




# Provenance tools for reproducible science

**Matthew B. Jones**

National Center for Ecological Analysis and Synthesis (NCEAS)

@metamattj

jones@nceas.ucsb.edu

0000-0003-0077-4738

# What is Provenance?

- **Provenance**: **origin** and **processing history** of data
  - *Audit trail*
  - *Attribution/credit*
  - R*eplication* and *reproducible science*
  - *Discovery* of data, methodologies, experiments

- Provenance continuum
  - Prose
  - Formal provenance traces
  - Fully executable environments

# Facilitate Reproducible Science

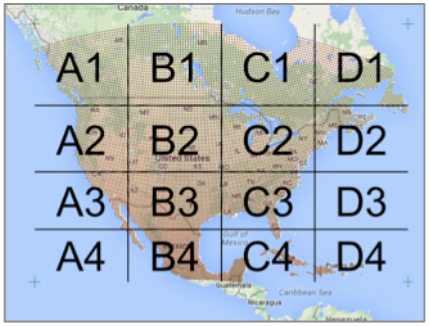
**Goal: Use provenance for reproducible science**

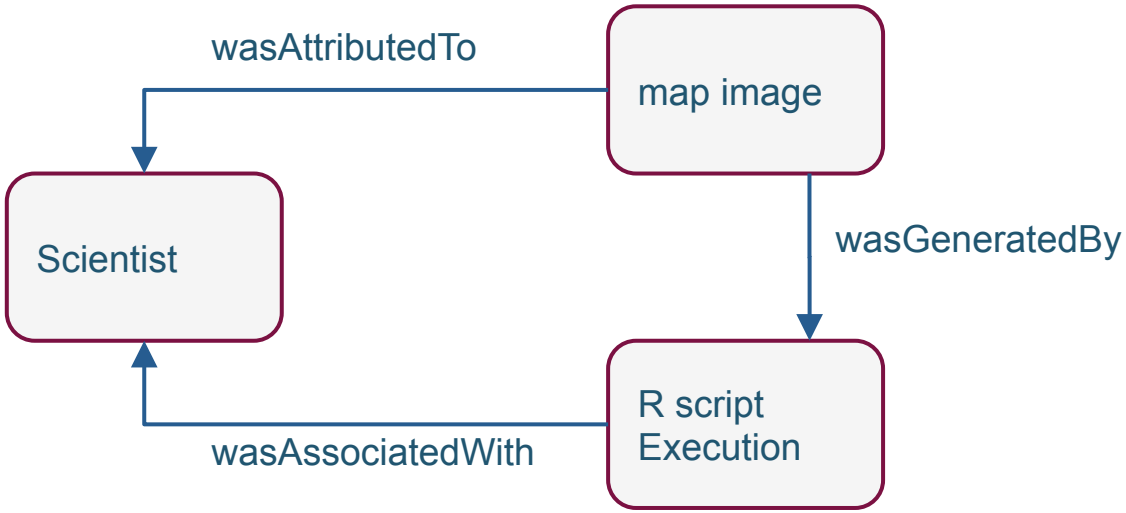
- Track **data derivation** history
- Track data **inputs** and **outputs** of analyses
- Track analysis and model **executions**
- Preserve and document software **workflows**
- Link all of these to publications

