https://arcticdata.io

@arcticdatactr

Best Practices: Data and Metadata Submission

Matthew B. Jones & Kathryn Meyer

jones@nceas.ucsb.edu
https://orcid.org/0000-0003-0077-4738
@metamattj

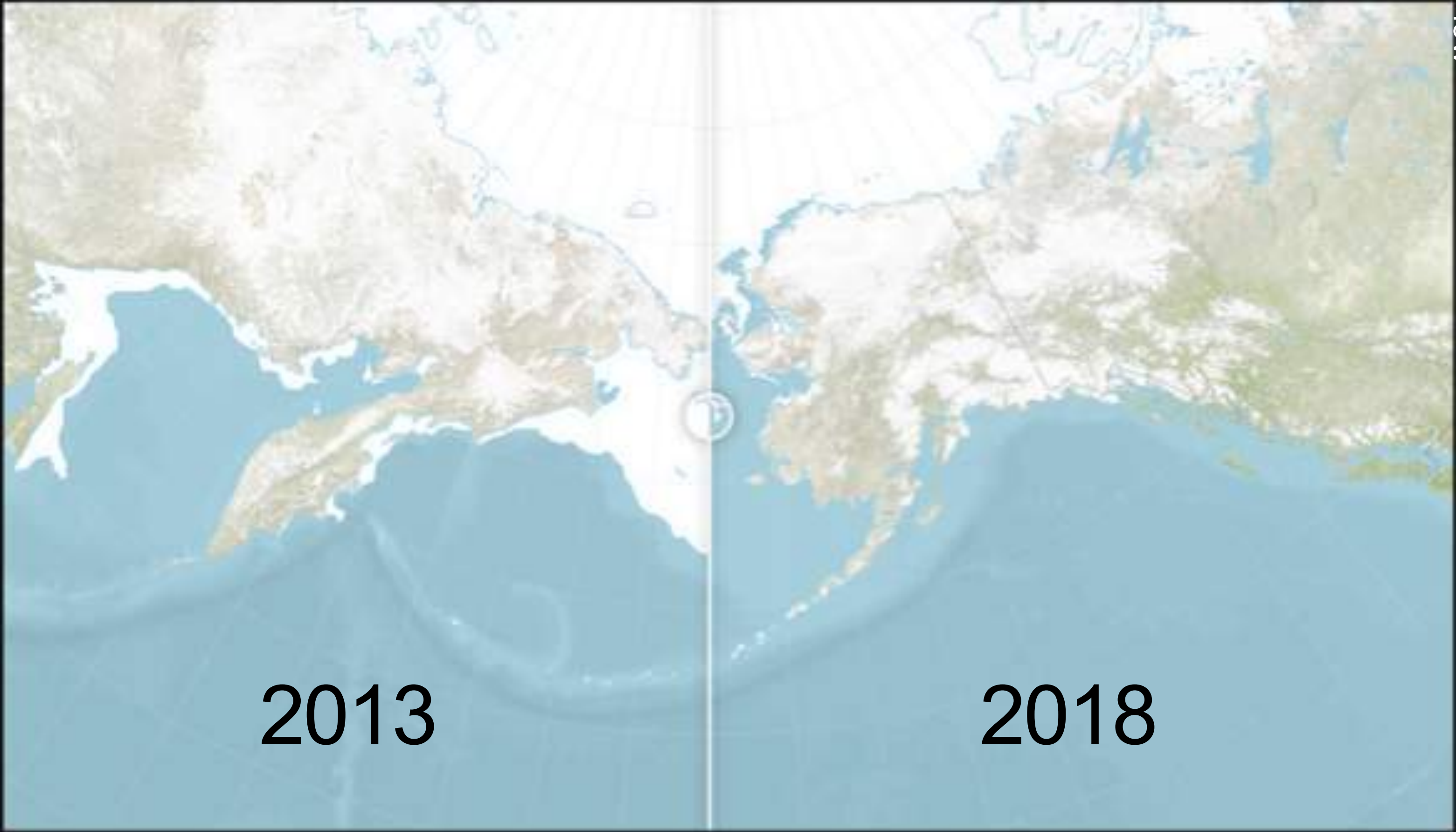
meyer@nceas.ucsb.edu
https://orcid.org/0000-0003-0200-0787

NSF Award #1546024

NCEAS

DataONE

POLAR 2018 Open Science Conference
21 June 2018



https://climate.nasa.gov/images-of-change?id=646#646-bering-sea-ice-at-record-low

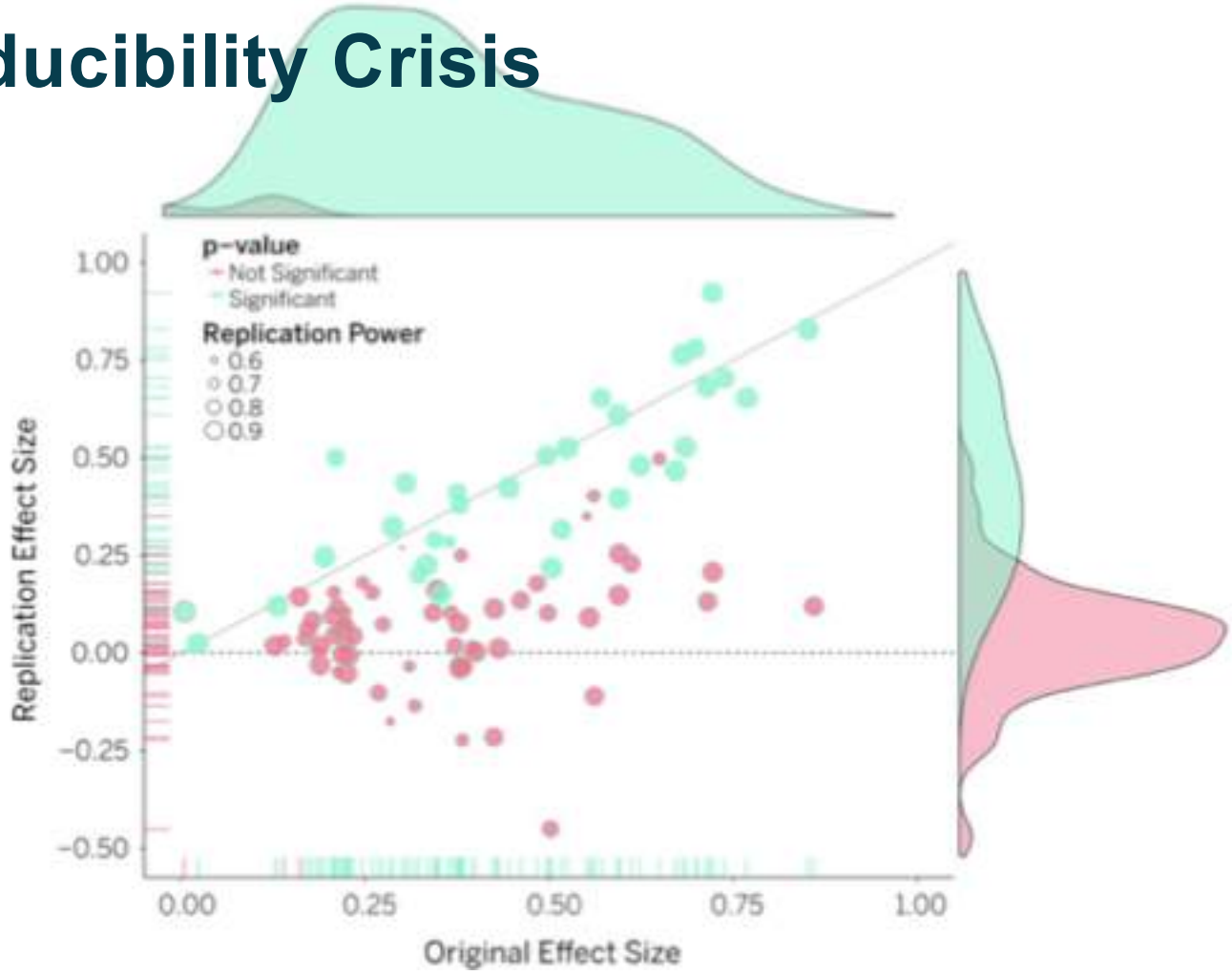
Reproducibility Crisis

Why Most Published Research Findings Are False.

