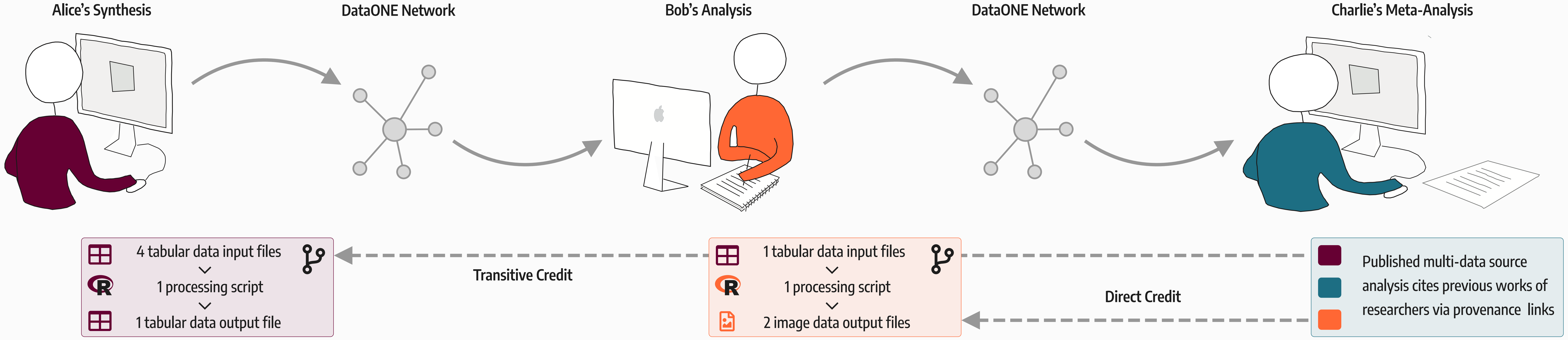
# Data re-use: Tools for producing and displaying data provenance across DataONE repositories

Christopher Jones[1], Yang Cao[2], Matthew B. Jones[1], Ben Leinfelder[1], Bertram Ludaescher[2], Paolo Missier[3], Peter Slaughter[1], Dave Vieglais[4], Lauren Walker[1]

## Building Trust in Sharing Data

In the process of synthesizing datasets to address pressing scientific questions, researchers make decisions that **must increase trust and decrease risk when incorporating data that they did not collect firsthand** into their analyses. With consistently larger amounts of data being made available on the web, metadata about those data are increasingly important for sharing, reusing, and reproducing scientific analyses. And with new synthetic data products being used to make management and policy decisions that affect us all, **understanding the lineage of these products**, including **the computer code** used to process input datasets, is the **essence of building that trust**. Providing appropriate **attribution to previous scientists** in the processing chain is also critical to building that trust.



4 tabular data input files
1 processing script
1 tabular data output file

Transitive Credit

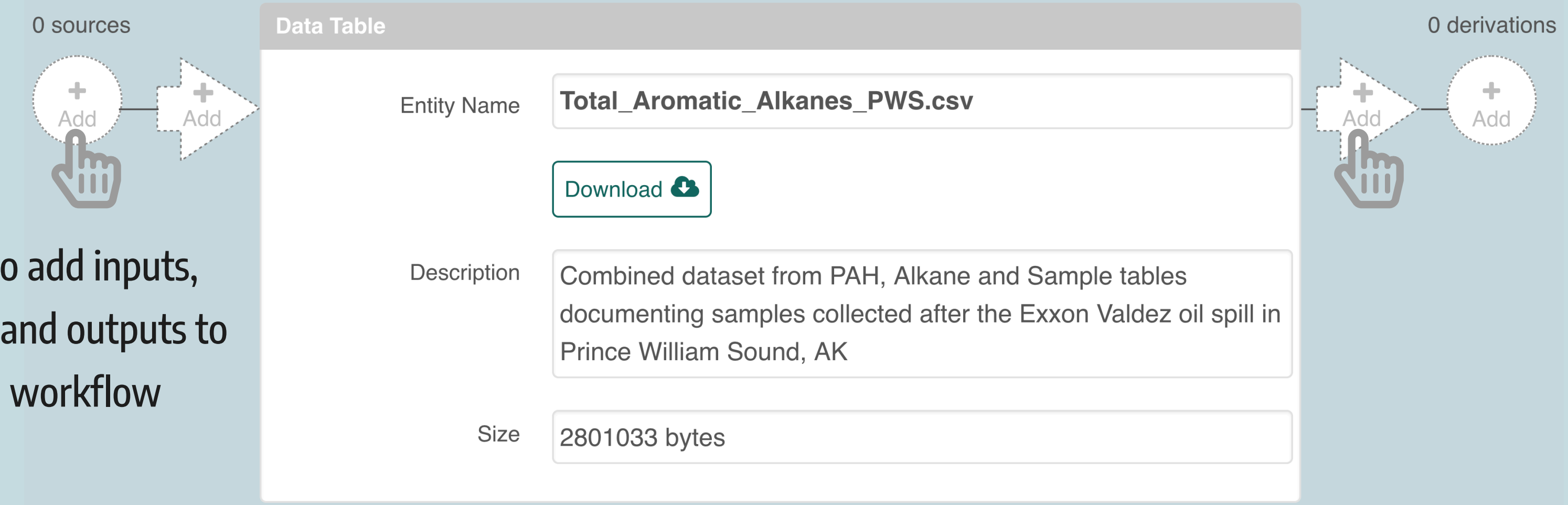
1 tabular data input files
1 processing script
2 image data output files