**Goal: Easy provenance, in the tools scientists use**

# Using a common model

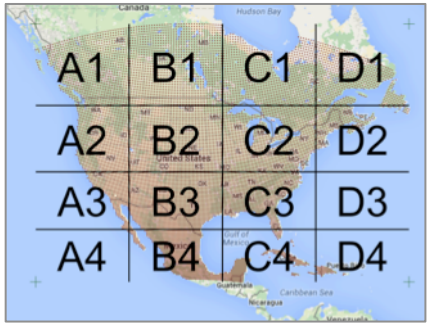
- Example: Map from CSV data

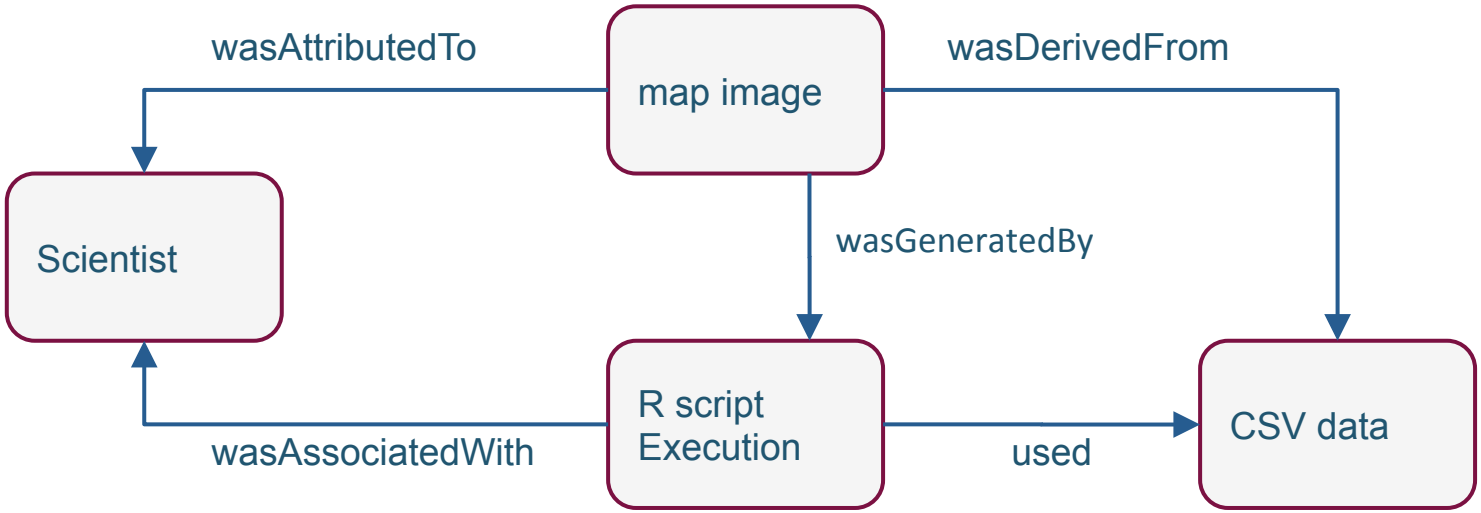


# Using a common model

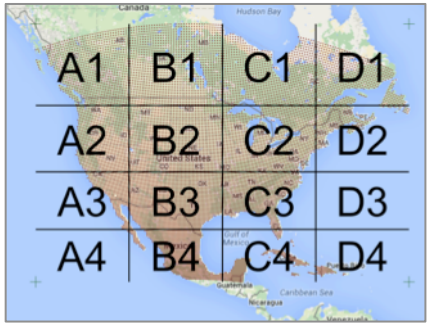
- Example: Map from CSV data



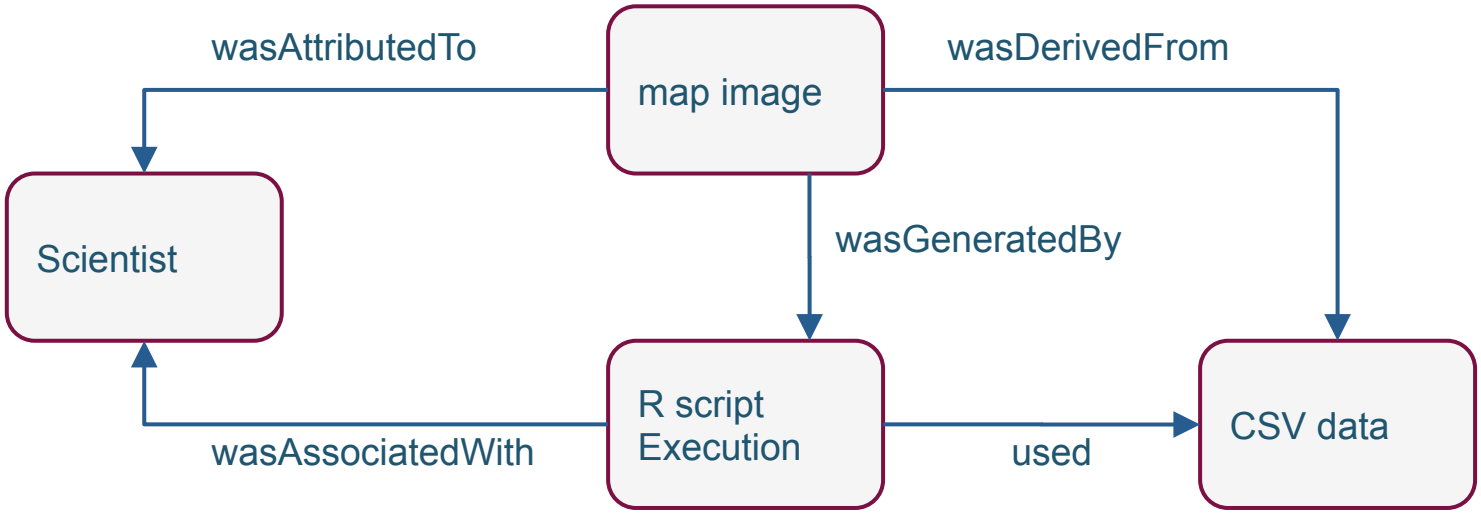
# Using a common model

- Example: Map from CSV data




ProvONE

< “map image” prov:wasDerivedFrom “CSV data” >



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.

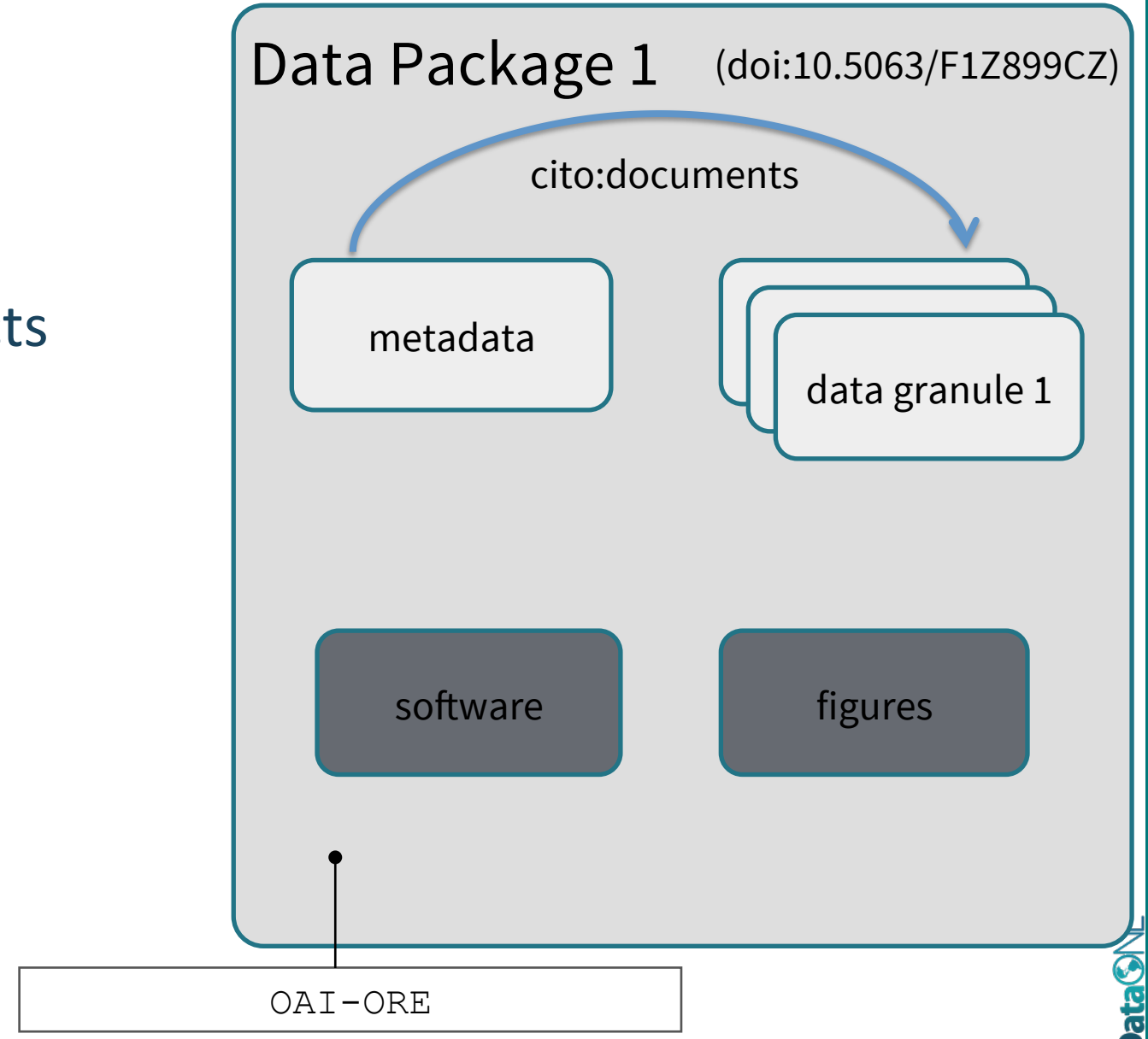
Copy Citation

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

| Name | | File type | Size | Downloads | Download All |
|---|---|---|---|---|---|
| Metadata: metadata.xml | | EML v2.1.1 | 140 KB | 158 views | Download |
| Total_Aromatic_Alkanes_PWS.csv | More info | text/csv | 3 MB | 6 downloads | Download |
| CollectionMethods.csv | More info | text/csv | 793 B | 3 downloads | Download |
| Non-EVOS_SINs.csv | More info | text/csv | 3 KB | 1 download | Download |
| PAH.csv | More info | text/csv | 5 MB | 1 download | Download |
| Alkane.csv | More info | text/csv | 4 MB | 1 download | Download |
| Sample.csv | More info | text/csv | 1 MB | 1 download | Download |
| Total_PAH_and_Alkanes_GoA_Hydrocarbons_Clean.R | More info | application/R | 5 KB | 14 downloads | Download |
| hcdbSites.R | More info | application/R | 3 KB | 4 downloads | Download |
| hcdbSampleLocs.png | More info | image/png | 99 KB | 176 downloads | Download |
| hcdbSamplesGOA.png | More info | image/png | 57 KB | 173 downloads | Download |
| DataDownload.R | More info | application/R | 2 KB | 4 downloads | Download |

