



NSF

ARCTIC
Data
Center

<https://arcticdata.io>

 @arcticdatactr

Best Practices: Data and Metadata Submission

Matthew B. Jones



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

NSF Award #1546024



DataONE



Computational Reproducibility

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

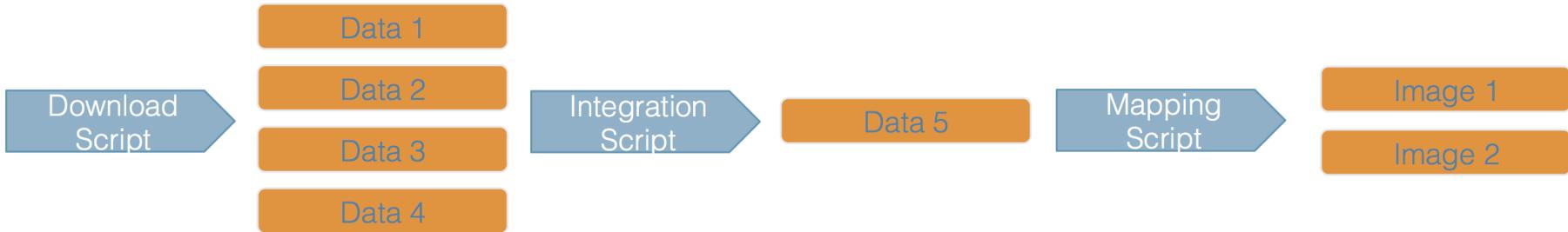


Metadata

Software

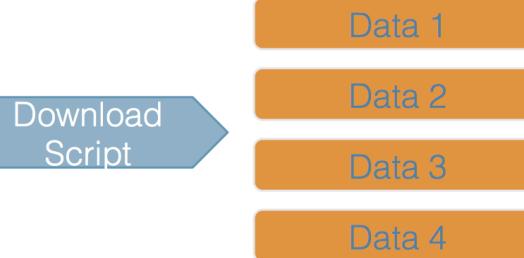


Computational Workflows

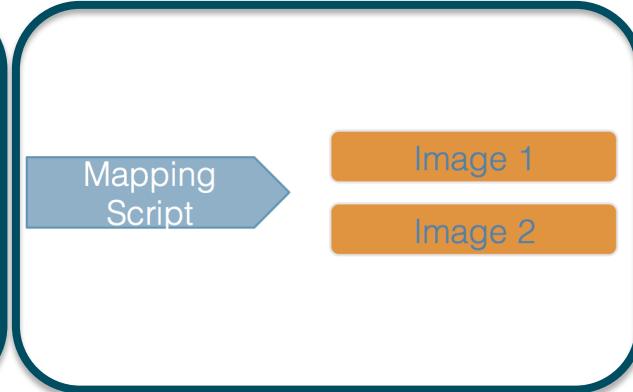




Data Packages



Raw data package



Derived data package



[Home](#) / [Search](#) / [Metadata](#)

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.