Ioannidis, John P A. 2005.

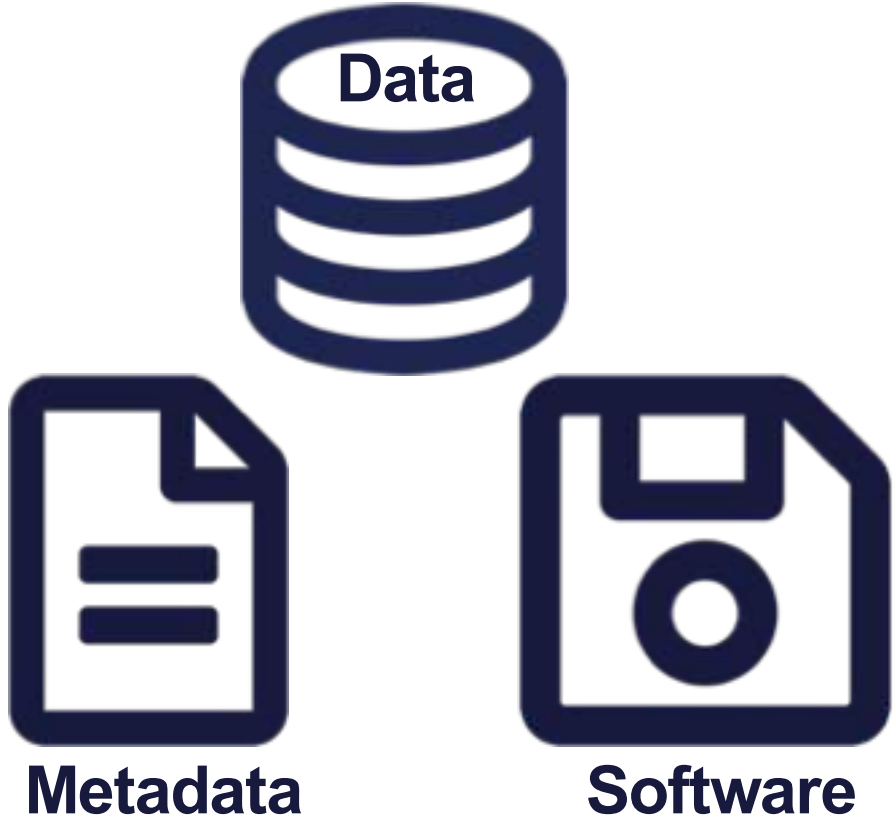
PLoS Medicine 2 (8): e124.

https://doi.org/10.1371/journal.pmed.0020124

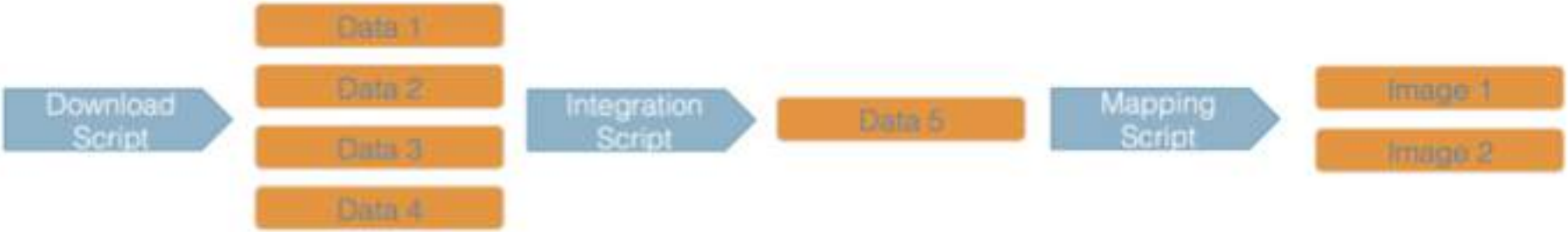


Computational Reproducibility

- Preservation enables:
  - Understanding
  - Evaluation
  - Reuse
- Future You!



Computational Workflows



Data Packages

Raw data package

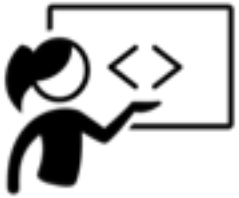
Derived data package

NSF Arctic Data Center

Data archive

Data discovery portal

Support services



Training
& Outreach

Data rescue

Icons by bezier master, Bigg Boss Icon Store, Dinosoft Labs, Landan Lloyd, Mazil, Nicolas Vicent, sachan for the Noun Project

Search

Search phrase

Filter by:

- Data attribute
- Creator
- Year
- Identifier
- Taxon
- Location

DATASETS 1 TO 25 OF 5,024

1 2 3 ... 201 Next

Sort by: Most recent

Zhangxian Ouyang. 2017. **Underway fCO2 measurement in the western Arctic Ocean CHINARE2010, CHINARE2012, and CHINARE2014, 2010, 2012, 2014.** Arctic Data Center. urn:uuid:51cc87e5-da30-4316-8cdf-21a671d61240.

Baoshan Chen. 2018. **Chinese Arctic Research Expedition 2012 (CHINARE12) cruise western Arctic Ocean carbonate data.** Arctic Data Center. doi:10.18739/A25C83.

Baoshan Chen. 2018. **Chinese Arctic Research Expedition 2010 (CHINARE10) cruise western Arctic Ocean carbonate data.** Arctic Data Center. doi:10.18739/A2S05T.

Jacqueline M. Grebmeier and Lee W. Cooper. 2016. **Collaborative Research: The Distributed Biological Observatory (DBO)-A Change Detection Array in the Pacific Arctic Region.** Arctic Data Center. urn:uuid:2b0f44fb-810e-4c30-a63f-9dd2f3f00a7.

Jacqueline M. Grebmeier and Lee W. Cooper. 2018. **Benthic macroinfaunal samples collected from the CCGS Sir Wilfrid Laurier, Northern Bering Sea to Chukchi Sea, 2012.** Arctic Data Center. doi:10.18739/A2FS1H.

Kang Wang, Irina Overeem, Elchin Jafarov, Gary Clow, Vladimir Romanovsky, et al. 2018. **A synthesis dataset of near-surface permafrost conditions for Alaska, 1997-2016.** Arctic Data Center. doi:10.18739/A2KS5S.

Rainer Amon. 2017. **Estimating fluxes of greenhouse gasses along the Yenisei River, Siberia, 2016.** Arctic Data Center. doi:10.18739/A2VW0W.

Lauren Andrews. 2018. **GPS-derived data from the Pâkitsoq Region, Western Greenland Ice Sheet during the 2011 summer melt season.** Arctic Data Center. doi:10.18739/A2FS0H.