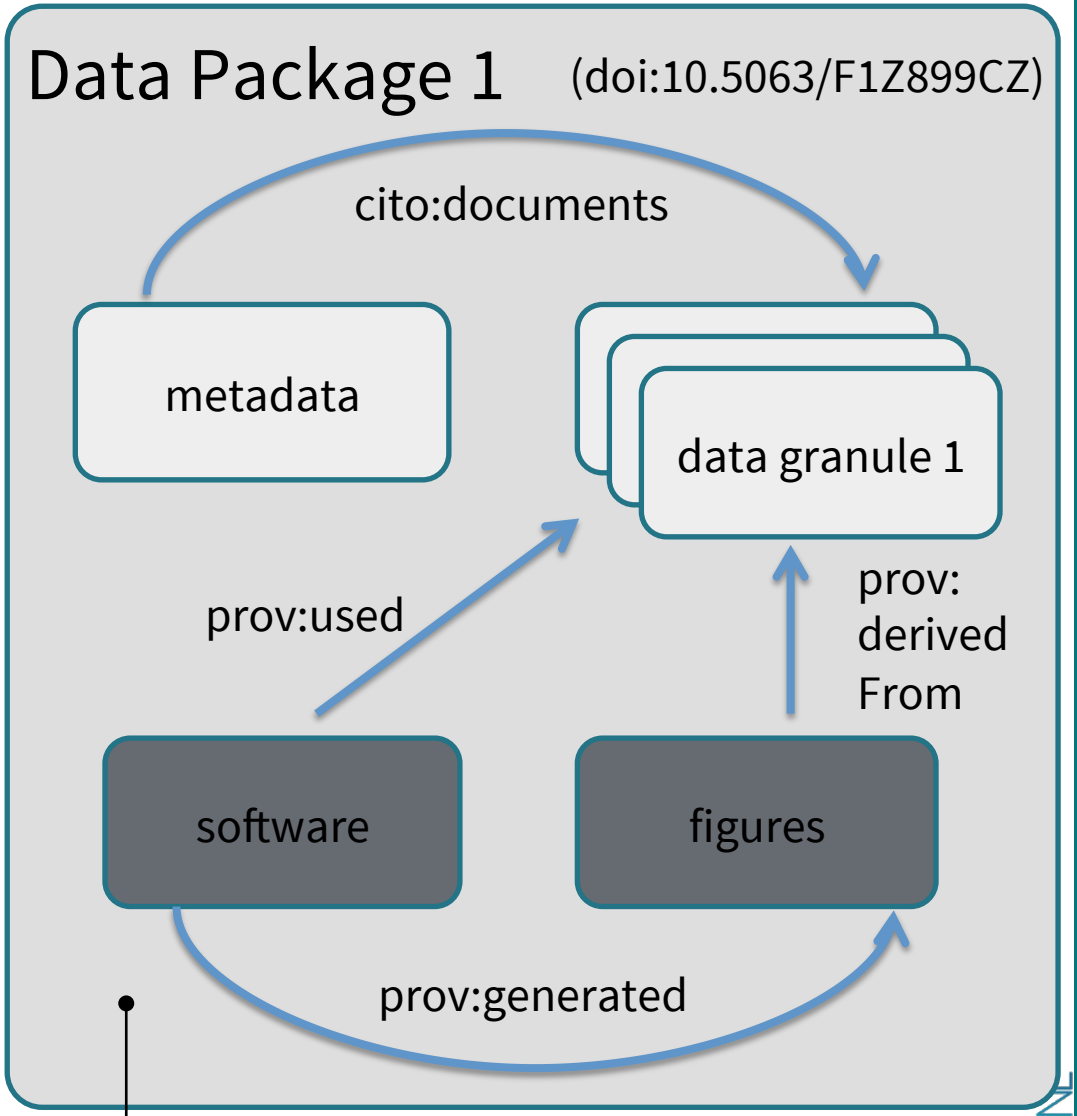
# Data package sans provenance

- Metadata
- Data files
- Other products
  - Photos
  - Graphs
  - Movies
  - …

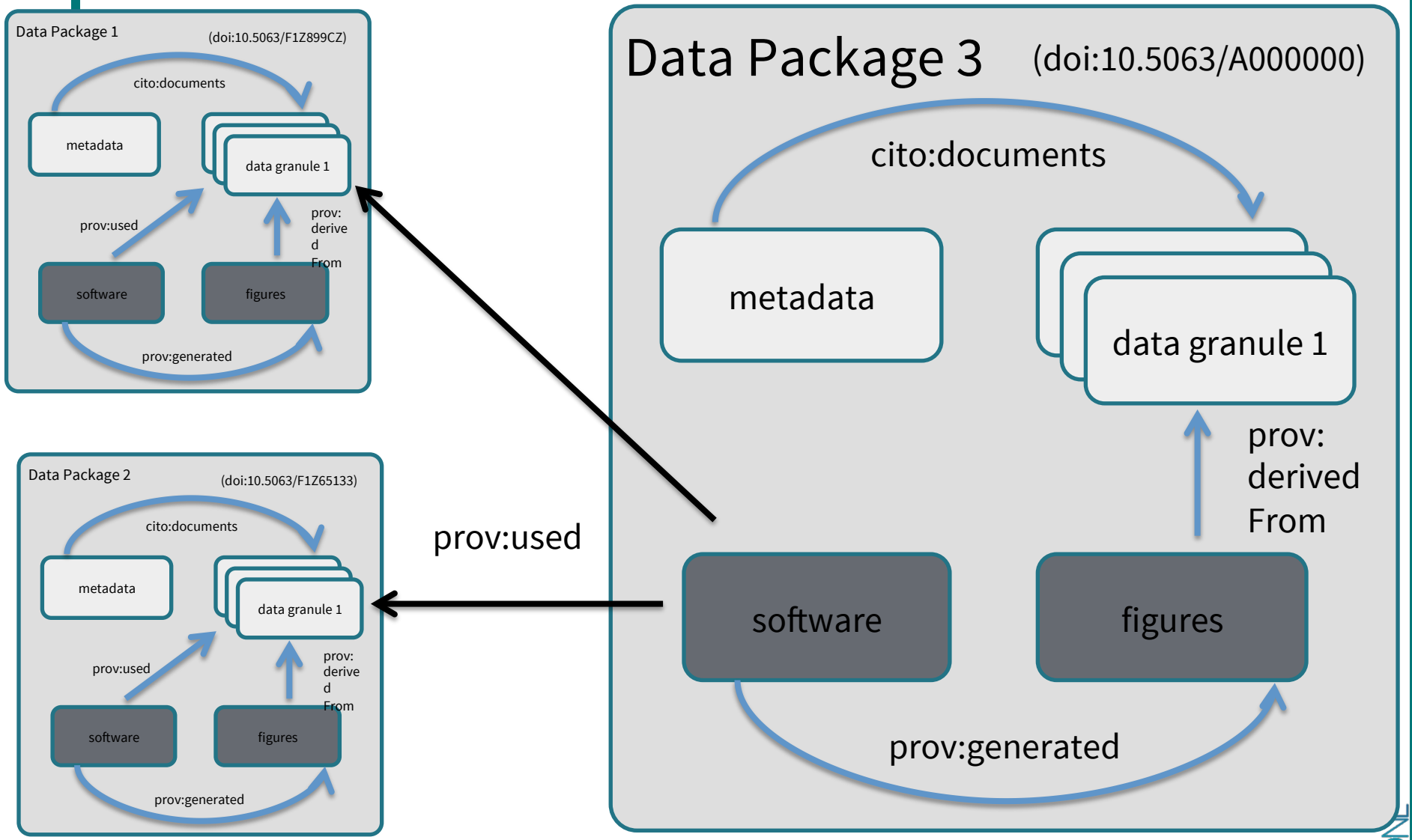


# Provenance relationships

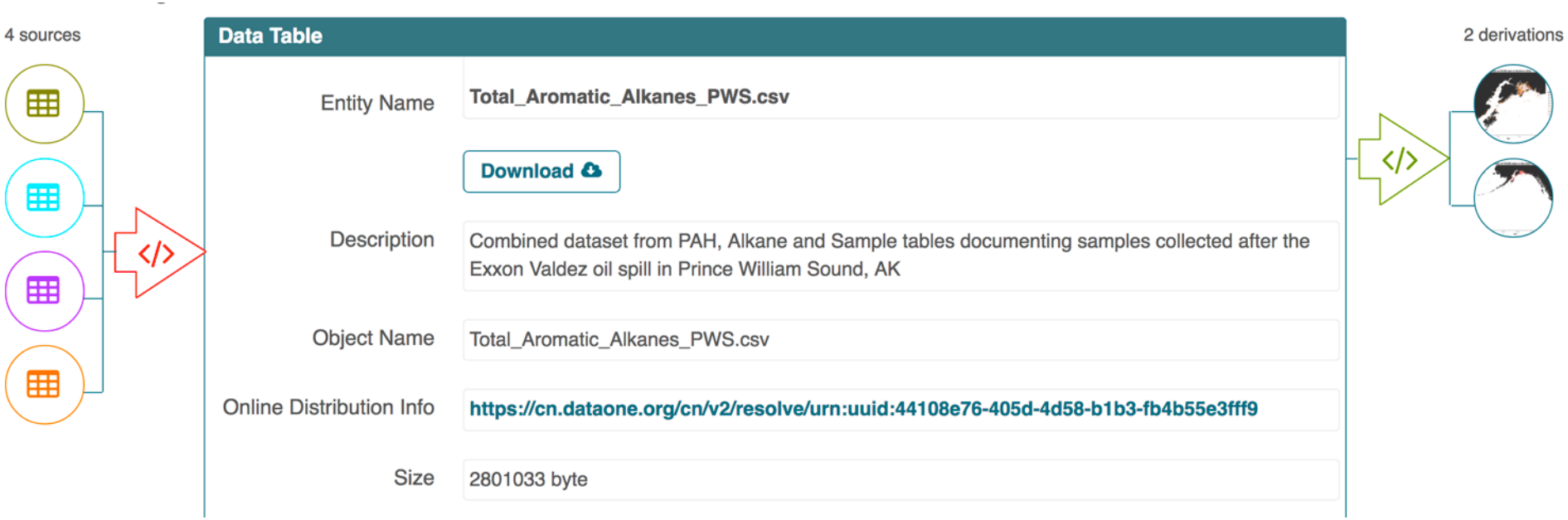
- prov:used
- prov:generated
- prov:derivedFrom



# Provenance across packages



# Provenance in DataONE




- https://search.dataone.org/index.html#view/urn:uuid:3249ada0-afe3-4dd6-875e-0f7928a4c171

# Data Table, Image, and Other Data Details

0 