DataONE: A Data Federation with Provenance Support

Yang Cao[1], Christopher Jones[2], Víctor Cuevas-Vicenttín[3], Matthew B. Jones[2], Bertram Ludäscher[1], Timothy McPhillips[1], Paolo Missier[4], Christopher Schwalm[5], Peter Slaughter[2], Dave Vieglais[6], Lauren Walker[2], Yaxing Wei[7]

[1]University of Illinois, Urbana-Champaign, [2]National Center for Ecological Analysis and Synthesis, UCSB, [3]Universidad Popular Autónoma del Estado de Puebla, Mexico, [4]School of Computing Science, Newcastle University, UK, [5]Woods Hole Research Center, Falmouth, MA, [6]University of Kansas, Lawrence, [7]Environmental Sciences Division, Oak Ridge National Lab, TN

Introduction

Scientific workflow ***provenance*** can help computational scientists better understand and share their work with others while maintaining attribution. *Prospective* provenance describes the data transformation process or workflow; *retrospective* provenance describes the processing history and data derivations of a workflow.

DataONE (Data Observation Network for Earth) is a federated data network for Earth observational data. Its infrastructure consists of: *Member Nodes*, *Coordinating Nodes*, and an *Investigator Toolkit*. **DataONE Search** is a web application that lets users seamlessly and efficiently discover publicly accessible data packages within the DataONE federated network of Member Node repositories.

DataONE enables provenance-based browsing. We present provenance and search features with an example involving three scientists who interact through a Member Node. DataONE provenance systems enable reproducible research and facilitate proper attribution of scientific results transitively across generations of derived data products.

DataONE Provenance

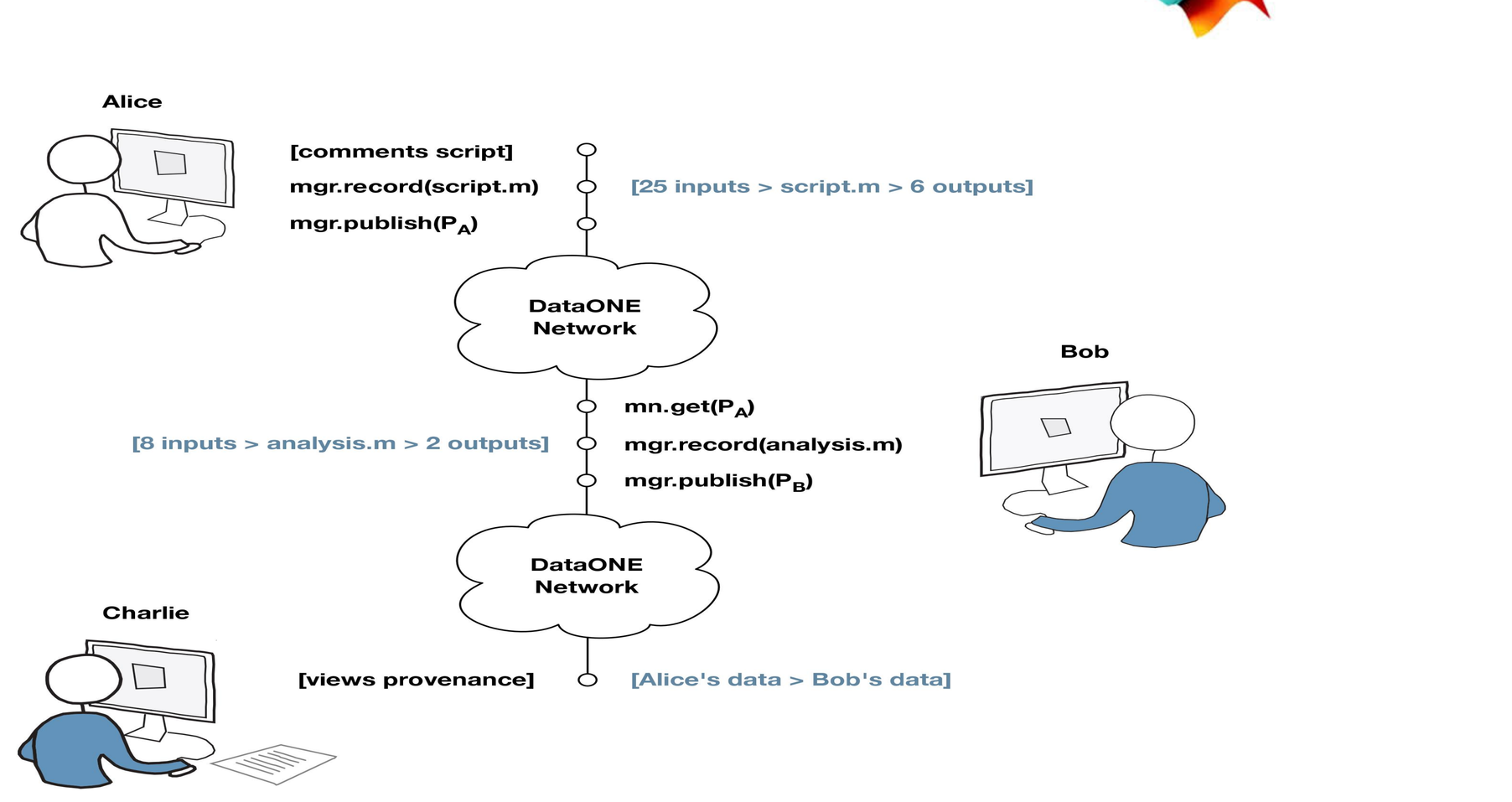
Phase II Goal: Facilitate reproducible science