Operations Metrics

5,100+
DATA SETS

2,000
CREATORS

700K+
DATA FILES

6,000+
USERS

26 TB
DATA

10K+
DOWNLOADS/MO

Anna-Maria Virkkala and Miska Luoto. 2018. Arctic Chamber Metadata, 2000-2018. Arctic Data Center. doi:10.18739/A28C6Q.

Copy Citation Quality report

Files in this dataset Package: resource_map_doi:10.18739/A28C6Q

Name		File type	Size	Downloads	Download All
Metadata: science_metadata.xml		EML v2.1.1	33 KB	50 views	Download
Virkkala_ArcticChamber_2018.csv	More info	text/csv	191 KB	12 downloads	Download

General

Identifier	doi:10.18739/A28C6Q
Abstract	This data summarizes the metadata of terrestrial Arctic or sub-Arctic CO2 flux chamber studies published in the 21st century. It provides descriptive information regarding the studies in general (title, keywords, authors), sites (coordinates, region), measurements (chamber size, measurement device, measurement period, fluxes), and measured plots (species, vegetation type). We aim to update the table every few years to keep track of the current state and distribution of chamber studies.

Citing Data

Anna-Maria Virkkala and Miska Luoto. 2018. Arctic Chamber Metadata, 2000-2018. Arctic Data Center. doi:10.18739/A28C6Q.

https://doi.org/10.18739/A28C6Q

Practical Reproducibility

NSF Arctic Data Center

Preserve the data

Preserve the software workflow

Document what you did

Describe how to interpret it all

knb

DataONE

Data and Metadata Guidelines

A Data Life Cycle

A Data Life Cycle

Guidelines

https://arcticdata.io/submit/

- Who Must Submit?
- Organizing Data
- File Formats
- Large Data Packages
- Metadata
- Data Identifiers
- Provenance
- Licensing and Distribution

Troms Fylke

Who Must Submit? https://arcticdata.io/submit/#who-must-submit

- **Arctic Research Opportunities (ARC):**
  - Complete metadata and all data and derived products
  - Within 2 years of collection or before end of award

- **Arctic Observing Network (AON):**
  - Complete metadata and all data
  - Real-time data made public immediately
  - QA'ed data within 6 months of collection

Who Must Submit? 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

Organizing Data

- Understand basics of “tidy” data models
- Design and create effective data tables

- **Benefits of tidy data systems**
- Powerful search and filtering
- Handle large, complex data sets
- Enforce data integrity
- Decrease errors from redundant updates

Not Tidy: Multiple Tables

AtlasGroveCOMPLETE.xls

Table 1

species	tree	main trunks kg	reiterated trunks kg	limbs kg	branches kg	leaves kg
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9
SESE	Buena Vista	128833.0	3486.5	0.0	8552.1	518.4
SESE	Demeter	155896.0	11085.6	3204.3	10054.1	768.7
SESE	Epimetheus	226987.0	12915.7	1797.2	13585.2	1029.4
SESE	Iluvatar	349586.6	65003.9	12315.6	13987.0	1461.8
SESE	Kronos	134154.1	12204.4	7232.7	5036.1	597.3
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6	762.2
SESE	Pleiades II	235838.8	11183.4	4306.0	11306.5	877.7
SESE	Prometheus	239414.0	25228.9	1612.6	12458.2	1086.0
SESE	Rhea	147710	487.5	730.1	5524.2	691.2
SESE	Zeus	241366	885.5	1620.4	19104.7	954.3
SESE	3	76	0.0	0.0	87.6	41.4
SESE	4	6312.0	356.0	73.5	214.1	43.8
SESE	5	206.0	0.0	0.0	8.7	2.5
SESE	6E	18697.4	0.0	0.0	1055.2	66.3
SESE	6W	14651.5	7.7	0.0	626.3	49.6
SESE	11	614.4	0.0	0.0	28.1	17.0
SESE	12	232.1	0.0	0.0	11.2	10.3
SESE	18	15632.0	0.0	0.0	946.3	106.8
SESE	19	11805.5	0.0	0.0	770.1	80.3
SESE	20	309.5	0.0	0.0	12.5	5.9
SESE	22	25618.3	0.0	0.0	1504.0	120.2
SESE	23	463.7	0.0	0.0	18.9	4.5
SESE	25	87.7	0.0	0.0	4.1	1.3
SESE	30	512.1	1.8	0.0	18.7	8.7

Table 2

type	species	dry masses (kg) main trunk	reiteration	limb	branch	leaf	TOTAL	% total
tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
tree	PSME	135815	0	0	8338	961	145114	3.3876
tree	THSE	31799	0	0	6343	864	39006	0.9105
tree	ACMA	4444	0	0	925	264	5634	0.1315
tree	UMCA	2921	0	0	937	273	4131	0.0964
shrub	RUSP	0	0	0	1974	686	2660	0.0620
fern	POMU	[illegible]	0	0	[illegible]	1271	1271	0.0296
shrub	VAOV	0	0	0	52	26	552	0.0129
shrub	COCO	0	0	0	24	6	289	0.0067
fern	POSC	0	0	0	107	89	196	0.0045
tree	RHPU	100	0	0	44	18	162	0.0037
herb	OXOR	0	0	0	0	112	112	0.0026
shrub	VAPA	0	0	0	94	4	99	0.0023
tree	PISI	0	0	0	1	0	1	0.0000
tree	CHLA	0	0	0	1	0	1	0.0000
shrub	GASH	0	0	0	0	0	0	0.0000
shrub	SACA	0	0	0	0	0	0	0.0000
		3744390	213247	53714	250519	21767	4283638	

Table 3

	main trunk	reiteration	limb	branch	leaf	total	proportion geophytic
SESE geo	3569312	213247	53714	230945	17192	4084409	1.00
SESE epi	0	0	0	0	0	0	
PSME geo	135815	0	0	8338	961	145114	1.00
PSME epi	0	0	0	0	0	0	
TSHE geo	31740	0	0	8332	960	38932	0.99
TSHE epi	59	0	0	12	4	74	
ACMA geo	4444	0	0	925	264	5634	1.00
ACMA epi	0	0	0	0	0	0	

Not Tidy: Inconsistent observations

AtlasGroveCOMPLETE.xls

species	tree	main trunks kg	reiterated trunks kg	limbs kg	branches kg	leaves kg
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9
SESE	Buena Vista	128533.0	3486.5	0.0	8552.1	518.4
SESE	Demeter	155896.0	110[illegible]	3204.3	10054.1	768.7
SESE	Epimetheus	226987.0	12915[illegible]	1797.2	13585.2	
SESE	Iluvatar	349586.6	65003.9	[illegible]5.6	13987.0	
SESE	Kronos	134154.1	12204.4	72[illegible]	5036.1	
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6	
SESE	Pleiades II	235838.8	11183.4	4306.0	[illegible]06.5	
SESE	Prometheus	239414.0	25228.9	1612.6	12[illegible]2	
SESE	Rhea	143710.4	487.8	730.1	5524.2	
SESE	Zeus	243365.7	2885.5	1620.4	19104.7	
SESE	3	1761.3	0.0	0.0	87.6	
SESE	4	6312.0	356.0	73.5	214.1	
SESE	5	206.0	0.0	0.0	8.7	
SESE	6E	18697.4	0.0	0.0	1055.2	
SESE	6W	14651.5	7.7	0.0	626.3	
SESE	11	614.4	0.0	0.0	28.1	17.0
SESE	12	232.1	0.0	0.0	11.2	10.3
SESE	18	15632.0	0.0	0.0	946.3	106.8
SESE	19	11805.5	0.0	0.0	770.1	80.3
SESE	20	309.5	0.0	0.0	12.5	5.9
SESE	22	25618.3	0.0	0.0	1504.0	120.2
SESE	23	463.7	0.0	0.0	18.9	4.5
SESE	25	87.7	0.0	0.0	4.1	1.3
SESE	30	512.1	1.8	0.0	18.7	8.7