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-2.test.dataone.org/cn/v2/resolve/urn:uuid:7a555441-9efc-4857-89bc-2369ab729b56 |
| Size | 2801033 byte |

Text Format

| | |
|---|---|
| Number of Header Lines | 1 |
| Record Delimiter | #x0A |
| Attribute Orientation | column |
| **Simple Text** | |

0 derivations

Add Add

# Provenance the scripted way

Matlab DataONE Toolbox

Recordr R Library

# Provenance Functions

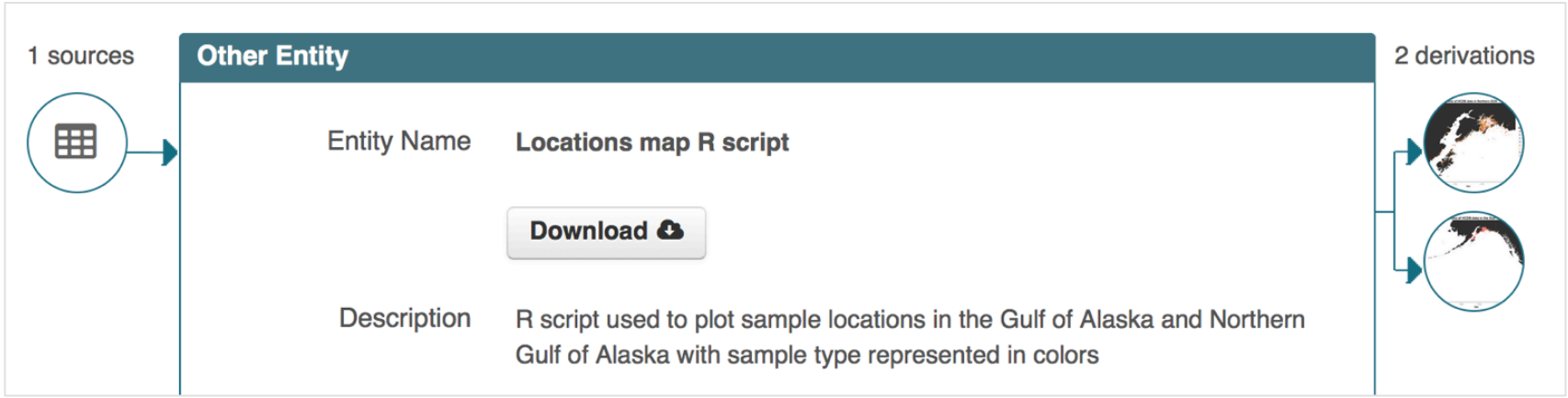
| **describeWorkflow()** |
| --- |
| record() |
| startRecord() |
| endRecord() |
| listRuns() |
| deleteRuns() |
| viewRun() |
| publish() |