- track data derivation history
- track data inputs and outputs of analyses
- track analysis and model executions
- preserve and document software

New DataONE Provenance Tools

- provenance indexing and search
- web user interface for browsing provenance
- R tool for generating provenance
- MATLAB tool for generating provenance

Provenance Use Case



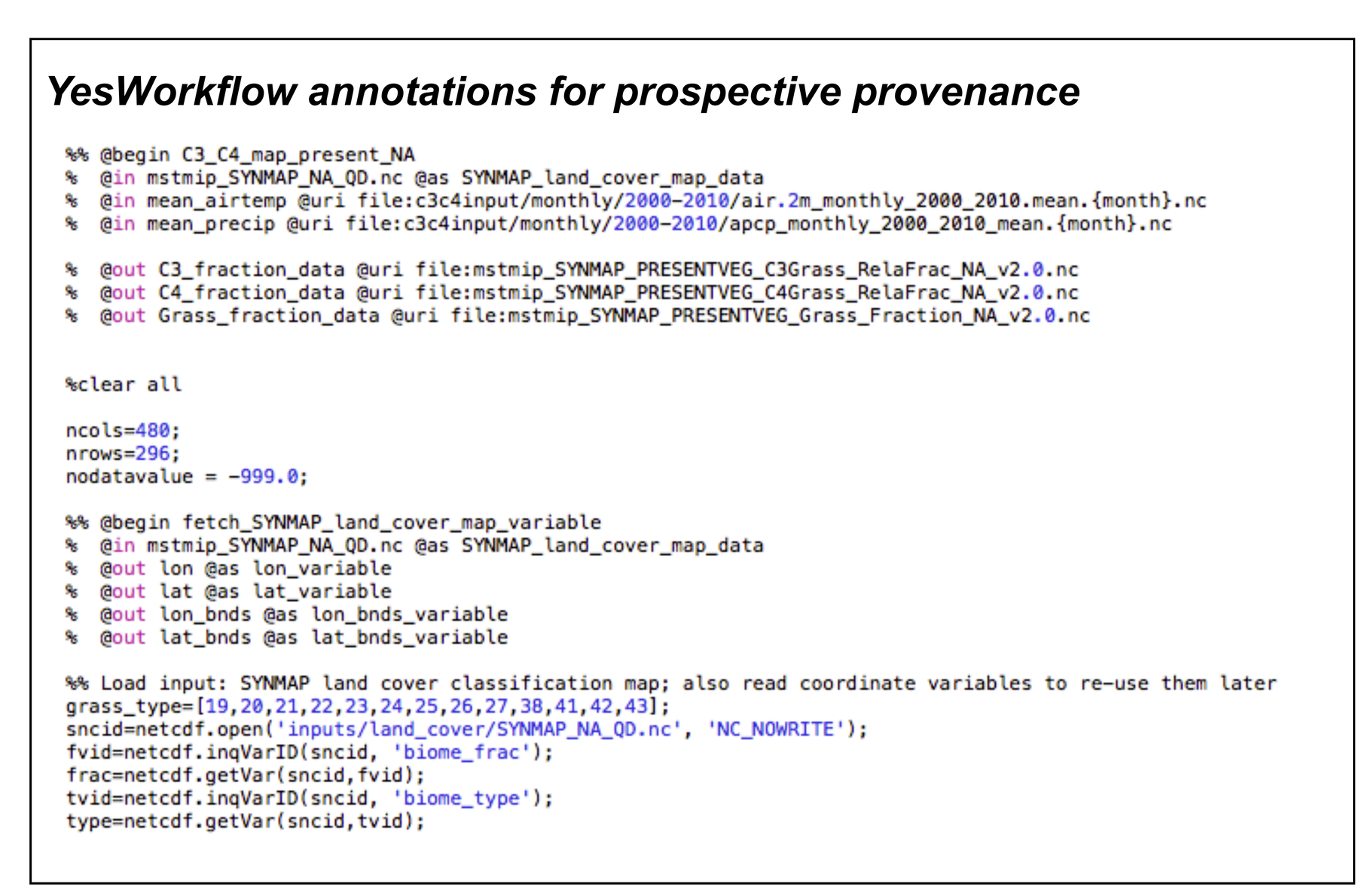
