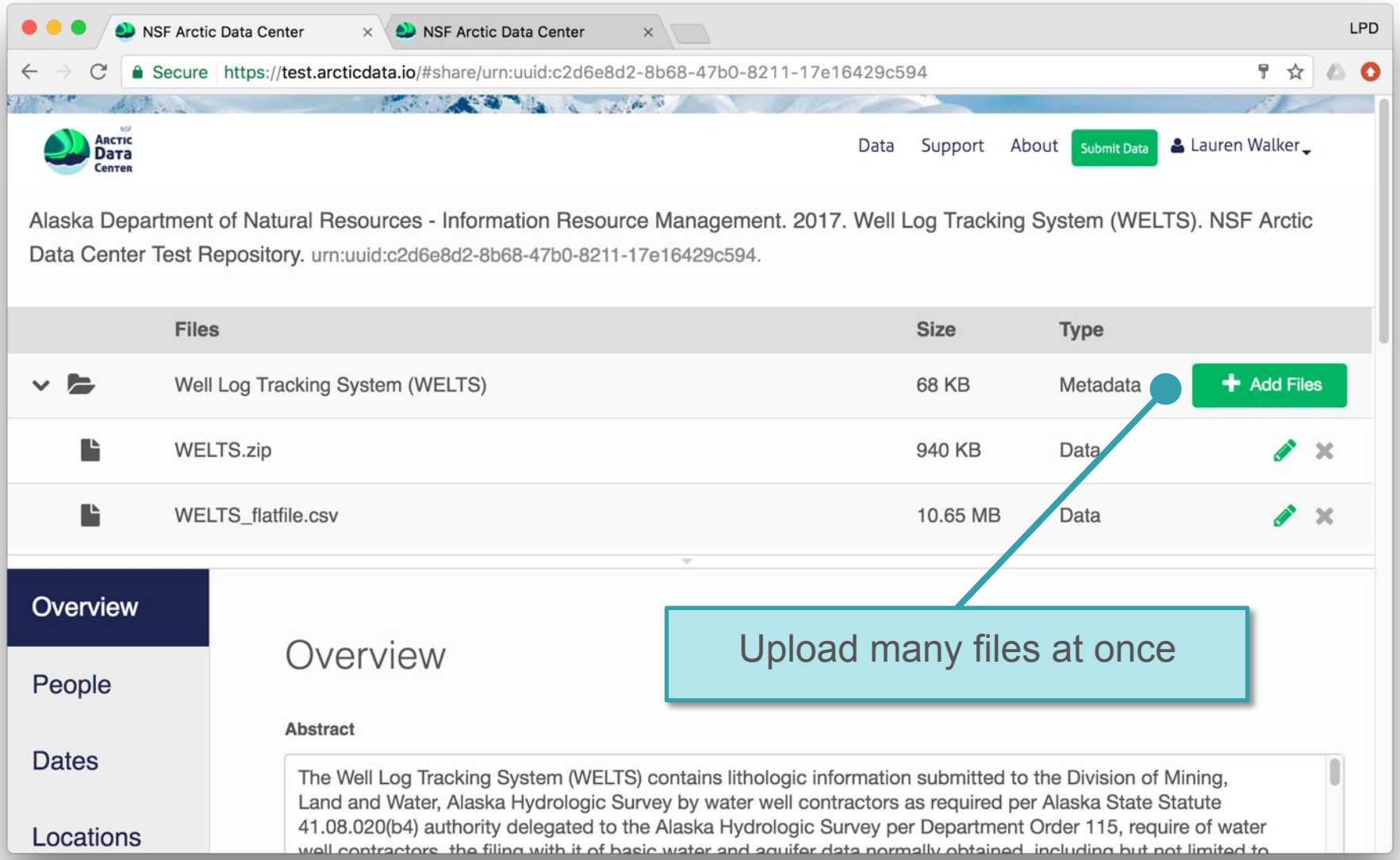
package	verified
arcticdata.1.1	✔
arcticdata.2.1	✔
arcticdata.3.1	—

