ACCELERATING SYNTHESIS SCIENCE THROUGH REPRODUCIBLE SCIENCE PRACTICES

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

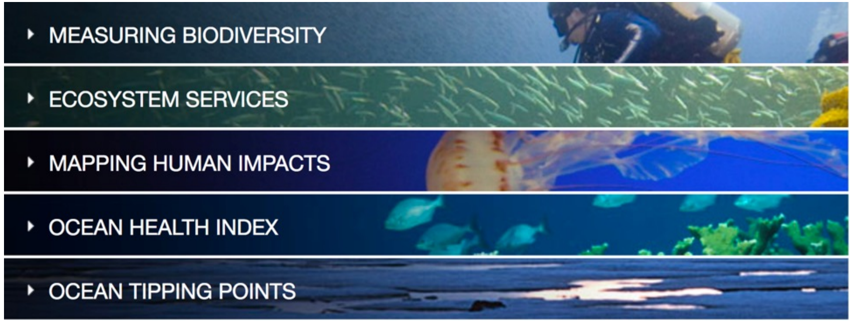
@metamattj
jones@nceas.ucsb.edu
https://orcid.org/0000-0003-0077-4738

Ecological Synthesis

Marine Systems

- ESTUARINE AND MARINE NURSERIES
- RECRUITMENT PATTERNS
- DEEP SEA BIODIVERSITY
- ECOSYSTEM-BASED MANAGEMENT
- MARINE PROTECTED AREAS

Understanding Ocean Health



- MEASURING BIODIVERSITY
- ECOSYSTEM SERVICES
- MAPPING HUMAN IMPACTS
- OCEAN HEALTH INDEX
- OCEAN TIPPING POINTS

Threats and Population Declines

- SEAGRASS ECOSYSTEMS
- CORAL REEFS
- MARINE MAMMALS
- SEA TURTLES
- FISHING
- CLIMATE CHANGE

Climate and Ecosystems

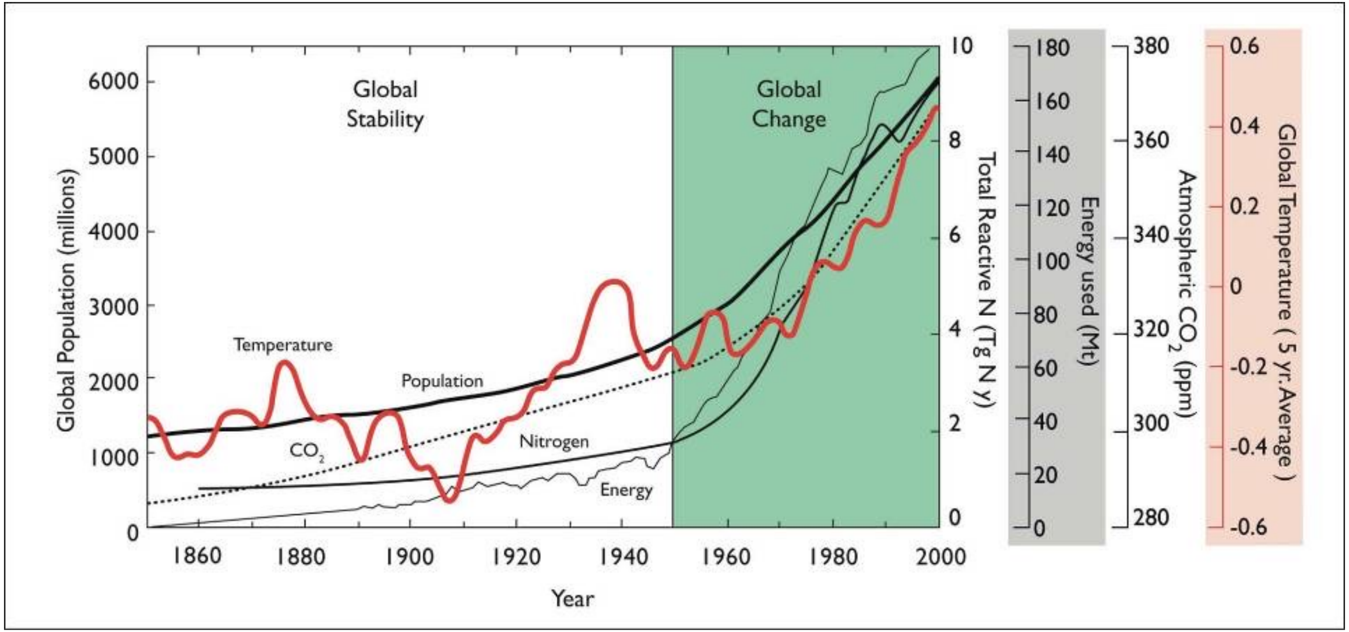
- ARCTIC ECOSYSTEMS
- FIRE REGIMES
- FORESTS
- FRESHWATER AND WETLAND ECOSYSTEMS
- NET PRIMARY PRODUCTIVITY
- SOIL AND NUTRIENT CYCLING
- PERMAFROST

Reproducible Science

Climate Change
Fisheries
Sustainabiity
Subsistence

Science
Governance
Regulation
Policy

Trust in Science



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

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

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

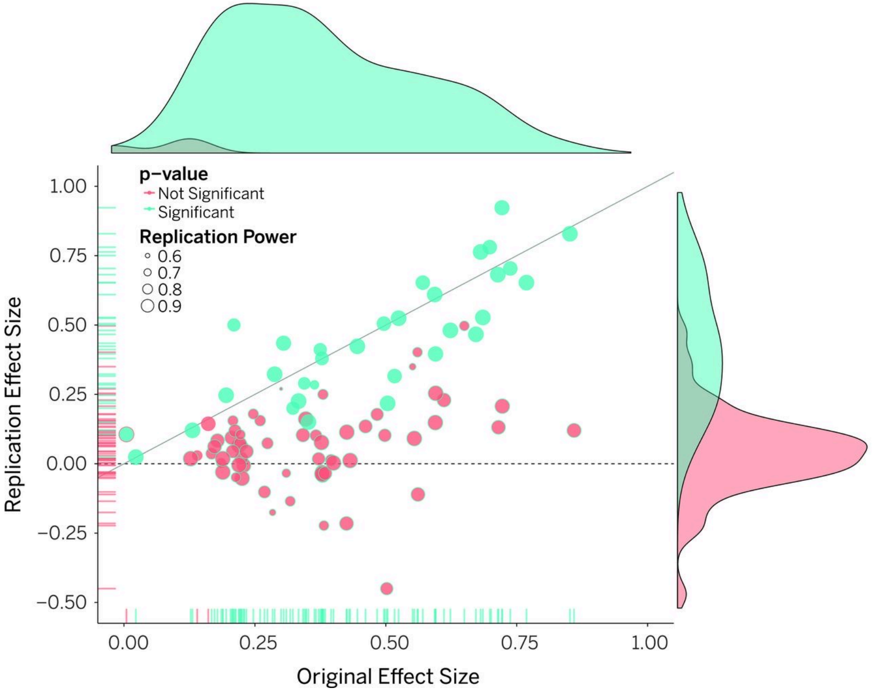
Reproducibility Crisis

“Most research findings are false for most research designs and for most fields”

Ioannidis, 2005

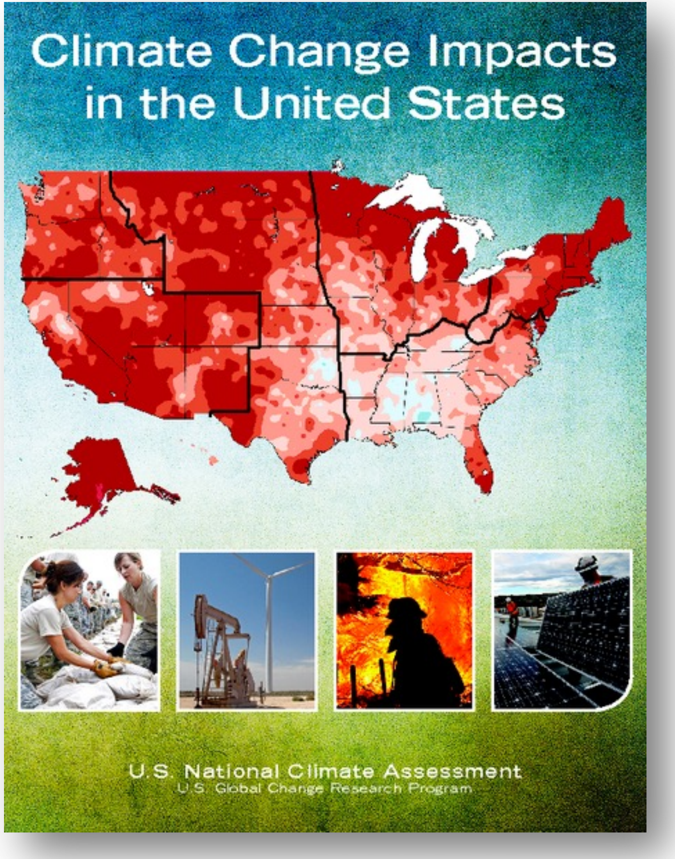
“Most replication effects were smaller than original results”

Open Science Collaboration, 2015



doi:10.1126/science.aac4716

National Climate Assessment


Climate Change Impacts in the United States
U.S. National Climate Assessment
U.S. Global Change Research Program

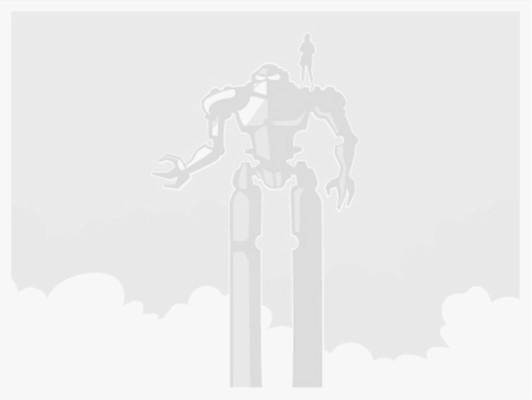
"This report is the result of a **three-year** analytical effort by a team of **over 300 experts**, overseen by a broadly constituted Federal Advisory Committee of **60 members**. It was developed from information and analyses gathered in over 70 workshops and listening sessions held across the country."