| set() |
| --- |
| get() |
| saveConfig() |
| loadConfig() |
| listConfig() |

See: Run Manager API document

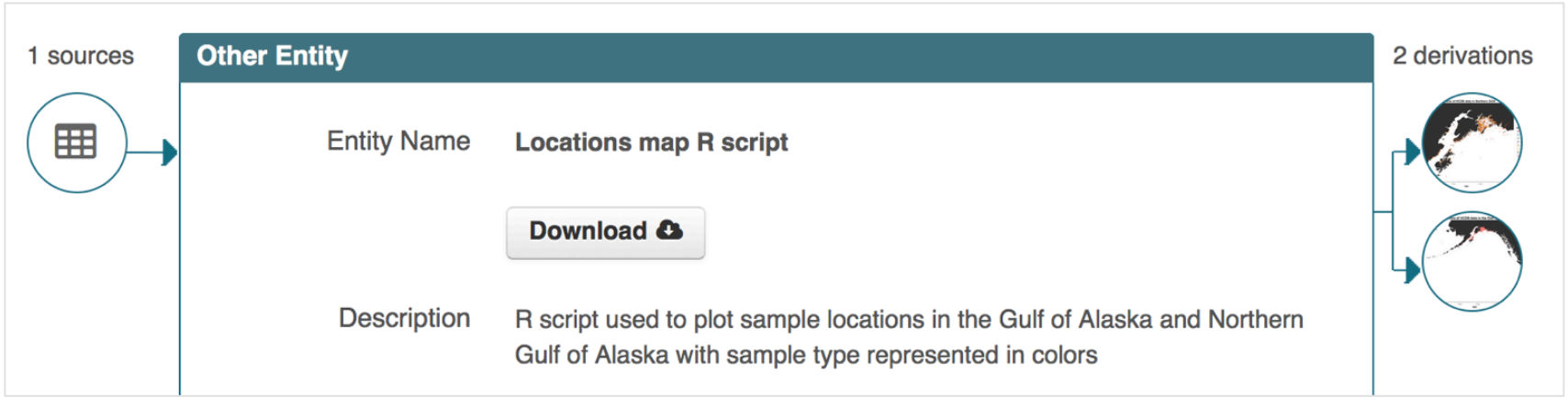
# R: Asserting provenance

```
1  library(datapack)

2  source <- new("DataObject", format="text/csv",
                 filename=”sample.csv"))

3  dp <- addMember(dp, source, metadata)

4  dp <- describeWorkflow(dp, sources=source)
```



# R: recording provenance

```
1  # Generate map of locations by type
2  library(recordr)
3  recordr <- new("Recordr")
4  pkg <- record(recordr, "./hcdbSites.R", "loc-by-type-png")
```



# R: managing script runs

```
> listRuns(recordr)

Script        StartTime   EndTime   Published    Tag                RunID

hcdbSites.R   18:53:09    18:53:09  unpublished  loc-by-type-png    C85A ...
```

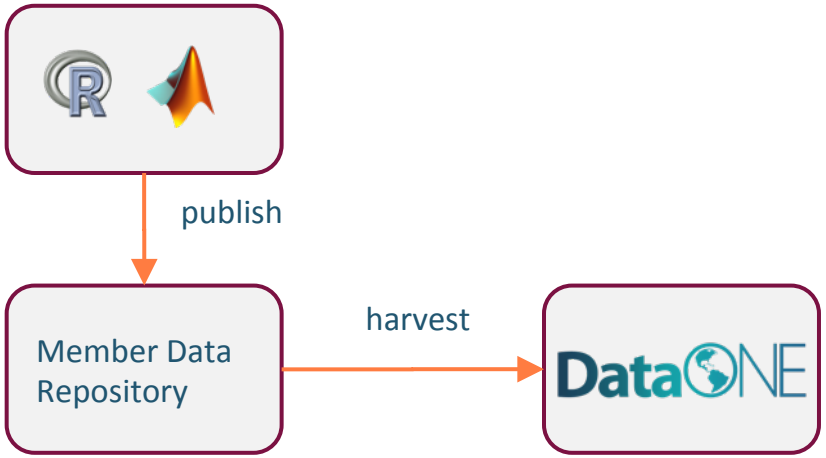
```
> deleteRun(recordr, "loc-by-type-png")

C85A188-B72E-49F1-AEF4-7BFC24DA186B
```

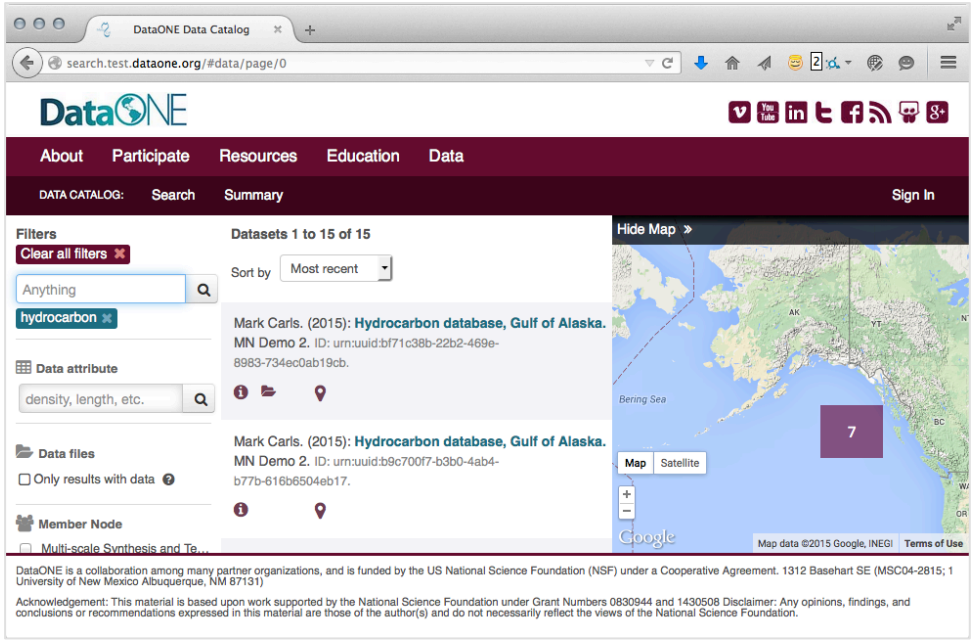
```
> viewRun(recordr, "loc-by-type-png")

… details about the run listed here ...
```

```
> publishRun(recordr, "loc-by-type-png")

C85A188-B72E-49F1-AEF4-7BFC24DA186B
```

