Provenance and DataONE: Facilitating Reproducible Science

Bertram Ludäscher, Chris Jones, Lauren Walker

GRADUATE SCHOOL OF LIBRARY AND INFORMATION SCIENCE
The iSchool at Illinois

UCSB
UNIVERSITY OF CALIFORNIA SANTA BARBARA

NCEAS

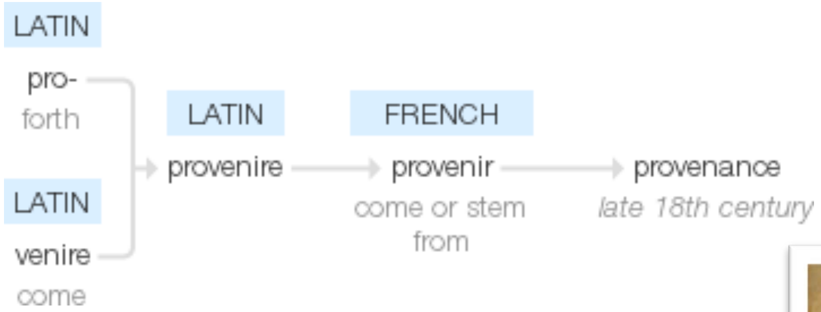
Outline

1. **Overview** on Provenance (*Bertram*)
2. **Searching** and **Navigating** Provenance (*Lauren*)
3. **Further Details**, *"look behind the scenes"* (*Chris*)

Acknowledgments & special thanks to:

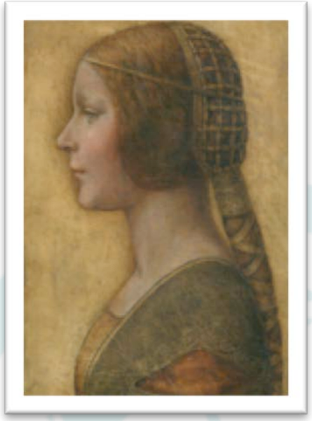
- NSF, DataONE CI Team, WG members (Phases I & II), others contributors (YW)

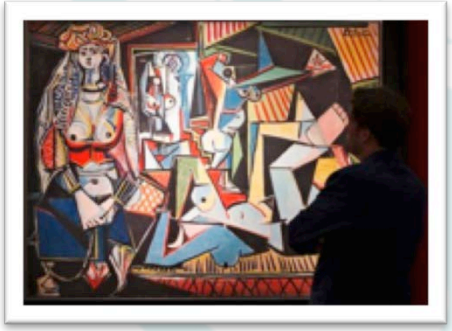
Provenance



1. The place of **origin** or **earliest known history** of something
2. The **beginning** of something’s existence
3. A **record of ownership** of a work of art or an antique, used as a guide to **authenticity** or **quality**

