



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*
 - *Replication and reproducible science*
 - *Discovery of data, methodologies, experiments*
- Provenance continuum
 - Prose
 - Formal provenance traces
 - Fully executable environments

Facilitate Reproducible Science

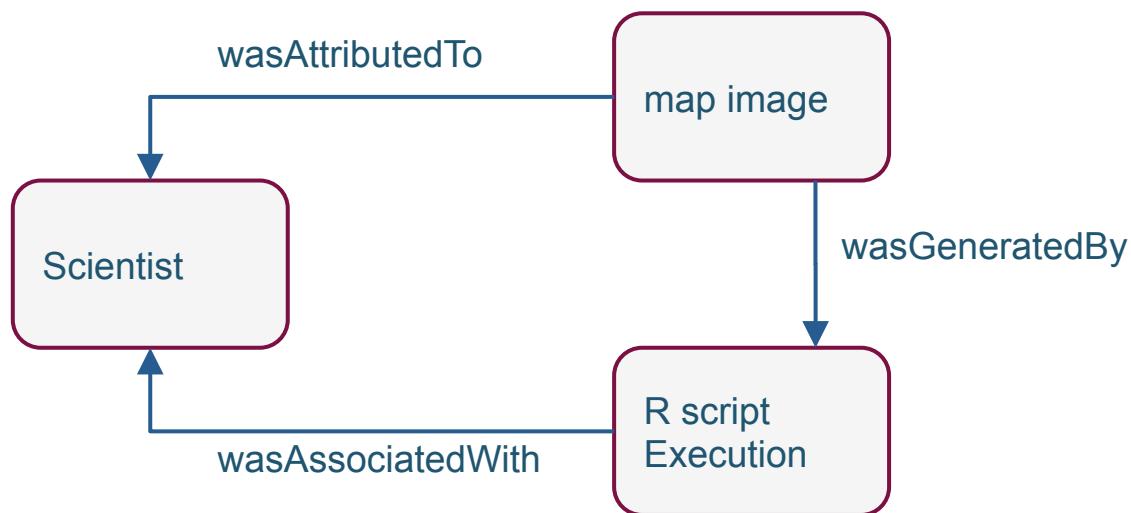
Goal: Use provenance for reproducible science

