ACCELERATING SYNTHESIS SCIENCE THROUGH REPRODUCIBLE SCIENCE PRACTICES

Matthew B. Jones
National Center for Ecological Analysis and Synthesis
University of California Santa Barbara

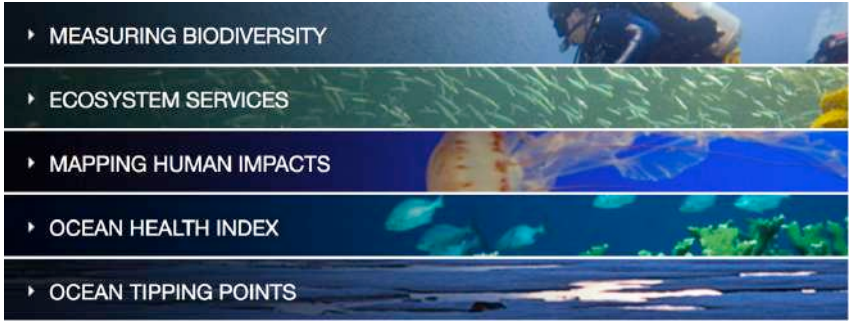
@metamattj
jones@nceas.ucsb.edu
https://orcid.org/0000-0003-0077-4738

Ecological Synthesis

NCEAS
National Center for Ecological Analysis and Synthesis

Marine Systems

Understanding Ocean Health



Threats and Population Declines

Climate and Ecosystems

Provenance

Citation

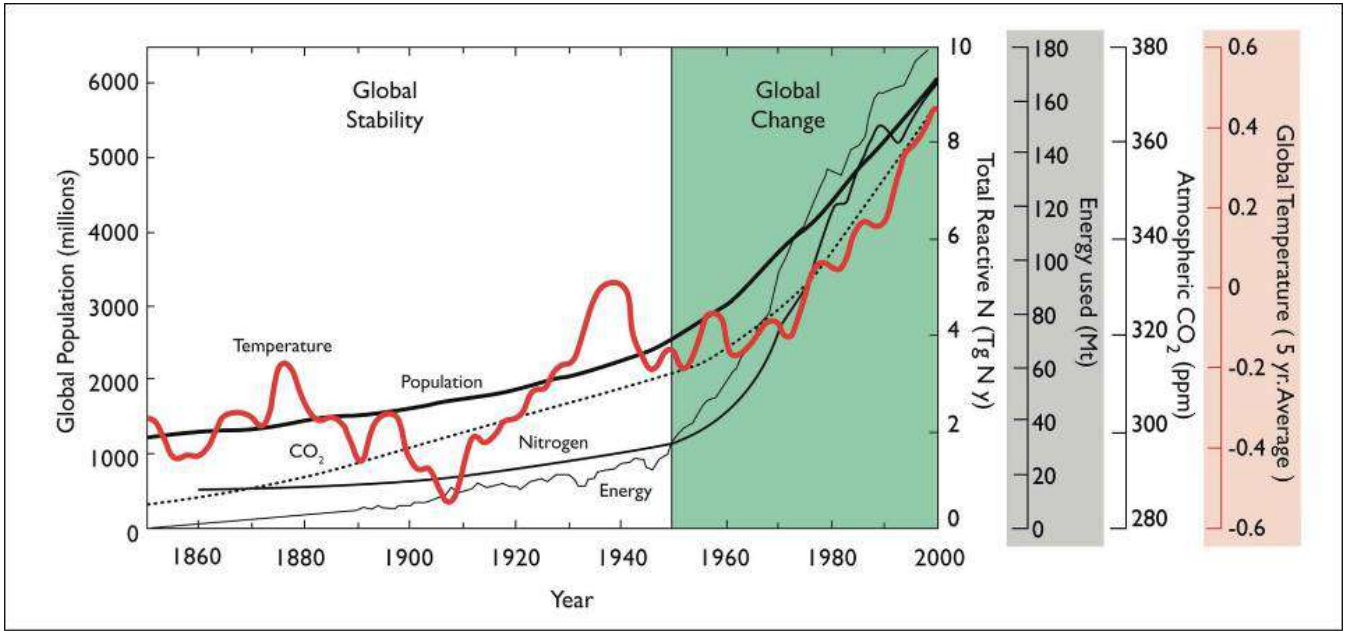
Synthesis

Reproducible Science

Climate Change
Fisheries
Sustainabiity
Subsistence

Science
Governance
Regulation
Policy

Trust in Science



What **data**?
What **methods**?
What **parameter settings**?

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

Smith et al. (2009) Ecology doi:10.1890/08-1815.1

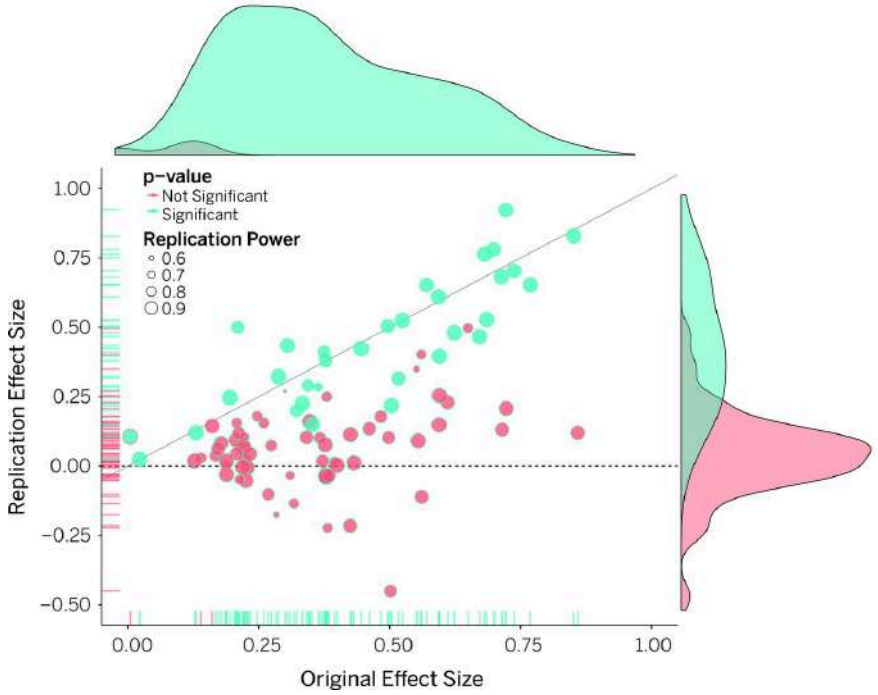
Reproducibility Crisis

"Most research findings are false for most research designs and for most fields"

Ioannidis, 2005

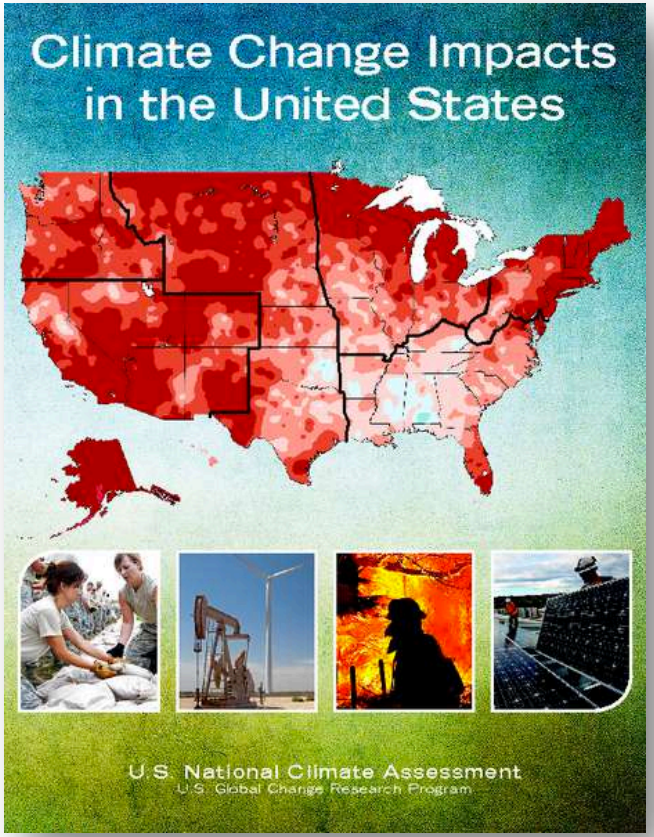
"Most replication effects were smaller than original results"

Open Science Collaboration, 2015



doi:10.1126/science.aac4716

National Climate Assessment



"This report is the result of a **three-year** analytical effort by a team of **over 300 experts**, overseen by a broadly constituted Federal Advisory Committee of **60 members**. It was developed from information and analyses gathered in over 70 workshops and listening sessions held across the country."

Computational Reproducibility

Facilitate transparency by **capturing** and **communicating** scientific workflows

Increase **trust in science**

Stand on the shoulders of giants (build on work that came before)

Give credit for that **secondary** usage enabling **easy attribution**

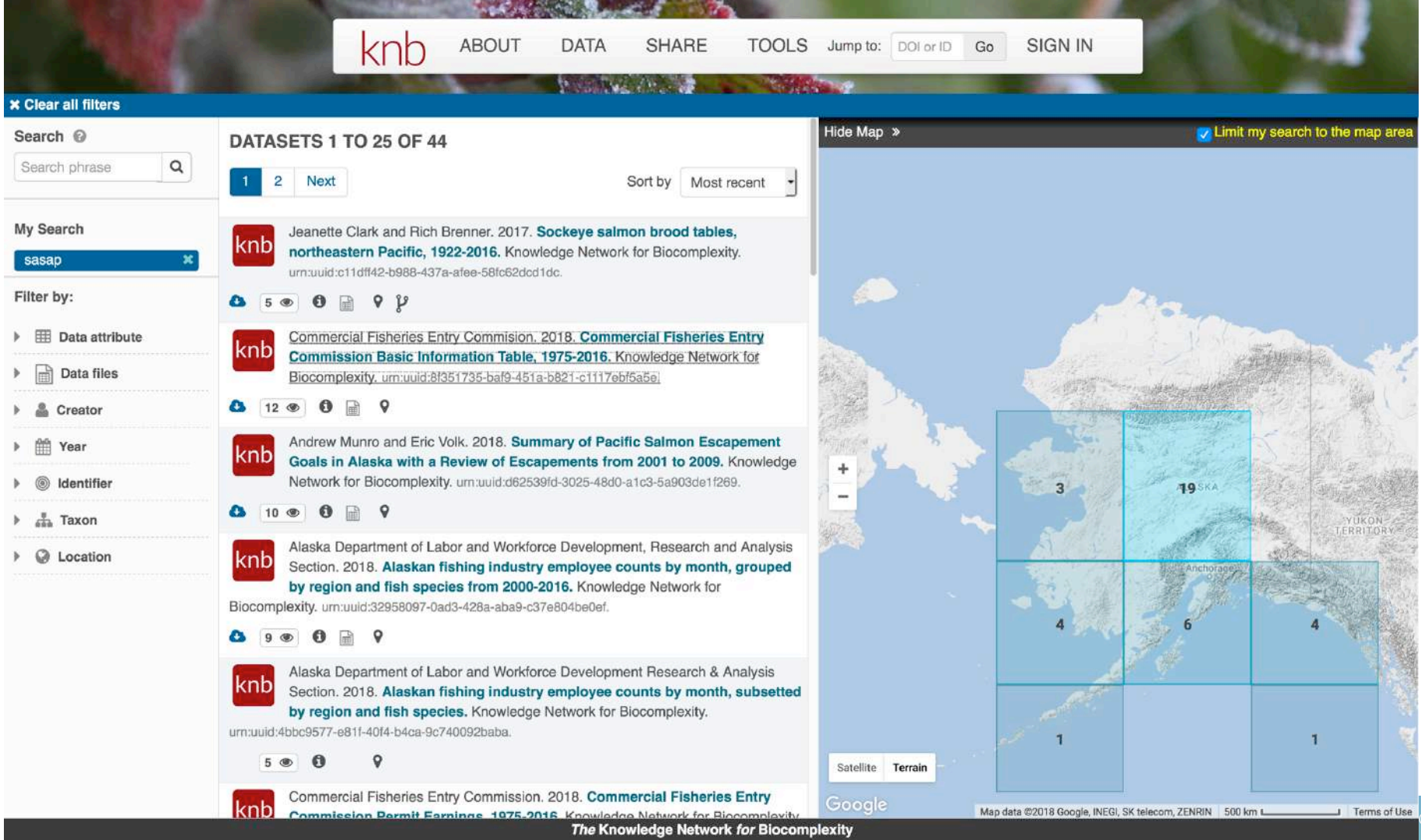
Practical Reproducibility

Preserve the data

Preserve the software workflow

Document what you did

Describe how to interpret it all


knb
ABOUT
DATA
SHARE
TOOLS
Jump to:
DOI or ID
Go
SIGN IN
× Clear all filters
Search
Search phrase
My Search
sasap
Filter by:
Data attribute
Data files
Creator
Year
Identifier
Taxon
Location
DATASETS 1 TO 25 OF 44
1
2
Next
Sort by
Most recent
Jeanette Clark and Rich Brenner. 2017. Sockeye salmon brood tables, northeastern Pacific, 1922-2016. Knowledge Network for Biocomplexity.
urn:uuid:c11dff42-b988-437a-afee-58fc62dcd1dc.
Commercial Fisheries Entry Commision. 2018. Commercial Fisheries Entry Commission Basic Information Table, 1975-2016. Knowledge Network for Biocomplexity. urn:uuid:8f351735-baf9-451a-b821-c1117ebf5a5e.
Andrew Munro and Eric Volk. 2018. Summary of Pacific Salmon Escapement Goals in Alaska with a Review of Escapements from 2001 to 2009. Knowledge Network for Biocomplexity. urn:uuid:d62539fd-3025-48d0-a1c3-5a903de1f269.
Alaska Department of Labor and Workforce Development, Research and Analysis Section. 2018. Alaskan fishing industry employee counts by month, grouped by region and fish species from 2000-2016. Knowledge Network for Biocomplexity. urn:uuid:32958097-0ad3-428a-aba9-c37e804be0ef.
Alaska Department of Labor and Workforce Development Research & Analysis Section. 2018. Alaskan fishing industry employee counts by month, subsetted by region and fish species. Knowledge Network for Biocomplexity.
urn:uuid:4bbc9577-e81f-40f4-b4ca-9c740092baba.
Commercial Fisheries Entry Commission. 2018. Commercial Fisheries Entry
Hide Map »
Limit my search to the map area
3
19
4
6
4
1
1
YUKON TERRITORY
Anchorage
Satellite
Terrain
Google
Map data ©2018 Google, INEGI, SK telecom, ZENRIN
500 km
Terms of Use