Related terms: lineage, genealogy, pedigree, …

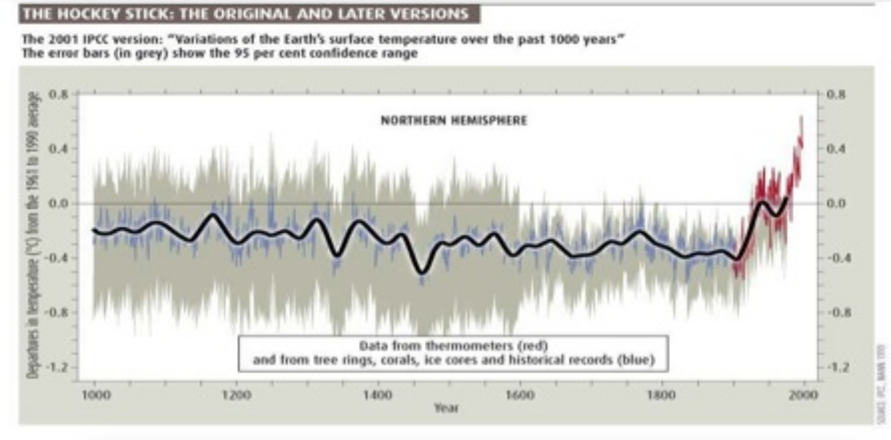




Computational Provenance

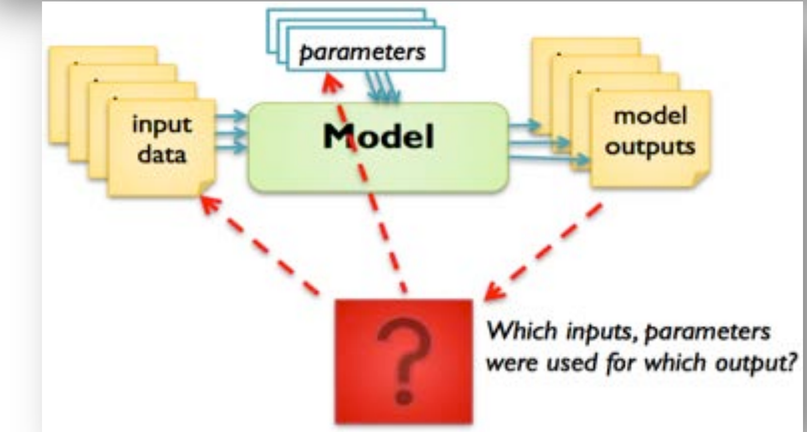
- **Origin** and **processing history** of an artifact
- usually: data products, figures, …
- sometimes: workflow & script evolution …
- **Provenance sightings**:
- Data science, eScience, CI, Big Data, computational science, 4th paradigm …
- Bio(diversity)informatics, ecoinformatics, geoinformatics, ..
- Computer science, library & information science, …
- Scientific workflows & scripts …
- Databases, programming languages, …
- Privacy vs. provenance, …

Reproducible Science

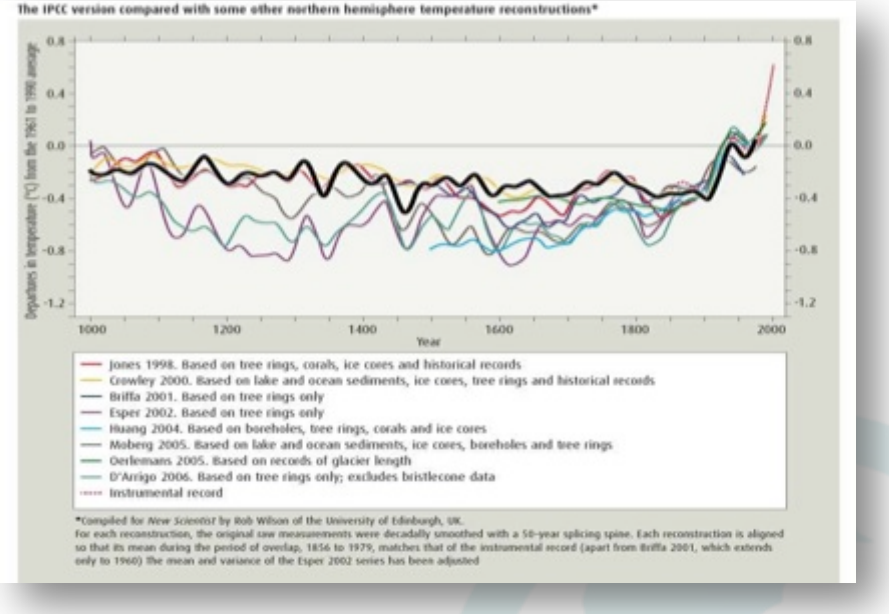


THE HOCKEY STICK: THE ORIGINAL AND LATER VERSIONS

The 2001 IPCC version: "Variations of the Earth's surface temperature over the past 1000 years"
The error bars (in grey) show the 95 per cent confidence range



Which inputs, parameters were used for which output?

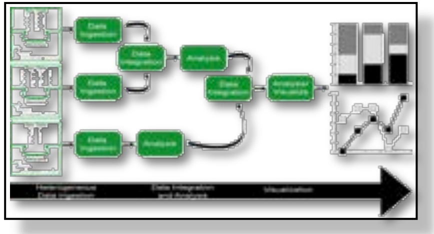


The IPCC version compared with some other northern hemisphere temperature reconstructions*

*Compiled for *New Scientist* by Rob Wilson of the University of Edinburgh, UK.
For each reconstruction, the original raw measurements were decadally smoothed with a 50-year splicing spine. Each reconstruction is aligned so that its mean during the period of overlap, 1856 to 1979, matches that of the instrumental record (apart from Briffa 2001, which extends only to 1960) The mean and variance of the Esper 2002 series has been adjusted

Capturing **provenance** is crucial for transparency, interpretation, debugging, …
=> *repeatable experiments,*
=> *reproducible science*

Scientific Workflows: ASAP!