Practical Reproducibility



Preserve the data

Preserve the software workflow

Document what you did

Describe how to interpret it all



NSF
**ARCTIC
Data
Center**

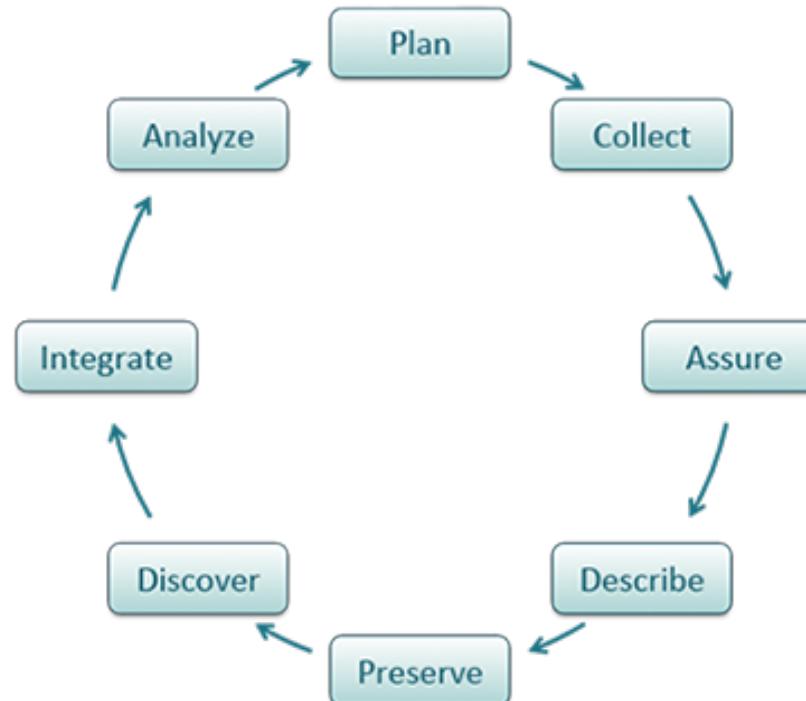




Data and Metadata Guidelines

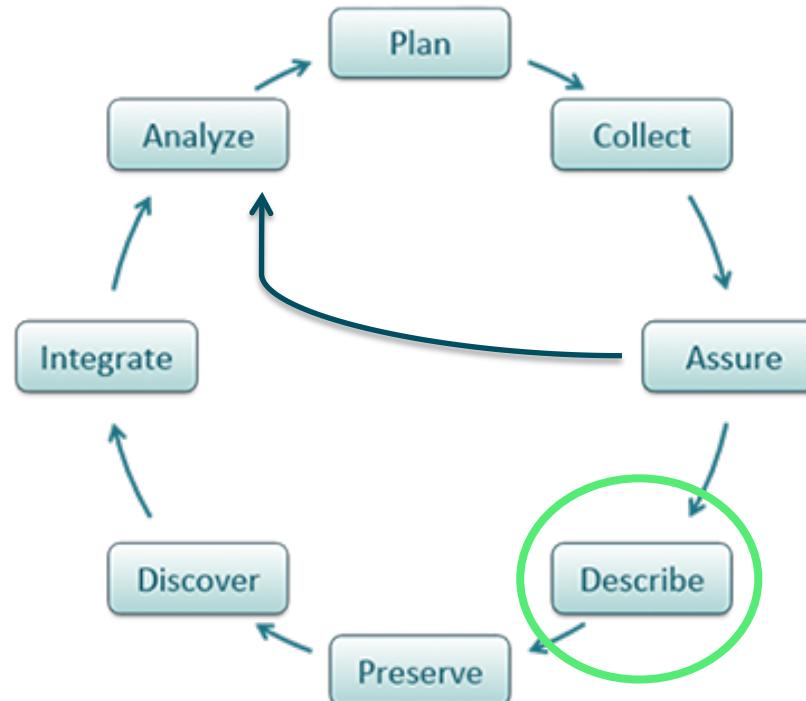


A Data Life Cycle





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





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

Table 1

species	tree	main trunks	reiterated trunks	limbs	branches	leaves
		kg	kg	kg	kg	kg
SESE	Atlas	255144.9	48020.6	5477.7	13433.2	1101.2
SESE	Ballantine	221966.4	7851.6	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	1481.8
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	147101.1	487.8	730.1	5524.2	691.2
SESE	Zeus	243679.1	1385.5	1620.4	19104.7	954.3
SESE	3	76.0	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	483.7	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

AtlasGroveCOMPLETE.xls

type	species	main trunk	reiteration	dry masses (kg)			TOTAL	% total
				limb	branch	leaf		
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	53	26	552	0.0129
shrub	COCO	0	0	0	84	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	4283636	

Table 2

	main trunk	reiteration	limb	branch	leaf	total	proportion	
							geo	phytic
SESE geo	3569312	213247	53714	230945	17192	4084409	1.00	
SESE epi	0	0	0	0	0	0	0	0
PSME geo	135815	0	0	8338	961	145114	1.00	
PSME epi	0	0	0	0	0	0	0	0
TSHE geo	31740	0	0	6332	360	38932	0.99	
TSHE epi	59	0	0	12	4	74	0	0
ACMA geo	4444	0	0	925	264	5634	1.00	
ACMA epi	0	0	0	0	0	0	0	0