Direct Credit

Published multi-data source analysis cites previous works of researchers via provenance links

## Producing Provenance Information

- **An online editor** allows researchers to **create provenance links** between inputs, scripts, and outputs



Entity Name: Total_Aromatic_Alkanes_PWS.csv

Description: Combined dataset from PAH, Alkane and Sample tables documenting samples collected after the Exxon Valdez oil spill in Prince William Sound, AK

Size: 2801033 bytes

Click to add inputs, scripts, and outputs to the workflow

- The **recordr** R library and the **matlab-dataone** toolbox : **Command-line client libraries** for creating, managing, and publishing data packages with embedded provenance information.

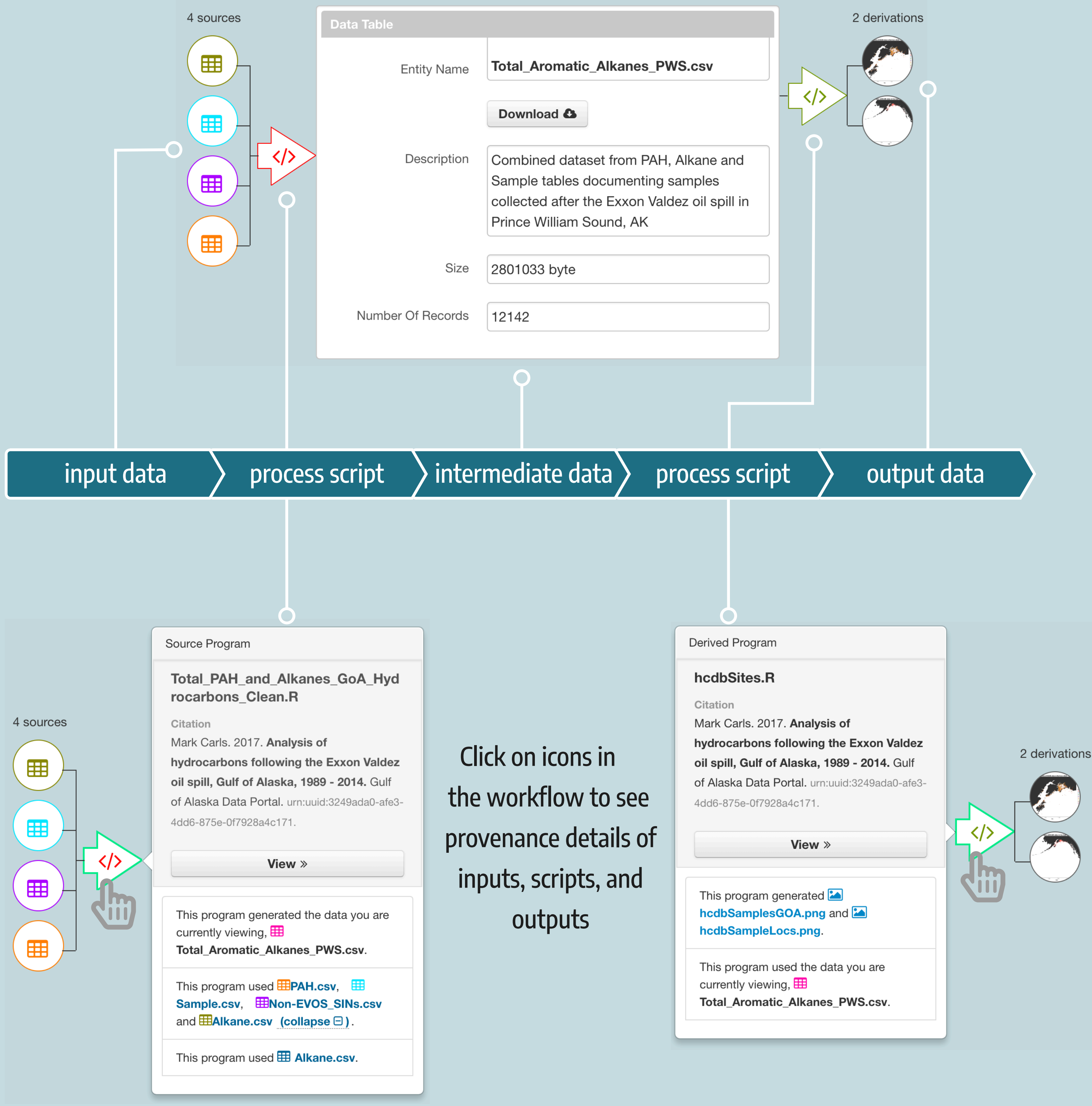
```
library(recordr)
rc <- new("Recordr")
% Record a script run
record(rc, system.file(“combine.R”, package="recordr"), tag="1”)
% List and view runs
listRuns(rc)
viewRuns(rc)
% Publish a run to a DataONE data repository
pkgId = mgr.publish(‘runNumber’, ‘1’);
```

```
import org.dataone.client.run.RunManager;
mgr = RunManager.getInstance();
% Record a script run
mgr.record(‘combine_data.m’, ‘run_01’);
% List all runs, and view a run’s details
mgr.listRuns();
mgr.viewRun(‘runNumber’, ‘1’);
% Publish a run to a DataONE data repository
pkgId = mgr.publish(‘runNumber’, ‘1’);
```