Computational Reproducibility

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

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

Increase **trust in science**



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

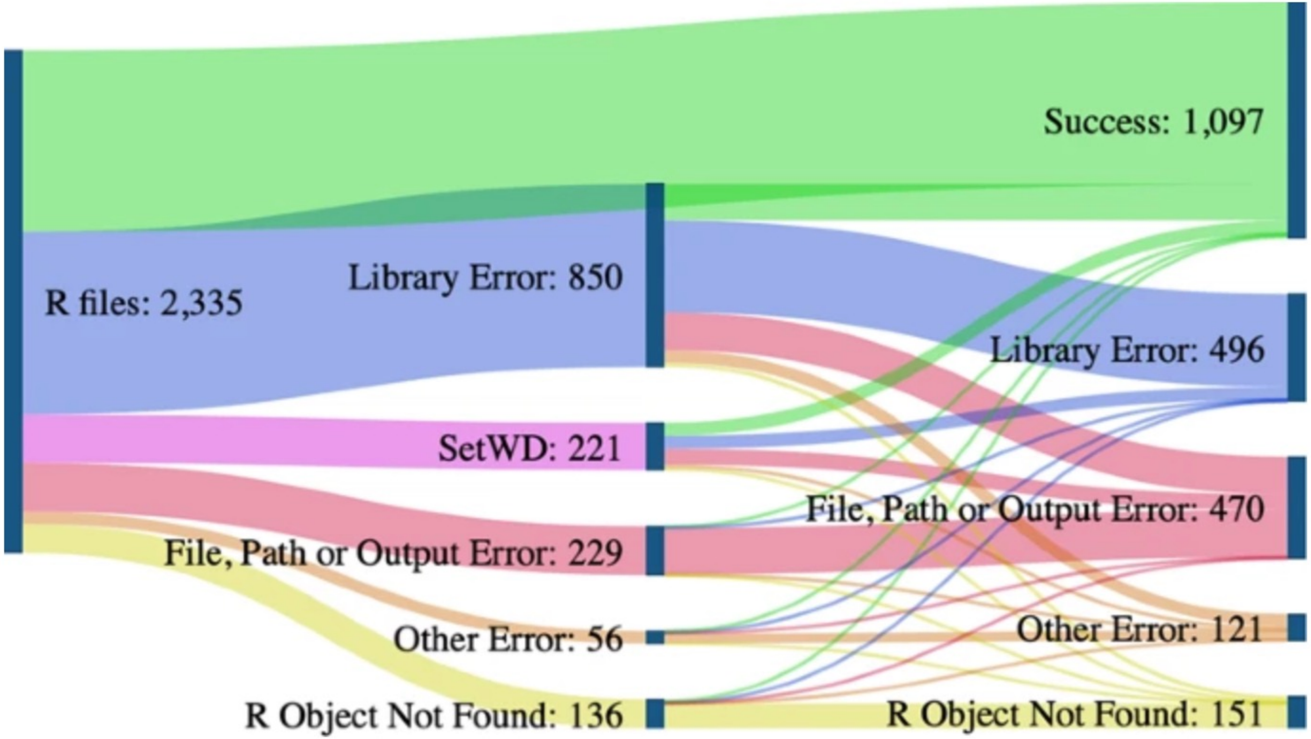
NSF Arctic Data Center

DataONE

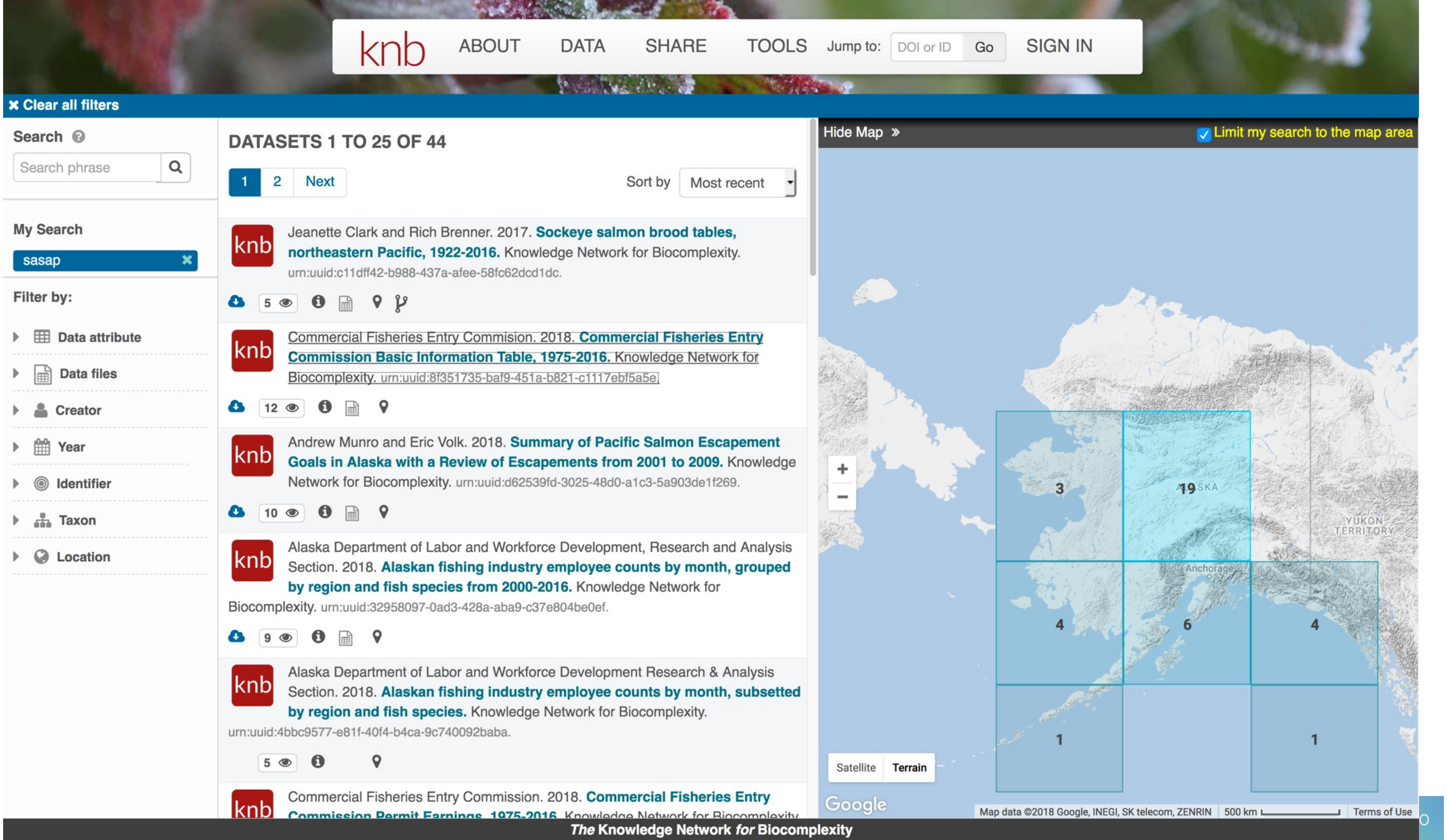
Harvard Dataverse: Reproducibility of R Code

- 26% ran without error
- 46% ran after cleaning

Trisovic, Ana, Matthew K. Lau, Thomas Pasquier, and Mercè Crosas. 2022. "**A Large-Scale Study on Research Code Quality and Execution**." *Scientific Data* 9 (1). https://doi.org/10.1038/s41597-022-01143-6.



Success rate and errors before and after code cleaning. To objectively determine the effects of code cleaning, we subset the results that have explicit "successes" and errors while excluding the ones with TLE values as the outcome. As a result, the count of files in this figure is lower than the total count.



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.

5

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.

12

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.

10

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.

9

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.

5

Commercial Fisheries Entry Commission. 2018. **Commercial Fisheries Entry Commission Permit Earnings, 1975-2016.** Knowledge Network for Biocomplexity.