- 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

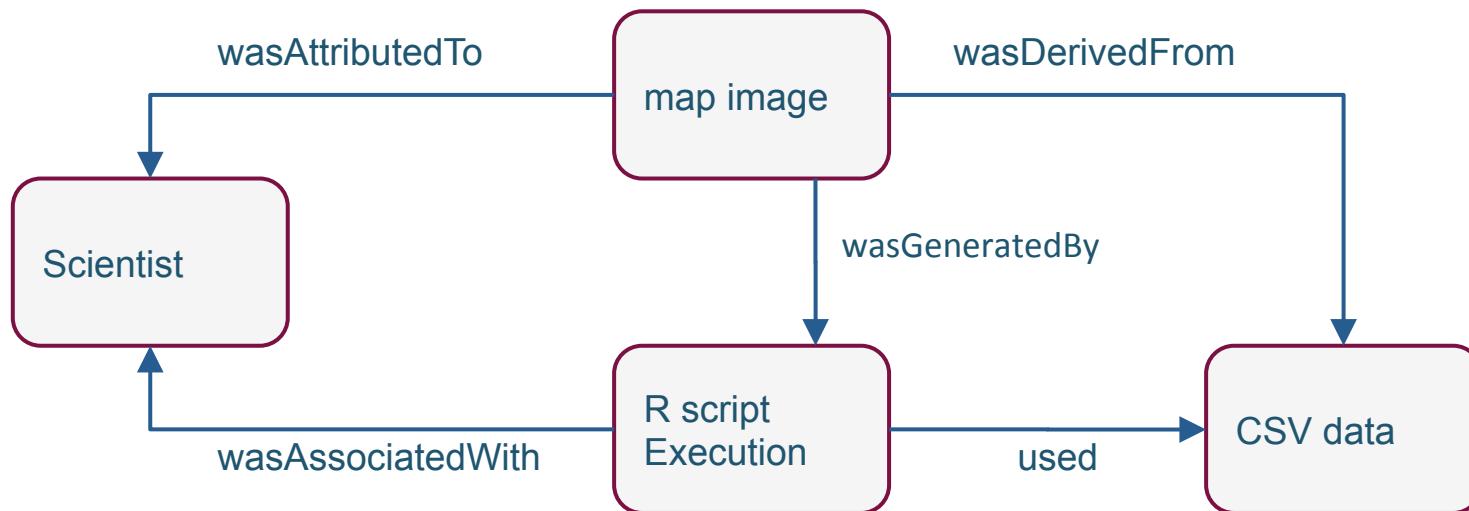
Using a common model

- Example: Map from CSV data



Using a common model

■ Example: Map from CSV data



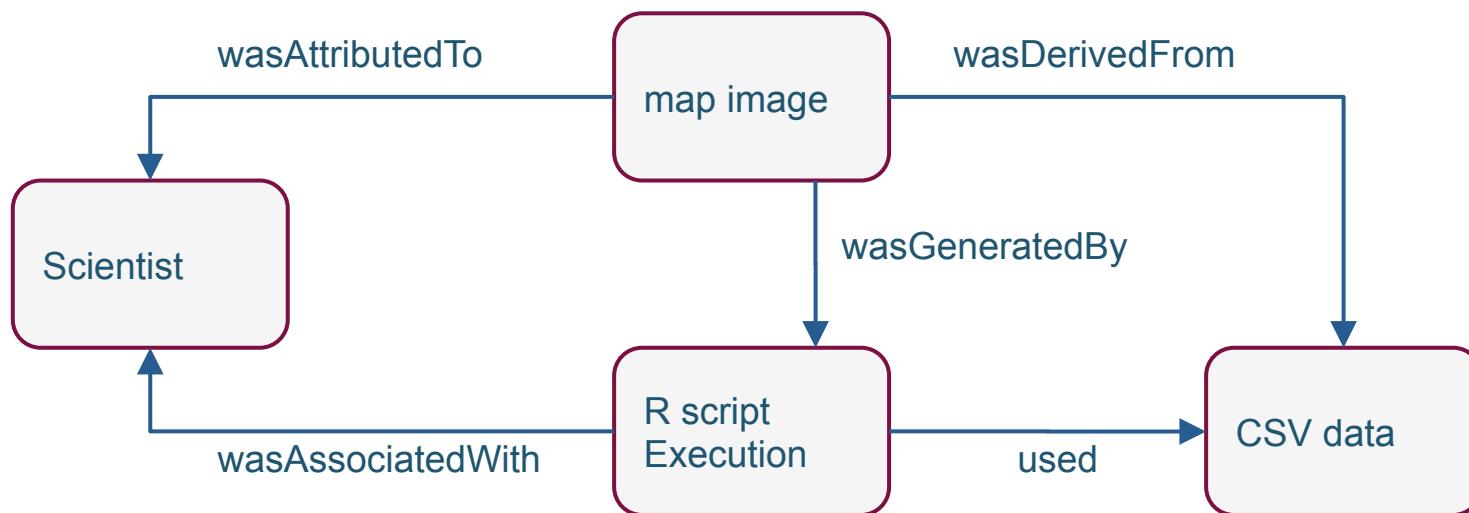
Using a common model

- 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](#)

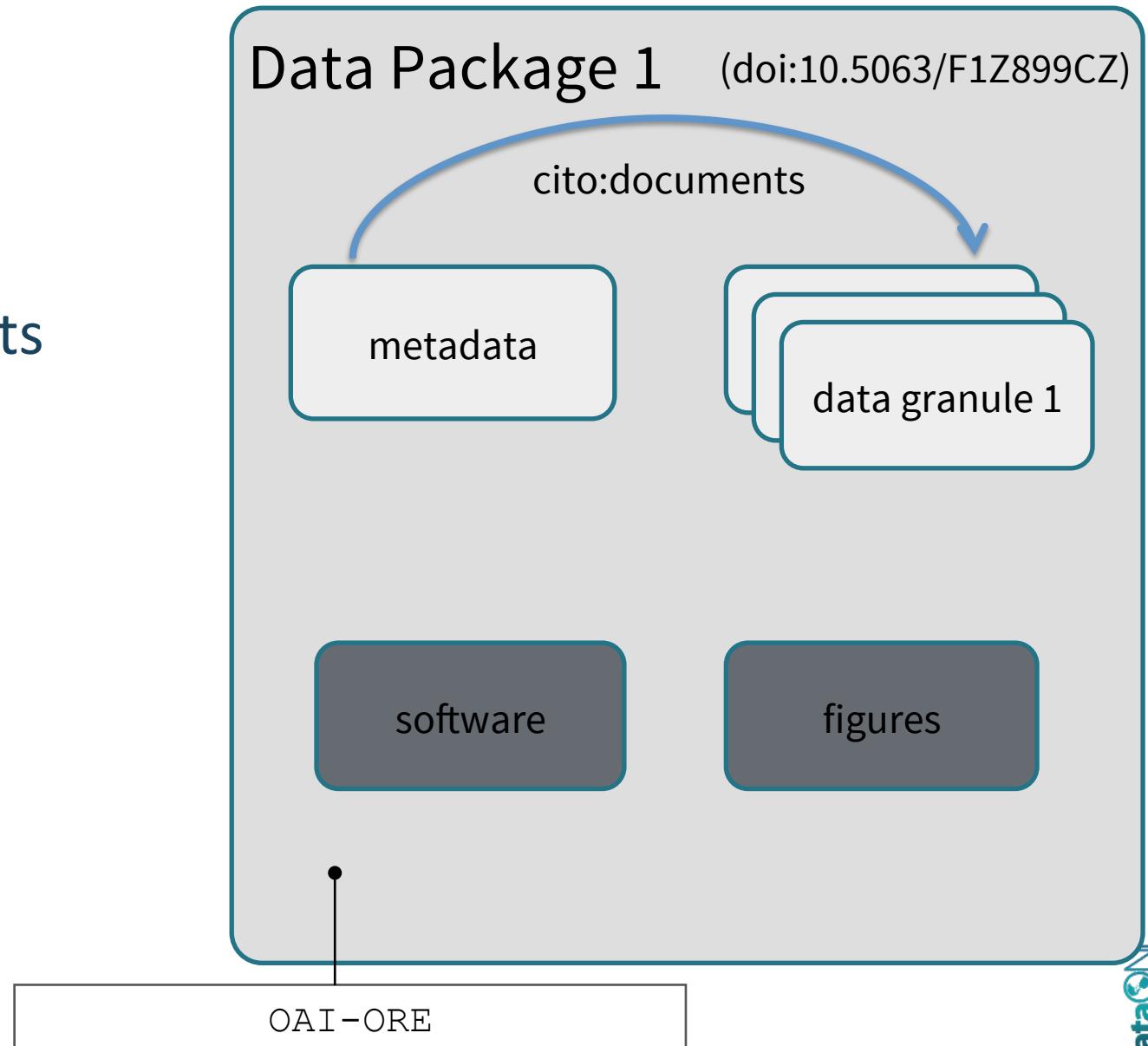
Files in this dataset Package: urn:uuid:1d23e155-3ef5-47c6-9612-027c80855e8d