Automation

wfs to **automate** computational aspects of science

Scaling (exploit and optimize ***machine*** *cycles*)

wfs should make use of **parallel compute resources**
wfs should be able handle **large data**

Abstraction, Evolution, Reuse *(**human** cycles)*

wfs should be easy to **(re-)use, evolve, share**

Provenance

wfs should capture **processing history**, **data lineage**
=> traceable data- and wf-evolution
=> Reproducible Science

Kepler

swift

TRIANA

VT

ASKALON

Taverna

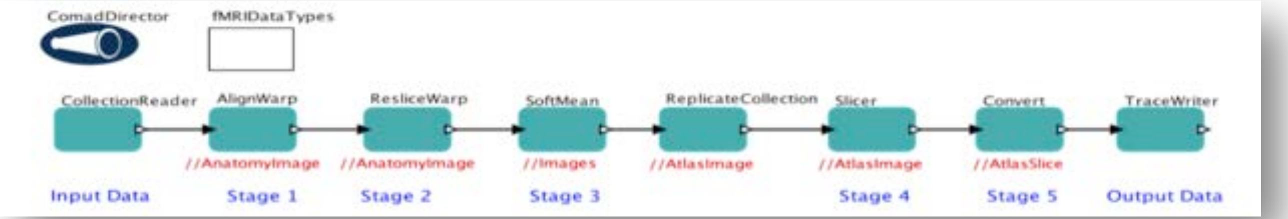
Common Uses of Provenance in Science

- **Audit trail**: trace data generation and possible errors
- **Attribution**: determine ownership and responsibility for data and scientific results
- **Data quality**: from quality of input data, computations
- **Discovery**: enable searching of data, methodologies and experiments
- **Replication**: facilitate repeatable derivation of data to maintain currency

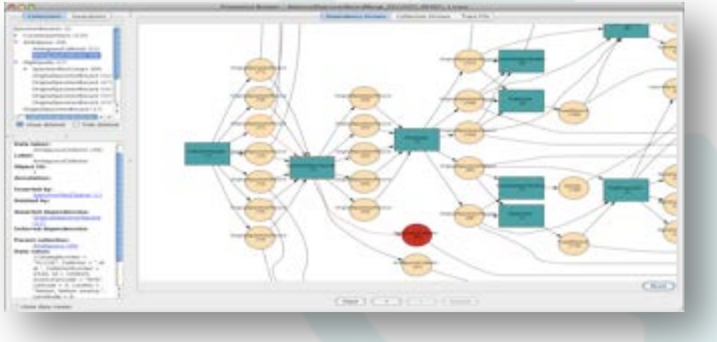
⇒ **Reproducible Science**

Kinds of Provenance

- **Prospective** Provenance
- method/workflow description (“***workflow-land***”)



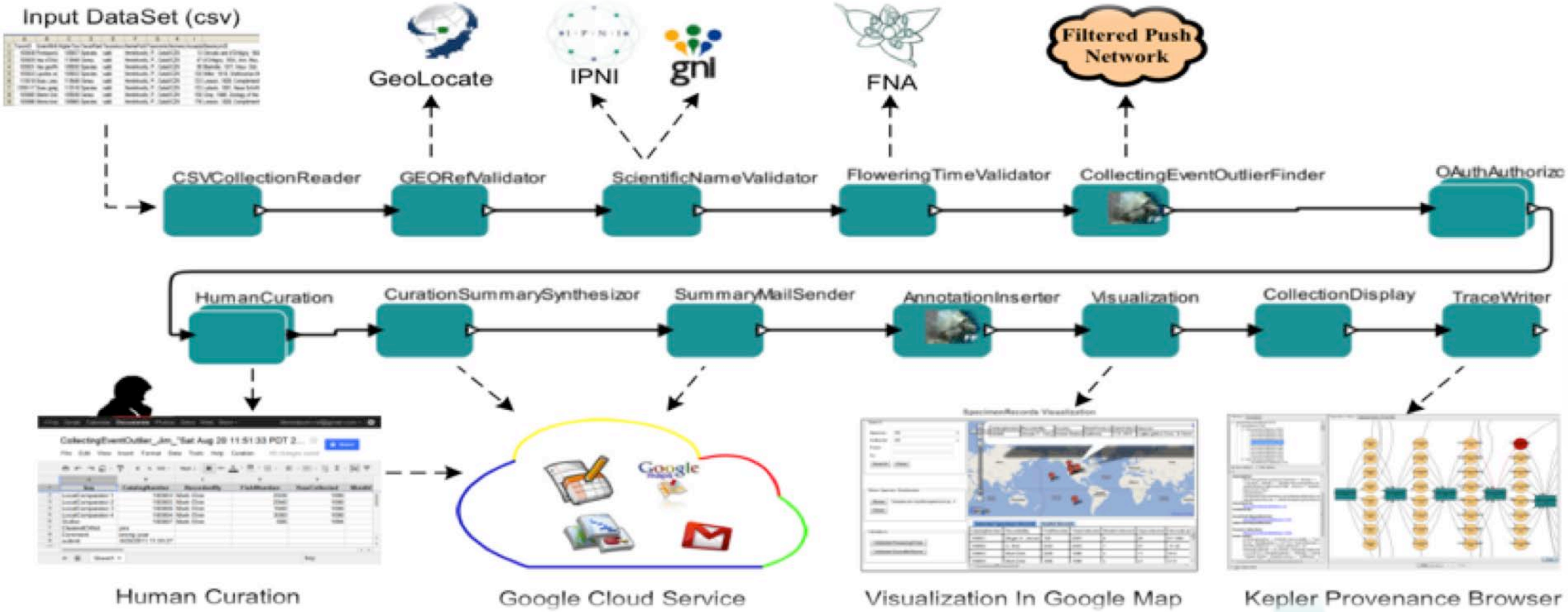
- **Retrospective** Provenance
- runtime provenance tracking (“***trace-land***”)