Table 3



Not Tidy: Inconsistent observations

AtlasGroveCOMPLETE.xls																	
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	
species	tree	main trunks	reiterated trunks	limbs	branches	leaves		type	species	main trunk	reiteration	dry masses (kg)	limb	branch	leaf	TOTAL	% total
SESE	Atlas	255144.9	48020.6	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	253248.4	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	1104.5	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							0	0	0	1271	1271	
SESE	Iluvatar	349586.6	65003.9	11215.6	13987.0							0	0	526	26	552	
SESE	Kronos	134154.1	12204.4	7237.7	5036.1							0	0	284	6	289	
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6							0	0	107	89	196	
SESE	Pleiades II	235838.8	11183.4	4306.0	1306.5							0	0	44	18	162	
SESE	Prometheus	239414.0	25228.9	1612.6	12938.2							0	0	0	112	112	
SESE	Rhea	143710.4	487.8	730.1	5524.2							0	0	94	4	99	
SESE	Zeus	243385.7	2885.5	1620.4	19104.7							0	0	1	0	1	
SESE	3	1761.3	0.0	0.0	87.6							0	0	1	0	1	
SESE	4	6312.0	356.0	73.5	214.1							0	0	0	0	0	
SESE	5	206.0	0.0	0.0	8.7							0	0	0	0	0	
SESE	6E	18697.4	0.0	0.0	1055.2							247	53714	250519	21767	4283636	
SESE	6W	14651.5	7.7	0.0	626.3	49.6										proportion	
SESE	11	614.4	0.0	0.0	28.1	17.0										geophytic	
SESE	12	232.1	0.0	0.0	11.2	10.3										1.00	
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	483.7	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											

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		
species	tree	main trunks	reiterated trunks	limbs	branches	leaves		type	species	main trunk	reiteration	dry mass	ses (kg)	branch	leaf	TOTAL	% total	
SESE	Atlas	255144.9	48020.6	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	0	1271	1271	0.0296	
SESE	Iluvatar	349586.6	65003.9	12315.6	13987.0	1481.8		shrub	VAOV	0	0	0	526	26	552	0.0129		
SESE	Kronos	134154.1	12204.4	7232.7	5036					0	0	0	284	6	289	0.0067		
SESE	Pleiades I	182385.2	3735.0	1935.2	10846					0	0	0	107	89	196	0.0045		
SESE	Pleiades II	235838.8	11183.4	4306.0	11306					0	0	0	44	18	162	0.0037		
SESE	Prometheus	239414.0	25228.9	1612.6	12456					0	0	0	0	112	112	0.0026		
SESE	Rhea	143710.4	487.8	730.1	5524					0	0	0	94	4	99	0.0023		
SESE	Zeus	243385.7	2885.5	1620.4	19104					0	0	0	1	0	1	0.0000		
SESE	3	1761.3	0.0	0.0	87					0	0	0	1	0	1	0.0000		
SESE	4	6312.0	356.0	73.5	214					0	0	0	0	0	0	0.0000		
SESE	5	206.0	0.0	0.0	8					0	0	0	0	0	0	0.0000		
SESE	6E	18897.4	0.0	0.0	1055					213247	53714	250519	21767	4283636				
SESE	6W	14651.5	7.7	0.0	626												proportion	
SESE	11	614.4	0.0	0.0	28												geophytic	
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	483.7	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												

All the same variable?
No.