	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

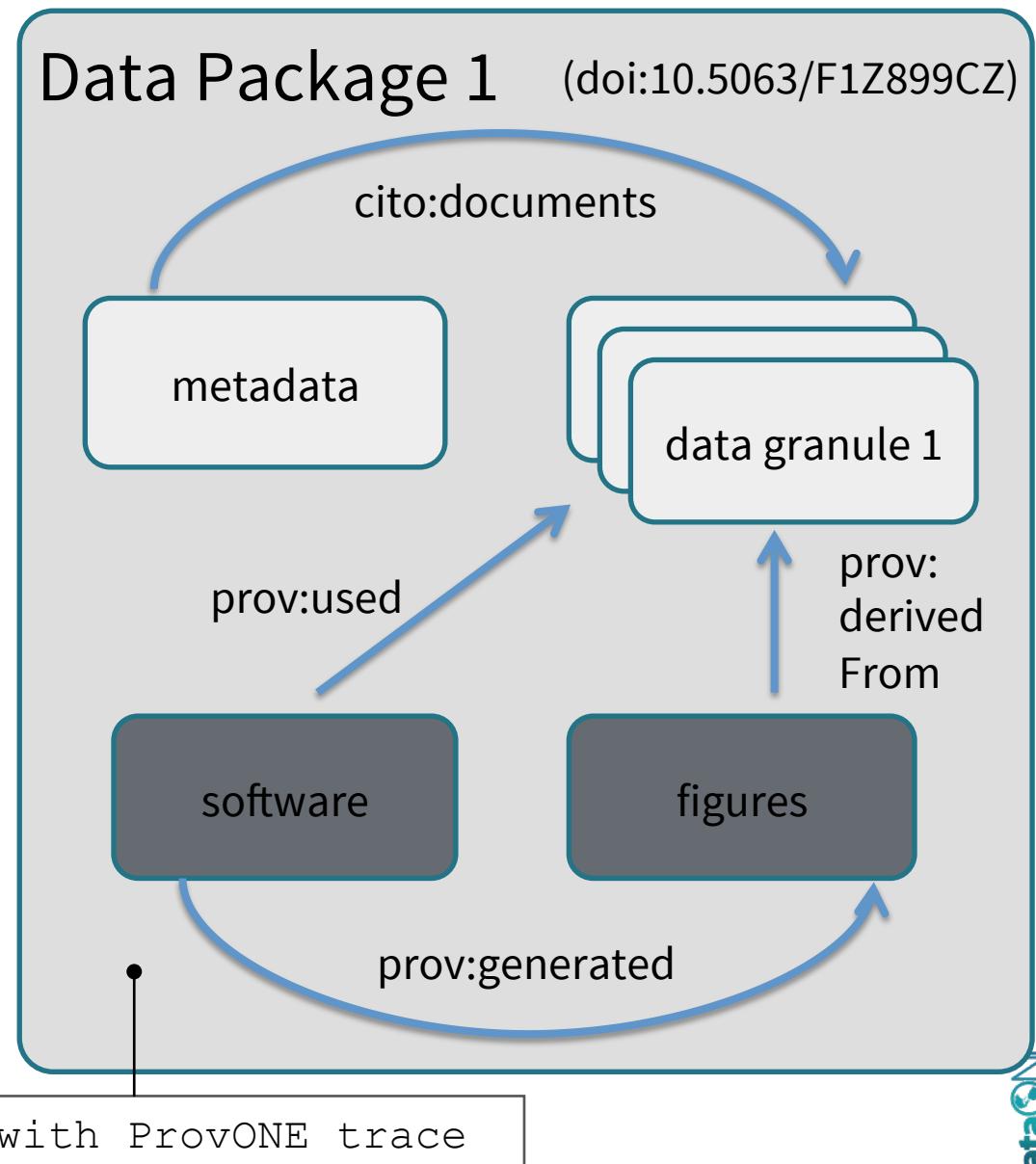


- Metadata
- Data files
- Other products
 - Photos
 - Graphs
 - Movies
 - ...

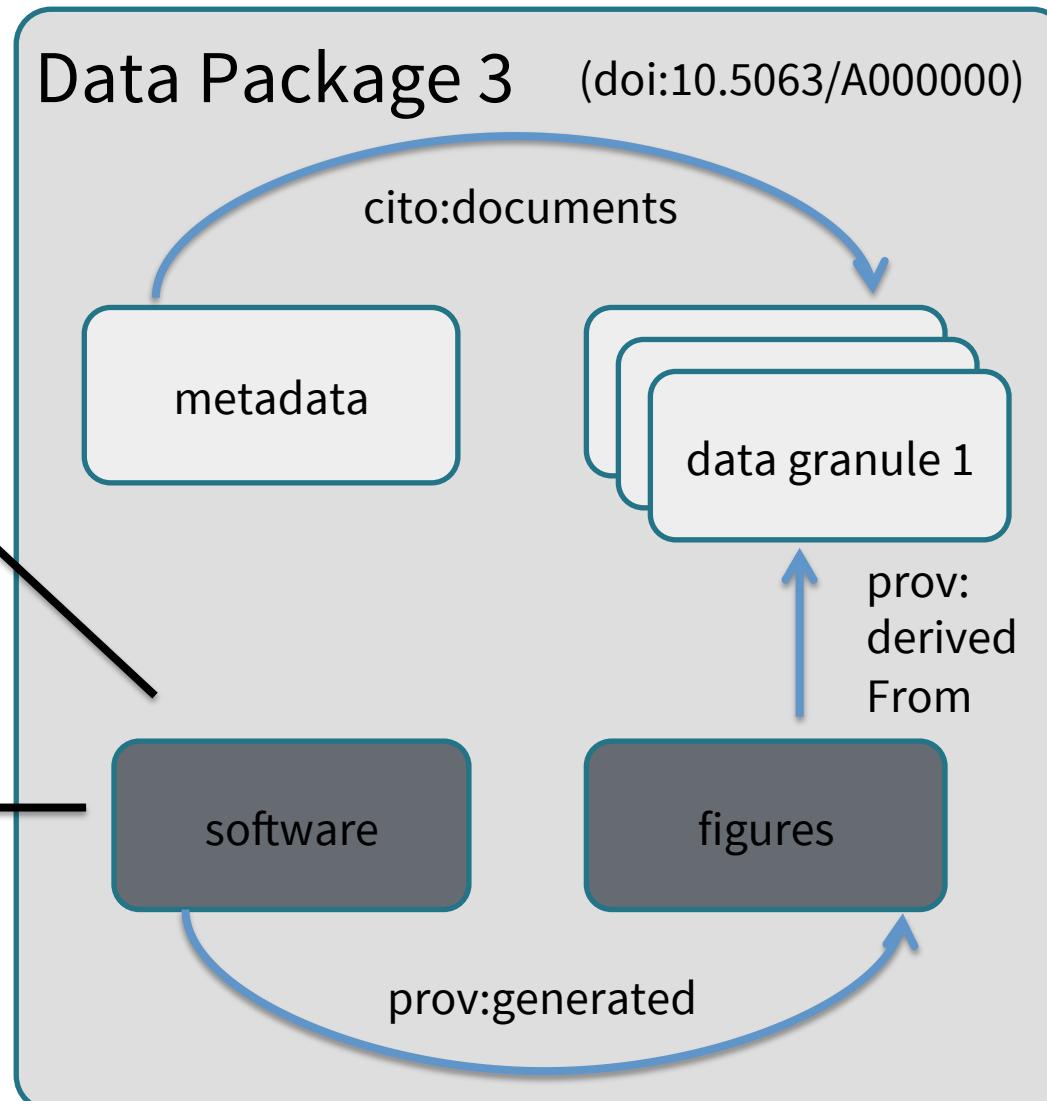
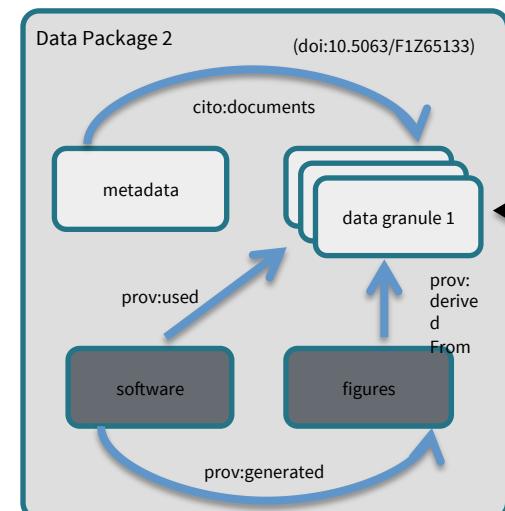
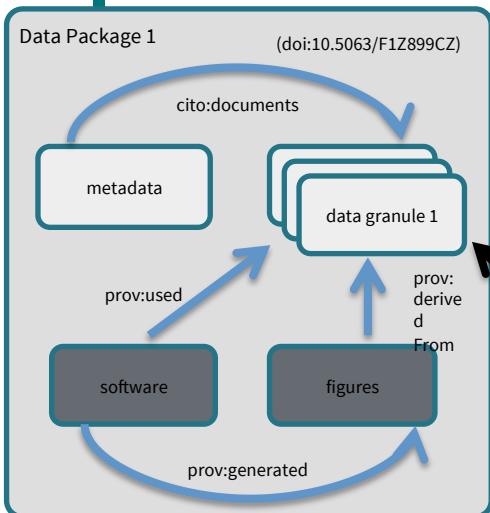