DataONE

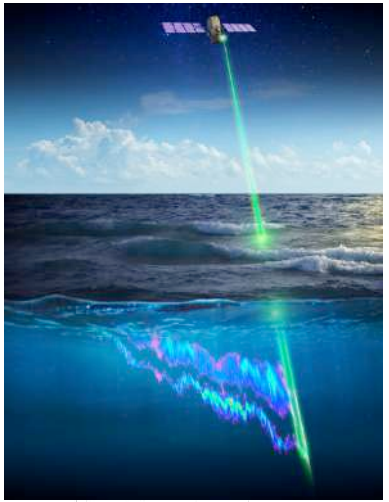
https://search.dataone.org

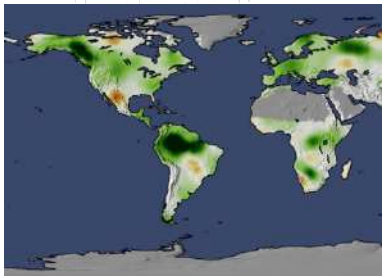
Global
Data Coverage

800K
Data Packages

40
Repositories

143K
Contributors





Reproducible Science

Provenance

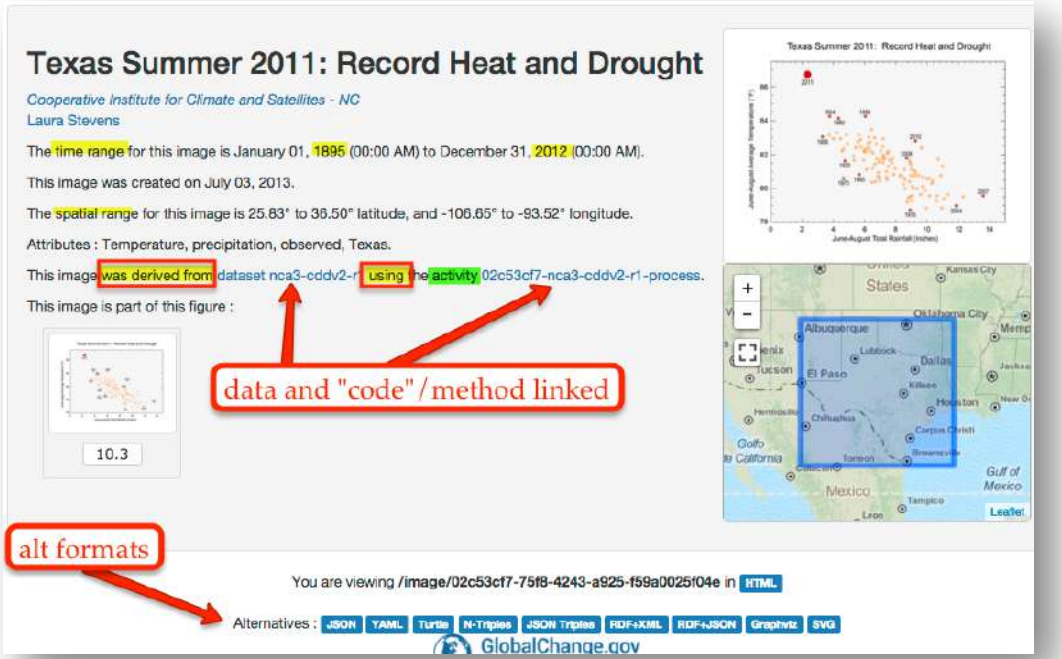
Citation

Synthesis

Computational Provenance

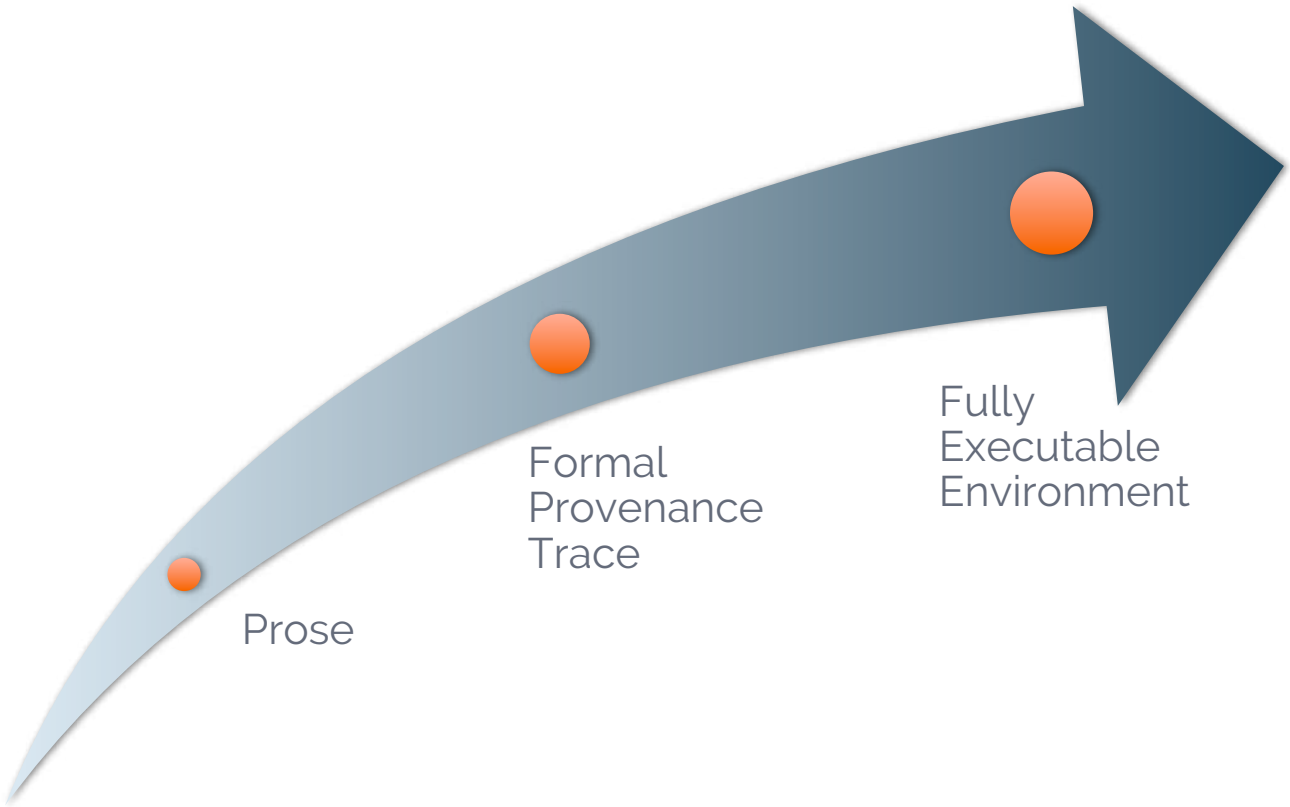
Origin, processing history of data

- Input data
- Workflow/scripts
- Output data
- Figures
- Understand methods, dataflow, and dependencies



Provenance

Origin and processing history of artifacts



Provenance in DataONE

Phase II Goal: Facilitate reproducible science

- Track **data derivation** history
- Track data **inputs** and **outputs** of analyses
- Track analysis and model **executions**
- Preserve and document software **workflows**
- Link all of these to **publications**