NCEAS

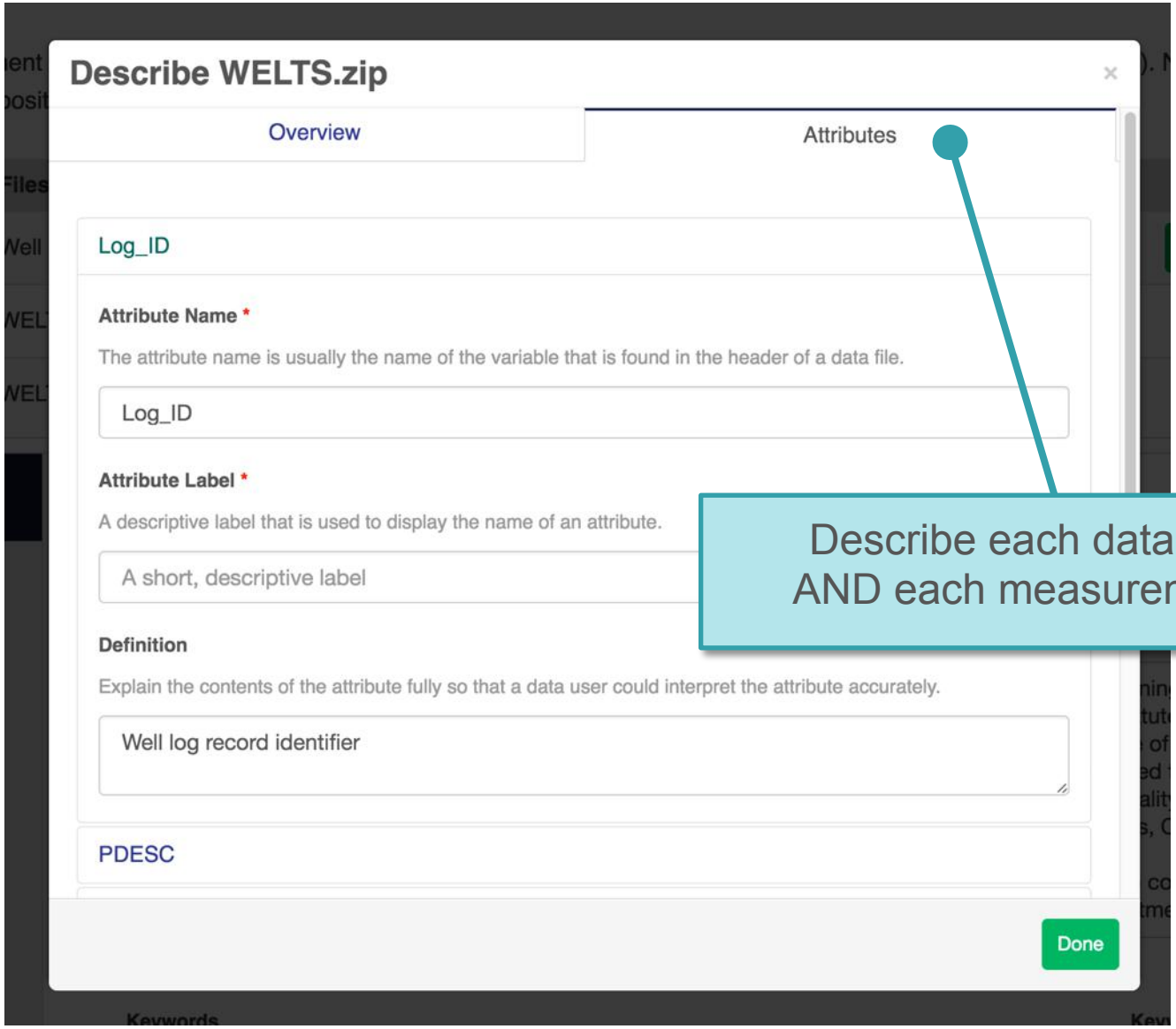
The New Web Submission System



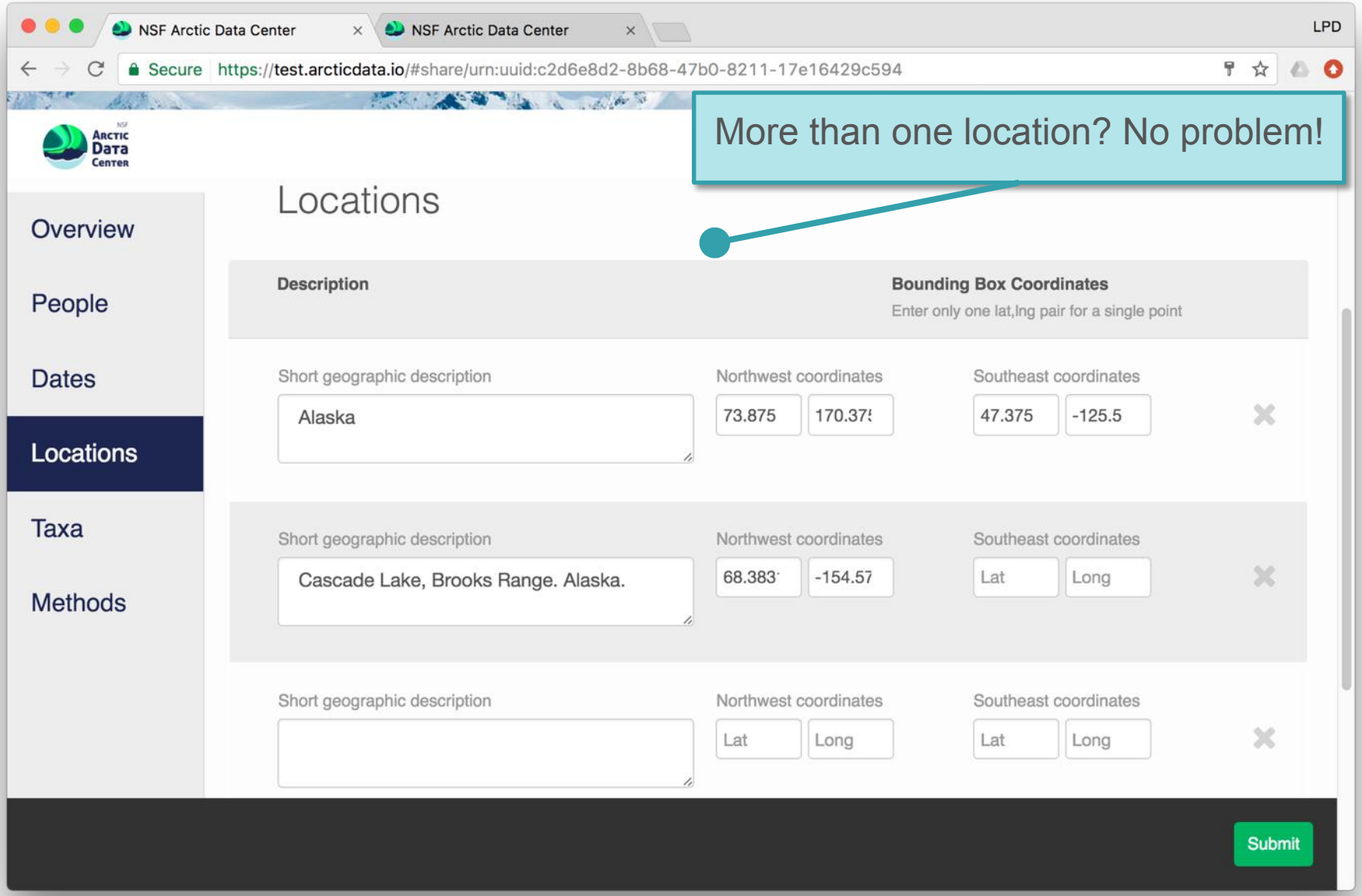
The New Web Submission System



The New Web Submission System



The New Web Submission System



Comparison

	Current	New	In future versions
We are able to quickly make improvements based on user feedback	✓ (sometimes)	✓	✓
Upload multiple files at once		✓	✓
Max. number of files that can be submitted	10	~100	1000+
Describe the characteristics of data files (file type, description of contents)		✓	✓
Describe the measurements inside each data file		✓	✓
Number of geographic regions that can be described	1	unlimited	unlimited
Provide ORCIDs for each person		✓	✓
Add nested folders			✓
Save a draft of your dataset and edit later			✓
Even more file and attribute-level metadata (geographic regions per data file, temporal coverage per data file, measurement precision, etc.)			✓

Agenda

Agenda	Day 1.	Day 2.
8:30-9:00	Welcome and introductions	Writing Good Data Management Plans
9:00-9:45	Arctic Data Center and NSF Standards and Policies	Writing Good Data Management Plans
9:45-10:00	Break	Break
10:00-12:00	Effective data modeling and management	Data packaging and file hierarchies
Noon-1:15	Lunch	Lunch
1:15-2:15	Authoring Quality metadata	Authoring large data sets
2:15-2:30	Break	Break
2:30-4:30	Authoring Quality metadata	Large data and Tracking data provenance
4:30-5:00	Question and Answer	Discussion