Provenance relationships

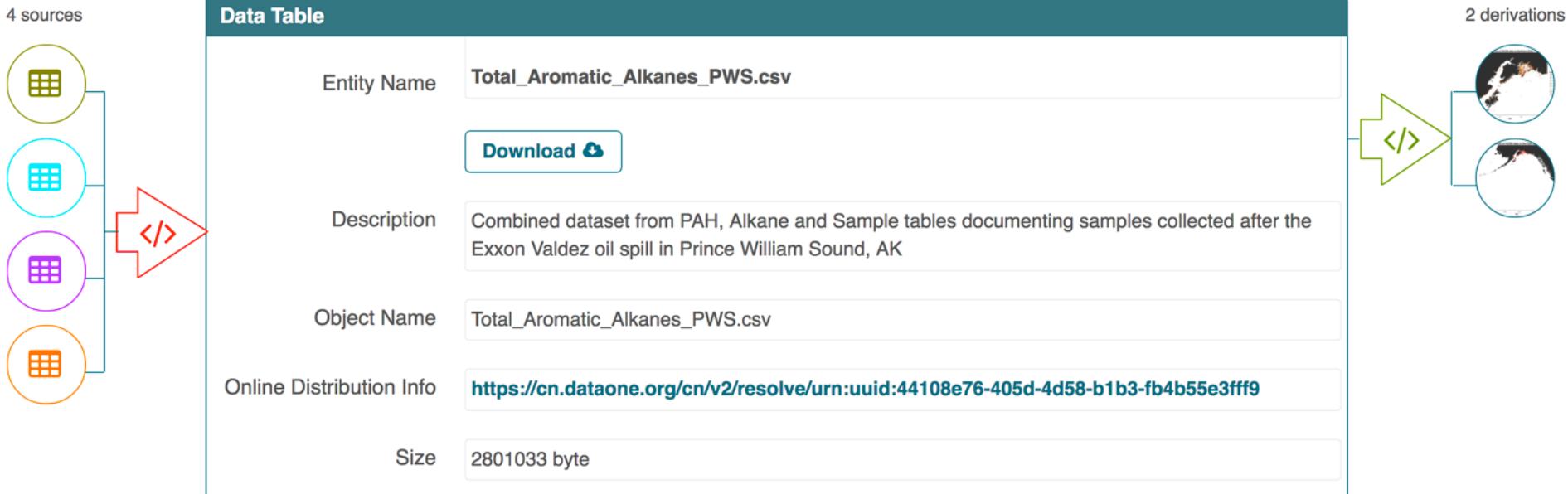
- prov:used
- prov:generated
- prov:derivedFrom



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



0 derivations



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	<table><tbody><tr><td>Number of Header Lines</td><td>1</td></tr><tr><td>Record Delimiter</td><td>#xA</td></tr><tr><td>Attribute Orientation</td><td>column</td></tr><tr><td colspan="2">Simple Text</td></tr></tbody></table>	Number of Header Lines	1	Record Delimiter	#xA	Attribute Orientation	column	Simple Text	
Number of Header Lines	1								
Record Delimiter	#xA								
Attribute Orientation	column								
Simple Text									

Provenance the scripted way



Matlab DataONE Toolbox



Recordr R Library



Provenance Functions

describeWorkflow()

record()

startRecord()

endRecord()

listRuns()

deleteRuns()

viewRun()

publish()

set()

get()

saveConfig()

loadConfig()