Provenance for Science Workflows

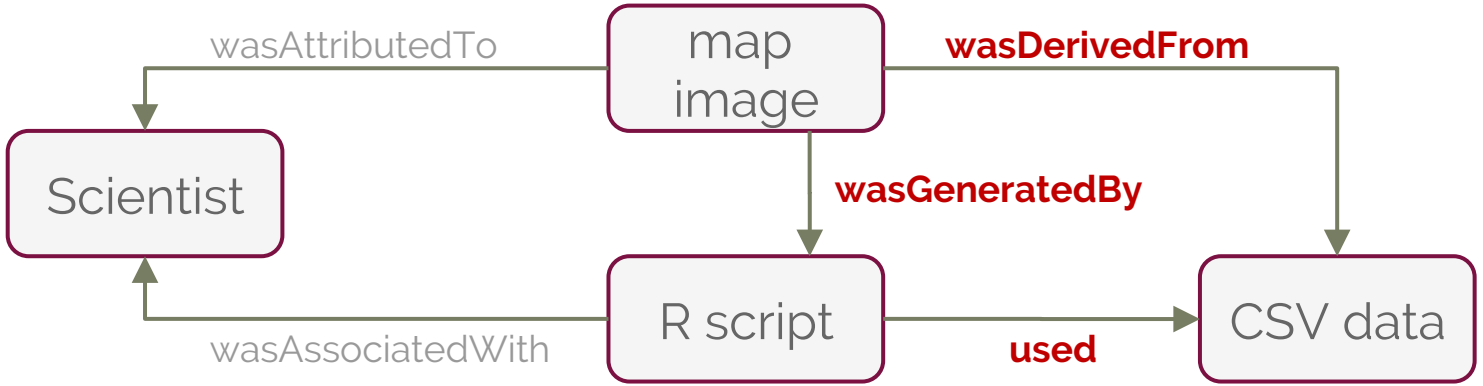
ProvONE – an extension of W3C PROV

See **purl.dataone.org/provone-v1-dev**

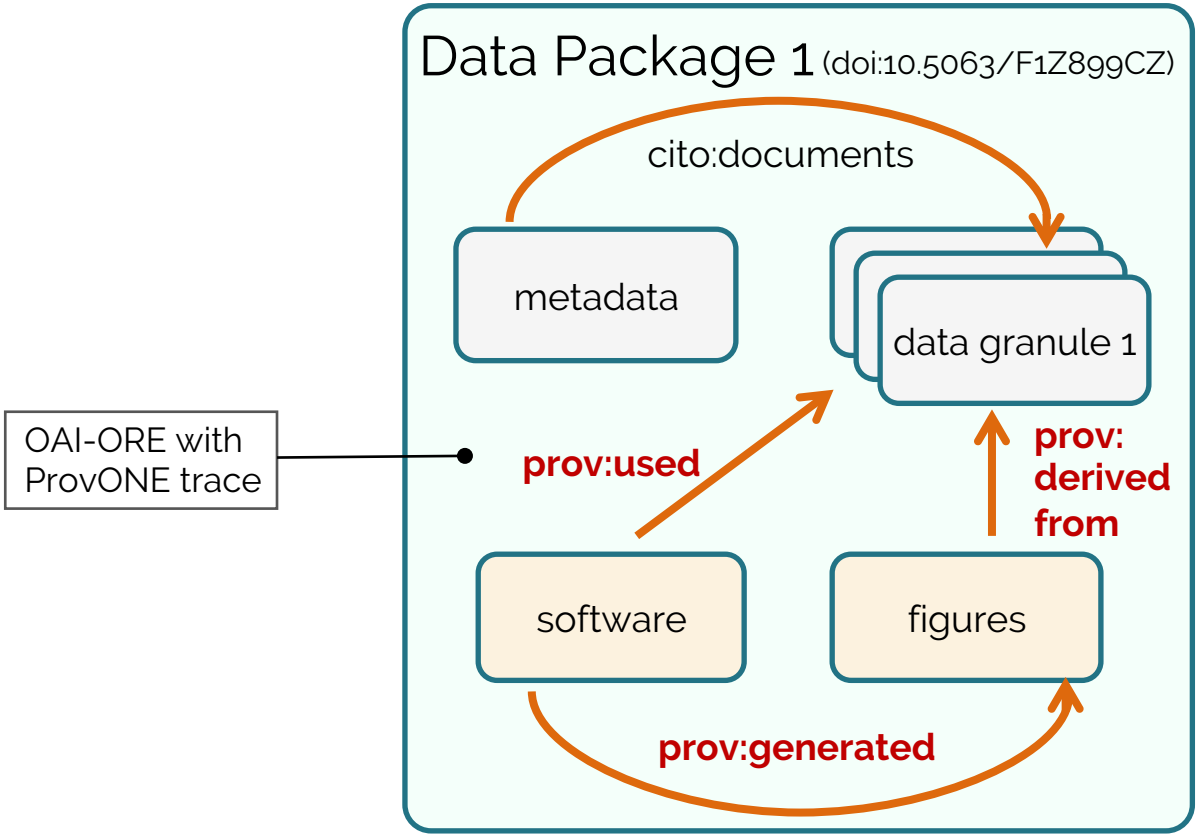
Provenance for Science Workflows

ProvONE – an extension of W3C PROV

See **purl.dataone.org/provone-v1-dev**

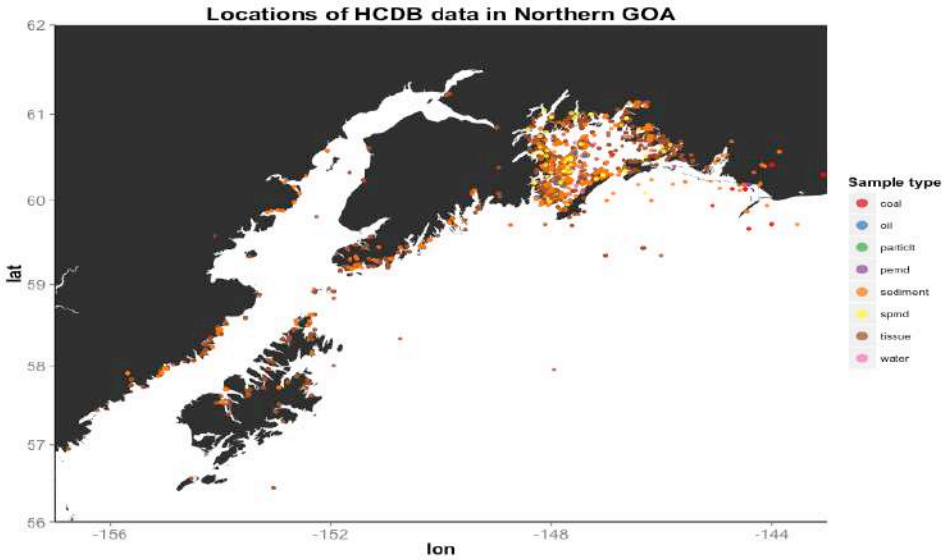


Data Package with Provenance

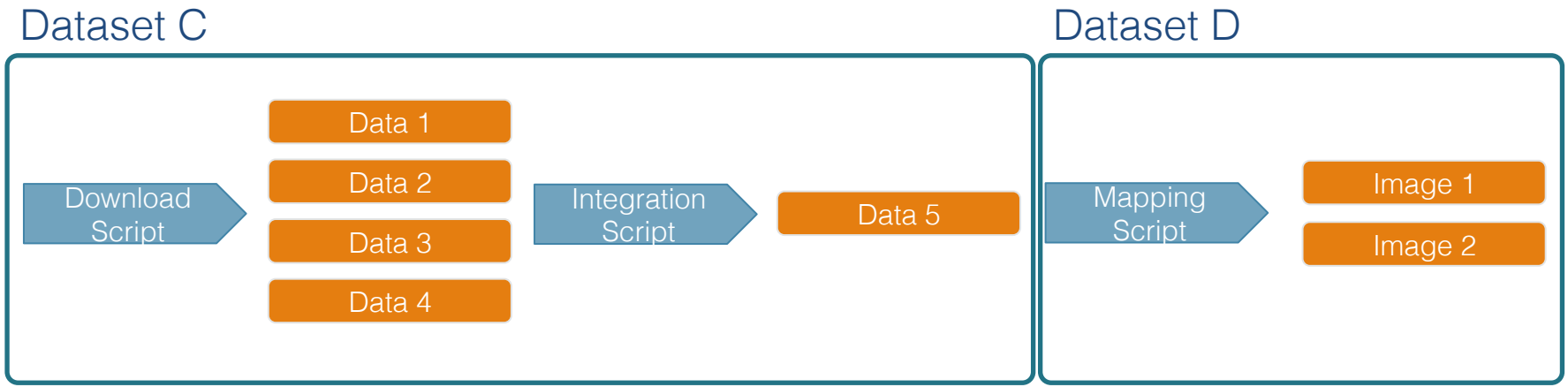


Hydrocarbon Data Example

Mark Carls. 2017. Analysis of hydrocarbons following the Exxon Valdez oil spill, Gulf of Alaska, 1989 - 2014. Arctic Data Center.



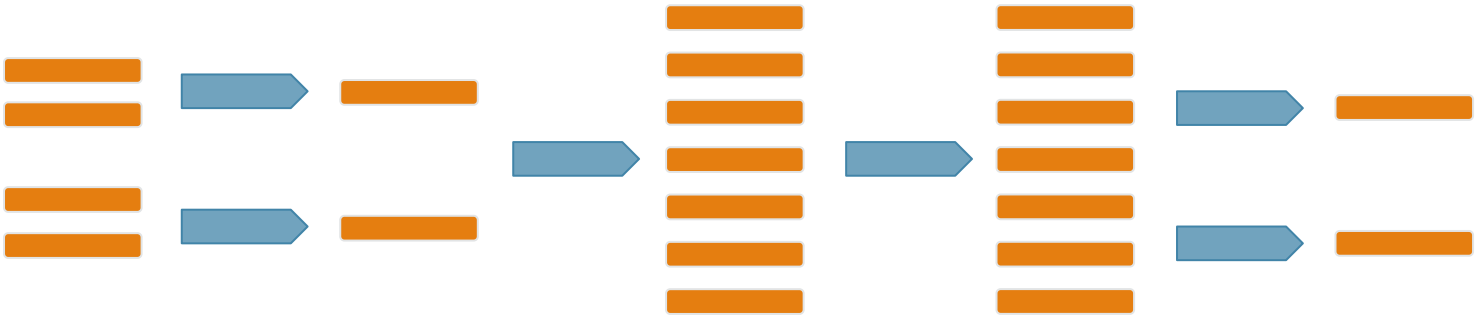
Publishing Data Workflows



Hydrocarbon Data Example

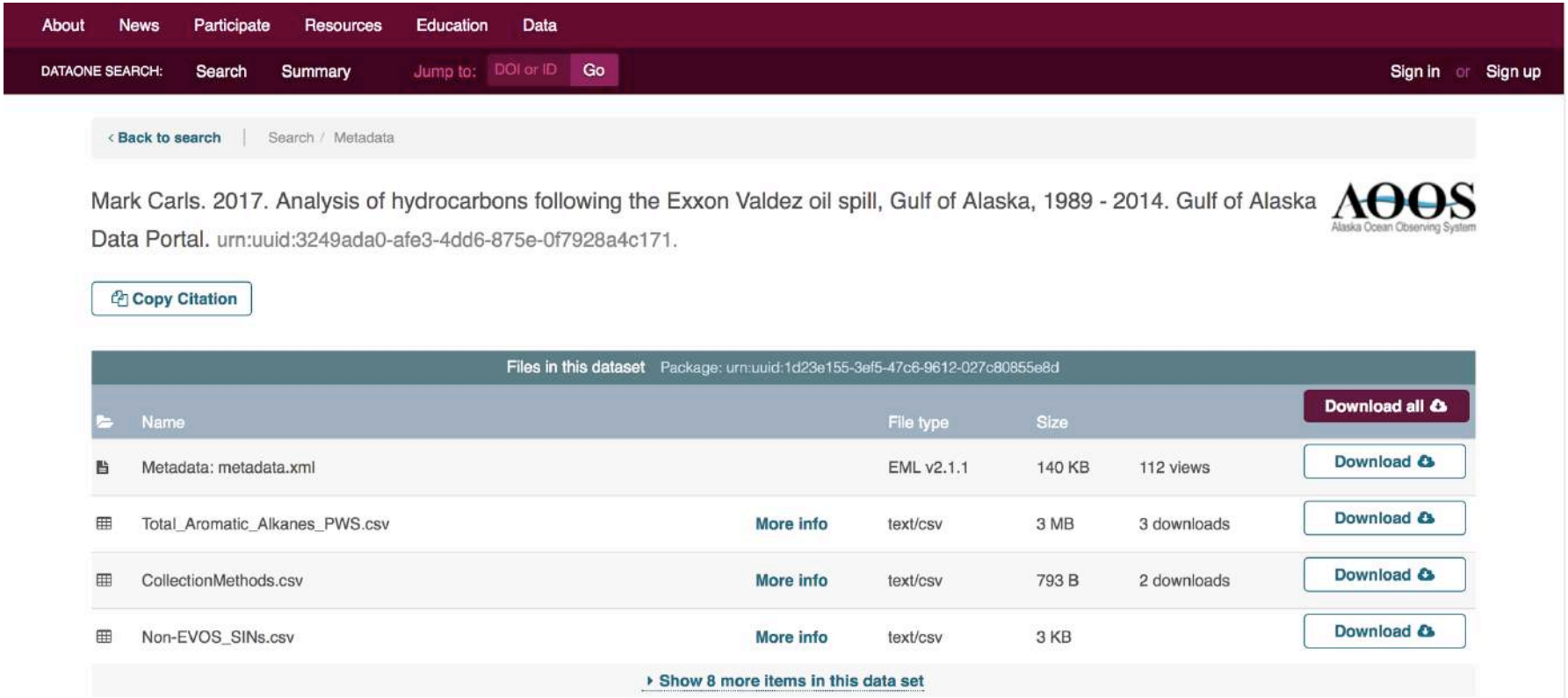
Complex Workflows

Simplified view of complex workflows

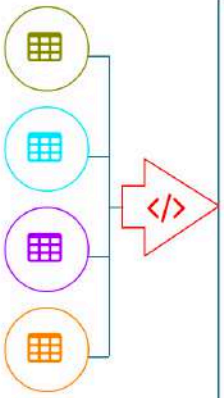


Provenance Display

DataONE Search



Data Table, Image, and Other Data Details

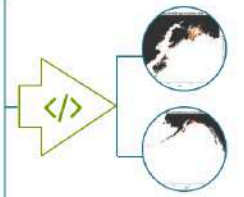


Data Table

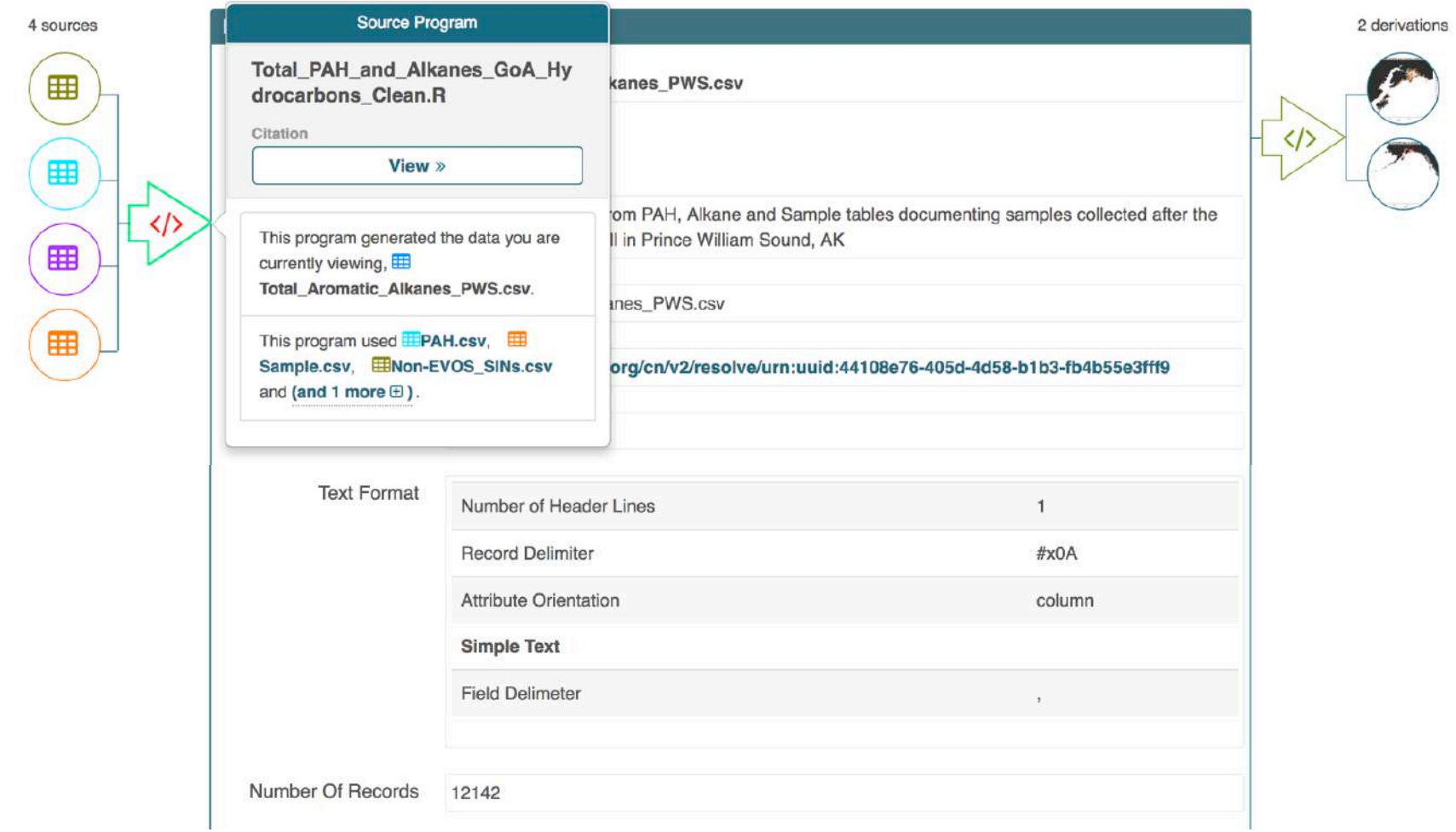
Entity Name	**Total_Aromatic_Alkanes_PWS.csv**
	Download
Description	Combined dataset from PAH, Alkane and Sample tables documenting samples collected after the Exxon Valdez oil spill in Prince William Sound, AK
Object Name	Total_Aromatic_Alkanes_PWS.csv
Online Distribution Info	**https://cn.dataone.org/cn/v2/resolve/urn:uuid:44108e76-405d-4d58-b1b3-fb4b55e3fff9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