Alice's Data Generation

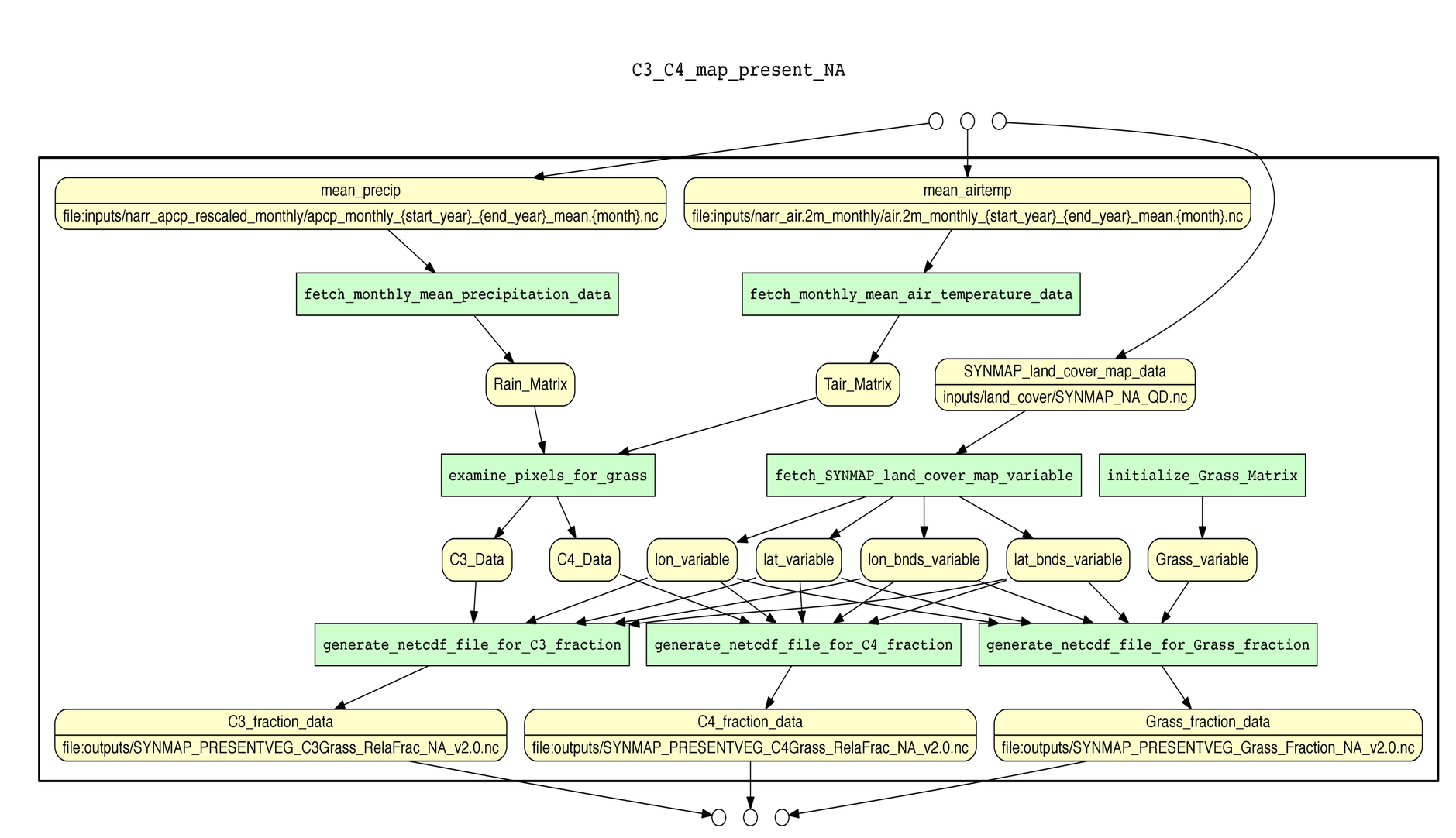
DataONE **Run Managers** capture **retrospective** provenance from R and MATLAB scripts, i.e., they:
• monitor file I/O events during a script's execution
• determine provenance relationships between files and an execution
• allow users to search, archive, and publish provenance data package