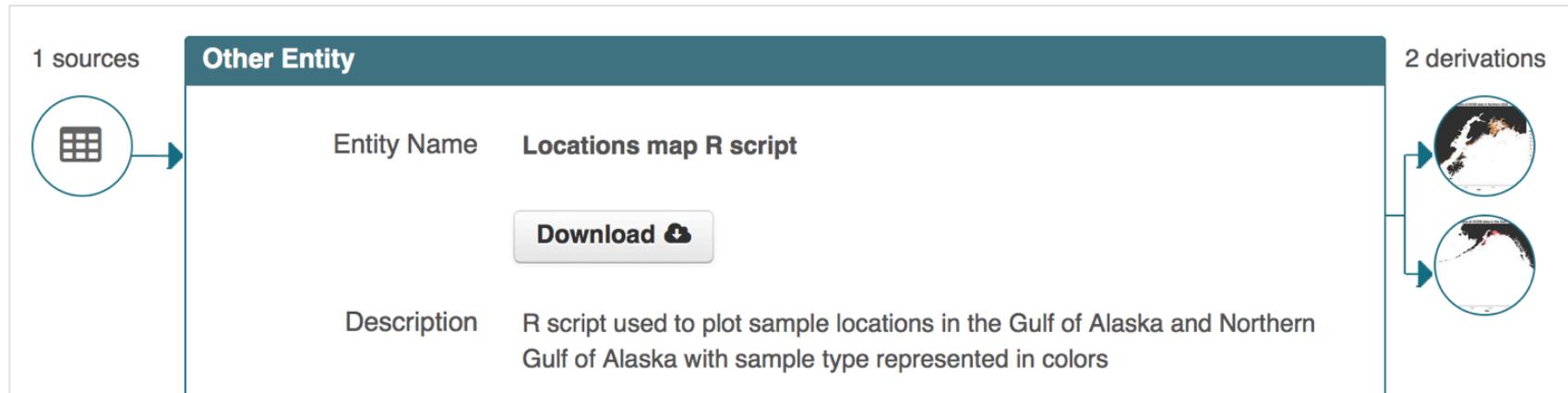
listConfig()

See: [Run Manager API](#)
document



R: Asserting provenance

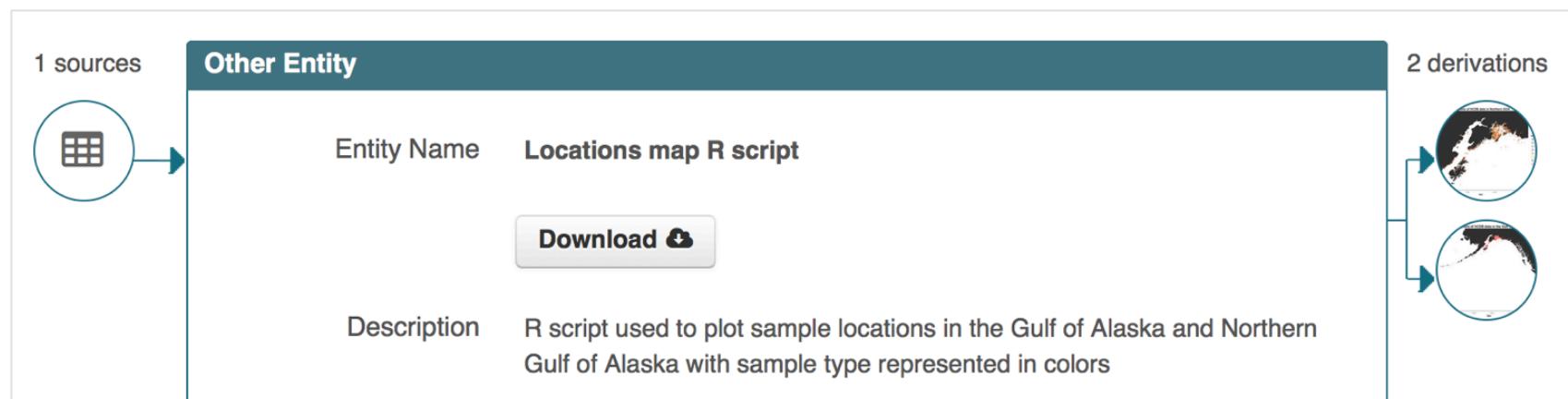
```
1 library(datapack)  
  
2 source <- new("DataObject", format="text/csv",  
2                 filename="sample.csv"))  
  
3 dp <- addMember(dp, source, metadata)  
  
4 dp <- describeWorkflow(dp, sources=source)
```