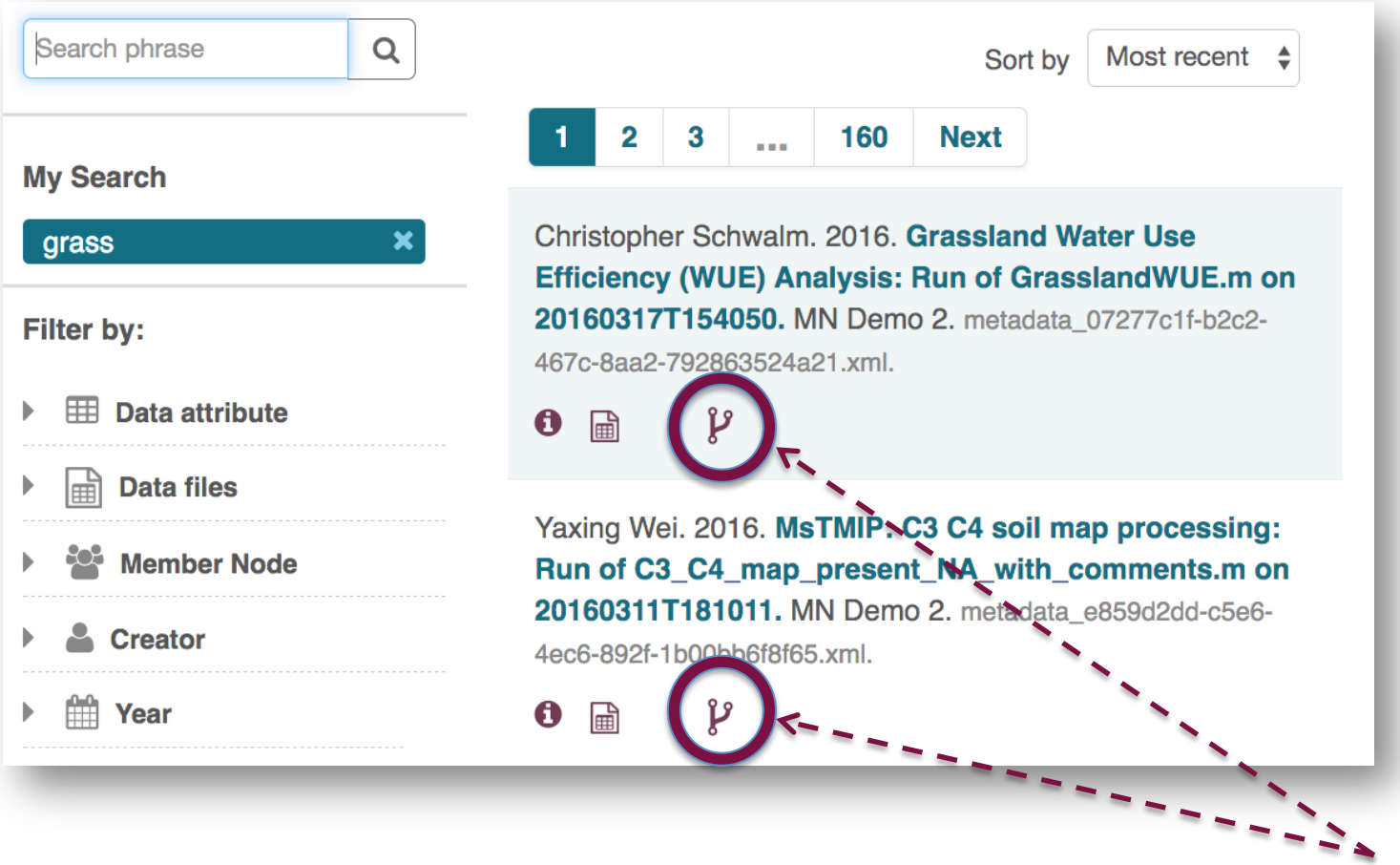
# Provenance Search and Browse

- Harvest and index provenance information



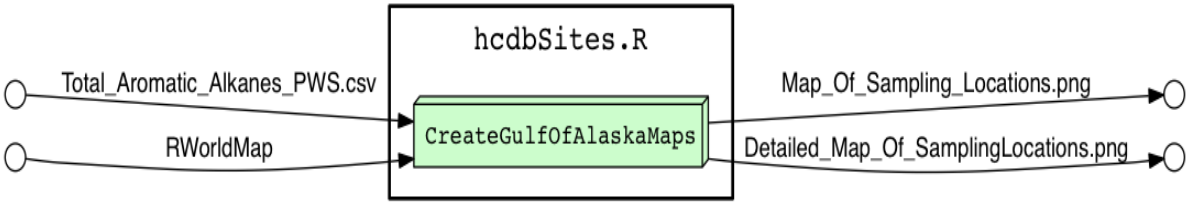
# Provenance in Action: Benefits & Impact