***The* Knowledge Network *for* Biocomplexity**

DataONE

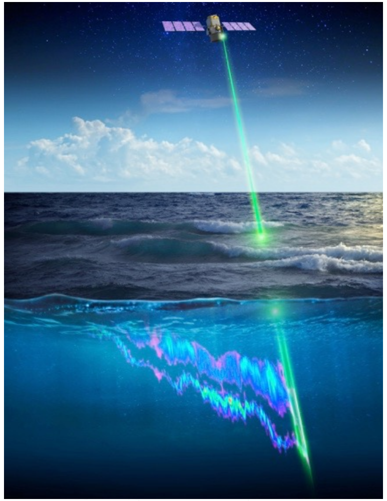
https://search.dataone.org

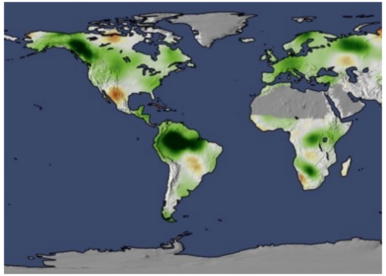
Global
Data Coverage

926K
Data Packages

58
Repositories

143K
Contributors

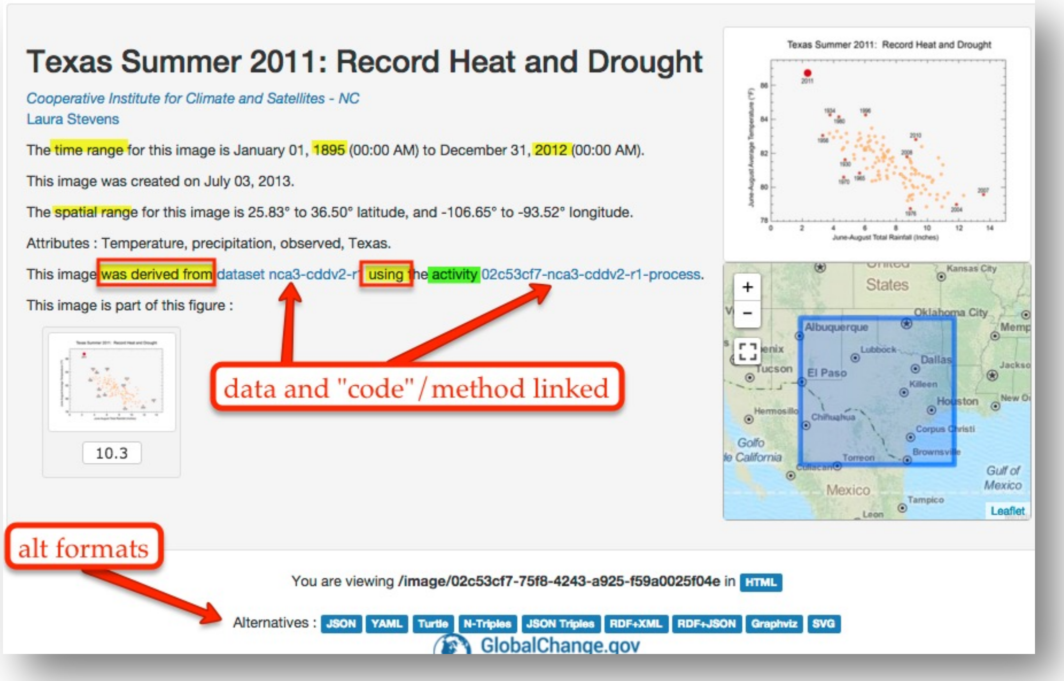




Computational Provenance

Origin, processing history of data

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



Texas Summer 2011: Record Heat and Drought

Cooperative Institute for Climate and Satellites - NC
Laura Stevens

The time range for this image is January 01, 1895 (00:00 AM) to December 31, 2012 (00:00 AM).

This image was created on July 03, 2013.

The spatial range for this image is 25.83° to 36.50° latitude, and -106.65° to -93.52° longitude.

Attributes : Temperature, precipitation, observed, Texas.

This image was derived from dataset nca3-cddv2-r1 using the activity 02c53cf7-nca3-cddv2-r1-process.

This image is part of this figure :

10.3

data and "code"/method linked

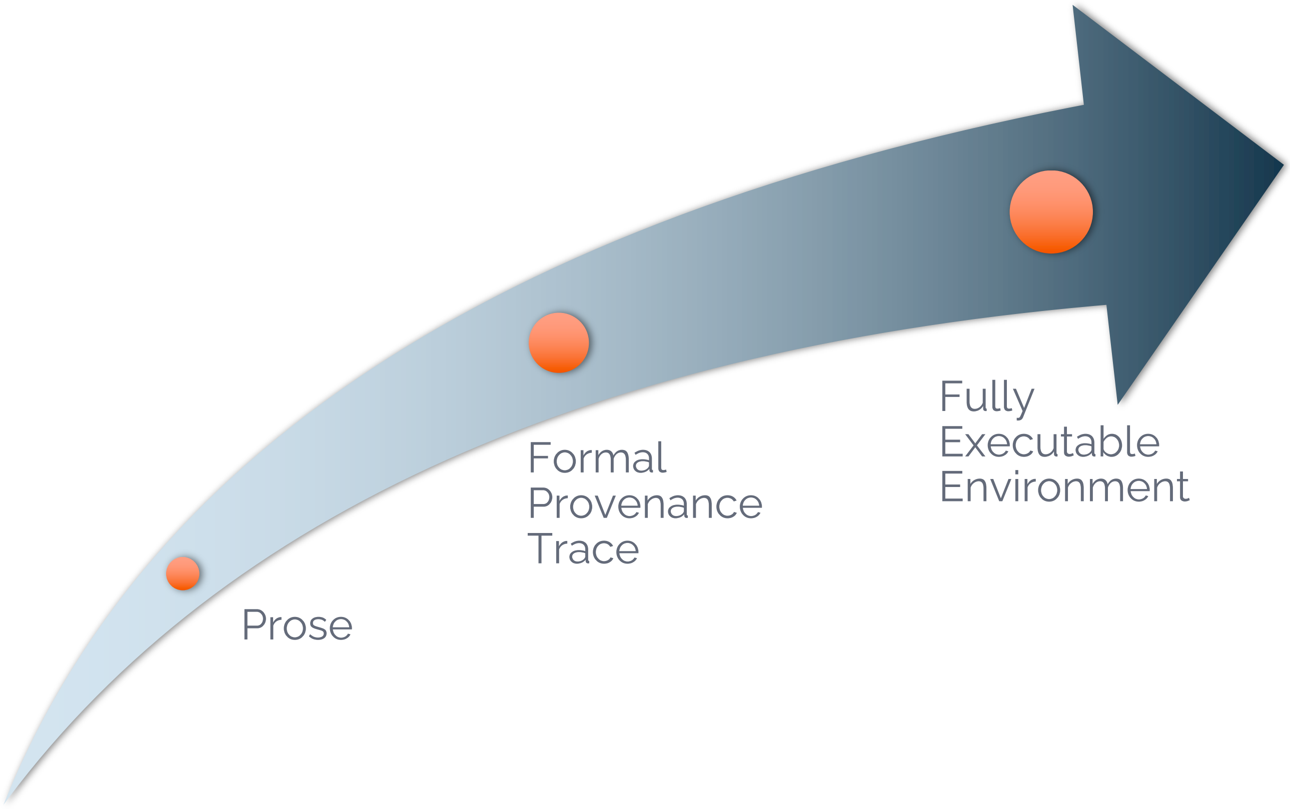
alt formats

You are viewing /image/02c53cf7-75f8-4243-a925-f59a0025f04e in HTML

Alternatives : JSON YAML Turtle N-Triples JSON Triples RDF+XML RDF+JSON Graphviz SVG

Provenance

Origin and processing history of artifacts



Prose

Formal Provenance Trace

Fully Executable Environment

Provenance in DataONE

Facilitate reproducible science

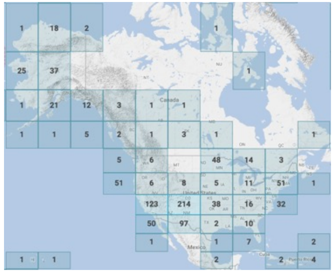
- Track **data derivation** history
- Track data **inputs** and **outputs** of analyses
- Track analysis and model **executions**
- Preserve and document software **workflows**
- Link all of these to **publications**