type	species	dry masses (kg) main trunk	reiteration	limb	branch	leaf	TOTAL	% total
tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
tree	PSME	135815	0	0	8338	961	145114	3.3876
tree	THSE	31799	0	0	6343	864	39006	0.9105
tree	ACMA	4444	0	0	925	264	5634	0.1315
tree	UMCA	2921	0	0	937	273	4131	0.0964
shrub	RUSP	0	0	0	1974	686	2660	0.0620
			0	0	0	1271	1271	0.0296
			0	0	526	26	552	0.0129
			0	0	284	6	289	0.0067
			0	0	107	89	196	0.0045
			0	0	44	18	162	0.0037
			0	0	0	112	112	0.0026
			0	0	94	4	99	0.0023
			0	0	1	0	1	0.0000
			0	0	1	0	1	0.0000
			0	0	0	0	0	0.0000
			0	0	0	0	0	0.0000
			247	53714	250519	21767	4283638	

		main trunk	reiteration	limb	branch	leaf	total	proportion geophy tic
SESE geo		3569312	213247	53714	230945	17192	4084409	1.00
SESE epi		0	0	0	0	0	0	
PSME geo		135815	0	0	8338	961	145114	1.00
PSME epi		0	0	0	0	0	0	
TSHE geo		31740	0	0	6332	860	38932	0.99
TSHE epi		59	0	0	12	4	74	
ACMA geo		4444	0	0	925	264	5634	1.00
ACMA epi		0	0	0	0	0	0	

All the same observation?
No.

Not Tidy: Inconsistent variables

AtlasGroveCOMPLETE.xls

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
		main trunks	reiterated trunks	limbs	branches	leaves						dry mass (kg)				
species	tree	kg	kg	kg	kg	kg		type	species	main trunk	reiteration	limb	branch	leaf	TOTAL	% total
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2		tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8		tree	PSME	135815	0	0	8338	961	145114	3.3876
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4		tree	THSE	31799	0	0	6343	864	39006	0.9105
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9		tree	ACMA	4444	0	0	925	264	5634	0.1315
SESE	Buena Vista	128833.0	3486.5	0.0	8552.1	518.4		tree	UMCA	2921	0	0	937	273	4131	0.0964
SESE	Demeter	155896.0	11085.6	3204.3	10054.1	768.7		shrub	RUSP	0	0	0	1974	686	2660	0.0620
SESE	Epimetheus	226987.0	12915.7	1797.2	13585.2	1029.4		fern	POMU	0	0	0	0	1271	1271	0.0296
SESE	Iluvatar	349586.6	65003.9	12315.6	13987				VACV		0	0	526	26	552	0.0129
SESE	Kronos	134154.1	12204.4	7232.7	503						0	0	284	6	289	0.0067
SESE	Pleiades I	182385.2	3735.0	1935.2	1084						0	0	107	89	196	0.0045
SESE	Pleiades II	235838.8	11183.4	4306.0	1130						0	0	44	18	162	0.0037
SESE	Prometheus	239414.0	25228.9	1612.6	1245						0	0	0	112	112	0.0026
SESE	Rhea	143710.4	487.8	730.1	552						0	0	94	4	99	0.0023
SESE	Zeus	243365.7	2885.5	1620.4	1910						0	0	1	0	1	0.0000
SESE	3	1761.3	0.0	0.0	8							0	1	0	1	0.0000
SESE	4	6312.0	356.0	73.5	21						0	0	0	0	0	0.0000
SESE	5	206.0	0.0	0.0							0	0	0	0	0	0.0000
SESE	6E	18697.4	0.0	0.0	105						213247	53714	250519	21767	4283638	
SESE	6W	14651.5	7.7	0.0	62											proportion
SESE	11	614.4	0.0	0.0	2						eration	limb	branch	leaf	total	geophy tic
SESE	12	232.1	0.0	0.0	11.2	10.3			SESE geo	3569312	213247	53714	230945	17192	4084409	1.00
SESE	18	15632.0	0.0	0.0	946.3	106.8			SESE epi	0	0	0	0	0	0	
SESE	19	11805.5	0.0	0.0	770.1	80.3			PSME geo	135815	0	0	8338	961	145114	1.00
SESE	20	309.5	0.0	0.0	12.5	5.9			PSME epi	0	0	0	0	0	0	
SESE	22	25618.3	0.0	0.0	1504.0	120.2			TSHE geo	31740	0	0	6332	860	38932	0.99
SESE	23	463.7	0.0	0.0	18.9	4.5			TSHE epi	59	0	0	12	4	74	
SESE	25	87.7	0.0	0.0	4.1	1.3			ACMA geo	4444	0	0	925	264	5634	1.00
SESE	30	512.1	1.8	0.0	18.7	8.7			ACMA epi	0	0	0	0	0	0	

All the same variable?

No.

Not Tidy: Marginal info

AtlasGroveCOMPLETE.xls

species	tree	main trunks kg	reiterated trunks kg	limbs kg	branches kg	leaves kg
SESE	Atlas	255144.9	46020.6	5477.7	13433.2	1101.2
SESE	Ballantine	221966.4	7651.6	5922.9	11210.0	1084.8
SESE	Bell	253246.4	5454.3	5792.6	48500.7	1043.4
SESE	Broken Top	130928.9	4805.2	1608.1	5137.4	729.9
SESE	Buena Vista	128833.0	3486.5	0.0	8552.1	518.4
SESE	Demeter	155896.0	11085.6	3204.3	10054.1	768.7
SESE	Epimetheus	226987.0	12915.7	1797.2	13585.2	1029.4
SESE	Iluvatar	349586.6	65003.9	12315.6	13987.0	1461.8
SESE	Kronos	134154.1	12204.4	7232.7	5036.1	597.3
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6	762.2
SESE	Pleiades II	235838.8	11183.4	4306.0	11306.5	877.7
SESE	Prometheus	239414.0	25228.9	1612.6	12458.2	1086.0
SESE	Rhea	143710.4	487.8	730.1	5524.2	691.2
SESE	Zeus	243365.7	2885.5	1620.4	19104.7	954.3
SESE	3	1761.3	0.0	0.0	87.6	41.4
SESE	4	6312.0	356.0	73.5	214.1	43.8
SESE	5	206.0	0.0	0.0	8.7	2.5
SESE	6E	18697.4	0.0	0.0	1055.2	66.3
SESE	6W	14651.5	7.7	0.0	626.3	49.6
SESE	11	614.4	0.0	0.0	28.1	17.0
SESE	12	232.1	0.0	0.0	11.2	10.3
SESE	18	15632.0				
SESE	19	11805.5				
SESE	20	309.5				
SESE	22	25618.3				
SESE	23	463.7				
SESE	25	87.7				
SESE	30	512.1				