- **Q**: Which one is more important?

Prospective Provenance

(A Data Curation Workflow: FilteredPush)



HARVARD UNIVERSITY HERBARIA

MUSEUM OF COMPARATIVE ZOOLOGY
HARVARD UNIVERSITY

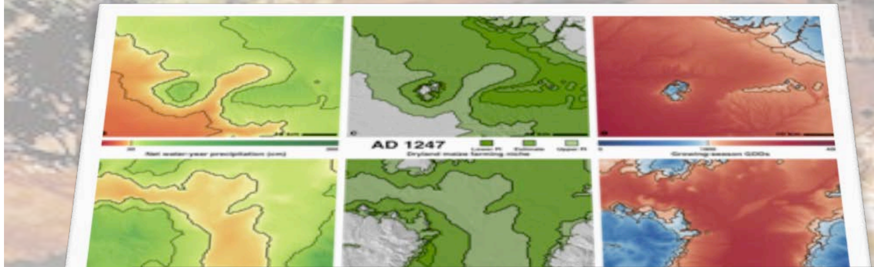
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

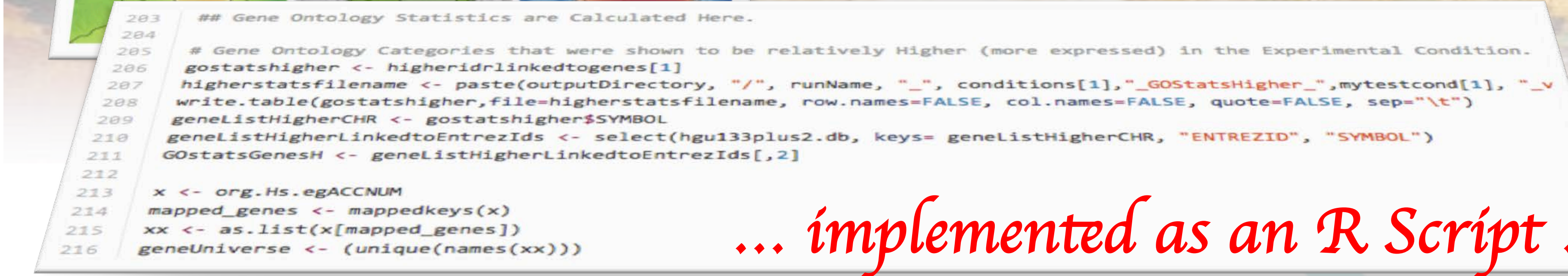
SKOPE: Synthesized Knowledge Of Past Environments

Bocinsky, Kohler *et al.* study rain-fed maize of **Anasazi**

Four Corners; AD 600–1500. **Climate change** influenced **Mesa Verde Migrations**; late 13th century AD. Uses **network of tree-ring chronologies** to **reconstruct a spatio-temporal climate** field at a fairly high resolution (~800 m) from AD 1–2000. Algorithm estimates joint information in tree-rings and a climate signal to identify “best” tree-ring chronologies for climate reconstructing.