Not Tidy: Marginal info

AtlasGroveCOMPLETE.xls

species	tree	main trunks	reiterated trunks	limbs	branches	leaves	type	species	main trunk	reiteration	limb	branch	leaf	TOTAL	% total
		kg	kg	kg	kg	kg			kg	kg	kg	kg	kg	kg	kg
SESE	Atlas	255144.9	48020.6	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	0	1271	1271
SESE	Iluvatar	349586.6	65003.9	12315.6	13987.0	1481.8	shrub	VAOV	0	0	0	526	26	552	0.0129
SESE	Kronos	134154.1	12204.4	7232.7	5036.1	597.3	shrub	COCO	0	0	0	284	6	289	0.0067
SESE	Pleiades I	182385.2	3735.0	1935.2	10846.6	762.2	fern	POSC	0	0	0	107	89	196	0.0045
SESE	Pleiades II	235838.8	11183.4	4306.0	11306.5	877.7	tree	RHPU	100	0	0	44	18	162	0.0037
SESE	Prometheus	239414.0	25228.9	1612.6	12458.2	1086.0	herb	OXOR	0	0	0	0	112	112	0.0026
SESE	Rhea	143710.4	487.8	730.1	5524.2	691.2	shrub	VAPA	0	0	0	94	4	99	0.0023
SESE	Zeus	243385.7	2885.5	1620.4	19104.7	954.3	tree	PISI	0	0	0	1	0	1	0.0000
SESE	3	1761.3	0.0	0.0	87.6	41.4	tree	CHLA	0	0	0	1	0	1	0.0000
SESE	4	6312.0	356.0	73.5	214.1	43.8	shrub	GASH	0	0	0	0	0	0	0.0000
SESE	5	206.0	0.0	0.0	8.7	2.5	shrub	SACA	0	0	0	0	0	0	0.0000
SESE	6E	18697.4	0.0	0.0	1055.2	66.3			3744390	213247	53714	250519	21767	4283636	
SESE	6W	14651.5	7.7	0.0	626.3	49.6									proportion
SESE	11	614.4	0.0	0.0	28.1	17.0									geophytic
SESE	12	232.1	0.0	0.0	11.2	10.3									1.0000
SESE	18	15632.0					SESE	SE epi	3569312	213247	53714	230945	17192	4084409	
SESE	19	11805.5					SE epi	0	0	0	0	0	0	0	
SESE	20	309.5					ME geo	135815	0	0	8338	961	145114	1.0000	
SESE	22	25618.3					ME epi	0	0	0	0	0	0	0	
SESE	23	483.7					HE geo	31740	0	0	6332	860	38932	0.99	
SESE	25	87.7					HE epi	59	0	0	12	4	74		
SESE	30	512.1					MA geo	4444	0	0	925	264	5634	1.0000	
							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)
- Any (meta)data is better than none
 - Microsoft Excel: common but proprietary
 - Export GIS data to ESRI shapefiles
 - Export MATLAB, IDL, etc. to NetCDF

**Always bet
on text!**



Csv

Xls

MP4

Avi

Html

Ppt

Gif

Txt

Eml

Tsv

Png

Zip

NetCDF

Pdf

XML

Jpeg

Docx



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



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



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



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



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



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



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

Research and Analysis Section. 2017. Resident vs Nonresident Workers Wages in the Alaskan Seafood and Fishing Processing Industry. KNB Test Node. urn:uuid:d52fa737-fdc1-4192-9c60-b2ad145aa7f9.

Files	Size	Type	Status
 Resident vs Nonresident Workers Wages in the Alaskan Seafood and Fishing Processing Industry	26 KB		
 AISFPOver.pdf	6 KB	Data	 
 processingWorkersWages4.csv	6 KB	Data	 
 ANSFPOver.pdf	6 KB	Data	 

Overview *

Overview

People

Title *

A title for this dataset. Include the topic, geographic location, dates, and if applicable, the scale of the data. Write out all abbreviations.

Resident vs Nonresident Workers Wages in the Alaskan Seafood and Fishing Processing Industry

Dates *

Abstract *

Provide a brief overview that summarizes the specific contents and purpose of this dataset.

Locations *

These data were taken from Alaska's Department of Labor and Workforce Development website (<http://live.laborstats.alaska.gov/seafood/>), Research and Analysis Section. The csv data file is extracted from the pdfs included in the data package. The data file contains the average wages of resident and nonresident workers in the Alaskan seafood and fishing processing industry from 2001-2015. The data are organized into 8 regions, and 1 'Statewide' region encompassing all 8 regions. For the Northern region data, the large jump in workers in 2013 was due to an employer previously in a different industry being recoded into the seafood processing industry.

Taxa

Methods



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

 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



Metadata Quality Improvement

Improve Metadata Quality at Level of Collection and Individual Data Set

The screenshot shows a web-based metadata quality report for a specific dataset. At the top, there's a navigation bar with the Arctic Data Center logo, menu items for Data, Support, About, and a prominent green 'Submit Data' button, along with a 'Sign in with Orcid' link.

The main content area displays a summary of 23 checks, with a large green circle indicating 12 passed, 3 warning, and 8 failed. Below this, four progress bars show completion levels for Identification (58%), Discovery (80%), Metadata (0%), and Interpretation (67%).