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

Number Of Records	12142

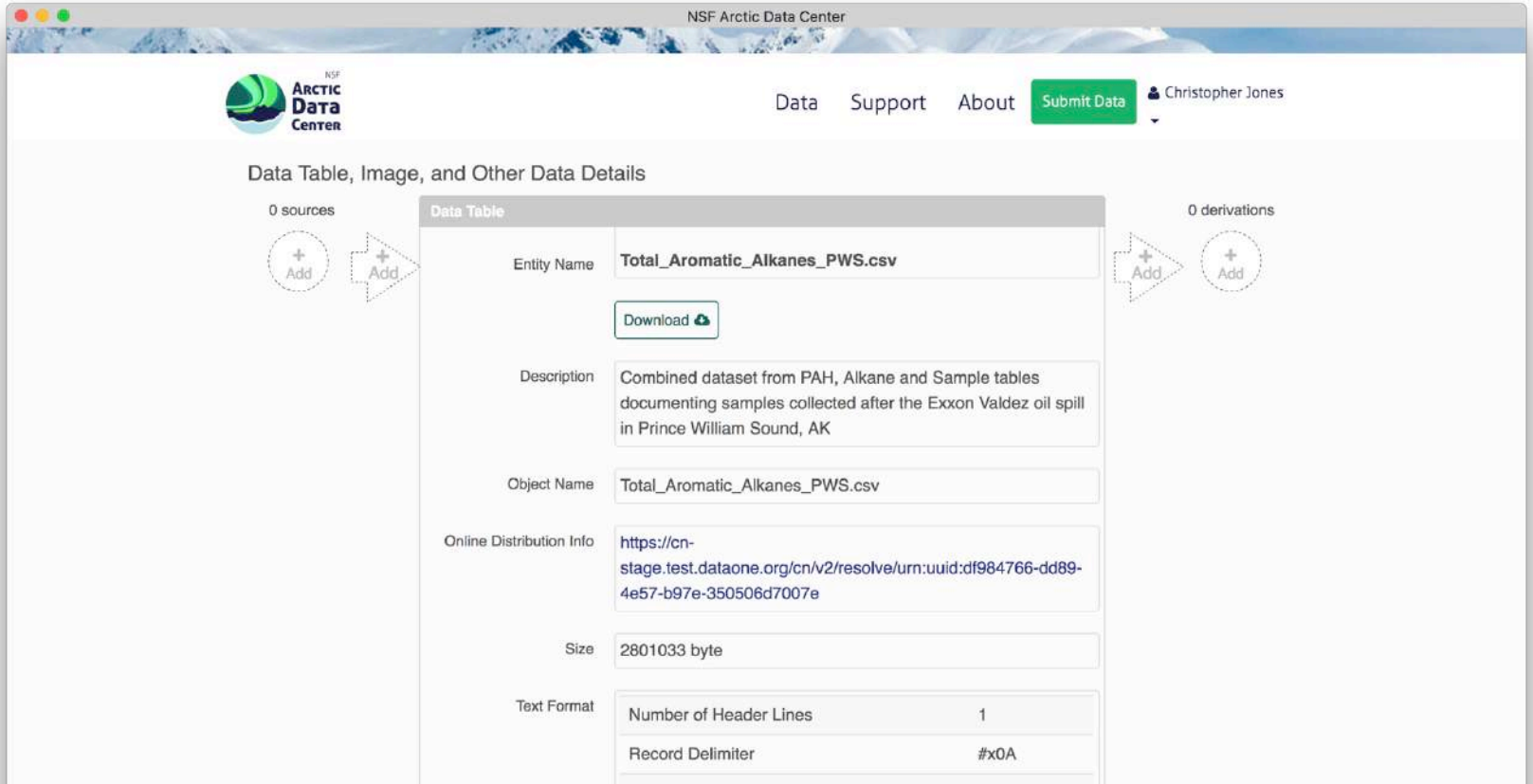


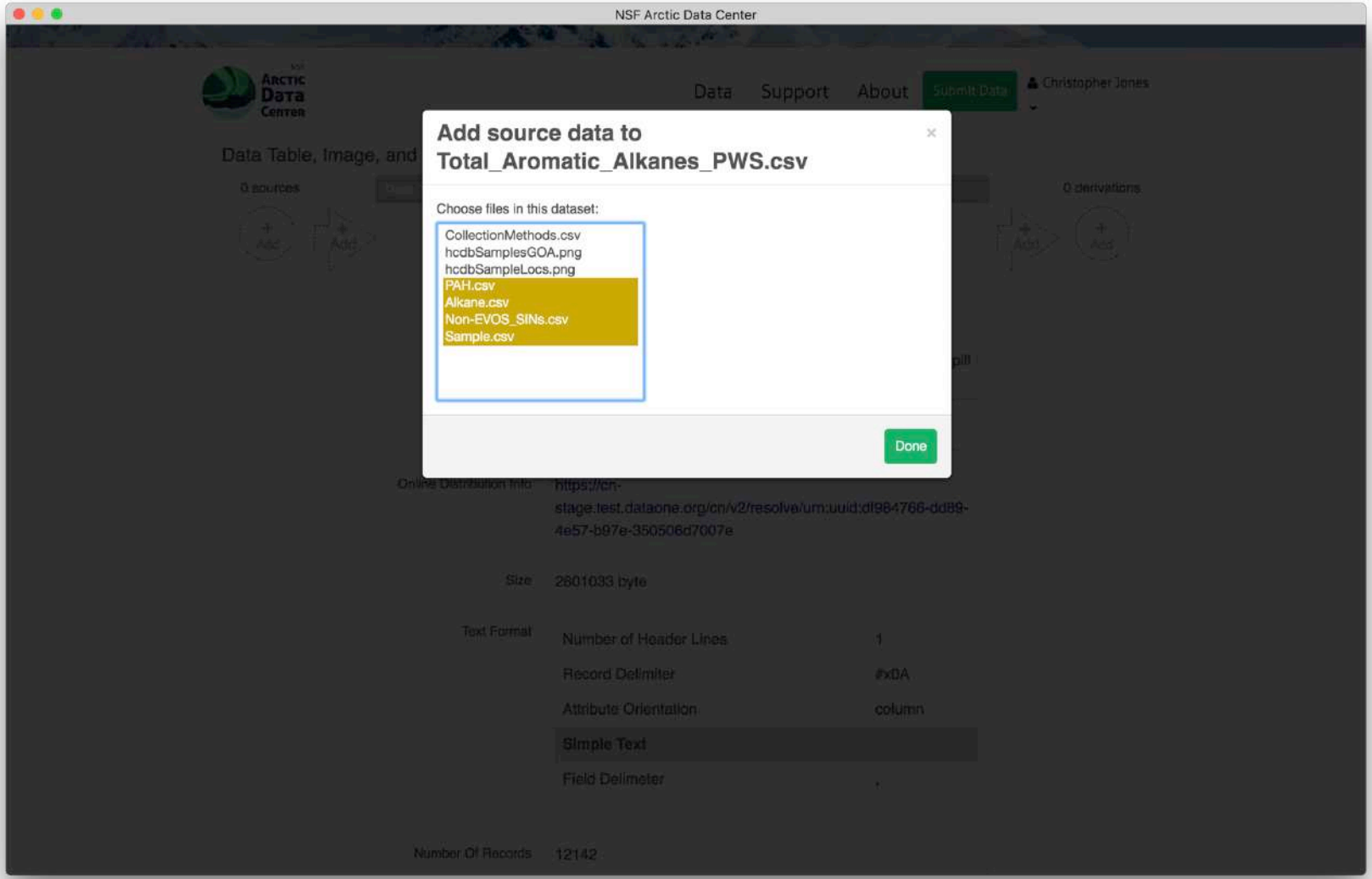
Data Table, Image, and Other Data Details

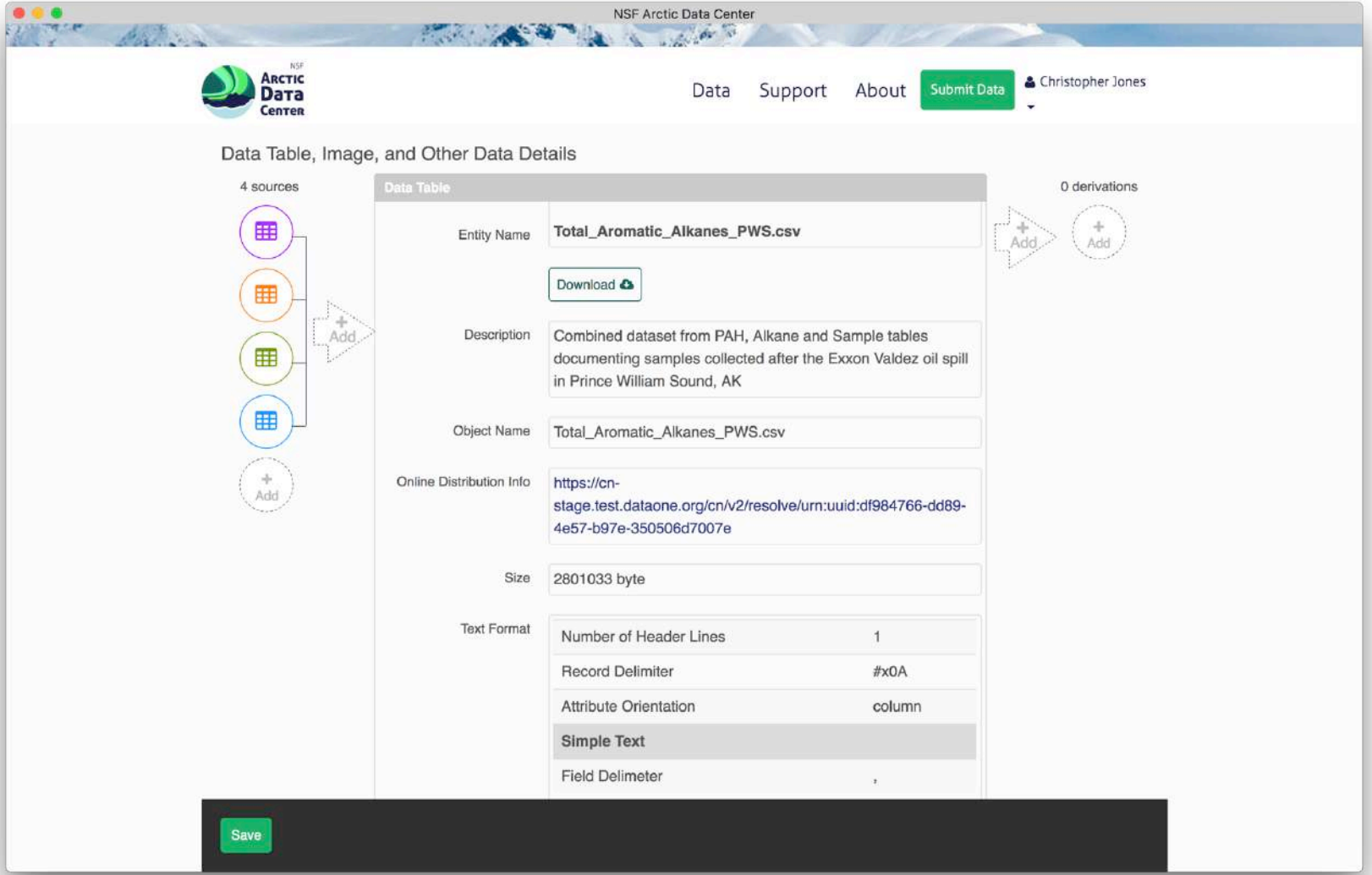


Web Provenance Editor

Deployed by Arctic Data Center




NSF Arctic Data Center
Data
Support
About
Submit Data
Christopher Jones
Add source data to
Total_Aromatic_Alkanes_PWS.csv
Choose files in this dataset:
CollectionMethods.csv
hcdbSamplesGOA.png
hcdbSampleLocs.png
PAH.csv
Alkane.csv
Non-EVOS_SINs.csv
Sample.csv
Done
Data Table, Image, and
0 sources
0 derivations
Online Distribution Info
https://cn-stage.test.dataone.org/cn/v2/resolve/urn:uuid:df984766-dd89-4e57-b97e-350506d7007e
Size
2801033 byte
Text Format
Number of Header Lines
1
Record Delimiter
#x0A
Attribute Orientation
column
Simple Text
Field Delimeter
,
Number Of Records
12142