type	species	dry masses (kg) main trunk	reiteration	limb	branch	leaf	TOTAL	% total
tree	SESE	3569312	213247	53714	230945	17192	4084409	95.3491
tree	PSME	135815	0	0	8338	961	145114	3.3876
tree	THSE	31799	0	0	6343	864	39006	0.9105
tree	ACMA	4444	0	0	925	264	5634	0.1315
tree	UMCA	2921	0	0	937	273	4131	0.0964
shrub	RUSP	0	0	0	1974	686	2660	0.0620
fern	POMU	0	0	0	0	1271	1271	0.0296
shrub	VAOV	0	0	0	526	26	552	0.0129
shrub	COCO	0	0	0	284	6	289	0.0067
fern	POSC	0	0	0	107	89	196	0.0045
tree	RHPU	100	0	0	44	18	162	0.0037
herb	OXOR	0	0	0	0	112	112	0.0026
shrub	VAPA	0	0	0	94	4	99	0.0023
tree	PISI	0	0	0	1	0	1	0.0000
tree	CHLA	0	0	0	1	0	1	0.0000
shrub	GASH	0	0	0	0	0	0	0.0000
shrub	SACA	0	0	0	0	0	0	0.0000
		3744390	213247	53714	250519	21767	4283638	

		main trunk	reiteration	limb	branch	leaf	total	proportion geophy tic
SESE		3569312	213247	53714	230945	17192	4084409	1.00
	SE epi	0	0	0	0	0	0	
	ME geo	135815	0	0	8338	961	145114	1.00
	ME epi	0	0	0	0	0	0	
	HE geo	31740	0	0	6332	860	38932	0.99
	HE epi	59	0	0	12	4	74	
	MA geo	4444	0	0	925	264	5634	1.00
	MA epi	0	0	0	0	0	0	

Marginal sums and totals

Data Modeling 101

id	date	site	elev	sp1code	sp1height	sp2code	sp2height
1	2017-10-10	1	3.7	DAPU	4.6	DAMA	4.5
2	2017-09-05	2	3.2	DAMA	3.5	DAPU	3.9

- Denormalized data (aka, not Tidy)
- Observations about different entities combined

Tidy Data (observe one entity per table)

- Species observations

id	date	site	spcode	height
1	2017-10-10	1	DAPU	4.6
2	2017-09-05	2	DAMA	3.5
3	2017-10-10	1	DAMA	4.5
4	2017-09-05	2	DAPU	3.9

- Site observations

site	name	elev	temp
1	Taku	3.7	21.2
2	Lituya	3.2	23.1

Tidy Data (Relational) Join Key

- Species observations

id	date	site	spcode	height
1	2017-10-10	1	DAPU	4.6
2	2017-09-05	2	DAMA	3.5
3	2017-10-10	1	DAMA	4.5
4	2017-09-05	2	DAPU	3.9

- Site observations

site	name	elev	temp
1	Taku	3.7	21.2
2	Lituya	3.2	23.1

Organizing Data: Best Practices

- **Some Simple Guidelines for Effective Data Management.**
  - Borer et al. 2009. Bulletin of the Ecological Society of America. https://doi.org/10.1890/0012-9623-90.2.205

- **Nine simple ways to make it easier to (re)use your data.**
  - White et al. 2013. Ideas in Ecology and Evolution 6. https://doi.org/10.4033/iee.2013.6b.6.f

Organizing Data: Best Practices

- **Scripts** for all data manipulation
  - Uncorrected raw data file
  - Document processing in scripts
- **Design to add rows, not columns**
  - Each column one variable
  - Each row one observation
- **Nonproprietary file formats**
  - Descriptive names, no spaces
  - Header line

File Formats

https://arcticdata.io/submit/#file-format-guidelines

- **Open Formats**
  - **Text** - support long term access and preservation
  - **Open binary** formats (NetCDF, HDF5)

Always bet on text!

- Any (meta)data is better than none
  - Microsoft Excel: common but proprietary
  - Export GIS data to ESRI shapefiles
  - Export MATLAB, IDL, etc. to NetCDF

Large Data Packages (> Terabytes)

- Talk to the data center early
- Tile data structures by subset
  - Spatial regions
  - Temporal windows
  - Measured variables
- Use efficient tools (NetCDF, HDF)
  - Compact data format
  - Parallel read/write libraries

Metadata Guidelines

Metadata: the Goal

- Target a typical researcher (maybe you!)
- 30+ years from now

- Goal
  - Understand
  - Interpret
  - Re-use

Metadata: the Goal

- **What** was measured?
- **Who** did it?
- **When** and **where**?
- **How**? (data structure & methods)
- **Why**? (science context)
- **Attribution** & **Licensing**

Metadata

Metadata: Bibliographic Details

- **Global Identifier (e.g., DOI)**
- **Descriptive title**
  - topic, geographic location, dates, and, if applicable, the scale of the data
- **Descriptive abstract**
  - brief overview of the specific contents and purpose of the data package.
- **Funding** information (award number and sponsor).
- **People and organizations**
  - **Creators** – who should be cited for the data set
  - Contacts
  - Contributors
  - Sponsors, and more

Metadata: Discovery Details

- **Geospatial coverage**
  - Field and laboratory sampling locations
  - including place names and precise coordinates
- **Temporal Coverage**
  - When measurements were made
  - To what time period do measurements apply
  - Might be calendar times, or geologic times
- **Taxonomic Coverage**
  - What species were measured
  - Taxonomy standards and procedures
- Other contextual information

Metadata: Interpretation Details

- Field and laboratory data **collection methods**
- Full **experimental and project design**, and relationship to data
- Full field and laboratory sample **processing methods**
- **Sampling quality control** procedures

- Analysis and modeling methods
  - **Provenance** information
  - **Hardware** and **software** used
    - including make, model, and version
  - **Computing quality control** procedures
    - testing, code review, etc.

Metadata: Data Structure and Contents

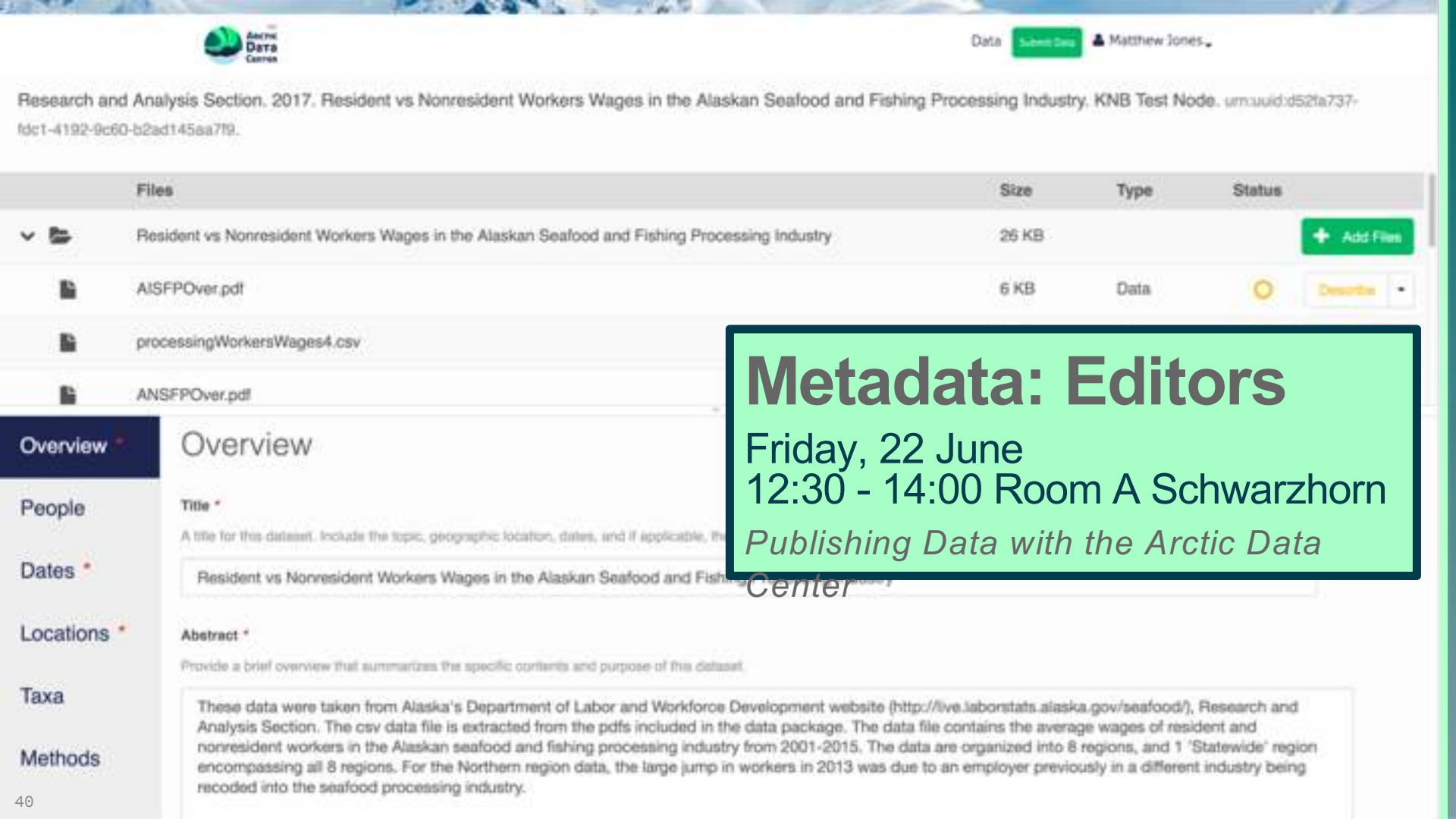
- **Data model description**
- **Data object descriptions (granules)**
  - Tables
  - Images
  - Matrices
  - Spatial layers, etc.
- **Variable information** (attributes/parameters)
  - Definitions / link to methods
  - Standardized measurement types
  - Units
  - Coded values
  - Missing value codes