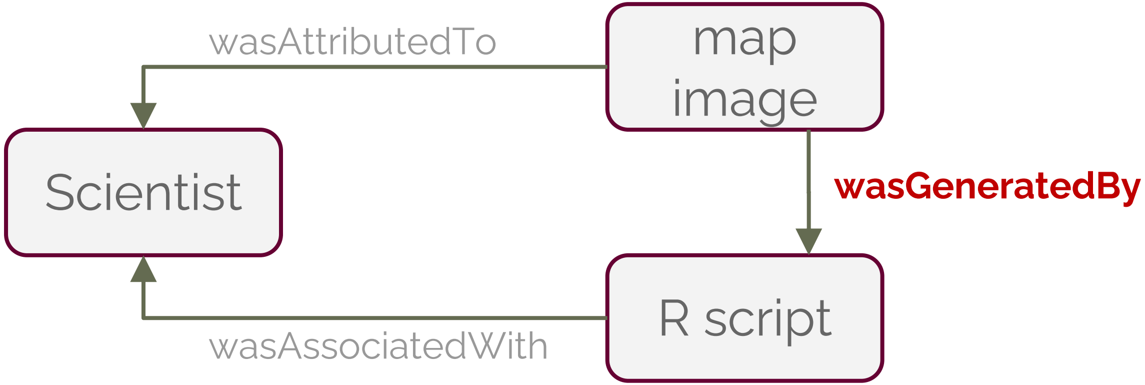
Provenance for Science Workflows

ProvONE

ProvONE – an extension of W3C PROV

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





wasAttributedTo

map image

Scientist

wasGeneratedBy

R script

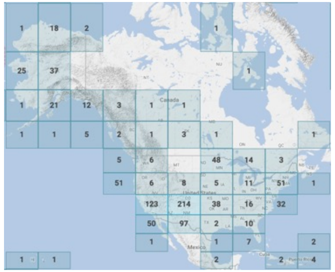
wasAssociatedWith

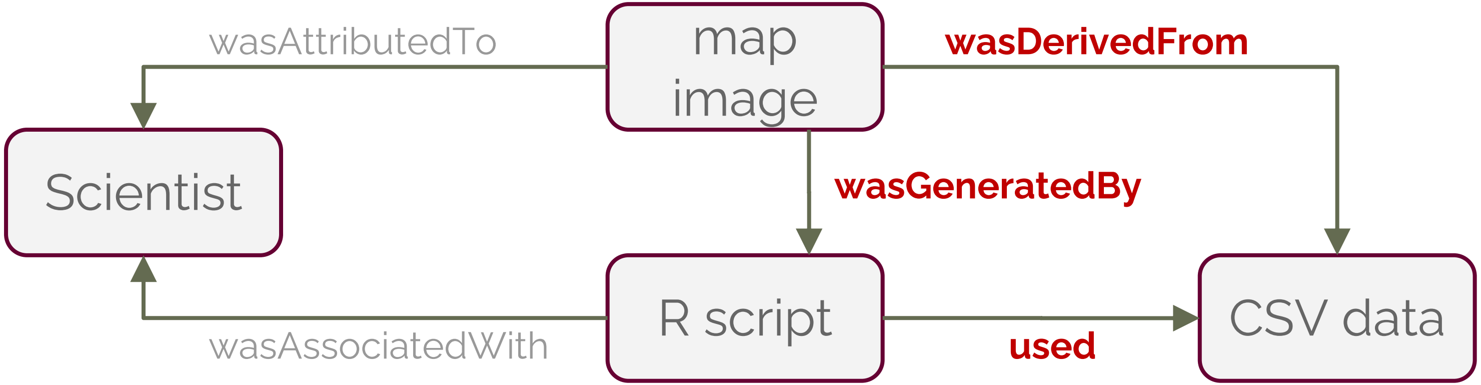
Provenance for Science Workflows

ProvONE

ProvONE – an extension of W3C PROV

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





map image

wasAttributedTo

Scientist

wasDerivedFrom

wasGeneratedBy

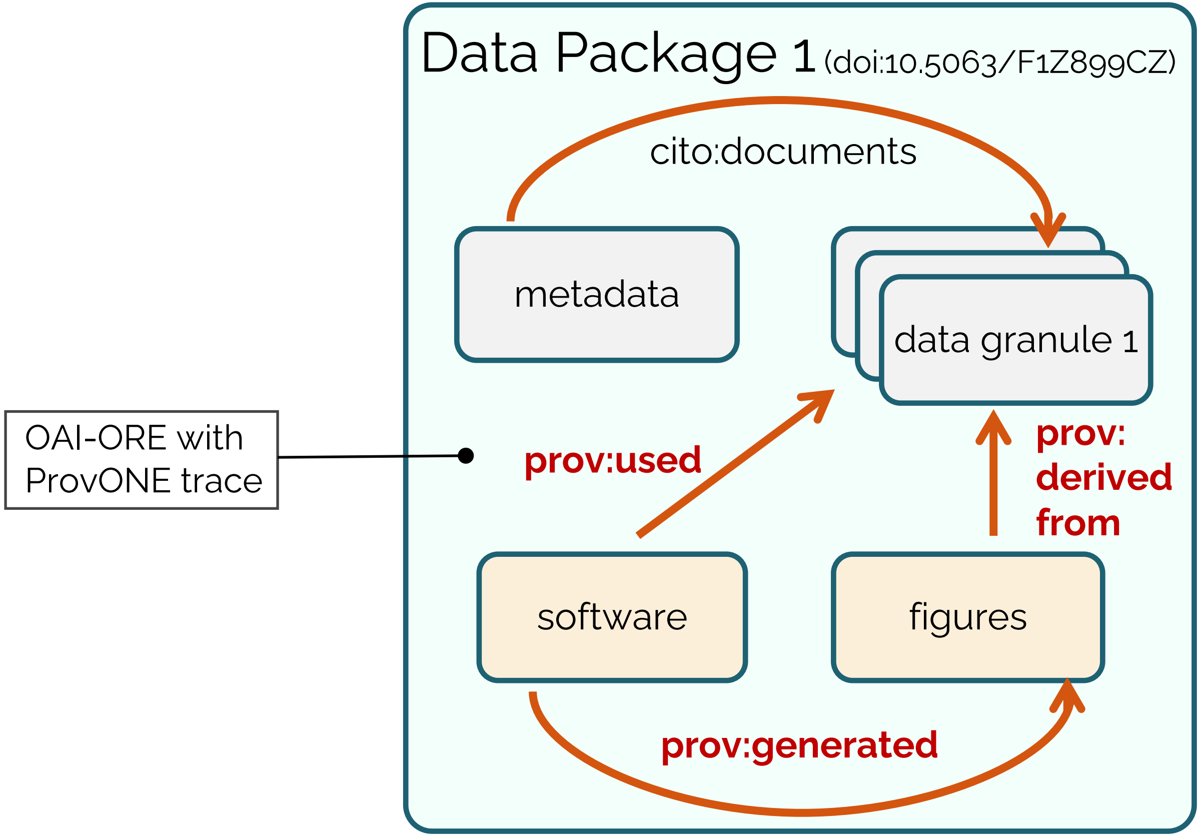
R script

wasAssociatedWith

used