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")
```

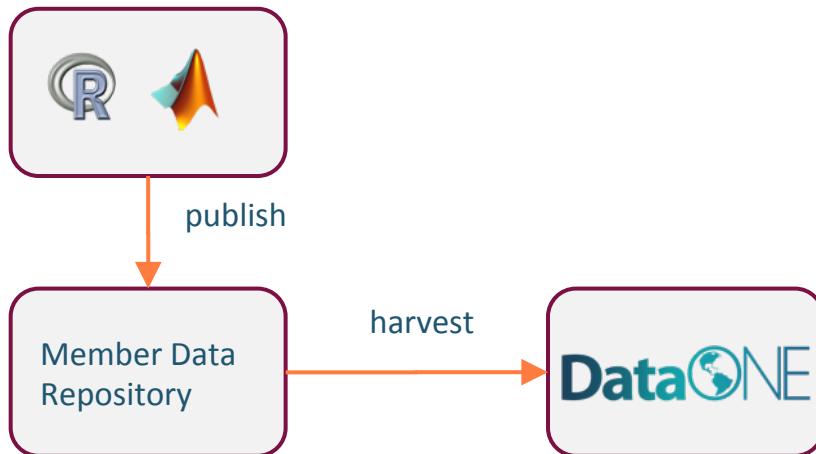
... details about the run listed here ...

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

C85A188-B72E-49F1-AEF4-7BFC24DA186B

Provenance Search and Browse

- Harvest and index provenance information



A screenshot of the DataONE Data Catalog interface. The top navigation bar includes links for About, Participate, Resources, Education, Data, and a Sign In button. Below the navigation is a search bar with the query "hydrocarbon" and a dropdown menu set to "Most recent". The main content area displays search results for datasets, with the first result being "Mark Carls. (2015): Hydrocarbon database, Gulf of Alaska. MN Demo 2, ID: urn:uuid:b71c38b-22b2-469e-8983-734ec0ab19cb." A map of the Gulf of Alaska is shown on the right, with a callout box highlighting the Bering Sea area. The bottom of the page contains footer text about DataONE's funding and disclaimer.

Filters
Clear all filters

Anything

hydrocarbon

Sort by Most recent

Datasets 1 to 15 of 15

Mark Carls. (2015): [Hydrocarbon database, Gulf of Alaska.](#) MN Demo 2, ID: [urn:uuid:b71c38b-22b2-469e-8983-734ec0ab19cb.](#)

Mark Carls. (2015): [Hydrocarbon database, Gulf of Alaska.](#) MN Demo 2, ID: [urn:uuid:b9c700f7-b3b0-4ab4-b77b-616b6504eb17.](#)

Data attribute

density, length, etc.

Data files Only results with data

Member Node Multi-scale Synthesis and Te...

Map Satellite Google Map data ©2015 Google, INEGI. Terms of Use

DataONE is a collaboration among many partner organizations, and is funded by the US National Science Foundation (NSF) under a Cooperative Agreement. 1312 Basehart SE (MSC04-2815; 1 University of New Mexico Albuquerque, NM 87131)
Acknowledgement: This material is based upon work supported by the National Science Foundation under Grant Numbers 0830944 and 1430508 Disclaimer: Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Provenance in Action: Benefits & Impact

The screenshot shows a search interface for "grass". The search bar contains "grass" and a magnifying glass icon. To the right are sorting options ("Sort by" dropdown set to "Most recent") and a page navigation bar with buttons for 1, 2, 3, ..., 160, and "Next".