Metadata: Rights and Attribution

- **Scientific rights and expectations**
  - **Citation format**
  - **Attribution expectations**
  - **Reuse rights**
    - Who may reuse data, and for what purposes
  - **Redistribution rights**
    - Who may copy and redistribute data and metadata
- **Legal terms and conditions**
  - **Licensing** terms

Metadata Standards

- Ecological Metadata Language (EML)
- Geospatial Metadata Standards
  - (ISO 19115*, ISO 19139)
- Biological Data Profile (BDP)
- Dublin Core
- Darwin Core
- PREMIS and METS
- … and the list goes on



Metadata: Editors

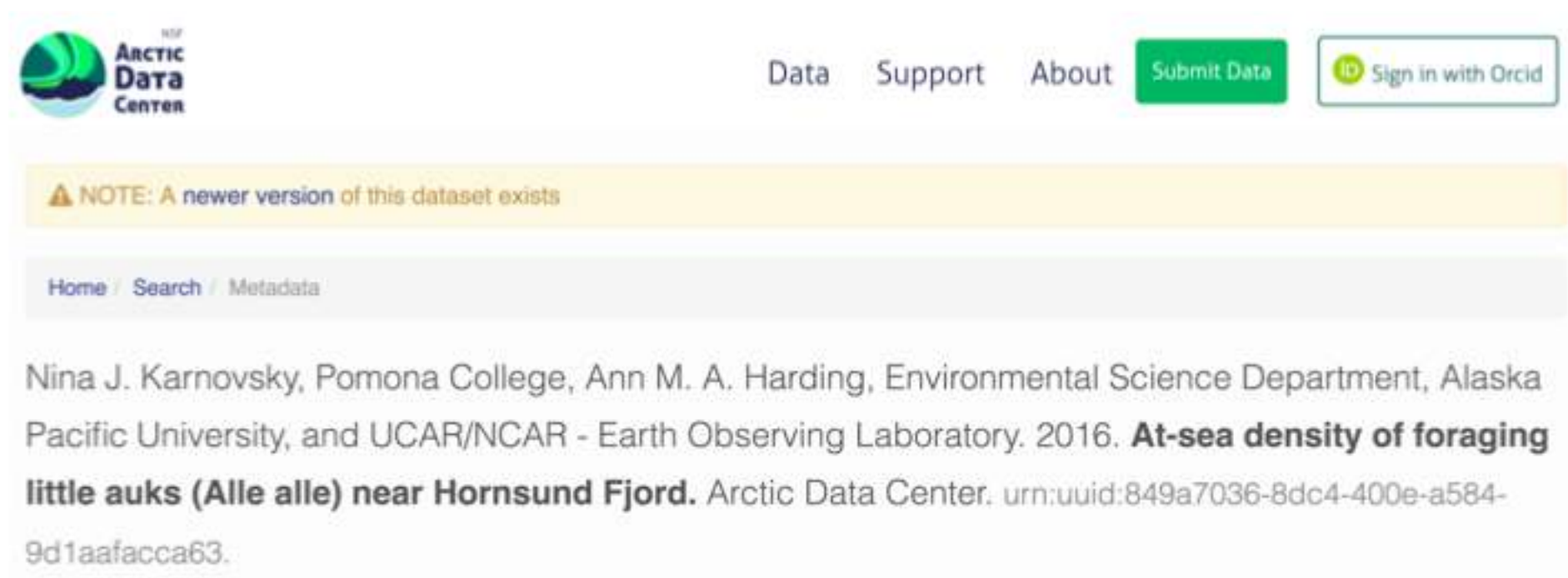
Friday, 22 June
12:30 - 14:00 Room A Schwarzhorn

Publishing Data with the Arctic Data Center

Data Identifiers

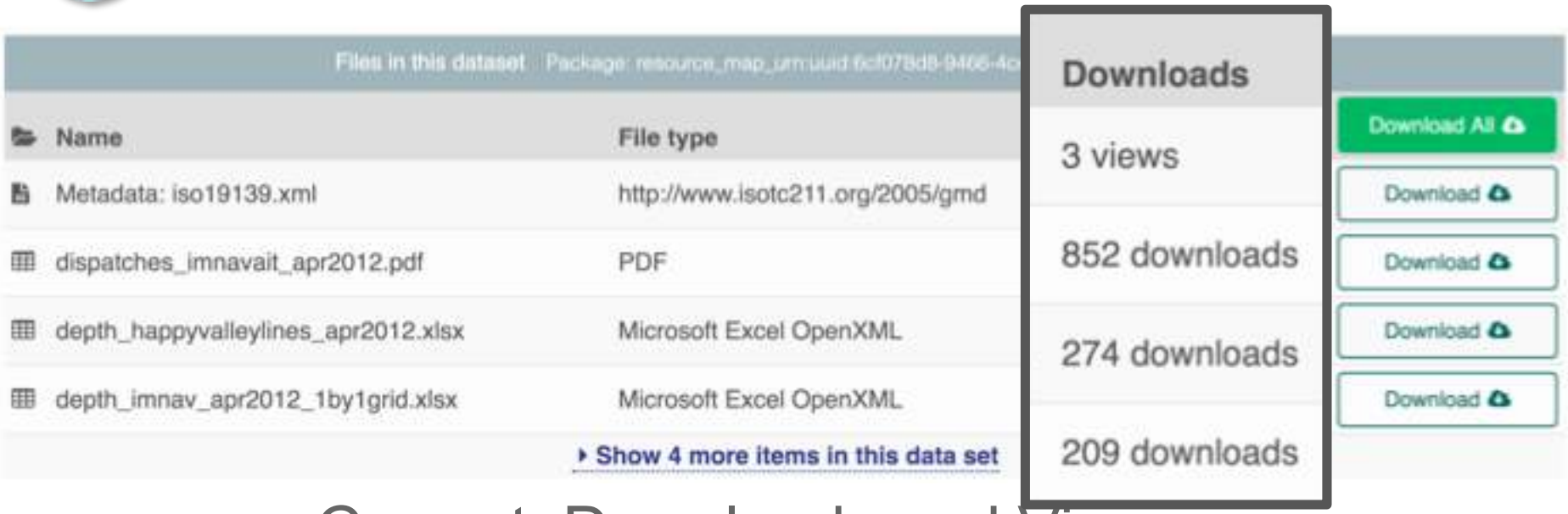
Nina J. Karnovsky and Ann M. A. Harding. 2016. At-sea density of foraging little auks (Alle alle) near Hornsund Fjord. Arctic Data Center. doi:10.5065/D6MK6B17.