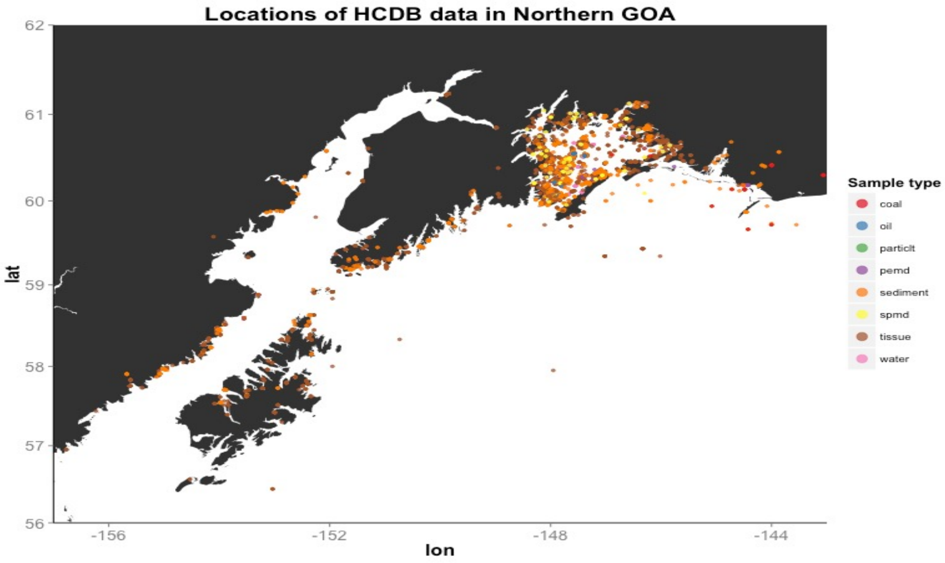
CSV data

Data Package with Provenance

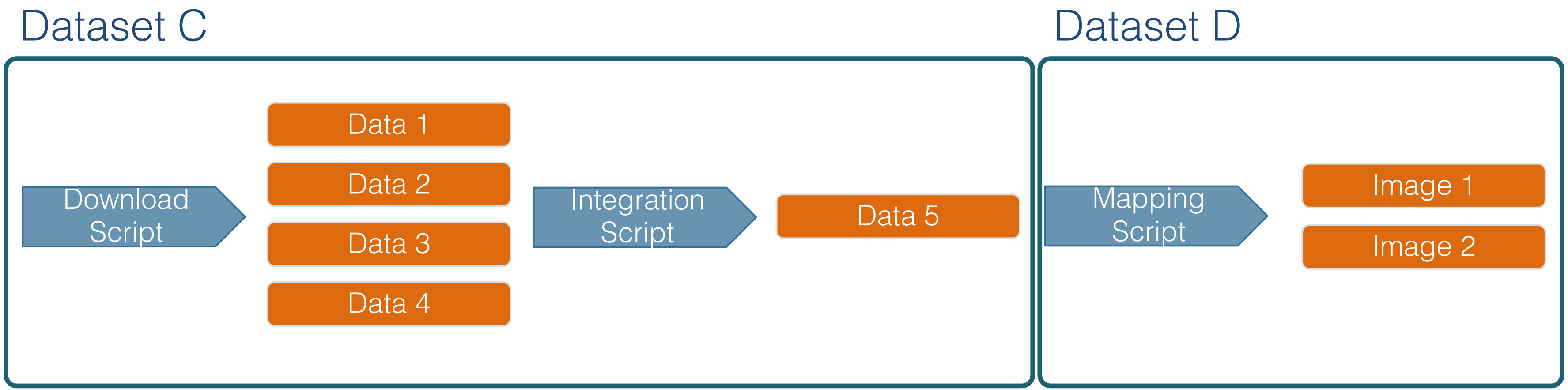


Hydrocarbon Data Example

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



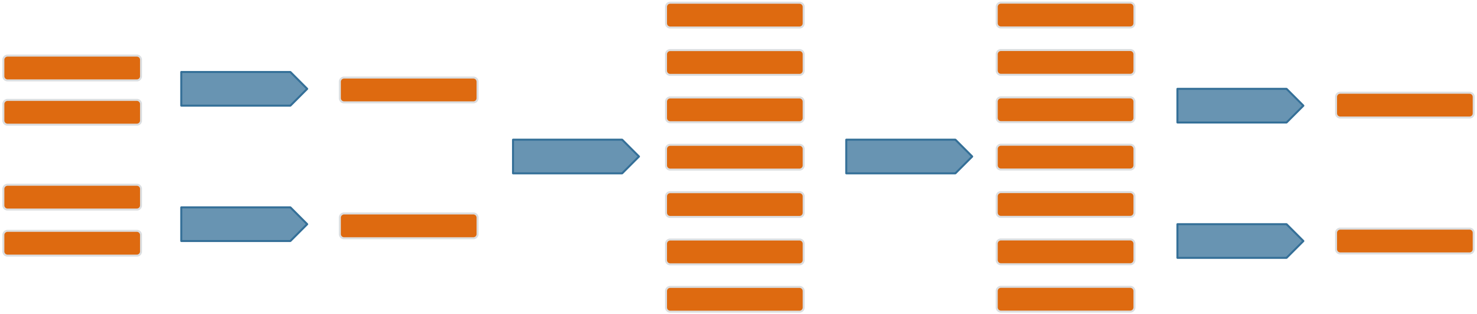
Publishing Data Workflows



Hydrocarbon Data Example

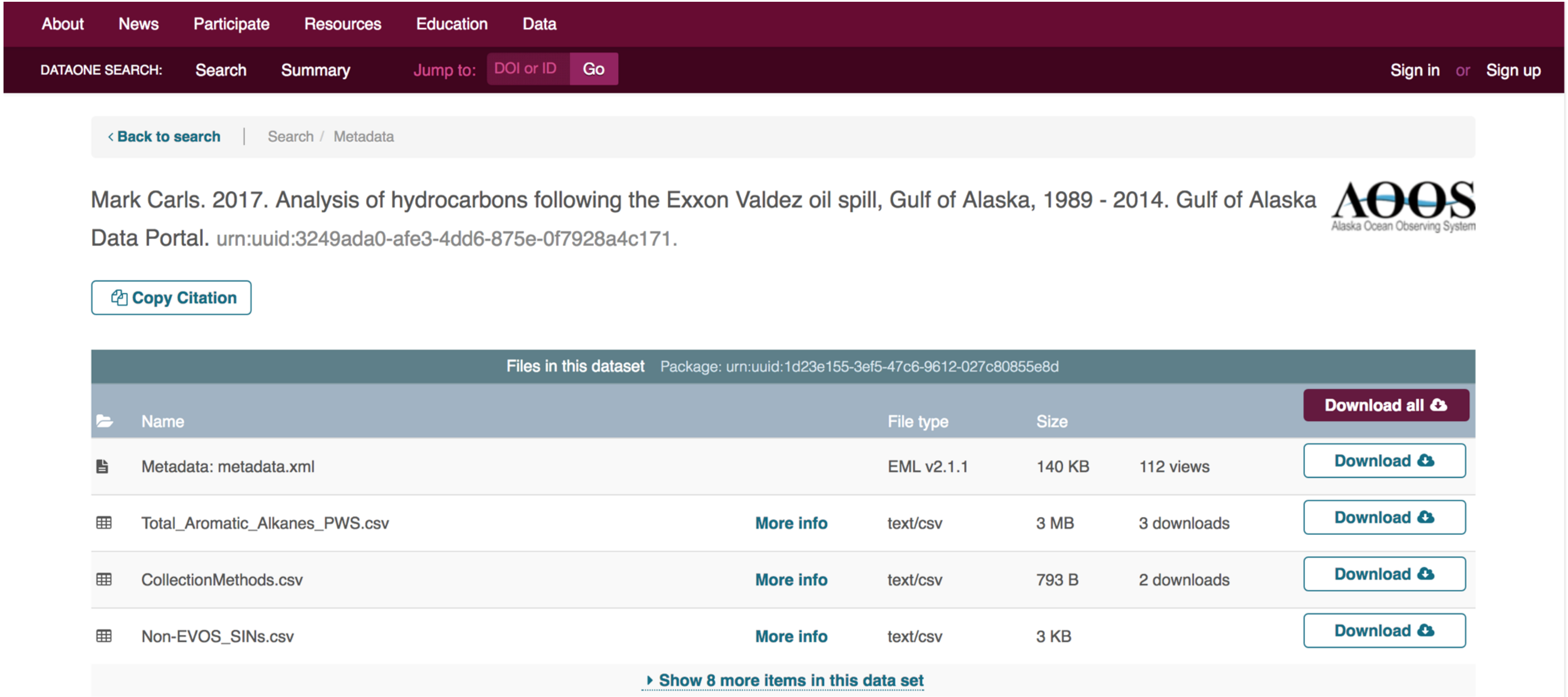
Complex Workflows

Simplified view of complex workflows



Provenance Display

DataONE Search

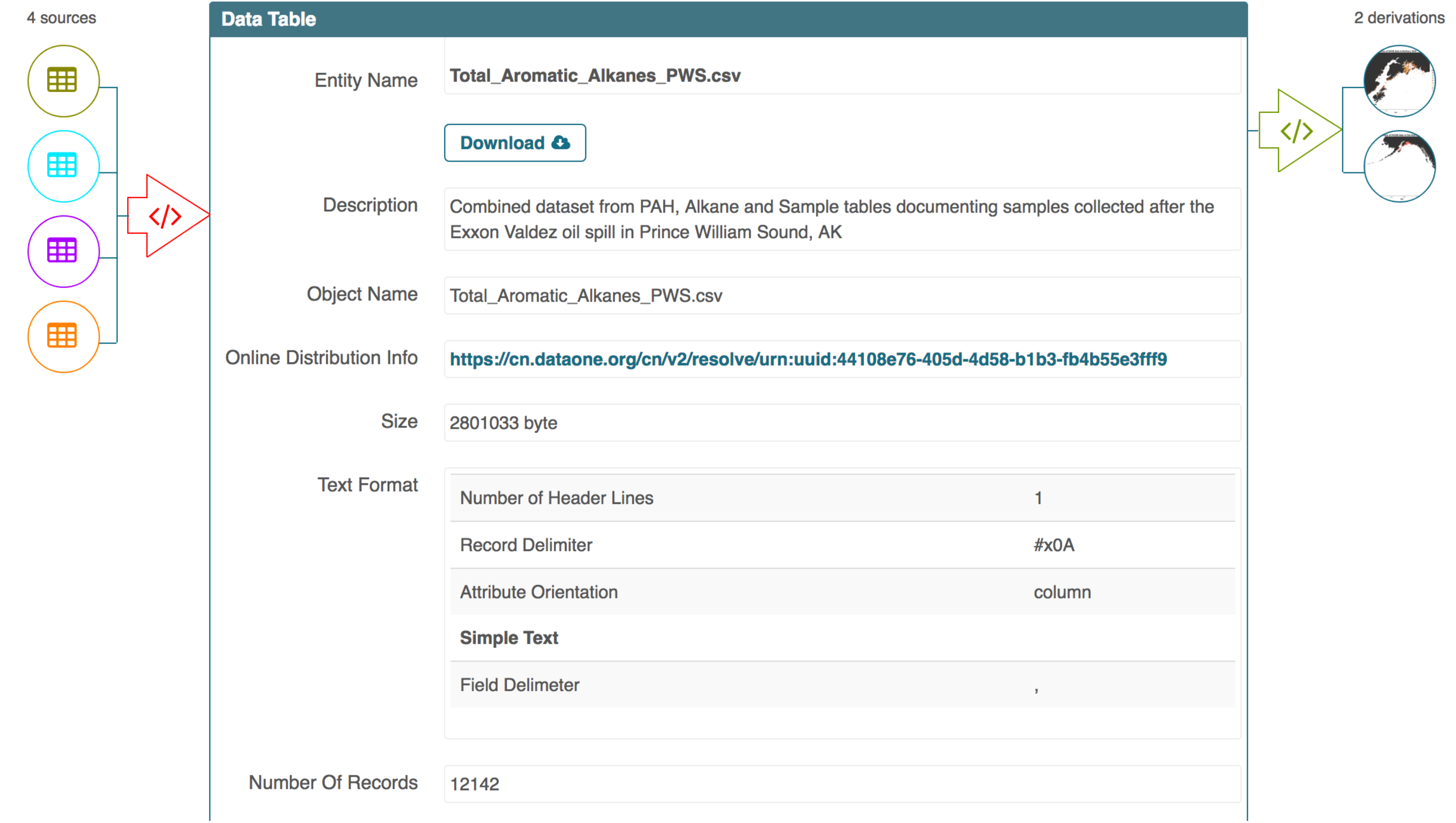


Mark Carls. 2017. Analysis of hydrocarbons following the Exxon Valdez oil spill, Gulf of Alaska, 1989 - 2014. Gulf of Alaska Data Portal. urn:uuid:3249ada0-afe3-4dd6-875e-0f7928a4c171.