My Search

Filter by:

- ▶ Data attribute
- ▶ Data files
- ▶ Member Node
- ▶ Creator
- ▶ Year

Search Results:

Christopher Schwalm. 2016. **Grassland Water Use Efficiency (WUE) Analysis: Run of GrasslandWUE.m on 20160317T154050.** MN Demo 2. metadata_07277c1f-b2c2-467c-8aa2-792863524a21.xml.

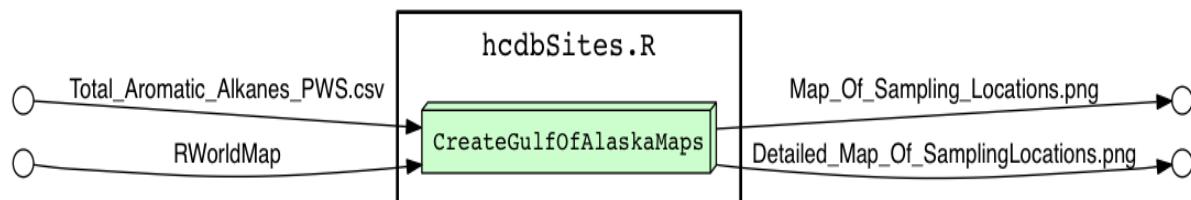
Yaxing Wei. 2016. **MsTMIP_C3 C4 soil map processing: Run of C3_C4_map_present_NA_with_comments.m on 20160311T181011.** MN Demo 2. metadata_e859d2dd-c5e6-4ec6-892f-1b00bb6f8f65.xml.

Two entries are shown, each with a detailed description and two circular icons with arrows pointing to them. The first entry has a purple circle around the "i" icon (info) and the second entry has a purple circle around the "i" icon. Dashed arrows point from these circles towards the bottom right corner of the slide, indicating the provenance of the data packages.

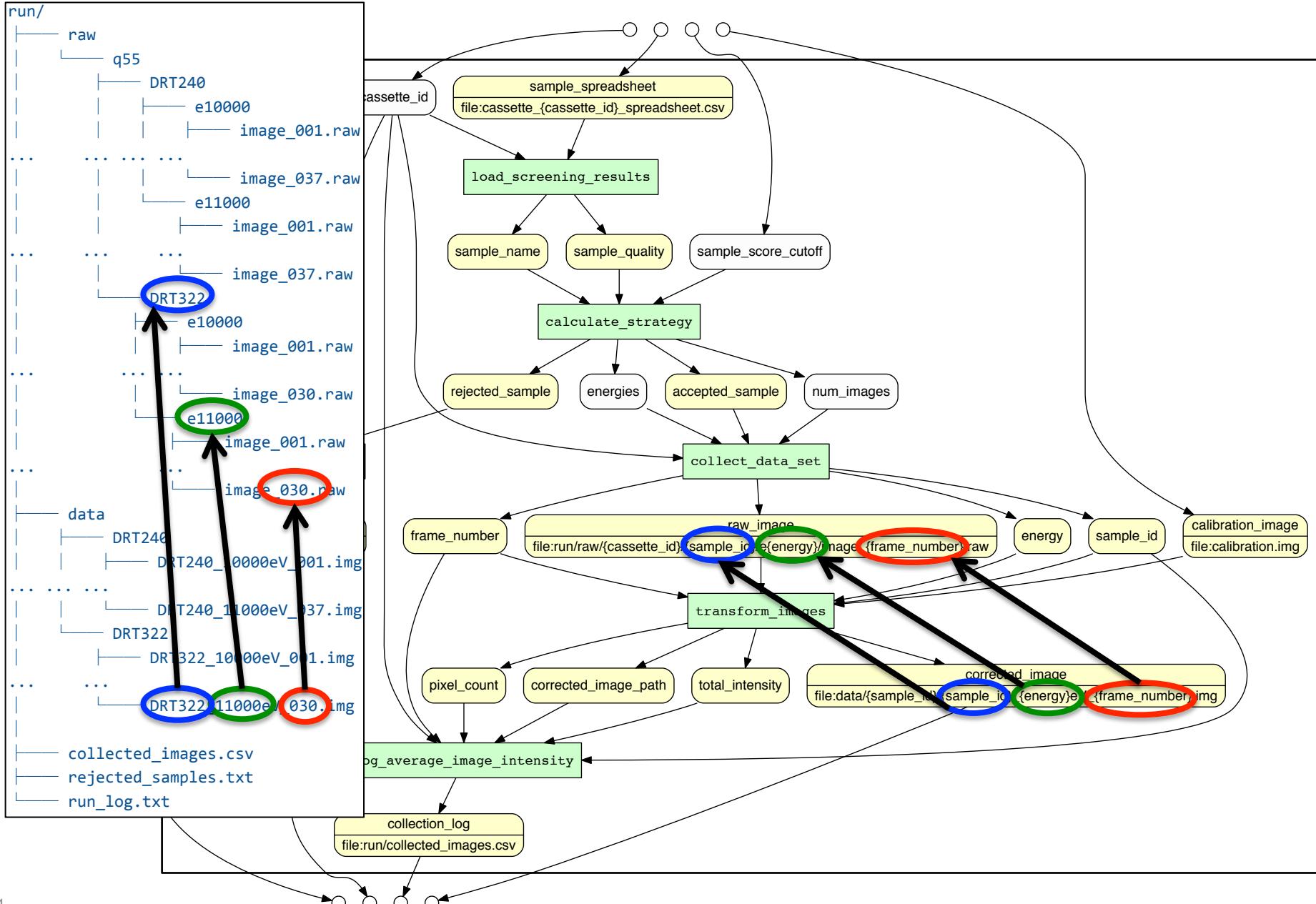
A DataONE search (here: “grass”) yields different packages with provenance

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

25 inputs

Other Entity

Entity Name: C3_C4_map_present_NA_with_comments.m

Data Object Type: text/plain

Physical Structure Description:

Object Name: C3_C4_map_present_NA_with_comments.m

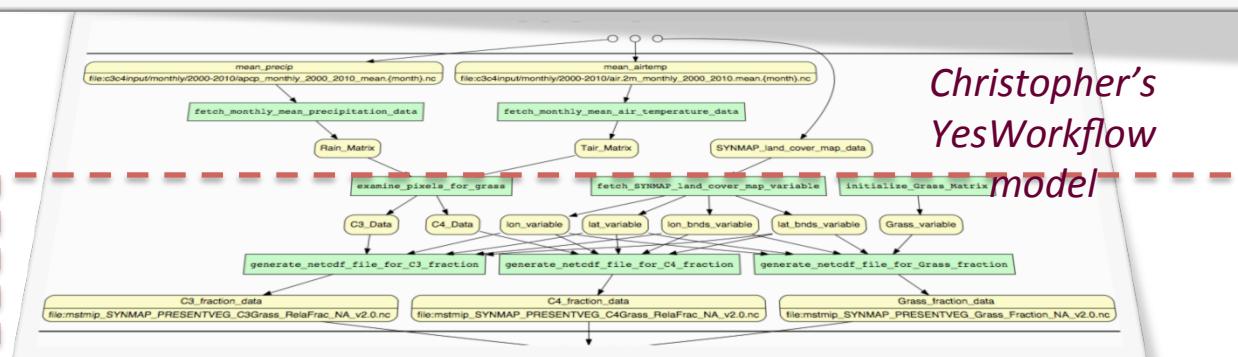
Size: 13962

Externally Defined Format: Format Name text/plain

Online Distribution Info: https://cn-sandbox-2.test.dataone.org/cn/v2/resolve/program_014c5a89-011b-4125-bdb5-a0475020e1a

(View more)

Yaxing's script with inputs & output products



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

8 inputs

Other Entity

Entity Name: GrasslandWUE.m

Data Object Type: text/plain

Physical Structure Description:

Object Name: GrasslandWUE.m

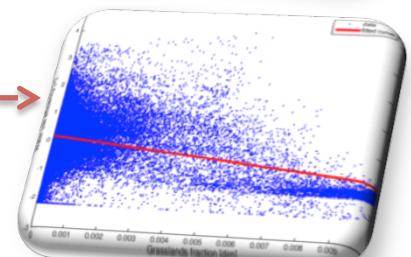
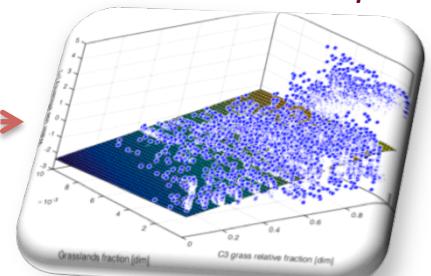
Size: 4443

Externally Defined Format: Format Name text/plain

Online Distribution Info: https://cn-sandbox-2.test.dataone.org/cn/v2/resolve/program_92511b50-c6b2-4949-9ee2-b176a46bd913

(View more)