NSF Arctic Data Center
NSF Arctic Data Center
Data
Support
About
Submit Data
Christopher Jones
Data Table, Image, and Other Data Details
4 sources
Add
Add
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-stage.test.dataone.org/cn/v2/resolve/urn:uuid:df984766-dd89-4e57-b97e-350506d7007e
Size
2801033 byte
Text Format
Number of Header Lines
1
Record Delimiter
#x0A
Attribute Orientation
column
Simple Text
Field Delimeter
,
0 derivations
Add
Add
Save

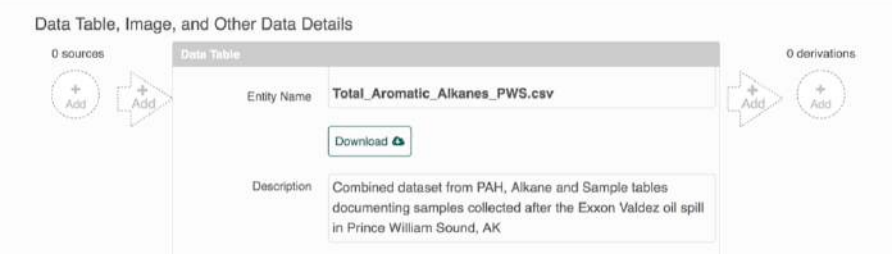
Provenance Editing

Matlab DataONE Toolbox

Recordr R Library

YesWorkflow Tool

MetacatUI
Web Provenance Editor



Reproducible Science

Provenance

Citation

Synthesis

Credit where credit is due

Indexing and exposing data citations in international data repository networks

Alfred P. Sloan Foundation

University of California CDL California Digital Library

DataCite FIND, ACCESS, AND R

DataONE

Force11 Data Citation Principles

1. Importance of data citation
2. **Credit and Attribution**
3. **Evidence**
4. Unique Identification
5. Access
6. **Persistence**
7. **Specificity** and Verifiability
8. Interoperability and Flexibility

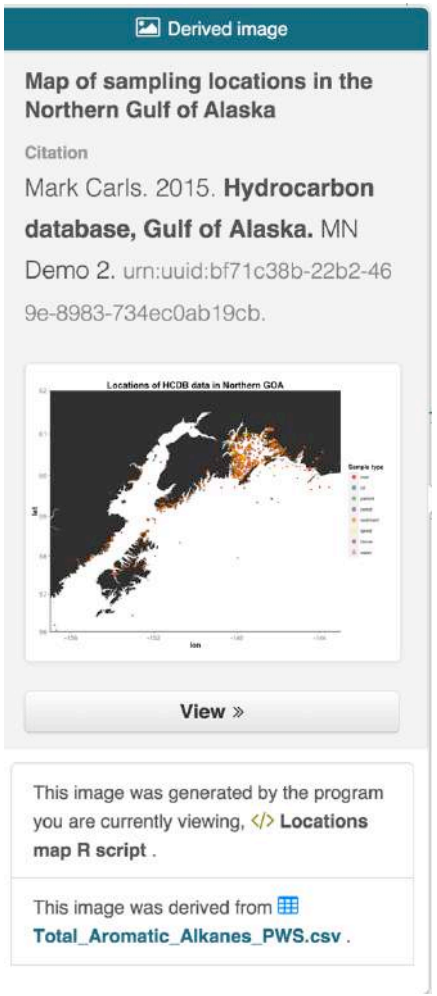
Joint Declaration of Data Citation Principles, https://doi.org/10.25490/a97f-egyk

Transitive Credit

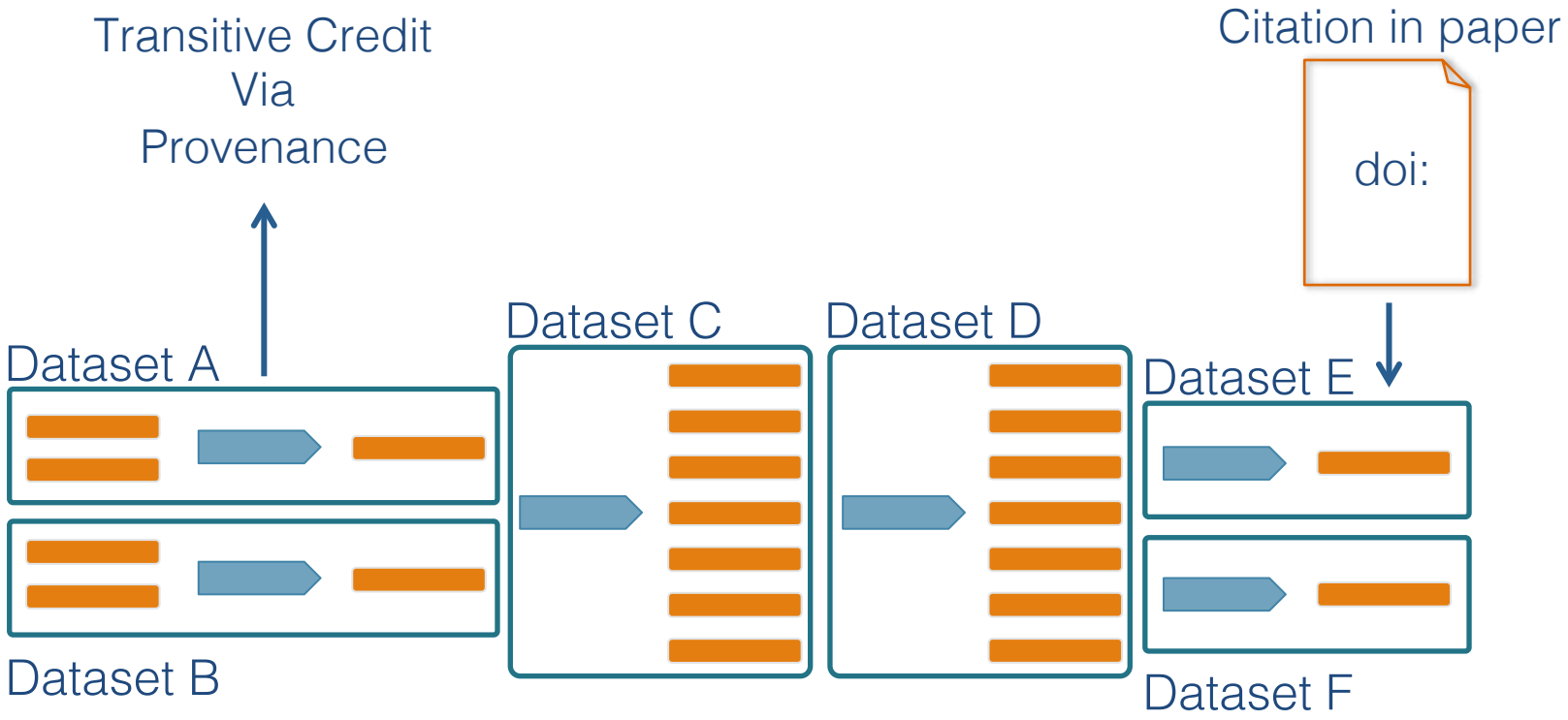
When a user cites a pub, we know:

- **Which data** produced it
- **What software** produced it
- What was **derived** from it
- **Who to credit** down the attribution stack

See: Katz & Smith. 2014. **Implementing Transitive Credit with JSON-LD**. arXiv:1407.51



Citing multi-generational workflows



Evolution of the Living Paper

Scholarly Publications

1st Gen	**Prose**		
2nd Gen	**Prose**	**+ Data**	
3rd Gen	**Prose**	**+ Data**	**+ Code**

DataONE

Prose + Data + Code + **Provenance**

WHOLETALE

Prose + Data + Code + **Provenance + Execution Environment**

Theoretical and Applied Climatology
November 2016, Volume 126, Issue 3–4, pp 699–703 | Cite as

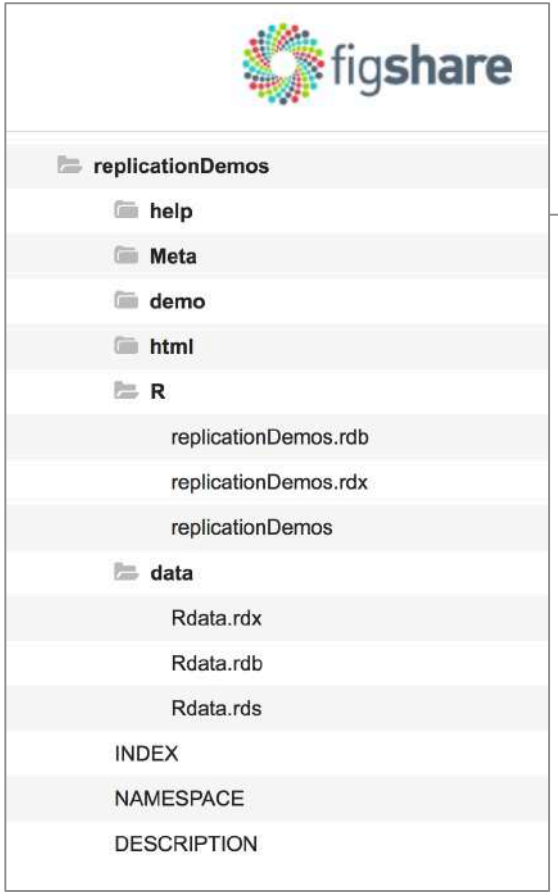
Learning from mistakes in climate research

Authors | Authors and affiliations

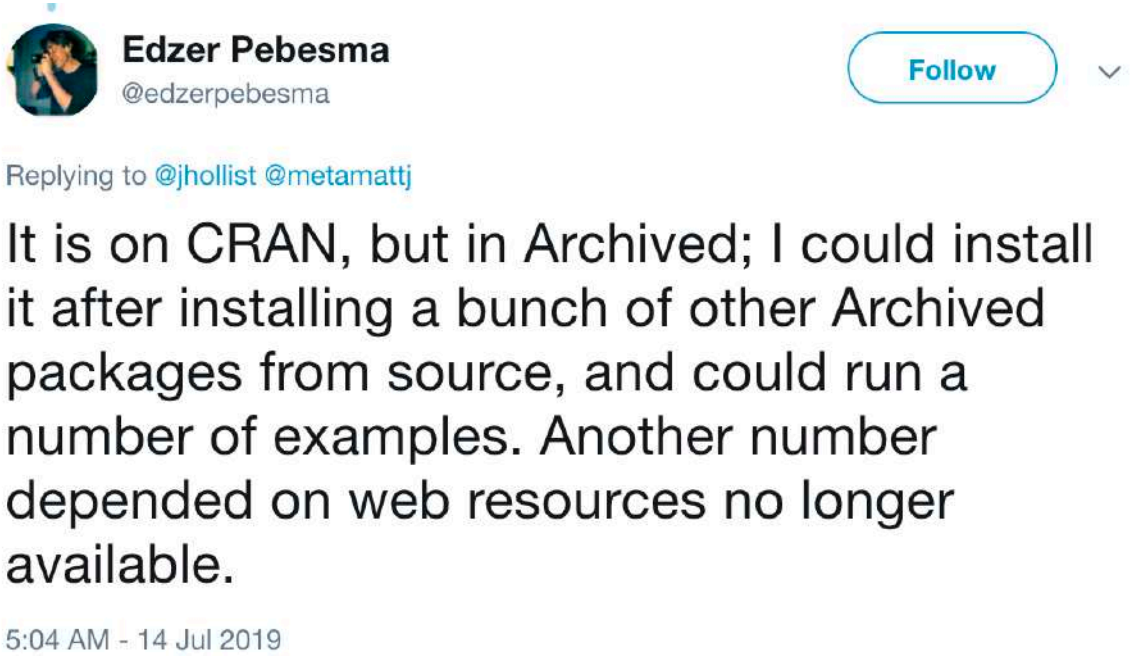
Rasmus E. Benestad, Dana Nuccitelli, Stephan Lewandowsky, Katharine Hayhoe, Hans Olav Hygen, Rob van Dorland, John Cook