- A python module is planned, and will implement the same RunManager API as the R and Matlab clients

## Understanding Provenance Information

- **Web-based views** allow researchers to traverse the lineage of data products at search.dataone.org
- Can visualize the flow of data **inputs**, processing **scripts**, intermediate products, and final **outputs**
- Intra-dataset and cross-dataset **provenance linkages** can be followed by clicking on linked icons



Entity Name: Total_Aromatic_Alkanes_PWS.csv

Description: Combined dataset from PAH, Alkane and Sample tables documenting samples collected after the Exxon Valdez oil spill in Prince William Sound, AK

Size: 2801033 byte

Number Of Records: 12142

Source Program

**Total_PAH_and_Alkanes_GoA_Hydrocarbons_Clean.R**

Citation

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.

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 **Alkane.csv** (collapse ⊟).

This program used **Alkane.csv**.

Click on icons in the workflow to see provenance details of inputs, scripts, and outputs

Derived Program

**hcdbSites.R**

Citation

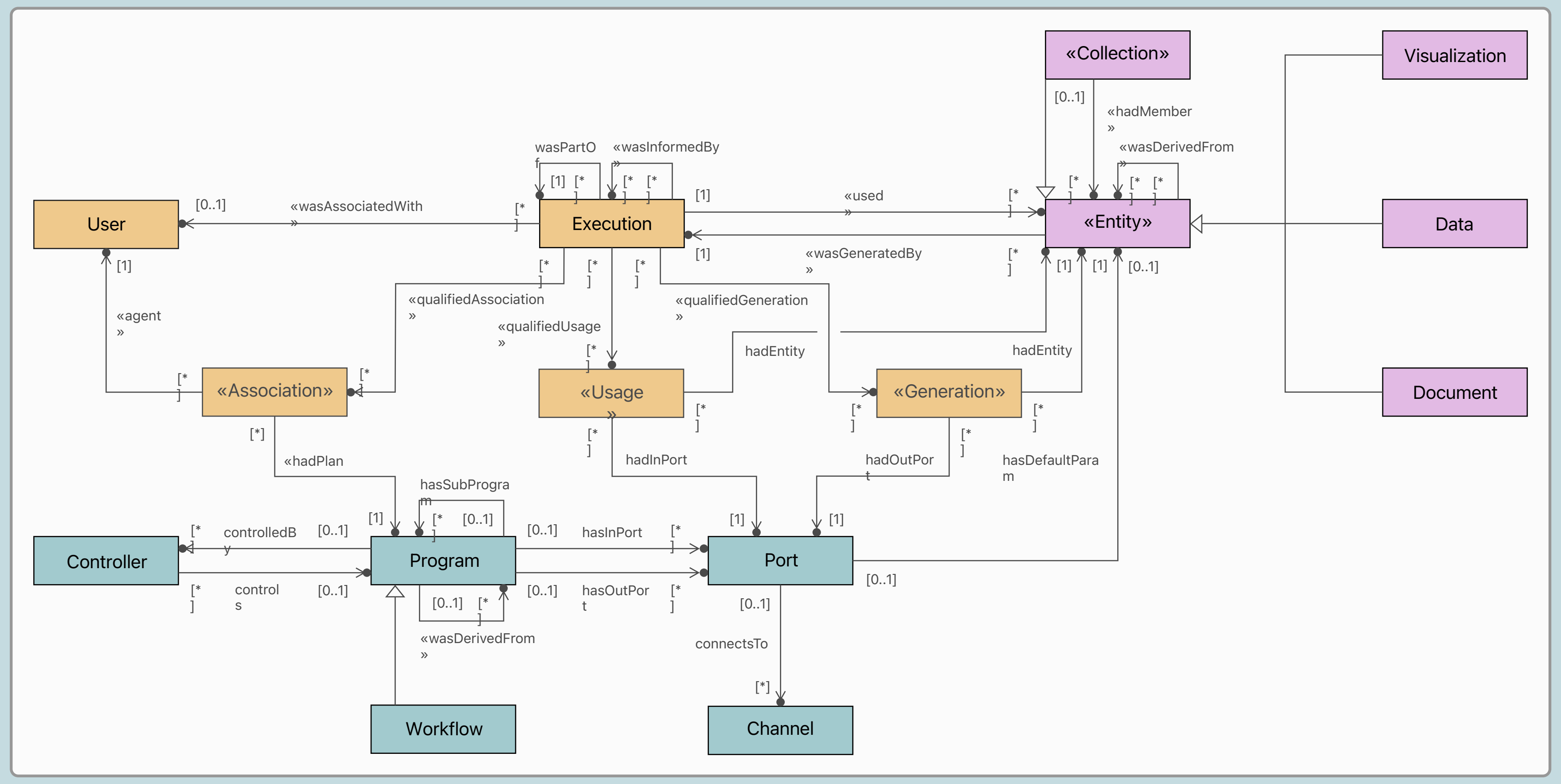
Mark Carls. 2017. **Analysis of hydrocarbons following the Exxon Valdez oil spill, Gulf of Alaska, 1989 - 2014.** Gulf of Alaska Data Portal. urn:uuid:3249ada0-afe3-4dd6-875e-0f7928a4c171.

This program generated **hcdbSamplesGOA.png** and **hcdbSampleLocs.png**.

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

## ProvONE Technical Model

- An OWL-based **ontology that extends the W3C PROV ontology** with classes specific to scientific workflows
- Is the next-generation of the community-based Open Provenance Model (OPM)
- Supports both **prospective** and **retrospective provenance** information (plans and executions)
- Available at https://purl.dataone.org/ontologies/provone-v1-dev (or https://github.com/DataONEorg/ontologies)



## Packaging Metadata

- Provenance relationships and other collection metadata are **serialized into RDF files** using the Open Archives Initiative Object Reuse and Exchange specification (**OAI-ORE**), with support for nested packages

**data package** =
**science metadata** +
**science data** +
**processing code** +
**provenance metadata**

[1] University of California, Santa Barbara; [2] University of Illinois at Urbana-Champaign; [3] University of Manchester, UK; [4] University of Kansas

This material is based upon work supported by the National Science Foundation under Grant Numbers 0830944, 1430508, and 1546024.