K. **Bocinsky**, T. **Kohler**, A 2000-year reconstruction of the rain-fed maize agricultural niche in the US Southwest. ***Nature Communications***. doi:10.1038/ncomms6618


AD 1247
Net water-year precipitation (cm)
Lower PI
Estimate
Upper PI
Dryland maize farming niche
Growing-season GDDs



```
 ## Gene Ontology Statistics are Calculated Here.

 # Gene Ontology Categories that were shown to be relatively Higher (more expressed) in the Experimental Condition.
 gostatshigher <- higheridrlinkedtogenes[1]
 higherstatsfilename <- paste(outputDirectory, "/", runName, "_", conditions[1],"_GOStatsHigher_",mytestcond[1], "_v
 write.table(gostatshigher,file=higherstatsfilename, row.names=FALSE, col.names=FALSE, quote=FALSE, sep="\t")
 geneListHigherCHR <- gostatshigher$SYMBOL
 geneListHigherLinkedtoEntrezIds <- select(hgu133plus2.db, keys= geneListHigherCHR, "ENTREZID", "SYMBOL")
 GOstatsGenesH <- geneListHigherLinkedtoEntrezIds[,2]

 x <- org.Hs.egACCNUM
 mapped_genes <- mappedkeys(x)
 xx <- as.list(x[mapped_genes])
 geneUniverse <- (unique(names(xx)))
```

... implemented as an R Script ...