Open Access | Original Paper
First Online: 20 August 2015

3.2k Shares
103k Downloads
18 Citations



Ships with an R package



Parsing Reproducibility

- **Empirical Reproducibility:**
  - traditional empirical experiments, e.g. at the bench/lab
- **Statistical Reproducibility**:
  - statistical methodology used permits generalizability of data inferences
- **Computational Reproducibility:**
  - transparency of computational steps that produce scientific findings

V. Stodden. (2013). *Resolving Irreproducibility in Empirical and Computational Research.* IMS Bulletin

Simplifying Computational Reproducibility in Whole Tale

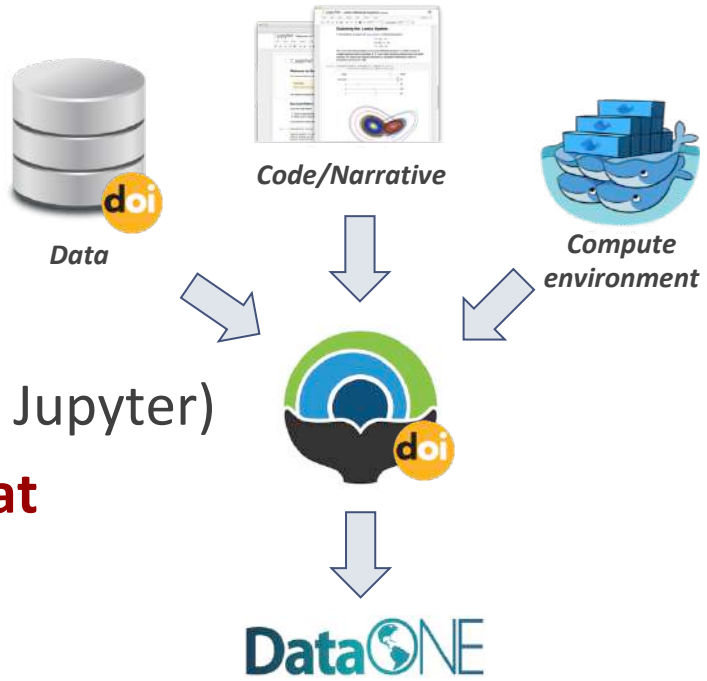
- Researchers can **easily package** and **share *tales***:
  - **Data**, **Code**, and **Compute Environment**
  - to **re-create the computational results** from a scientific study
  - achieving **computational reproducibility**
  - thus **"setting the default to reproducible."**

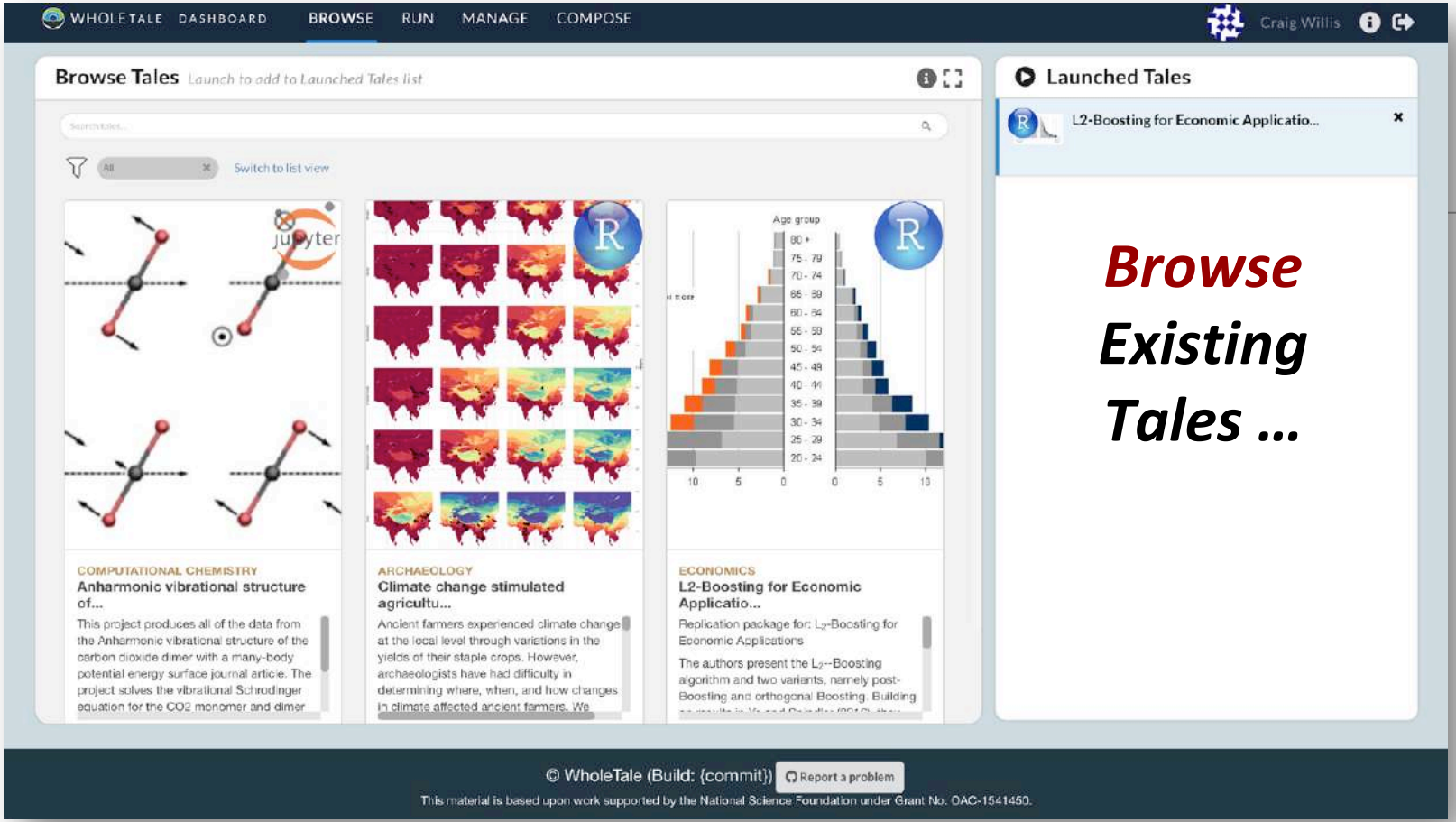
- Also empowers users to verify and extend results with **different** data, methods, and environments.

V. Stodden, D. H. Bailey, J. Borwein, R. J. LeVeque, W. Rider, and W. Stein. (2013). S*etting the Default to Reproducible: Reproducibility in Computational and Experimental Mathematics*, ICERM workshop (2013)

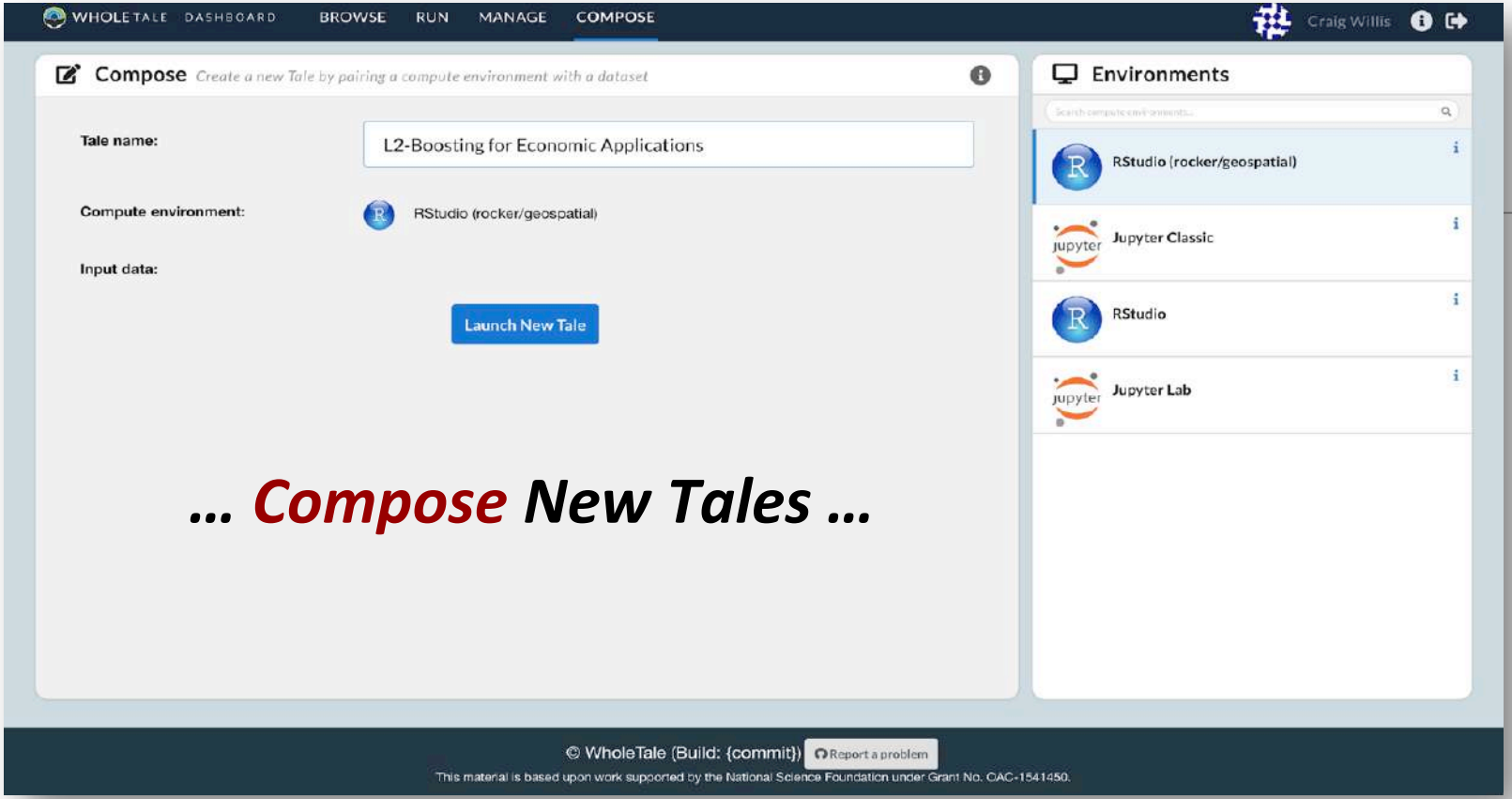
What exactly is (in) a Tale?

- **Tale** = executable **research object**, i.e.
  - **data** (references)
  - **+ code** (computational methods)
  - **+ narrative** (traditional science story)
  - **+ compute environment** (e.g. RStudio, Jupyter)
- Captured in a **standards-based tale format** complete with metadata