- DOI == Digital Object Identifier
- We assign a DOI to each published data set
- Researchers should cite data they use



- Each update has a unique identifier
- Cite the exact version used
- Newer versions are clearly indicated

Data Usage Metrics

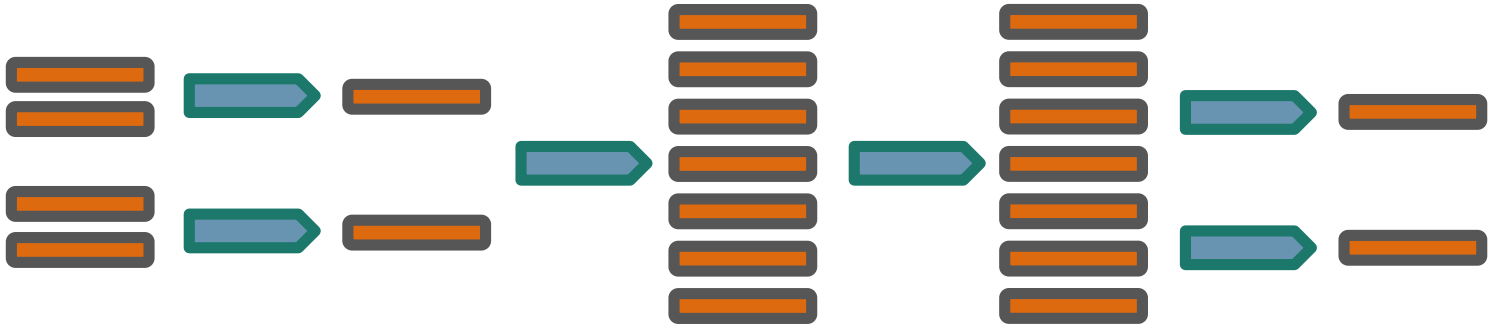


- Current: Downloads and Views
- Future: Citations

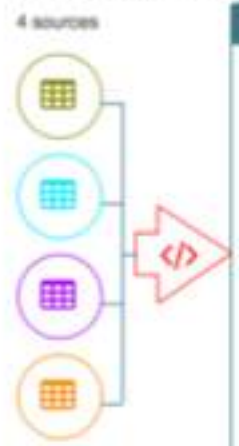
MAKE DATA COUNT

Provenance Metadata

- Simplified view of complex workflows



Data Table, Image, and Other Data Details

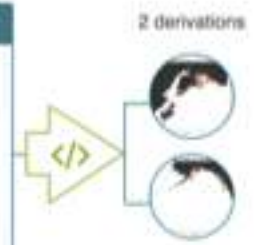


Data Table

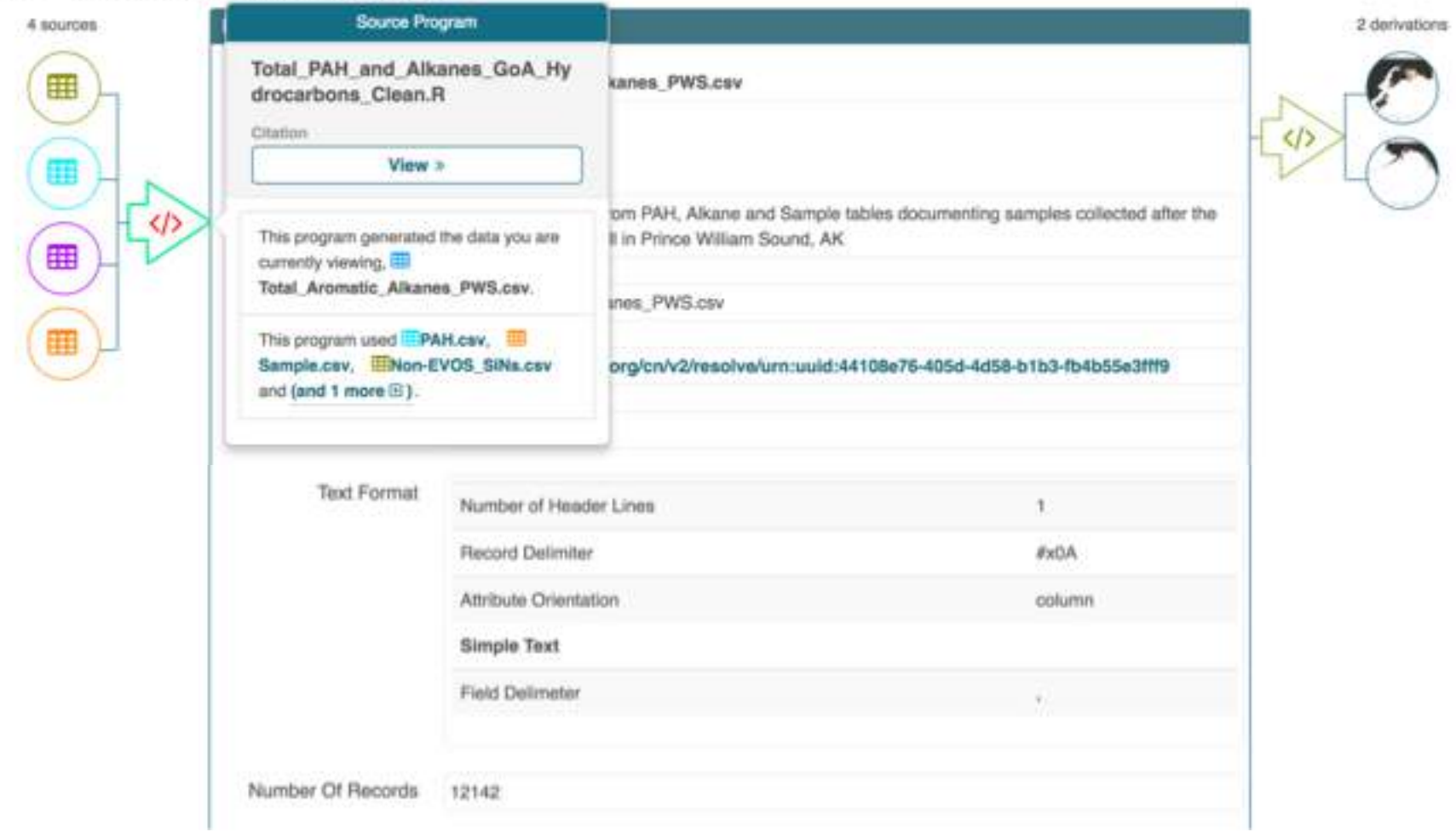
Entity Name	Total_Aromatic_Alkanes_PWS.csv
	Download
Description	Combined dataset from PAH, Alkane and Sample tables documenting samples collected after the Exxon Valdez oil spill in Prince William Sound, AK
Object Name	Total_Aromatic_Alkanes_PWS.csv
Online Distribution Info	https://cn.dataone.org/cn/v2/resolve/urn:uuid:44108e76-405d-4d58-b1b3-fb4b55e3fff9
Size	2801033 byte
Text Format	