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

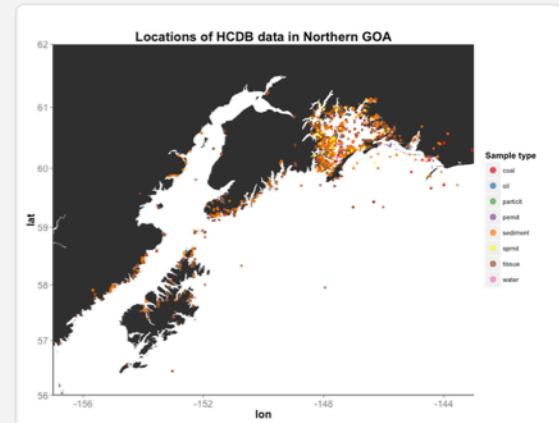
Making Provenance Easy!

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



docker



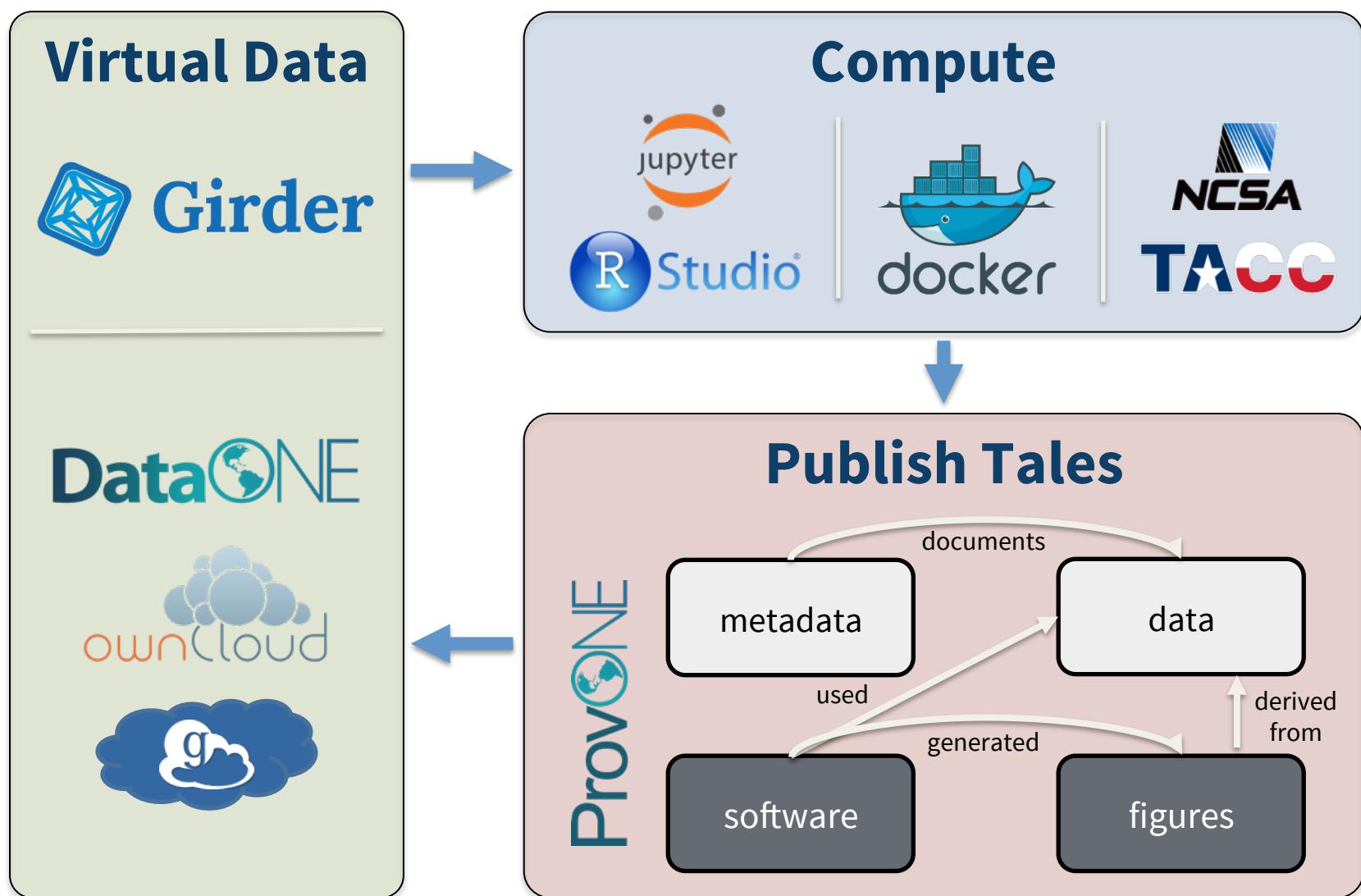
Reproducible Research



CISE DIBBS



WHOLETALE



Whole Tale Dashboard

WHOLE TALE

Dashboard

Matthew Jones

Home

Tales

Catalog

My Data

Compose

Status

Logout

Tales

i

Search...

New Tale

R

jupyter

SCIENCE Example Tale for Science
No description

Adam Adam

Launch

ECOLOGY Example Tale using R
This is an example Tale
It's using R Studio.

Kacper Kowalik

Launch

10 - ... (r)
11 summary(iris)
12 ...

Sepal.Length Sepal.Width Petal.Length Petal.Width Specie
Min. :4.300 Min. :1.000 Min. :1.000 Min. :0.100 setosa :50
1st Qu.:5.000 1st Qu.:1.280 1st Qu.:1.600 1st Qu.:0.800 versicolor :50
Median :5.800 Median :1.300 Median :4.350 Median :1.300 virginica :50
Mean :5.843 Mean :1.357 Mean :4.375 Mean :1.590
3rd Qu.:6.400 3rd Qu.:1.500 3rd Qu.:5.100 3rd Qu.:1.800
Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500

13 - ... (r)
14 - ... (r)
15 library(ggplot2)
16 ggplot(Sepal.Length, Petal.Length, data = iris, color = Species, size = Petal.Width)
17 ...

Petal.Length

Petal.Width

Species

Setosa
Versicolor
Virginica

Age group

Secondary education or more

Primary education

No education

er of women (thousands)

Number (thous.)

20 15 10 5 0 5 10 15 20

80 +
75 - 79
70 - 74
65 - 69
60 - 64
55 - 59
50 - 54
45 - 49
40 - 44
35 - 39
30 - 34
25 - 29
20 - 24

SCIENCE Jupyter Tale Demo
Lorem markdownum Cinyphius nullis inquit morbi puraque iterunque volentem me. Posita pestifera breve, non virtus certe ora medio patrio agere dextra. Mutabitur corpora Acesten, cum amanti equo si sic rursus ille? In ut mihi placidis, et erat: aut: pro fameque virum amplexans.

Ian Taylor

Launch

1 2 3 4 5

Kick-starting the discussion

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

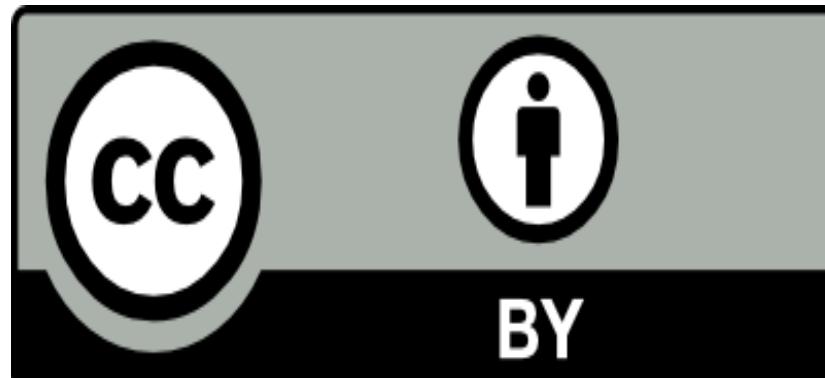


Acknowledgements



- Funding for DataONE from NSF

Award # 1430508



This presentation is made available under a
CC-BY 4.0 license.

<http://creativecommons.org/licenses/by/4.0/>