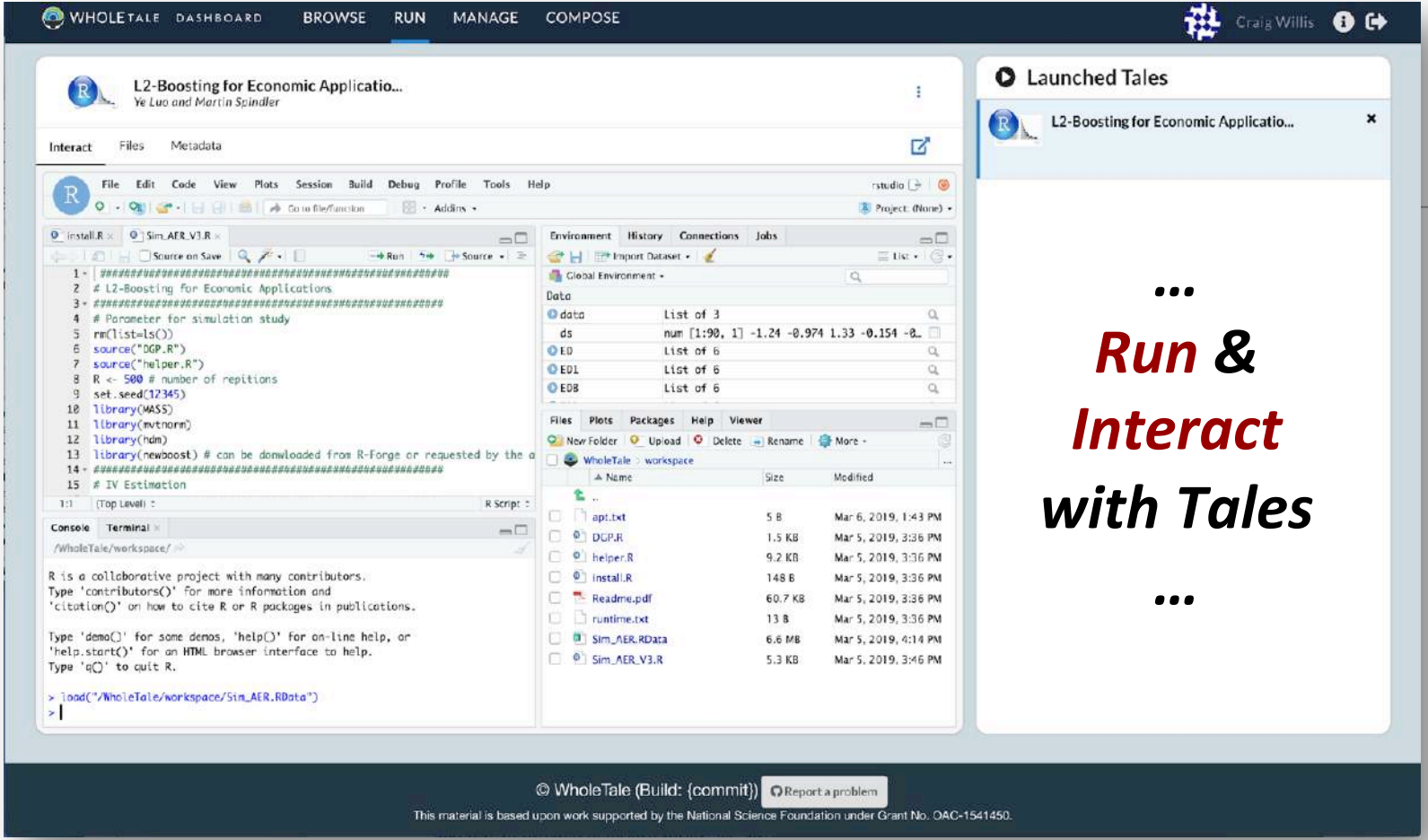


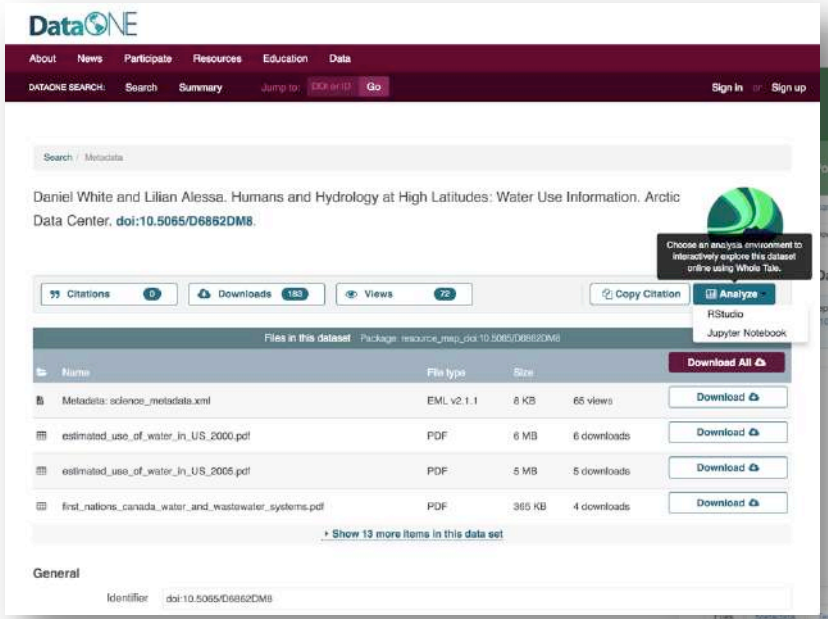


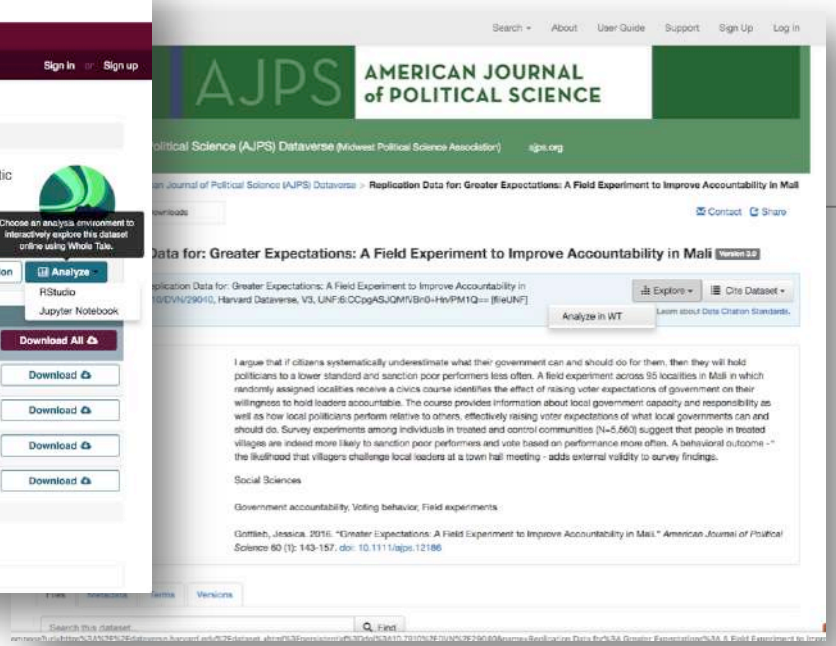
Browse Existing Tales …



… *Compose* *New Tales* …

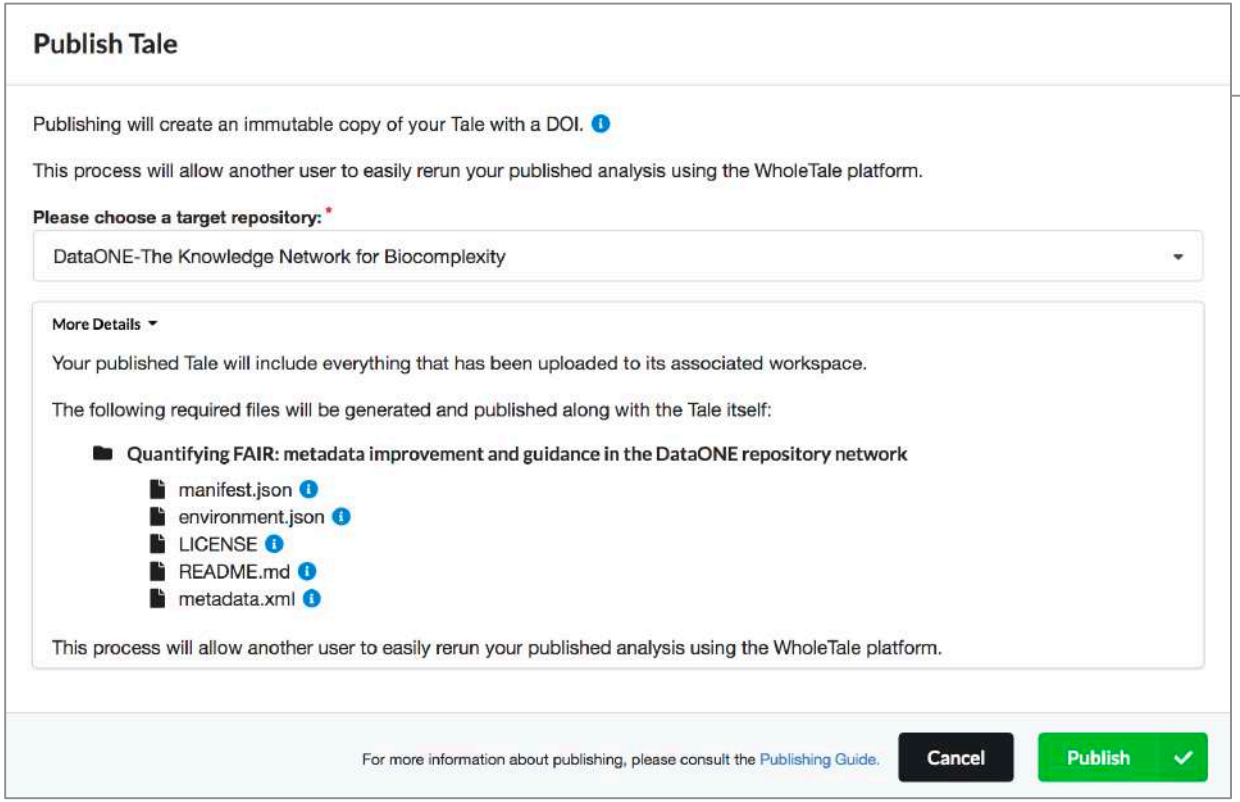

WHOLETALE DASHBOARD
BROWSE
RUN
MANAGE
COMPOSE
Craig Willis
L2-Boosting for Economic Applicatio...
Ye Luo and Martin Spindler
Interact
Files
Metadata
Launched Tales
L2-Boosting for Economic Applicatio...
...
Run &
Interact
with Tales
...
© WholeTale (Build: {commit})
Report a problem
This material is based upon work supported by the National Science Foundation under Grant No. OAC-1541450.





… Integrate Data Repos with Whole Tale!

- Enables **turnkey exploratory data analysis** on existing published datasets
- **DataONE** and **Dataverse** networks cover **> 90 major research repositories**!



… Publish Data, Code, and Environment

- Enables **full circle** reproducibility to **DataONE** repositories that accept API deposits

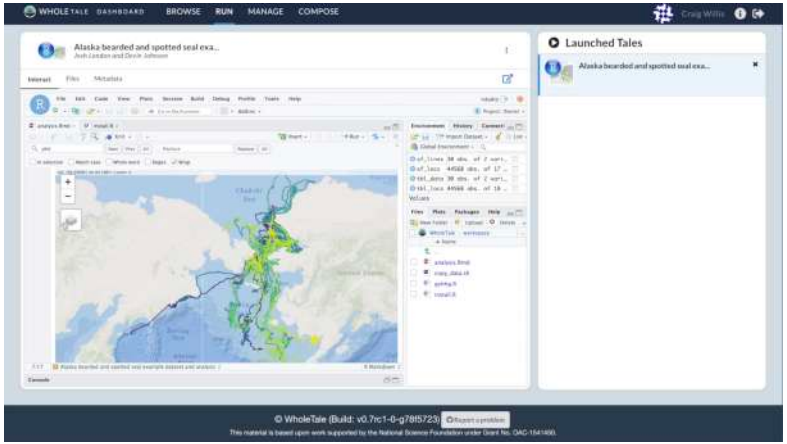
Whole Tale Forecast Demo

Demonstration of a model to predict the movement paths of seals using satellite telemetry data.

Based on analysis and models by:

Josh London and Devin Johnson

NOAA Marine Mammal Laboratory



https://youtu.be/Ml5d7r5OtCk

Reproducible Science

Provenance

Citation

Synthesis

State of Alaska's Salmon and People

NCEAS
National Center for Ecological Analysis and Synthesis

8 SASAP working groups

1: **Bio-physical State of Knowledge of Salmon Distribution & Habitat**
Leads: Peter Westley and Dan Rinella

2: **Sociocultural and Economic Dimensions of Salmon Systems**
Leads: Courtney Carothers, Jessica Black, Tobias Schworer

3: **Governance and Subsistence**
Leads: Steve Langdon, Taylor Brelsford, James Fall