Number of Header Lines	1
Record Delimiter	#x0A
Attribute Orientation	column
Simple Text	
Field Delimeter	,

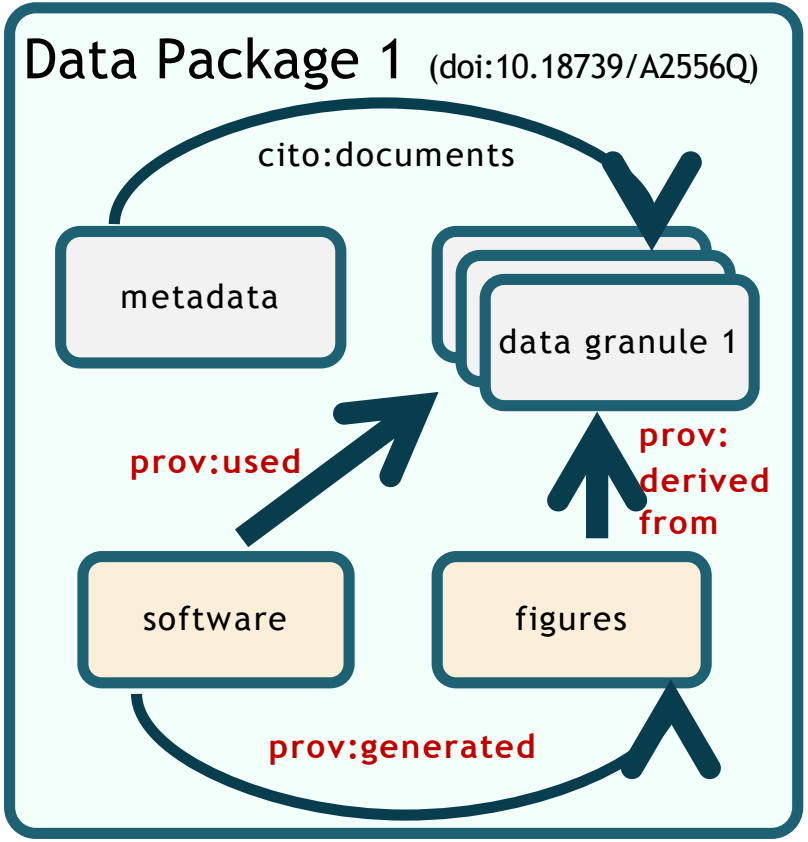
Number Of Records	12142



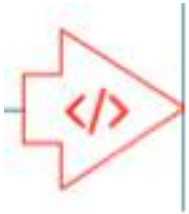
Data Table, Image, and Other Data Details



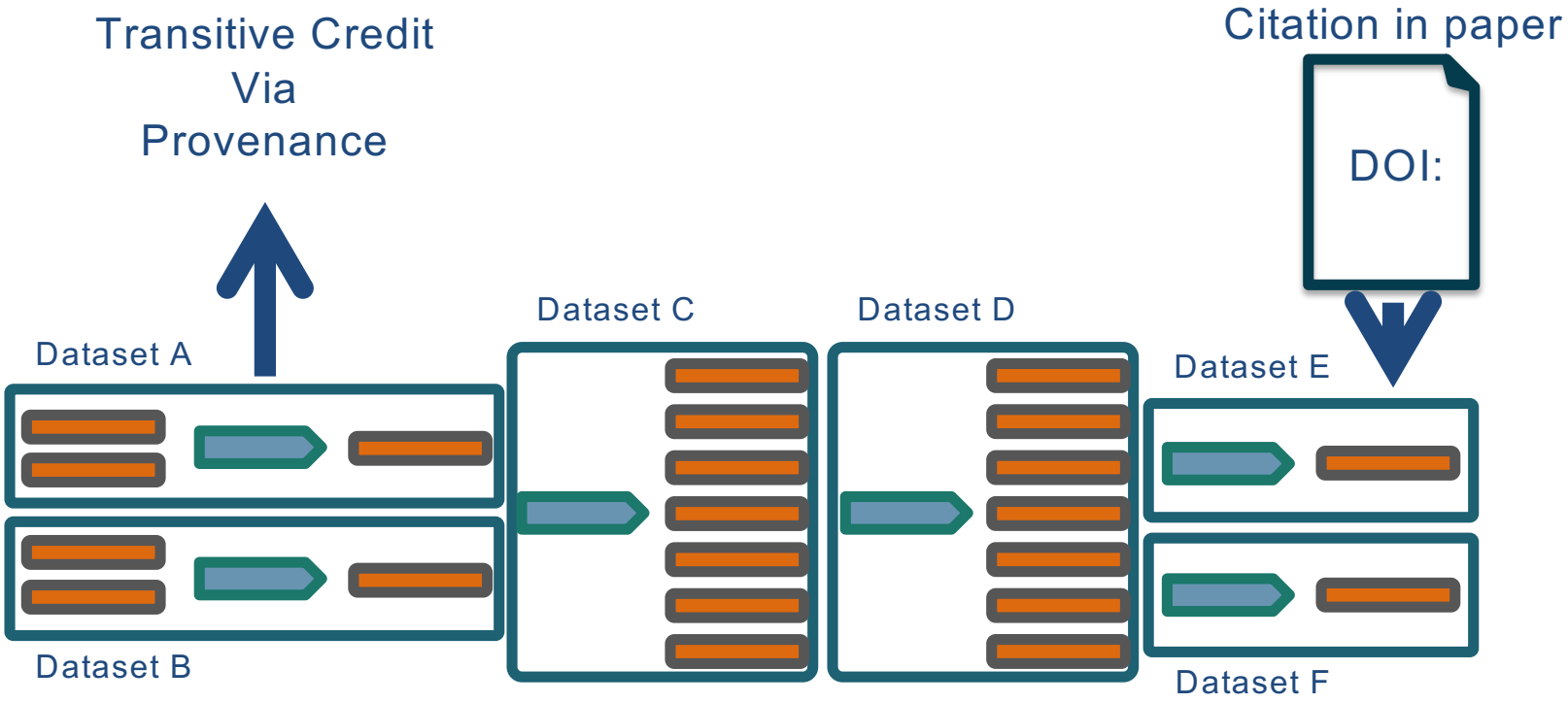
Data package with Provenance



Rmarkdown as Provenance



Citing multi-generational workflows



Licensing and Distribution

- **CC-0** Public Domain Dedication:

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

- **CC-BY** Creative Commons Attribution:

"… 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."*

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

Guidelines

https://arcticdata.io/submit/

- Who Must Submit?
- Organizing Data
- File Formats
- Large Data Packages
- Metadata
- Data Identifiers
- Provenance
- Licensing and Distribution

Troms Fylke

Arctic Data Center Support Team

support@arcticdata.io


Clark


Goldstein


Mullen


Chong


Meyer


Steves


Maier


Ochs


Train


Nguyen


Sun


Reevesman


Chen

Data Science Fellows

Student Interns

More POLAR 2018 Workshops

Friday, 22 June
12:30 - 14:00 Room A Schwarzhorn
Publishing Data with the Arctic Data Center

Friday, 22 June
18:30-21:30 Room Sanada C
Data and Drinks - Scientists & Data Managers

We're here to help!
Email: support@arcticdata.io
Website: https://arcticdata.io
@arcticdatactr

https://arcticdata.io