Below the progress bars, four sections list the check results:

- Passed:** Passed 12 checks out of 18. Good job! (Green background)
- Warning:** Warning for 3 checks. Please review these warnings. (Yellow background)
- Failed:** Failed 3 checks. Please correct these issues. (Red background)
- Informational:** 5 informational checks. These may include skips, errors and failures. (Light blue background)





Extensible quality checks

Check#	Check Name	Check	Type
M1	Descriptive Title	Title exists, > 7 words	Metadata
M2	Unique Attribute Names	Attribute names unique	Metadata
M3	Valid Units	Units assigned from controlled vocabulary	Metadata
M4	Schema valid	Metadata validates	Metadata
C1	Checksum matches	Data checksums match metadata	Congruency
C2	Data links live	All URLs return data	Congruency
D1	Duplicate data rows	Count duplicate rows	Data
...			

- Checks in Java, R, Python
- Categorized by function (discovery, re-use, ...)
- Operate across dialects (EML, CSDGM, ISO19139)



Recommendations

- Checks: like unit tests for recommendations
- Community Recommendations
 - Group of quality checks
 - Can be created by any community
 - Can include standard or custom checks

Recommendation	Checks
LTER Best Practice	M1, M2, C2, C3, D3, ...
ACDD	M2, M3, M4, C1, C2, D3, ...
Arctic Data Center	M3, M4, M5, C6, C8, D1, D2, D3, ...
...	

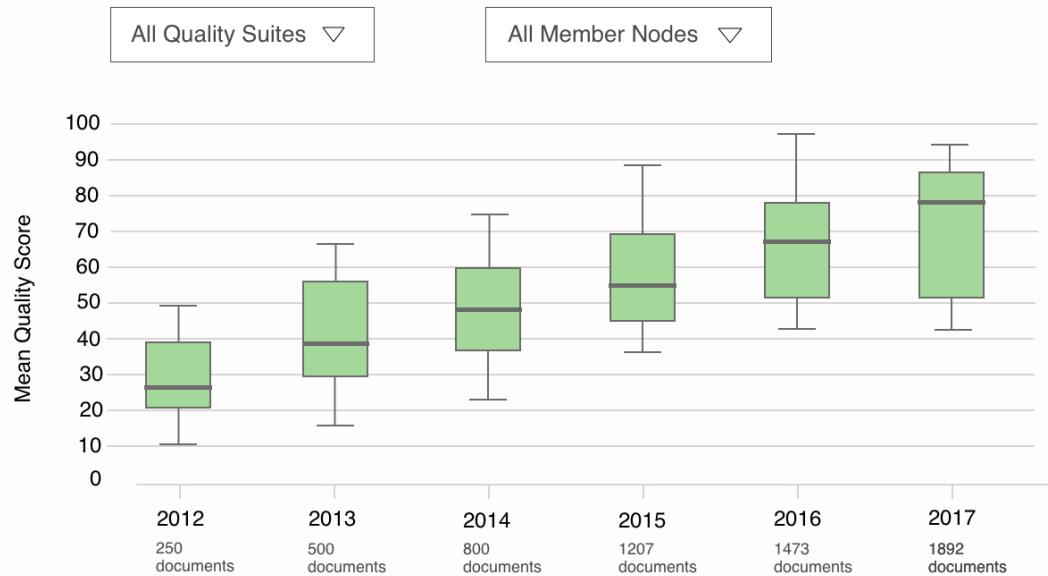
Longitudinal Quality Reports

Processing for millions of records

Metadata Quality

The quality score is a percentage of metadata quality checks that passed, out of the total number of checks. Scores are then averaged across all metadata documents.

The box plot to the right shows the quality scores for each time period. The dark line in each box is the median of the quality scores for that time period. Each time period shows the quality scores for all metadata that was publicly available during that period.