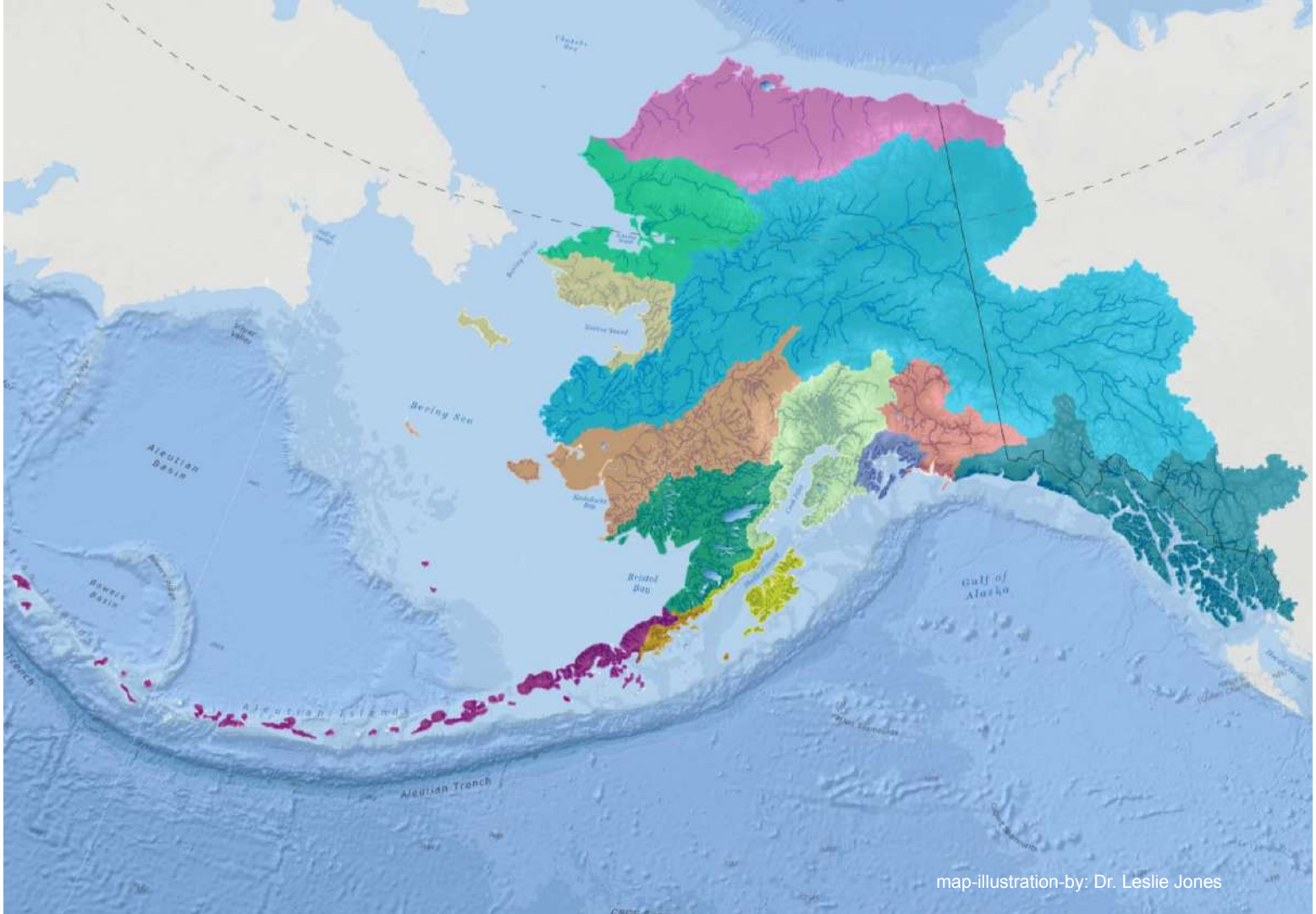
4: **Consistency, Causes, and Consequences of Declining Size and Age of Alaskan Salmon**
Leads: Eric P. Palkovacs, Peter Westley, Bert Lewis

5: **Well-Being and Alaska Salmon Systems**
Leads: Rachel Donkersloot, Jessica C. Black, Courtney Carothers

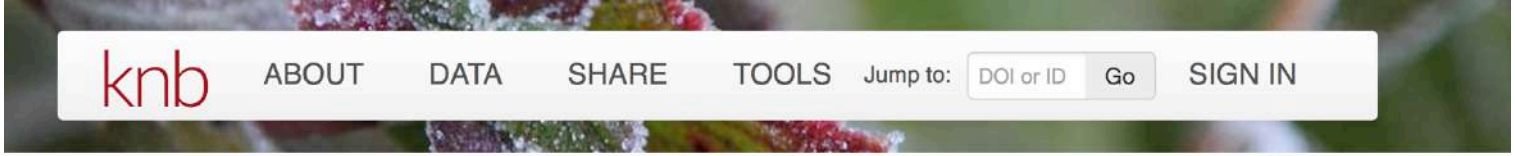
6: **Interacting Effects of Ocean Climate and At-Sea Competition on Alaskan Salmon**
Leads: Peter S. Rand, Robert W. Campbell, Kristen B. Gorman

7: **Using Participatory Modeling to Empower Community Engagement in Salmon Science**
Leads: Michael L. Jones

8: **Kenai Lowlands Salmon Research Synthesis and Design Tools for Integrated Watershed Management**
Leads: Coowe Walker, Mark Rains, Ryan King, Charles Simenstad, Dennis Whigham



map-illustration-by: Dr. Leslie Jones



h | Home / Search / Metadata

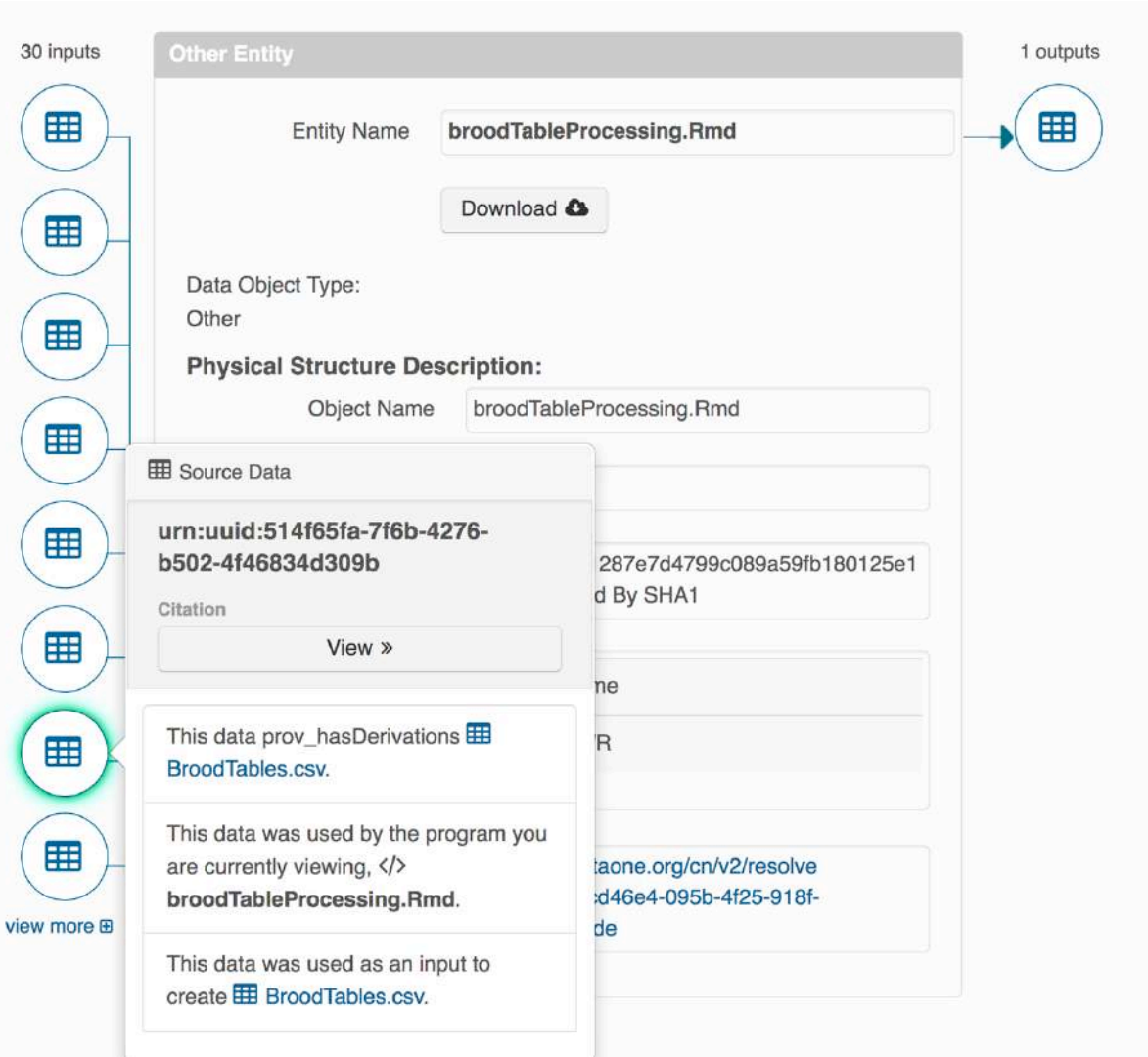
Jeanette Clark and Rich Brenner. 2017. Sockeye salmon brood tables, northeastern Pacific, 1922-2016. Knowledge Network for Biocomplexity. urn:uuid:c11dff42-b988-437a-afee-58fc62dcd1dc.

Copy Citation | Quality report

Files in this dataset Package: resource_map_urn:uuid:c11dff42-b988-437a-afee-58fc62dcd1dc

Name		File type	Size	Downloads	Download All
Metadata: broodTable_metadata.xml		EML v2.1.1	37 KB	5 views	Download
BroodTables.csv	More info	text/csv	449 KB	61 downloads	Download
StockInfo.csv	More info	text/csv	19 KB	2 downloads	Download
SourceInfo.csv	More info	text/csv	723 B	2 downloads	Download
broodTableProcessing.Rmd	More info	application/R	19 KB	3 downloads	Download
broodTableProcessing.html	More info	HTML	1 MB	9 downloads	Download

Show less


30 inputs
Other Entity
1 outputs
Entity Name
broodTableProcessing.Rmd
Download
Data Object Type:
Other
Physical Structure Description:
Object Name
broodTableProcessing.Rmd
Source Data
urn:uuid:514f65fa-7f6b-4276-b502-4f46834d309b
Citation
View »
287e7d4799c089a59fb180125e1
d By SHA1
ne
R
taone.org/cn/v2/resolve
d46e4-095b-4f25-918f-
de
This data prov_hasDerivations BroodTables.csv.
This data was used by the program you are currently viewing, </> broodTableProcessing.Rmd.
This data was used as an input to create BroodTables.csv.
view more

Rmarkdown as Provenance

01-brood-table-integration.Rmd

Knit · Insert · Run

```
## 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.

```{r, echo = FALSE}
stocks <- read.csv('data/original/StockInfo.csv', stringsAsFactors = F)
```

```{r, echo = FALSE}
datatable(stocks[, c('Stock.ID','Stock' ,'Region', 'Sub.Region')], rownames = FALSE,
caption = "Stock information")
```

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.

```{r, echo = FALSE}
salmon = makeIcon('images/salmon_tiny.png',
                  'images/salmon_big.png',
                  26, 14)

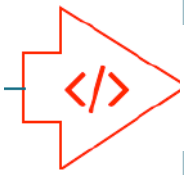
m <- leaflet(stocks) %>%
  setView(-median(stocks$Lon), median(stocks$Lat), zoom = 4) %>%
  addTiles() %>%
  addMarkers(~-Lon, ~Lat, icon = salmon)

m

```

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
[Phylonic](http://phylonic.org/)
```

37:72 Chunk 2 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.

Show 10 entries Search:

Stock 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.

SASAP

Group

Group Id: SASAP

4 years, 6 months Contributor since August 4, 2013

2 contributions

4,862 downloads

24 members

- Krista B Oke
  http://orcid.org/0000-0002-5415-3534
- Josh Baron
  http://orcid.org/0000-0002-4286-6959
- Rich Brenner
  http://orcid.org/0000-0001-7209-9757
- Jeanette Clark
  http://orcid.org/0000-0003-4703-1974

First 1 2 3 4 5 6 Last

DATASETS 1 TO 5 OF 60

1 2 3 ... 12 Next

Sort by Most recent

Alaska Department of Fish and Game, Division of Commercial Fisheries, Central Region. 2018. **Chinook age, sex, and length data from East Side Cook Inlet, Alaska, 1970-2012.** Knowledge Network for Biocomplexity. urn:uuid:16763faf-9ad6-4a95-bcfc-97d60957e499.

6

Jeanette Clark and Rich Brenner. 2017. **Sockeye salmon brood tables, northeastern Pacific, 1922-2016.** Knowledge Network for Biocomplexity. urn:uuid:c11dff42-b988-437a-afee-58fc62dcd1dc.

6

Alaska Department of Fish and Game. 2018. **Salmon age, sex, and length data from Lower Cook Inlet, Alaska, 1961-2014.** Knowledge Network for Biocomplexity. urn:uuid:99e94ab7-b822-458e-88b3-df0ed1964378.

7

Jared Kibele and Leslie Jones. 2018. **Glaciers in Alaska with subsetting by watershed and SASAP region.** Knowledge Network for Biocomplexity. urn:uuid:874e1ba2-48d2-4d31-b3fb-682aaf7e984b.

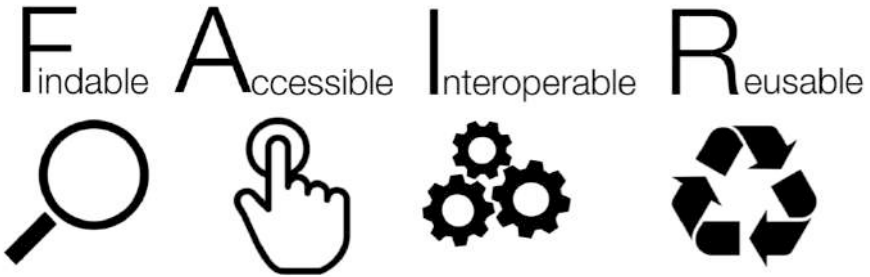
7

Jeanette Clark, Rich Brenner, and Bert Lewis. 2018. **Compiled age, sex, and length data for Alaskan salmon.** Knowledge Network for Biocomplexity. urn:uuid:63a9c8df-3543-44fe-a5d0-746469318f18.

10

Foundational Infrastructure

Providing ***findable***, ***accessible*** data with ***interoperable*** infrastructure enabling long term data ***reuse*** for synthesis



https://www.force11.org/fairprinciples