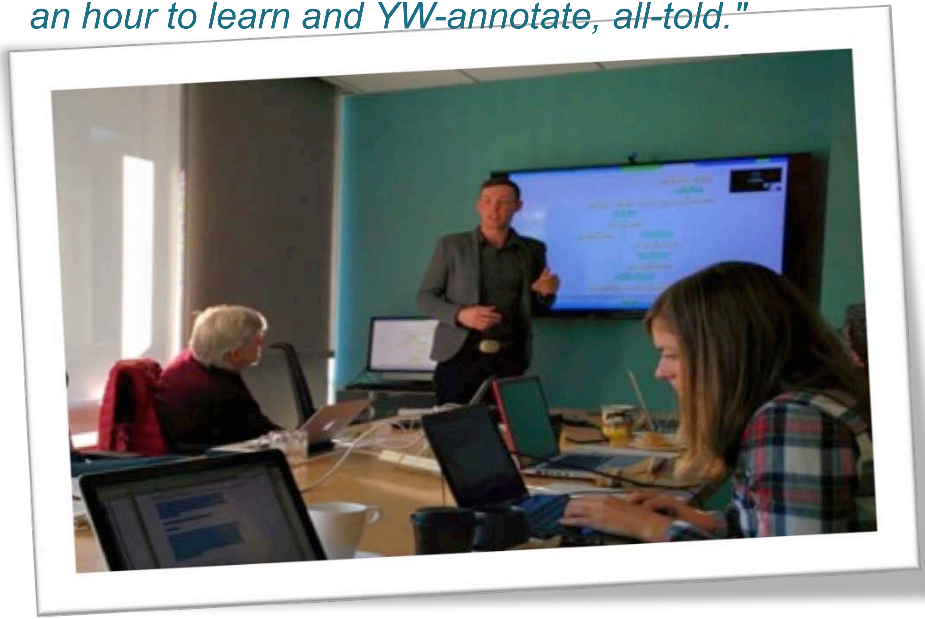
Paleoclimate Reconstruction …

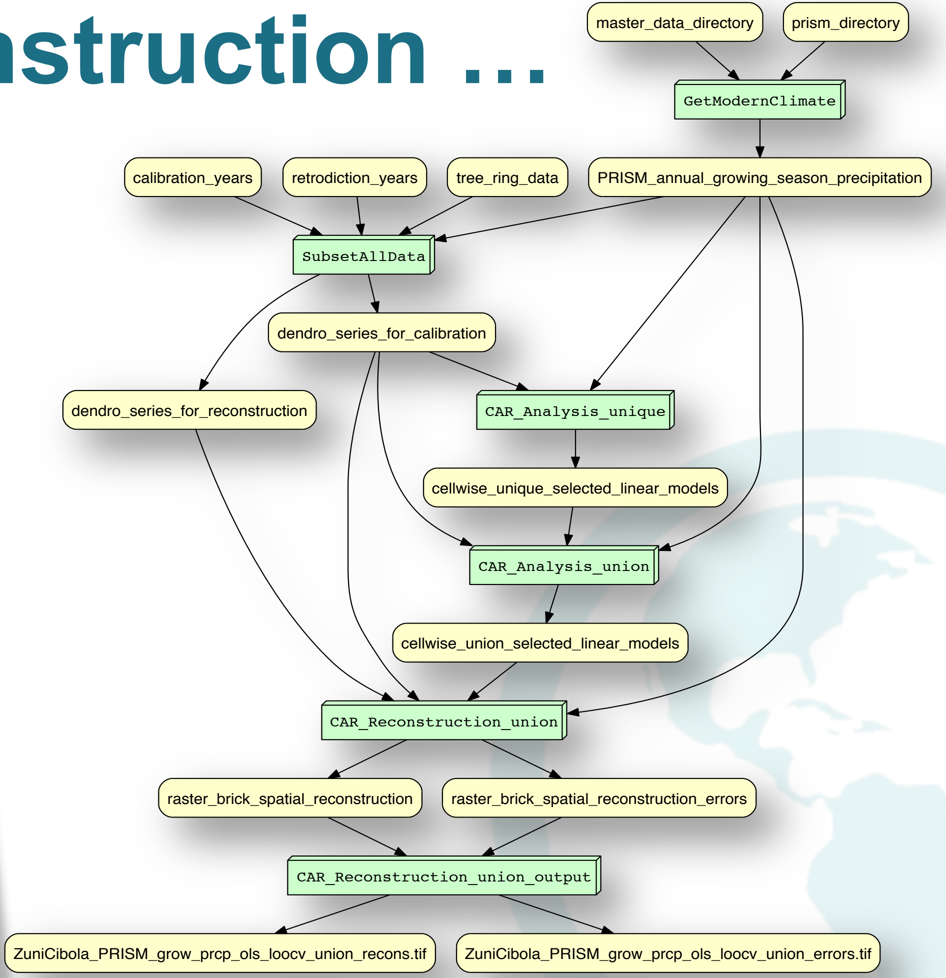
… explained using **YesWorkflow**

https://github.com/yesworkflow-org/

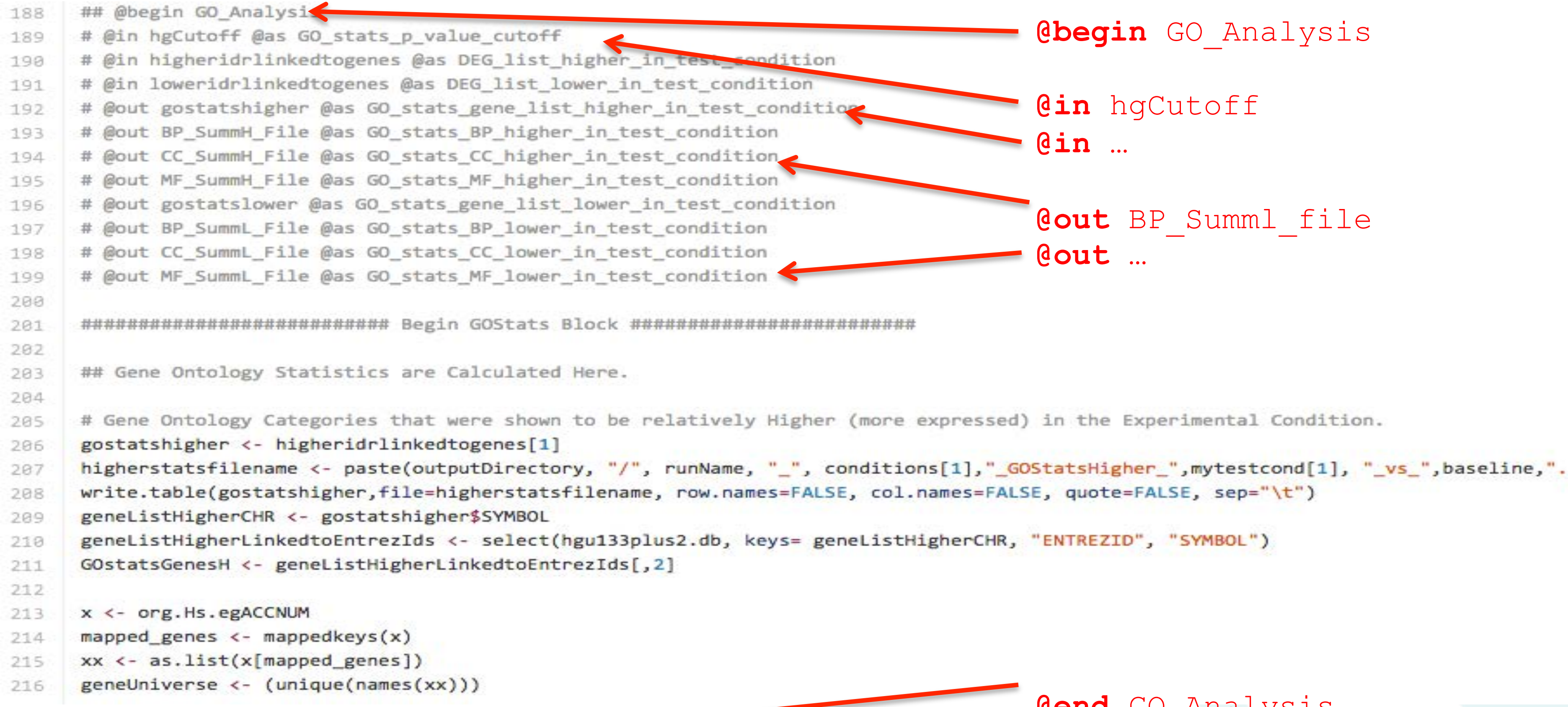
Kyle B., (computational) archeologist:

"It took me about 20 minutes to comment. Less than an hour to learn and YW-annotate, all-told."




master_data_directory
prism_directory
GetModernClimate
calibration_years
retrodiction_years
tree_ring_data
PRISM_annual_growing_season_precipitation
SubsetAllData
dendro_series_for_calibration
dendro_series_for_reconstruction
CAR_Analysis_unique
cellwise_unique_selected_linear_models
CAR_Analysis_union
cellwise_union_selected_linear_models
CAR_Reconstruction_union
raster_brick_spatial_reconstruction
raster_brick_spatial_reconstruction_errors
CAR_Reconstruction_union_output
ZuniCibola_PRISM_grow_prcp_ols_loocv_union_recons.tif
ZuniCibola_PRISM_grow_prcp_ols_loocv_union_errors.tif

User Comments: YW Annotations



```
## @begin GO_Analysis
# @in hgCutoff @as GO_stats_p_value_cutoff
# @in higheridrlinkedtogenes @as DEG_list_higher_in_test_condition
# @in loweridrlinkedtogenes @as DEG_list_lower_in_test_condition
# @out gostatshigher @as GO_stats_gene_list_higher_in_test_condition
# @out BP_SummH_File @as GO_stats_BP_higher_in_test_condition
# @out CC_SummH_File @as GO_stats_CC_higher_in_test_condition
# @out MF_SummH_File @as GO_stats_MF_higher_in_test_condition
# @out gostatslower @as GO_stats_gene_list_lower_in_test_condition
# @out BP_SummL_File @as GO_stats_BP_lower_in_test_condition
# @out CC_SummL_File @as GO_stats_CC_lower_in_test_condition
# @out MF_SummL_File @as GO_stats_MF_lower_in_test_condition

########################### Begin GOStats Block ##########################

## Gene Ontology Statistics are Calculated Here.

# Gene Ontology Categories that were shown to be relatively Higher (more expressed) in the Experimental Condition.
gostatshigher <- higheridrlinkedtogenes[1]
higherstatsfilename <- paste(outputDirectory, "/", runName, "_", conditions[1],"_GOStatsHigher_",mytestcond[1], "_vs_",baseline,".
write.table(gostatshigher,file=higherstatsfilename, row.names=FALSE, col.names=FALSE, quote=FALSE, sep="\t")
geneListHigherCHR <- gostatshigher$SYMBOL
geneListHigherLinkedtoEntrezIds <- select(hgu133plus2.db, keys= geneListHigherCHR, "ENTREZID", "SYMBOL")
GOstatsGenesH <- geneListHigherLinkedtoEntrezIds[,2]

x <- org.Hs.egACCNUM
mapped_genes <- mappedkeys(x)
xx <- as.list(x[mapped_genes])
geneUniverse <- (unique(names(xx)))
```

@begin GO_Analysis

@in hgCutoff

@in …

@out BP_Summl_file

@out …

@end GO_Analysis

. . .