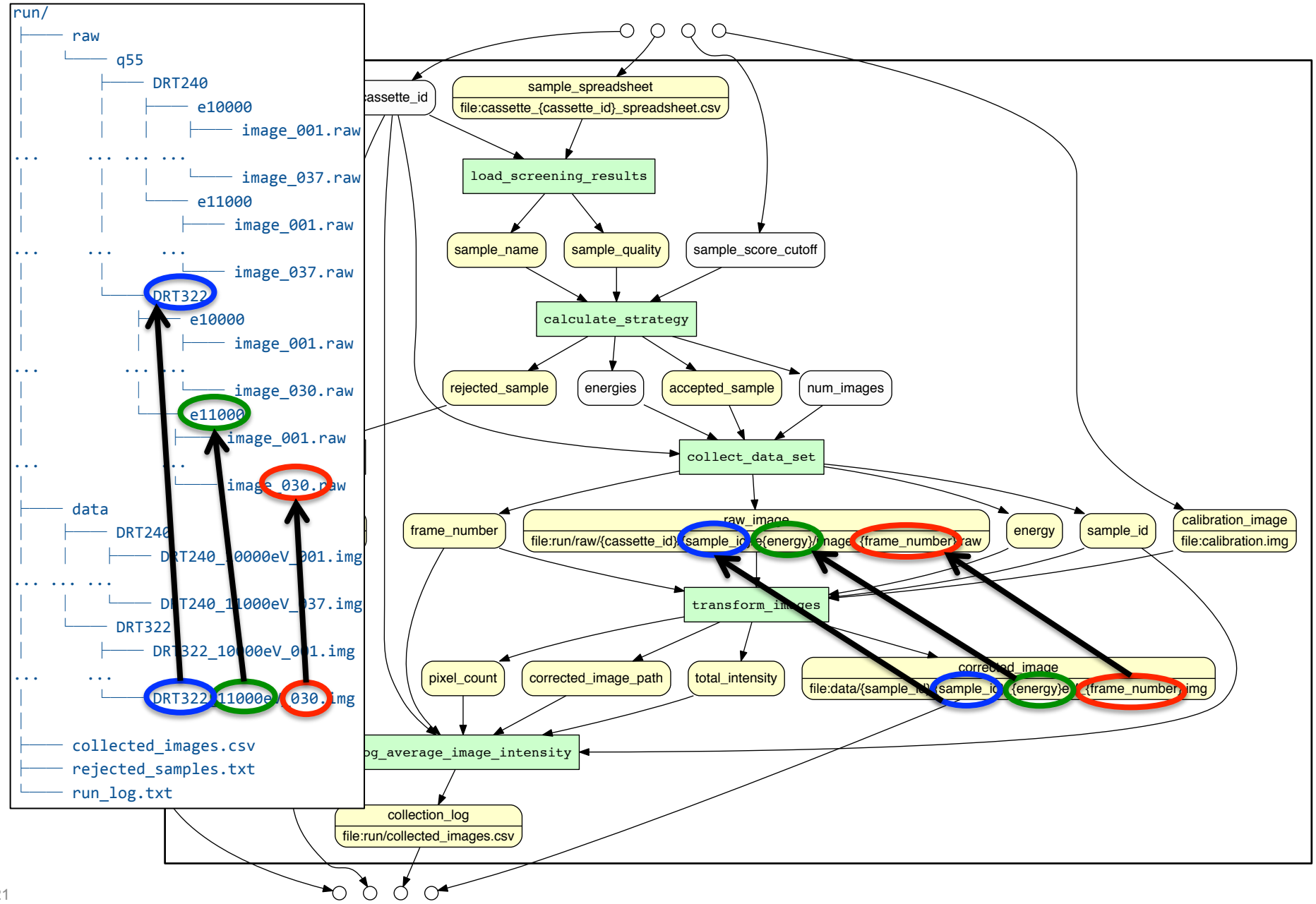

*A DataONE search (here: "grass") yields different packages with provenance*

# YesWorkflow (YW):
# Scripts as *prospective* provenance

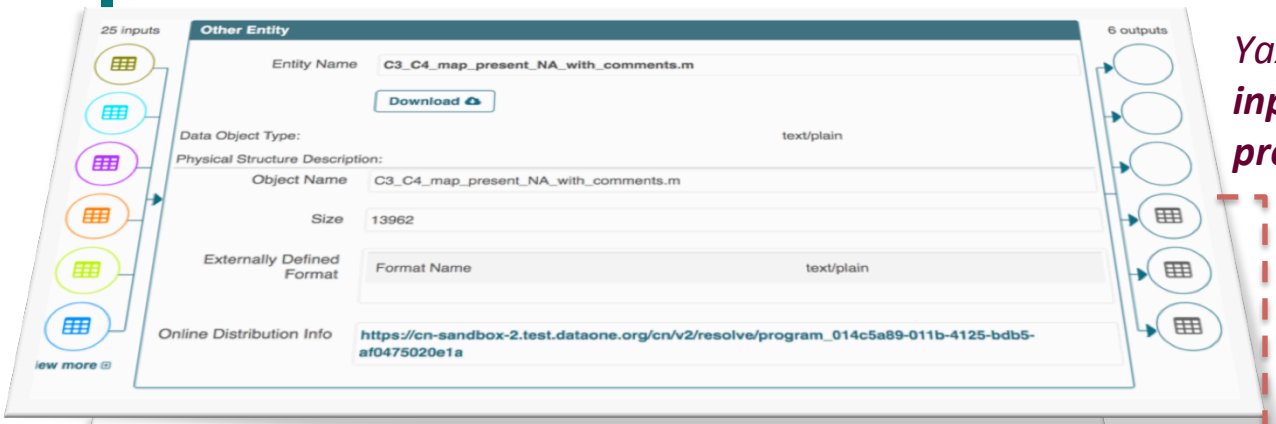
```
1   # @begin CreateGulfOfAlaskaMaps
2   # @in hcdb @as Total_Aromatic_Alkanes_PWS.csv
3   # @in world @as RWorldMap
4   # @out map @as Map_Of_Sampling_Locations.png
5   # @out detailMap @as Detailed_Map_Of_SamplingLocations.png
    ... mapping code is here ...
25  # @end CreateGulfOfAlaskaMaps
```



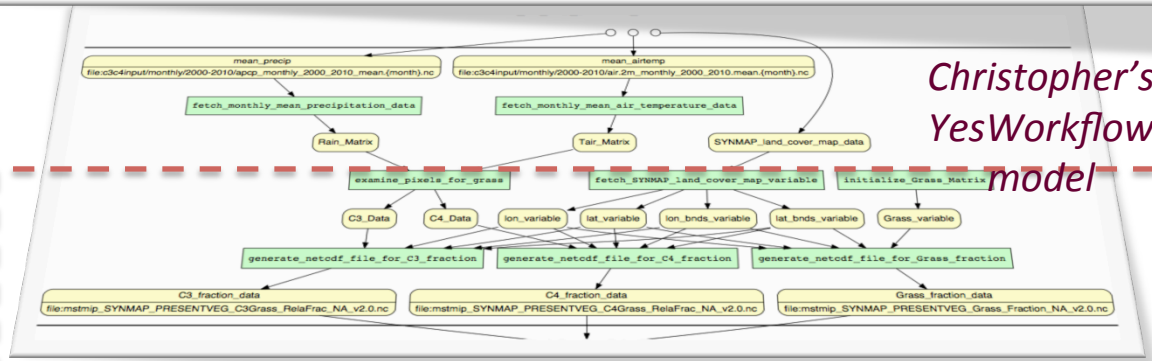
# Where is the raw image of the corrected image?



# Yes Workflow



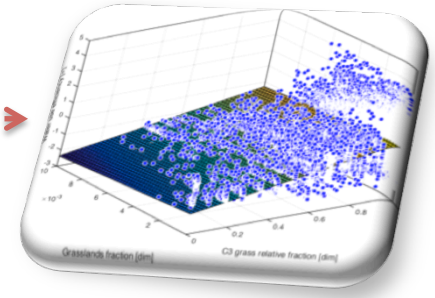

*Yaxing's **script** with **inputs** & **output products***

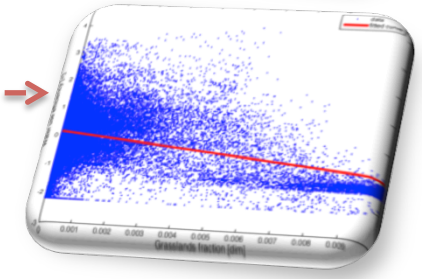



*Christopher's YesWorkflow model*

*Christopher's results can be traced back all the way to Yaxing's input*


*Christopher **using** Yaxing's outputs as inputs for his script*





# 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

Katz & Smith. 2014. Implementing Transitive Credit with JSON-LD. arXiv:1407.5117

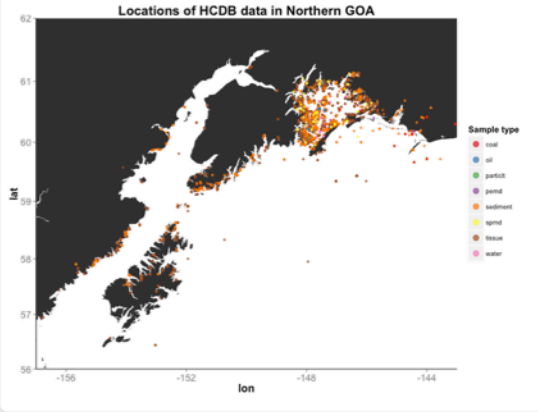
Making Provenance Easy!