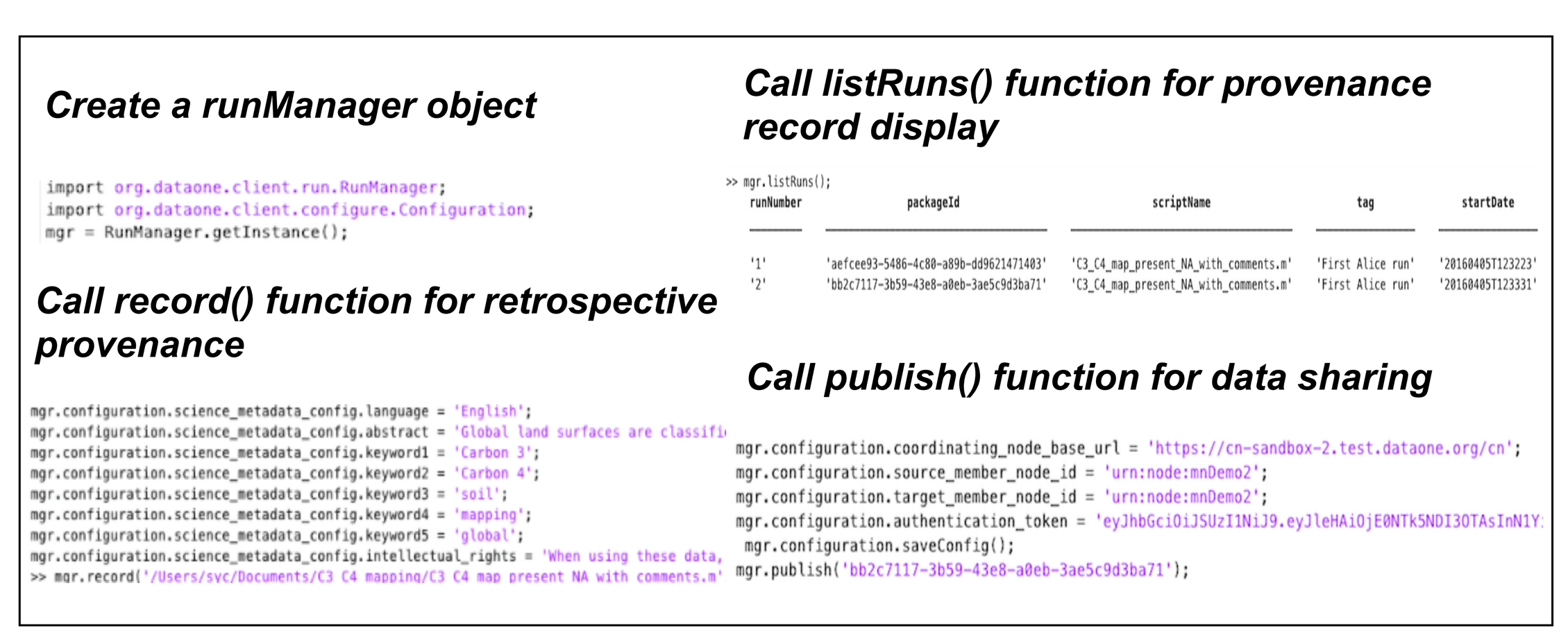
YesWorkflow (YW) is a script annotation tool for **prospective** provenance, which:

- enables users to mark up scripts with YesWorkflow annotations to reveal the computational steps and dataflows hidden in the scripts
- provides query capabilities for the prospective and retrospective provenance of the scripts

"Alice" annotates her script using YW. After running it, result files, script, prospective provenance, and retrospective provenance, represented in the ProvONE provenance model, are bundled into an OAI-ORE compliant data package and uploaded to the DataONE network.







Bob's Data Reuse

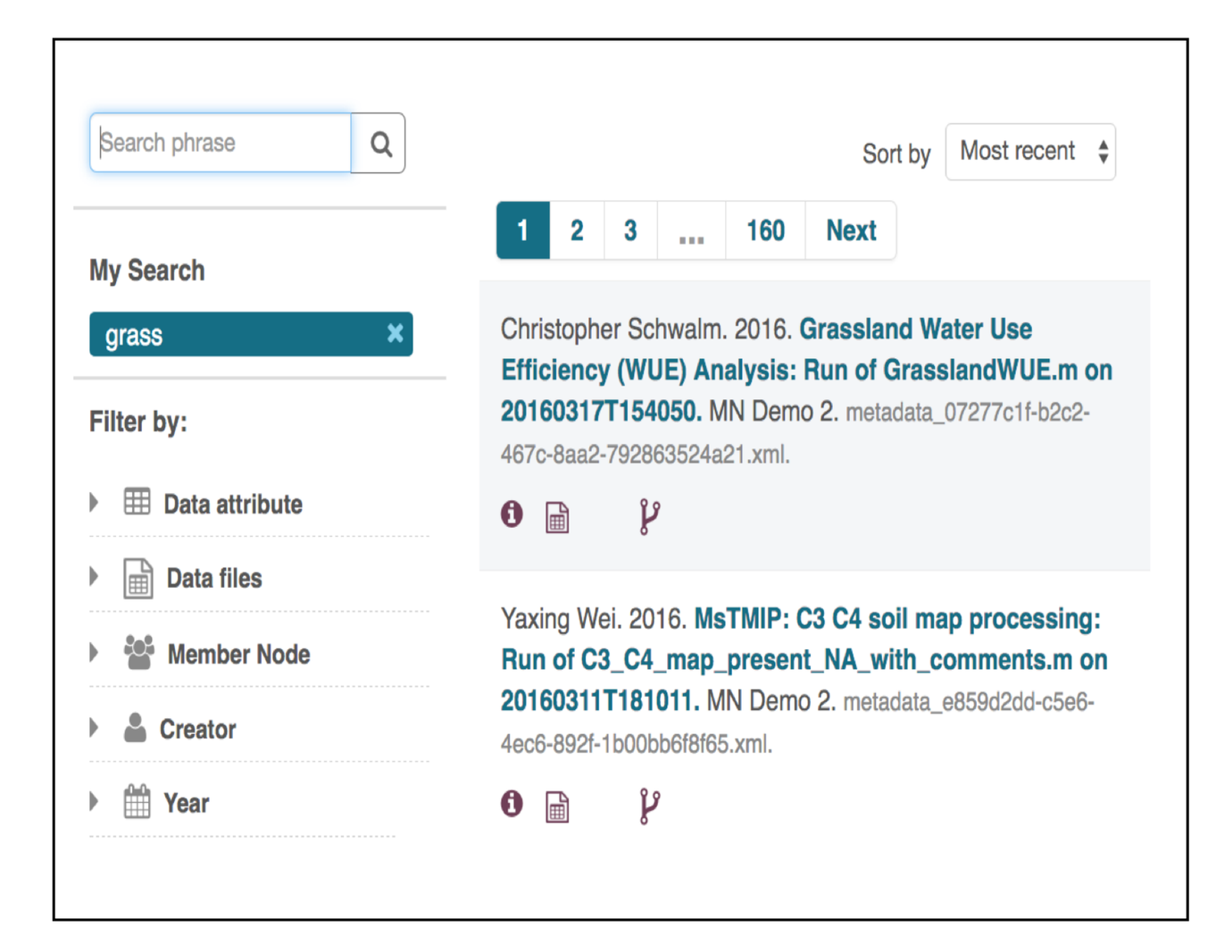
"Bob" discovers Alice's package and uses her data in his own analysis. In order to find Alice's published data package, Bob follows these steps:

1. Through a keyword search for "**grass**" data he finds Alice's data package.
2. He uses the data-Ids of Alice's outputs as input for his script, using MN.get(Id) for access.
3. He records retrospective script provenance using the MATLAB Run Manager.
4. He publishes a new data package with his results and their provenance.

Charlie's Visual Data Navigation

Charlie also searches for "grass" data and finds Bob's and Alice's data packages along with their provenance information. He also notes that Bob's inputs where obtained from Alice's outputs.

Demo Site: http://dataoneorg.github.io/provweek2016-demo/



DataONE Search result for keyword "grass". The fork icon indicate data packages having embedded provenance.

Charlie's view on the DataONE demo site: (1) A YesWorkflow model for Alice' soil processing script; (2) Data lineage from Bob's results back through his script to Alice's data package; (3) Two visualizations produced by Bob's water use efficiency analysis script.

ProvONE provenance statements in the Alice's data package

```
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix ns1: <http://purl.dataone.org/provone/2015/01/15/ontology#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix dc: <http://purl.org/dc/elements/1.1/> .
@prefix ore: <http://www.openarchives.org/ore/terms/> .
@prefix dcterms: <http://purl.org/dc/terms/> .
@prefix ns2: <http://www.w3.org/ns/prov#> .
@prefix ns3: <http://purl.org/spar/cito/> .
<https://cn-sandbox-
2.test.dataone.org/cn/v2/resolve/SYNMAP_PRESENTVEG_Grass_Fraction_NA_v2.0.
nc>
ore:isAggregatedBy "https://cn-sandbox-
2.test.dataone.org/cn/v1/resolve/resourceMap_e859d2dd-c5e6-4ec6-892f-1
b00bb6f8f65.rdf#aggregation" ;
ns2:wasDerivedFrom <https://cn-sandbox-
2.test.dataone.org/cn/v2/resolve/air.2m_monthly_2000_2010_mean.1.nc> ;
a ns1:Data ;
a ore:AggregatedResource ;
ns2:wasGeneratedBy <https://cn-sandbox-
2.test.dataone.org/cn/v2/resolve/execution_e859d2dd-c5e6-4ec6-892f-1
b00bb6f8f65> ;
ns3:isDocumentedBy <https://cn-sandbox-
2.test.dataone.org/cn/v2/resolve/metadata_e859d2dd-c5e6-4ec6-892f-1
b00bb6f8f65.xml> ;
dcterms:identifier
"SYNMAP_PRESENTVEG_Grass_Fraction_NA_v2.0.nc"^^<http://www.w3.org/2001/XML
Schema#string> .
```

Conclusions and Future Work

DataONE support for data discovery now includes provenance capabilities.

Future work includes:
(1) extensions to the Run Managers to support dependencies between multiple runs from one or more scripts by a single user (local operation);
(2) support for additional I/O functions; and
(3) finer grained provenance capture, e.g., via the noWorkflow tool (cf. "YesWorfklow & noWorkflow Demonstration, IPAW & TaPP 2016).

Acknowledgments. Supported by NSF awards ACI-1430508 and NSF ABI-1262458.

NCEAS · Oak Ridge National Laboratory · KU The University of Kansas · NSF · Newcastle University · Woods Hole Research Center

Graduate School of Library and Information Science, The iSchool at Illinois · Illinois 1867 University of Illinois at Urbana-Champaign · NCSA