Data Usage Metrics

Files in this dataset Package: resource_map_urn:uuid:6cf078d8-9466-4ce0-8f3e-0a2a2a2a2a2a

Name	File type
Metadata: iso19139.xml	http://www.isotc211.org/2005/gmd
dispatches_imnavait_apr2012.pdf	PDF
depth_happyvalleylines_apr2012.xlsx	Microsoft Excel OpenXML
depth_imnav_apr2012_1by1grid.xlsx	Microsoft Excel OpenXML

[Show 4 more items in this data set](#)

Downloads

3 views

852 downloads

274 downloads

209 downloads

[Download All](#)

[Download](#)

[Download](#)

[Download](#)

[Download](#)

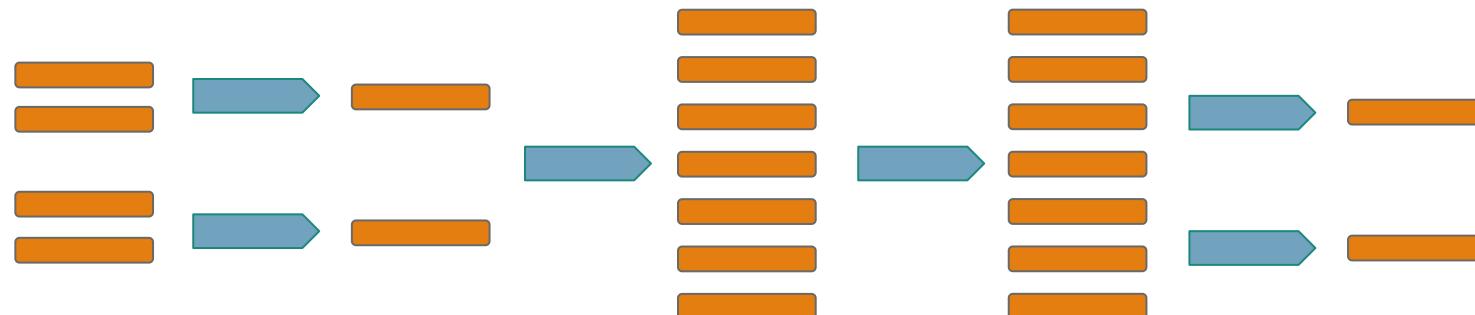
- Current: Downloads and Views
- Future: Citations





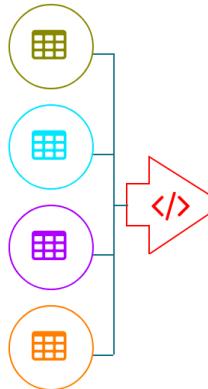
Provenance Metadata

- Simplified view of complex workflows



Data Table, Image, and Other Data Details

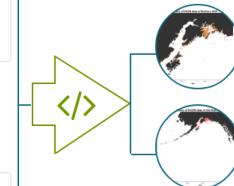
4 sources



Data Table

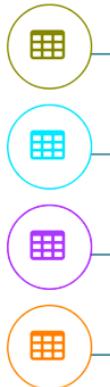
Entity Name	Total_Aromatic_Alkanes_PWS.csv										
	Download										
Description	Combined dataset from PAH, Alkane and Sample tables documenting samples collected after the Exxon Valdez oil spill in Prince William Sound, AK										
Object Name	Total_Aromatic_Alkanes_PWS.csv										
Online Distribution Info	https://cn.dataone.org/cn/v2/resolve/urn:uuid:44108e76-405d-4d58-b1b3-fb4b55e3fff9										
Size	2801033 byte										
Text Format	<table><tr><td>Number of Header Lines</td><td>1</td></tr><tr><td>Record Delimiter</td><td>#x0A</td></tr><tr><td>Attribute Orientation</td><td>column</td></tr><tr><td>Simple Text</td><td></td></tr><tr><td>Field Delimiter</td><td>,</td></tr></table>	Number of Header Lines	1	Record Delimiter	#x0A	Attribute Orientation	column	Simple Text		Field Delimiter	,
Number of Header Lines	1										
Record Delimiter	#x0A										
Attribute Orientation	column										
Simple Text											
Field Delimiter	,										
Number Of Records	12142										

2 derivations



Data Table, Image, and Other Data Details

4 sources



Source Program

Total_PAH_and_Alkanes_GoA_Hydrocarbons_Clean.R

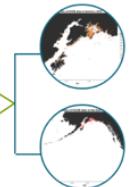
Citation

[View »](#)

This program generated the data you are currently viewing, Total_Aromatic_Alkanes_PWS.csv.

This program used PAH.csv, Sample.csv, Non-EVOS_SINs.csv and (and 1 more .