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.



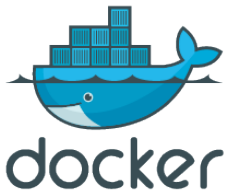

View »

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** .

# What about executable provenance?

- Computational environment
  - Hardware
  - Operating system
  - Installed software versions
- Virtual machines and containers
  - Represent a full machine
  - Lightweight description
  - Fully executable





DataONE

# Reproducible Research

CISE DIBBS

WHOLETALE

## Virtual Data

Girder

DataONE

ownCloud

## Compute

jupyter

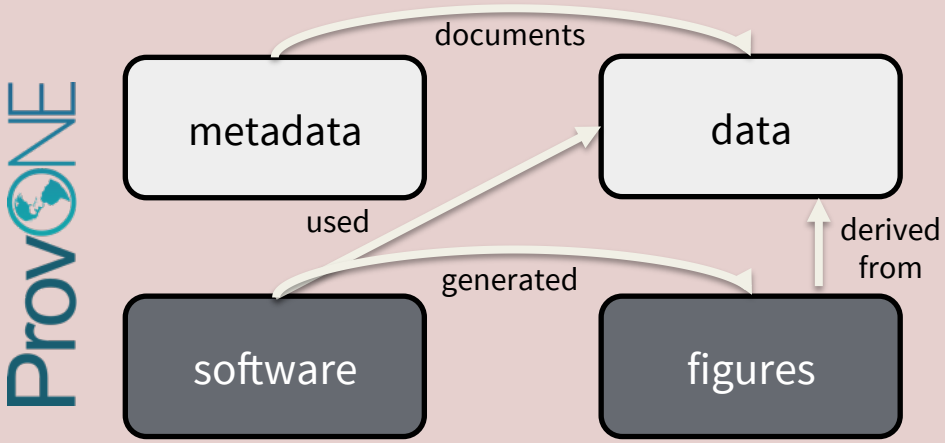
RStudio

docker

NCSA

TACC

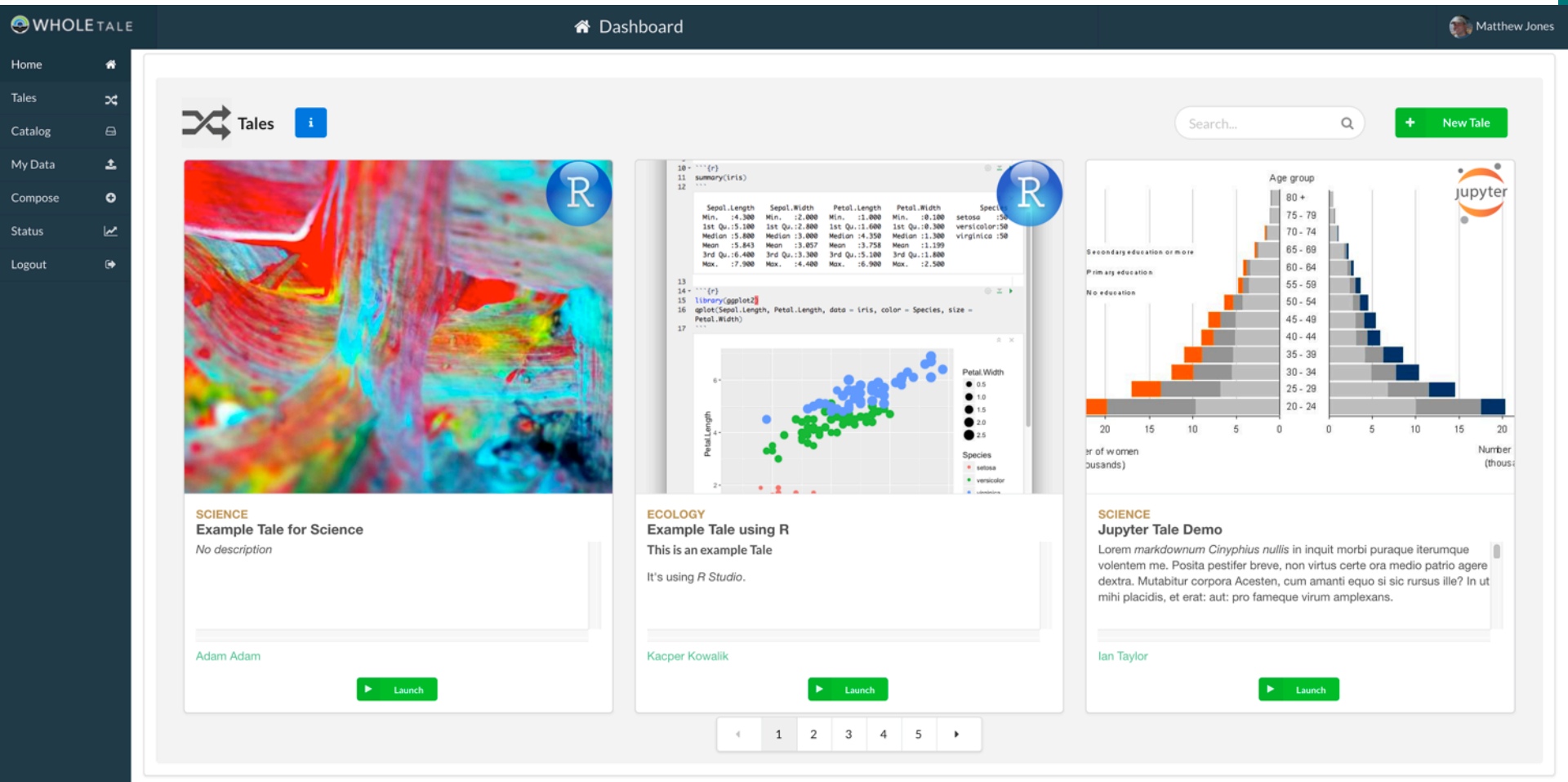
## Publish Tales






DataONE

# Whole Tale Dashboard



# Kick-starting the discussion

- How do we merge these effectively?
  - Detailed ProvONE and YW traces
  - Tales from WholeTale
- How complicated is too complicated?


DataONE

# Acknowledgements


- Funding for DataONE from NSF

Award # 1430508