Files in this dataset Package: urn:uuid:1d23e155-3ef5-47c6-9612-027c80855e8d

Name		File type	Size		Download all
Metadata: metadata.xml		EML v2.1.1	140 KB	112 views	Download
Total_Aromatic_Alkanes_PWS.csv	More info	text/csv	3 MB	3 downloads	Download
CollectionMethods.csv	More info	text/csv	793 B	2 downloads	Download
Non-EVOS_SINs.csv	More info	text/csv	3 KB		Download

Data Table, Image, and Other Data Details

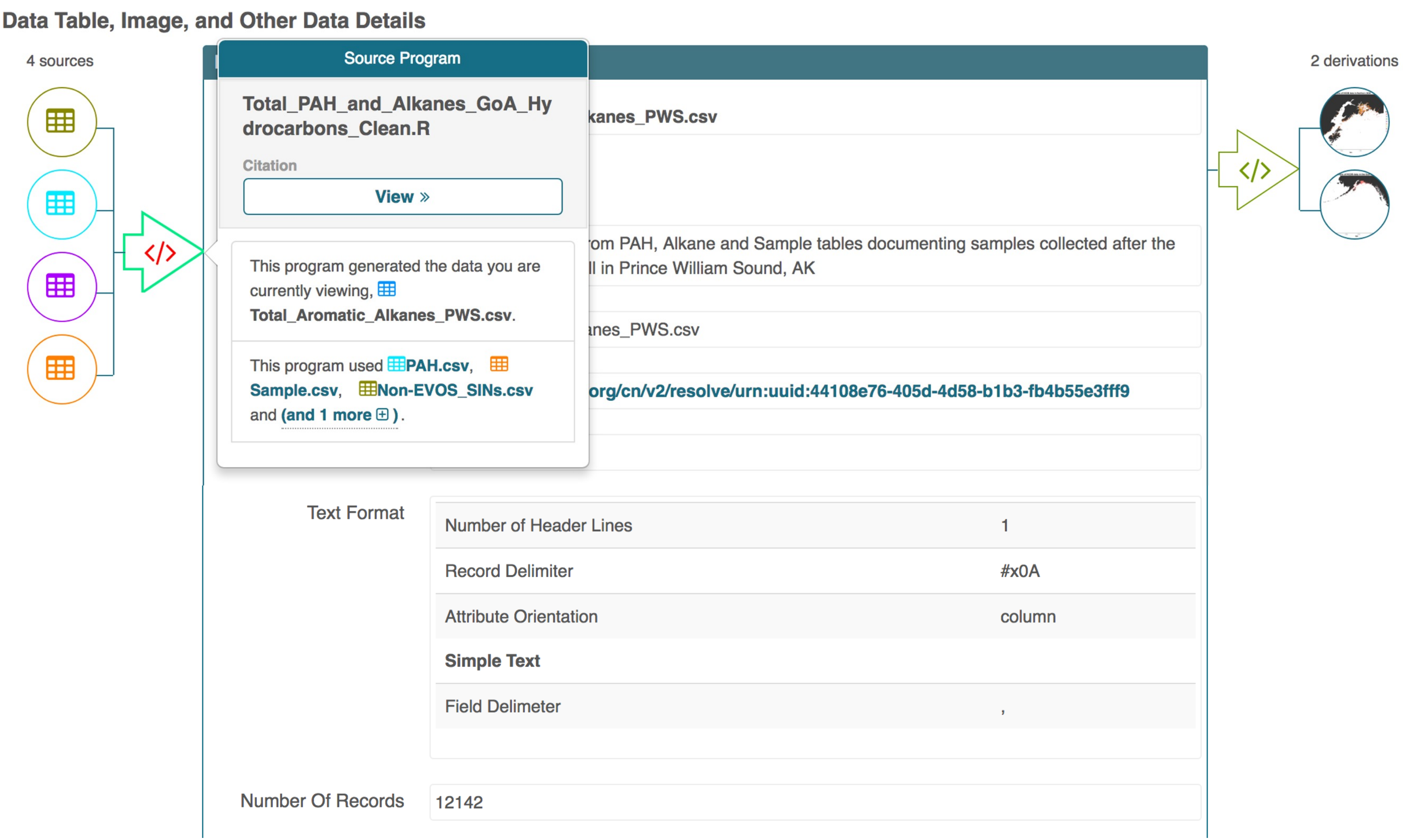


Data Table

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 Records	12142

Text Format:

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



Data Table, Image, and Other Data Details

4 sources

2 derivations

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 ⊞)**.

kanes_PWS.csv

rom PAH, Alkane and Sample tables documenting samples collected after the ll in Prince William Sound, AK

anes_PWS.csv

org/cn/v2/resolve/urn:uuid:44108e76-405d-4d58-b1b3-fb4b55e3fff9

Text Format

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

Number Of Records 12142

Credit where credit is due

Indexing and exposing data citations in international data repository networks

MAKE
DATA
COUNT

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

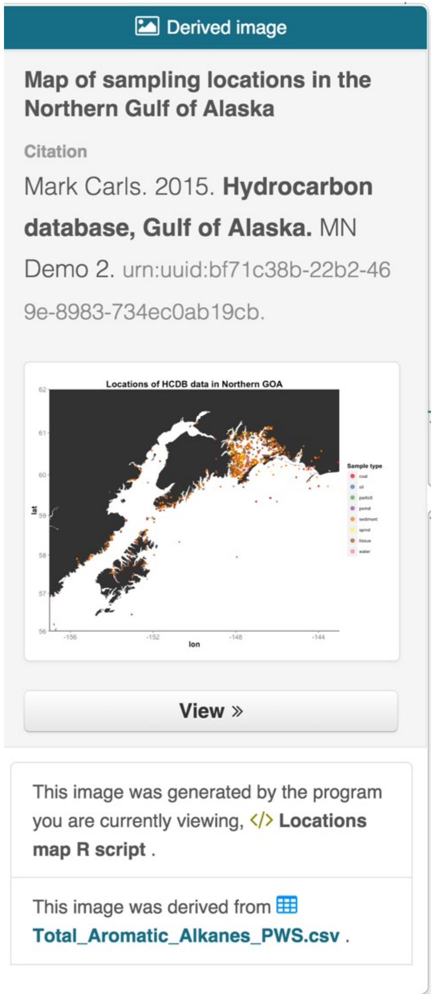
Joint Declaration of Data Citation Principles, https://doi.org/10.25490/a97f-egyk