2 derivations



Alkanes_PWS.csv

from PAH, Alkane and Sample tables documenting samples collected after the oil spill in Prince William Sound, AK

Alkanes_PWS.csv

<http://doi.org/cn/v2/resolve/urn:uuid:44108e76-405d-4d58-b1b3-fb4b55e3fff9>

Text Format

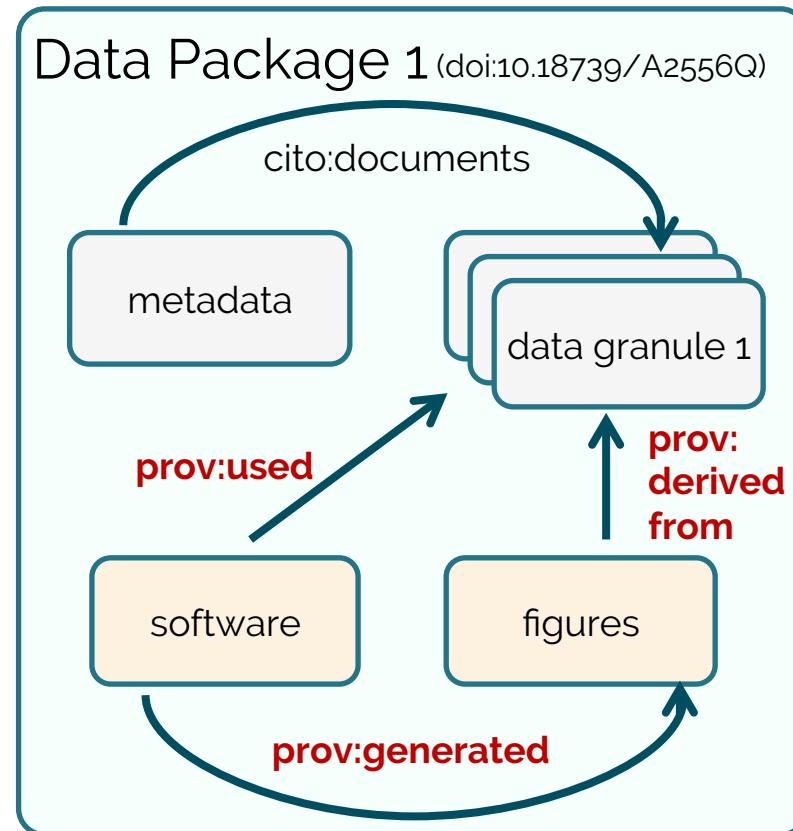
Number of Header Lines	1
Record Delimiter	#xA
Attribute Orientation	column
Simple Text	
Field Delimiter	,

Number Of Records

12142



Data package with Provenance



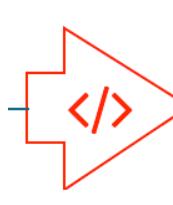


Rmarkdown as Provenance

```
01-brood-table-integration.Rmd < Knit Insert Run
```

31
32 ## Datasets
33
34 As part of the SASAP project, brood tables for 48 Sockeye salmon stocks were collected.
Table 2.1 shows a list of these stocks, along with other regional and location
information.
35
36 ````{r, echo = FALSE}
stocks <- read.csv('data/original/StockInfo.csv', stringsAsFactors = F)
37
38 ````
39
40 ````{r, echo = FALSE}
41 datatable(stocks[, c('Stock.ID','Stock' , 'Region', 'Sub.Region')], rownames = FALSE,
caption = "Stock information")
42
43
44 These stocks range geographically from Washington to Alaska. Although temporal coverage
varies by stock, many of the brood tables were updated in 2016, and some have
reconstructions dating back to 1922.
45
46 Figure 2.1 indicates the approximate location of the salmon stocks in Table 2.1.
47
48 ````{r, echo = FALSE}
49 salmon = makeIcon('images/salmon_tiny.png',
50 'images/salmon_big.png',
51 26, 14)
52
53 m <- leaflet(stocks) %>%
54 setView(-median(stocks\$Lon), median(stocks\$Lat), zoom = 4) %>%
55 addTiles() %>%
56 addMarkers(~Lon, ~Lat, icon = salmon)
57
58 m
59
60
61
62 Figure 2.1: Location of stocks used in this data integration. Salmonid icon by Servien
(vectorized by T. Michael Keesey)
[CC-BY-SA](https://creativecommons.org/licenses/by-sa/3.0/), available at
[PhyloIcon](http://phyloicon.org/)

37:72 R Markdown



2.2 Datasets

As part of the SASAP project, brood tables for 48 Sockeye salmon stocks were collected. Table 2.1 shows a list of these stocks, along with other regional and location information.