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



Derived image

Map of sampling locations in the Northern Gulf of Alaska

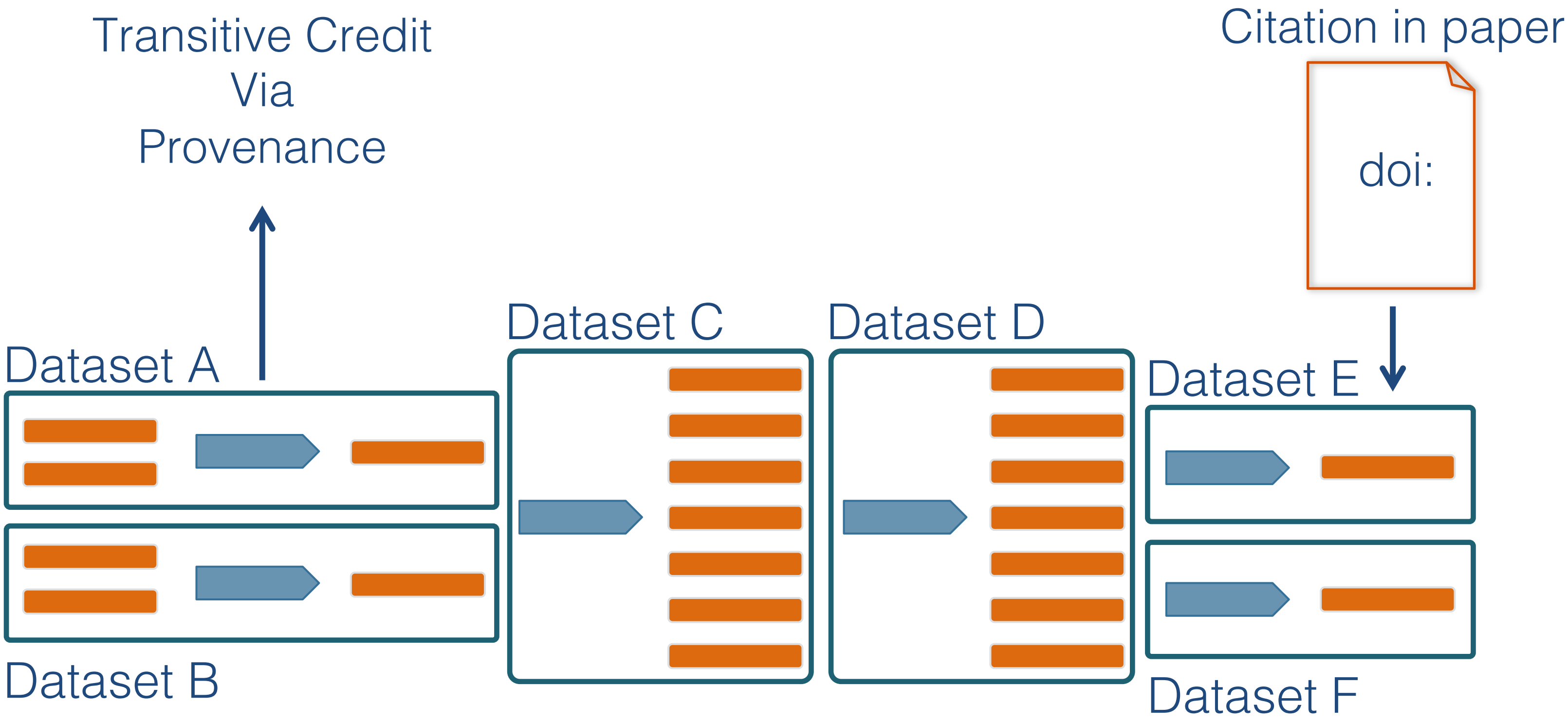
Citation

Mark Carls. 2015. **Hydrocarbon database, Gulf of Alaska.** MN Demo 2. urn:uuid:bf71c38b-22b2-469e-8983-734ec0ab19cb.

This image was generated by the program you are currently viewing, </> **Locations map R script** .

This image was derived from **Total_Aromatic_Alkanes_PWS.csv** .

Citing multi-generational workflows



Transitive Credit
Via
Provenance

Citation in paper

doi:

Dataset A

Dataset B

Dataset C

Dataset D

Dataset E

Dataset F

Evolution of the Living Paper

Scholarly Publications

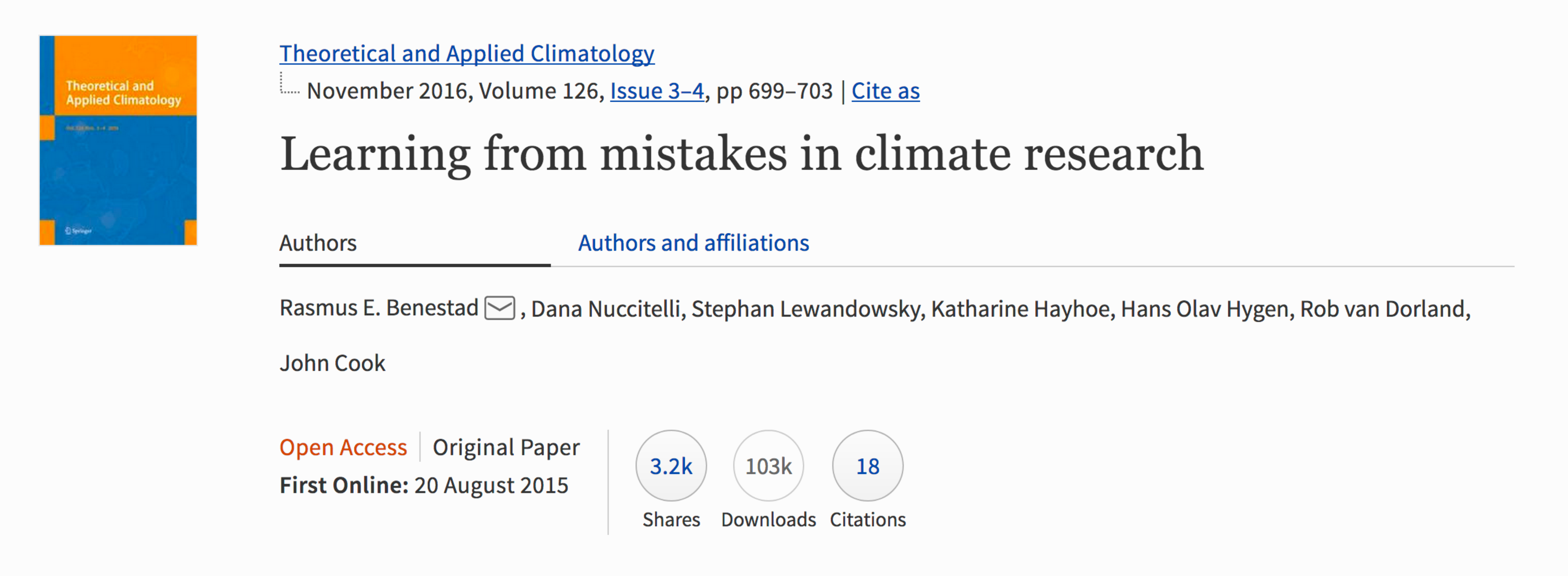
1st Gen	**Prose**		
2nd Gen	**Prose**	**+ Data**	
3rd Gen	**Prose**	**+ Data**	**+ Code**

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

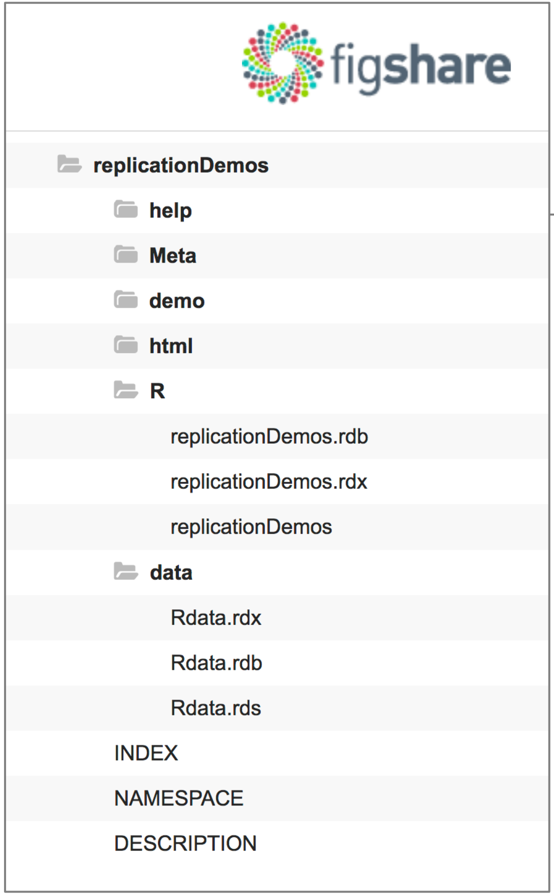
Authors | Authors and affiliations

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

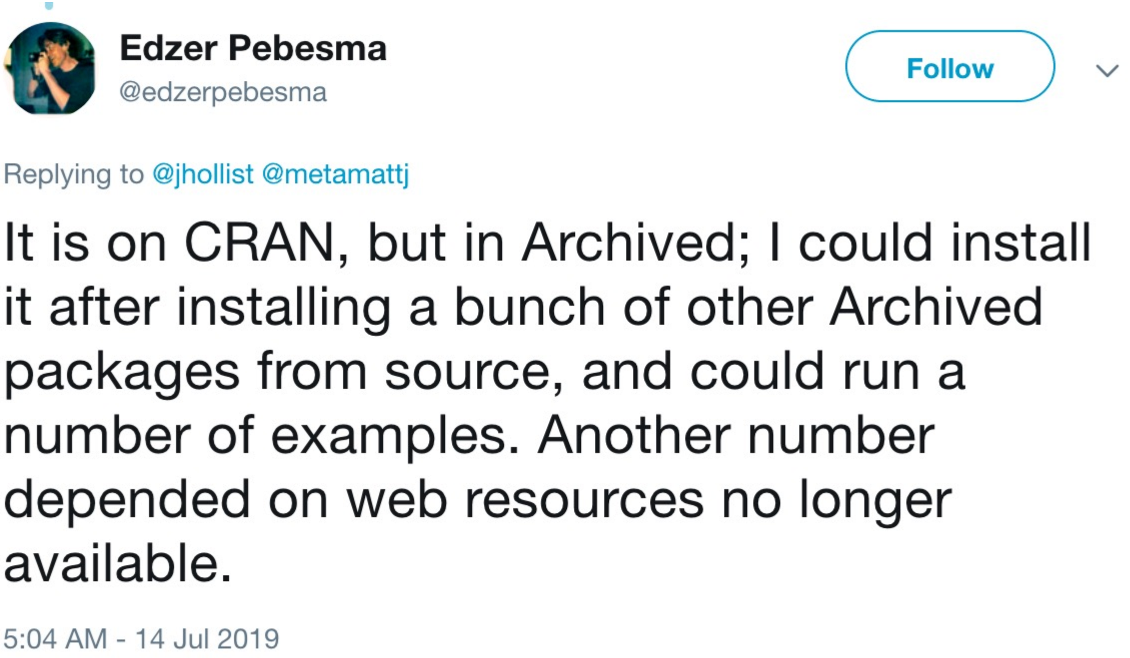
3.2k	103k	18
Shares	Downloads	Citations

Ships with an R package



figshare

- replicationDemos
  - help
  - Meta
  - demo
  - html
  - R
    - replicationDemos.rdb
    - replicationDemos.rdx
    - replicationDemos
  - data
    - Rdata.rdx
    - Rdata.rdb
    - Rdata.rds
  - INDEX
  - NAMESPACE
  - DESCRIPTION



Edzer Pebesma
@edzerpebesma

Replying to @jhollist @metamattj

It is on CRAN, but in Archived; I could install it after installing a bunch of other Archived packages from source, and could run a number of examples. Another number depended on web resources no longer available.

5:04 AM - 14 Jul 2019

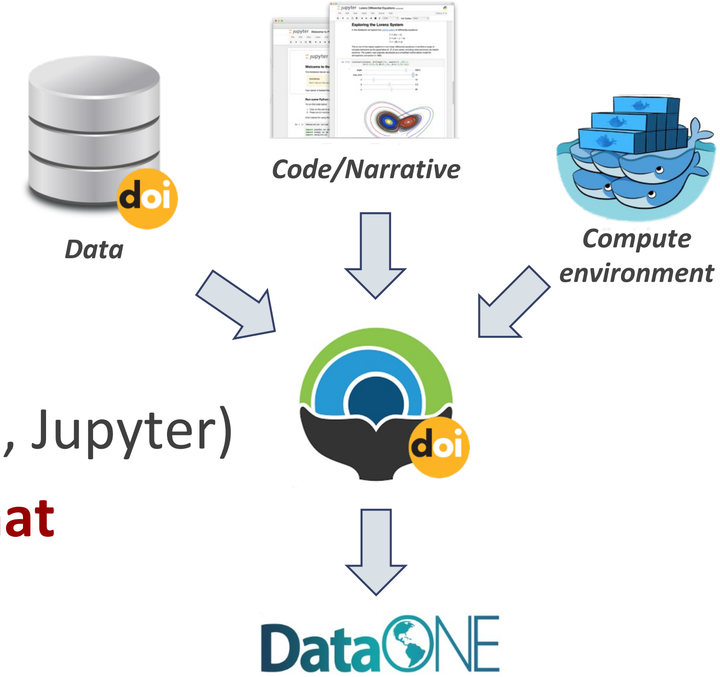
Parsing Reproducibility

- **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

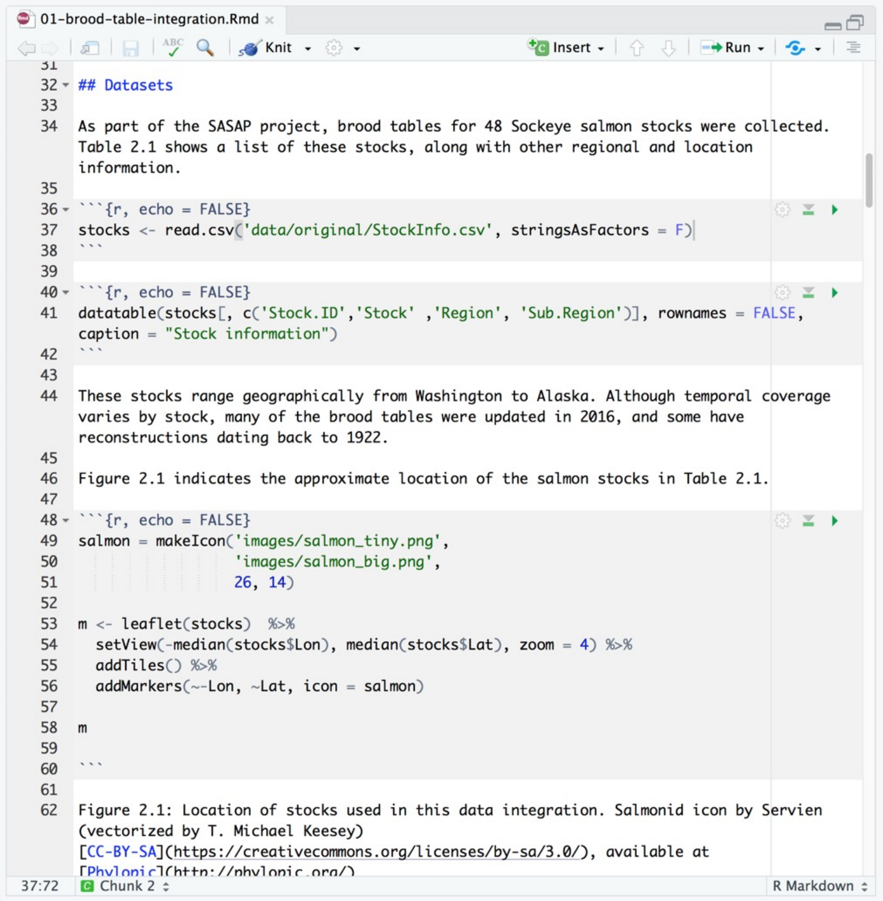
V. Stodden. (2013). *Resolving Irreproducibility in Empirical and Computational Research.* IMS Bulletin

What exactly is (in) a Tale?

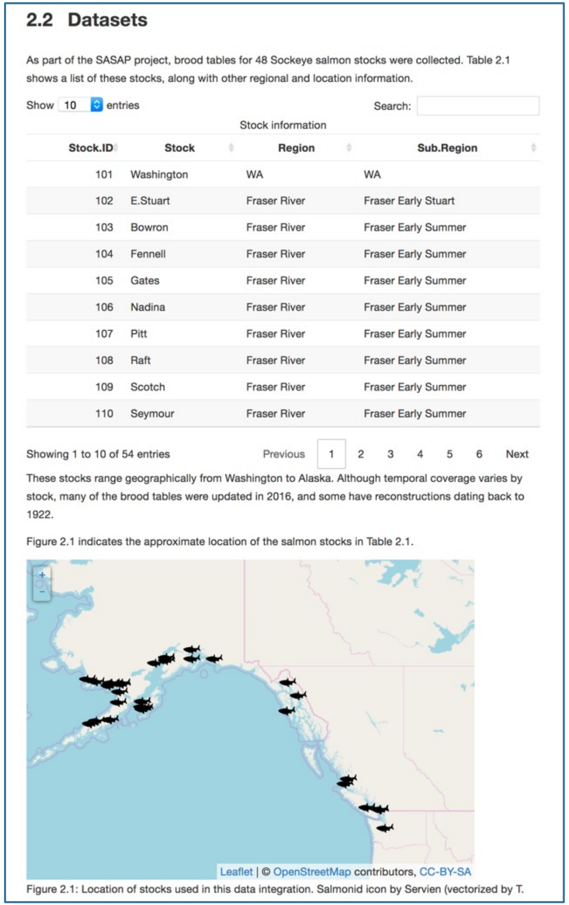
- **Tale** = executable **research object**, i.e.
  - **data** (references)
  - **+ code** (computational methods)
  - **+ narrative** (traditional science story)
  - **+ compute environment** (e.g. RStudio, Jupyter)
- Captured in a **standards-based tale format** complete with metadata



Quarto/Rmarkdown as Provenance



~~~
## 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
[Phylopic](http://phylopic.org/)
~~~



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 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

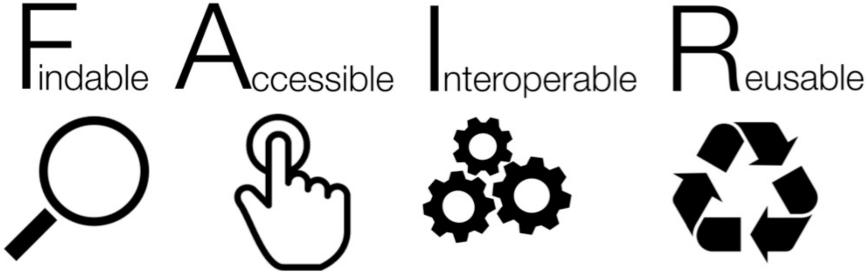
These stocks range geographically from Washington to Alaska. Although temporal coverage varies by stock, many of the brood tables were updated in 2016, and some have reconstructions dating back to 1922.

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.

Foundational Infrastructure

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



https://www.force11.org/fairprinciples