Stock.ID	Stock	Region	Sub.Region
101	Washington	WA	WA
102	E.Stuart	Fraser River	Fraser Early Stuart
103	Bowron	Fraser River	Fraser Early Summer
104	Fennell	Fraser River	Fraser Early Summer
105	Gates	Fraser River	Fraser Early Summer
106	Nadina	Fraser River	Fraser Early Summer
107	Pitt	Fraser River	Fraser Early Summer
108	Raft	Fraser River	Fraser Early Summer
109	Scotch	Fraser River	Fraser Early Summer
110	Seymour	Fraser River	Fraser Early Summer

Showing 1 to 10 of 54 entries Previous 1 2 3 4 5 6 Next

These stocks range geographically from Washington to Alaska. Although temporal coverage varies by stock, many of the brood tables were updated in 2016, and some have reconstructions dating back to 1922.

Figure 2.1 indicates the approximate location of the salmon stocks in Table 2.1.

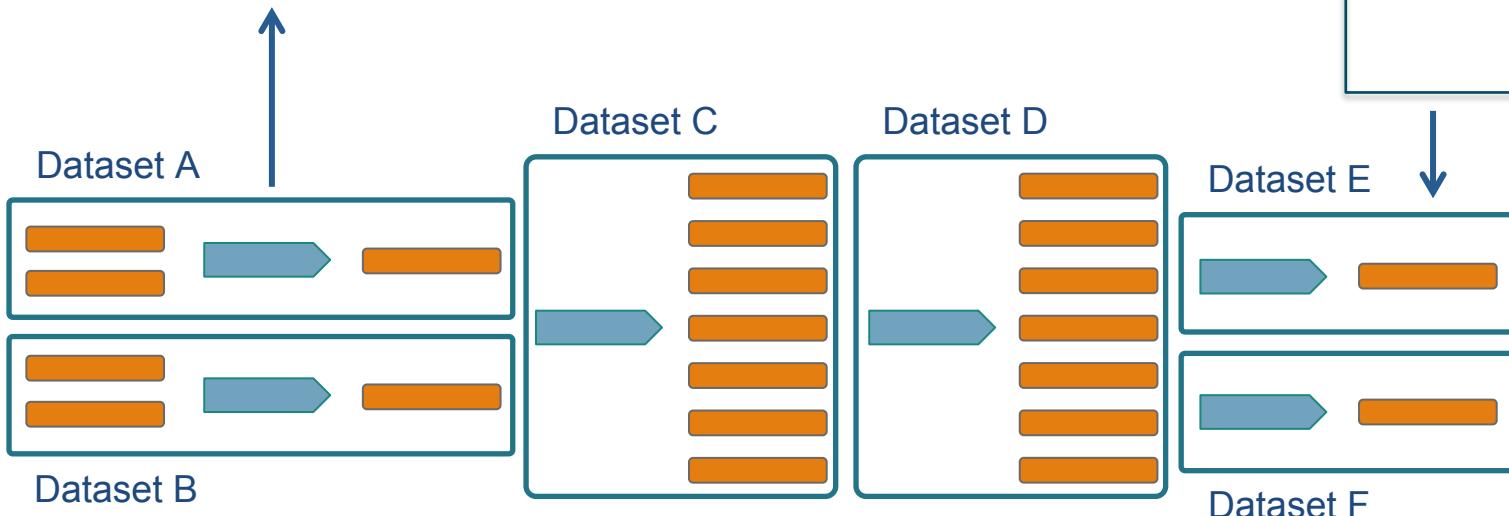


Figure 2.1: Location of stocks used in this data integration. Salmonid icon by Servien (vectorized by T.



Citing multi-generational workflows

Transitive Credit
Via
Provenance





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.”*



Guidelines

<https://arcticdata.io/submit/>

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





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



NSF

**Arctic
Data
Center**

<https://arcticdata.io>