Multi-Scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP)

North American Carbon Program
MsTMIP
NASA
MULTI-SCALE SYNTHESIS AND TERRESTRIAL MODEL INTERCOMPARISON PROJECT

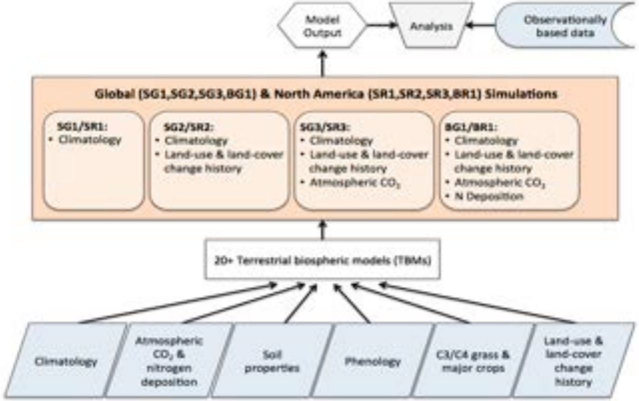

Model Output
Analysis
Observationally based data
Global (SG1,SG2,SG3,BG1) & North America (SR1,SR2,SR3,BR1) Simulations
SG1/SR1:
• Climatology
SG2/SR2:
• Climatology
• Land-use & land-cover change history
SG3/SR3:
• Climatology
• Land-use & land-cover change history
• Atmospheric CO2
BG1/BR1:
• Climatology
• Land-use & land-cover change history
• Atmospheric CO2
• N Deposition
20+ Terrestrial biospheric models (TBMs)
Climatology
Atmospheric CO2 & nitrogen deposition
Soil properties
Phenology
C3/C4 grass & major crops
Land-use & land-cover change history

Fig. 1. Schematic of the Multi-Scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP) framework. Global simulations (SG1,SG2, SG3,BG1) are run at 0.5° by 0.5° resolution; North American simulations (SR1, SR2, SR3, BR1) are run at 0.25° by 0.25° resolution).

The North American Carbon Program Multi-Scale Synthesis and Terrestrial Model Intercomparison Project
D. N. Huntzinger1, C. Schwalm2, A. M. Michalak3, K. Schaefer4,5, A. W. King6, Y. Wei6, A. Jacobson4,7, S. Liu6, R. B. Cook6, W. M. Post6, G. Berthier8, D. Hayes6, M. Huang9, A. Ito10, H. Lei11,12, C. Lu13, J. Mao6, C. H. Peng14,15, S. Peng8, B. Poulter8, D. Riccuito6, X. Shi6, H. Tian13, W. Wang16, N. Zeng17, F. Zhao17, and Q. Zhu15

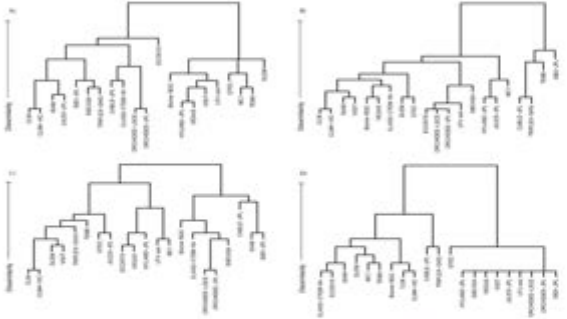
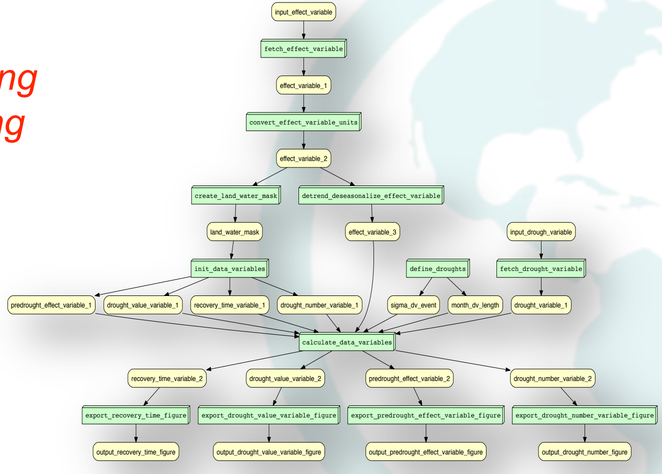


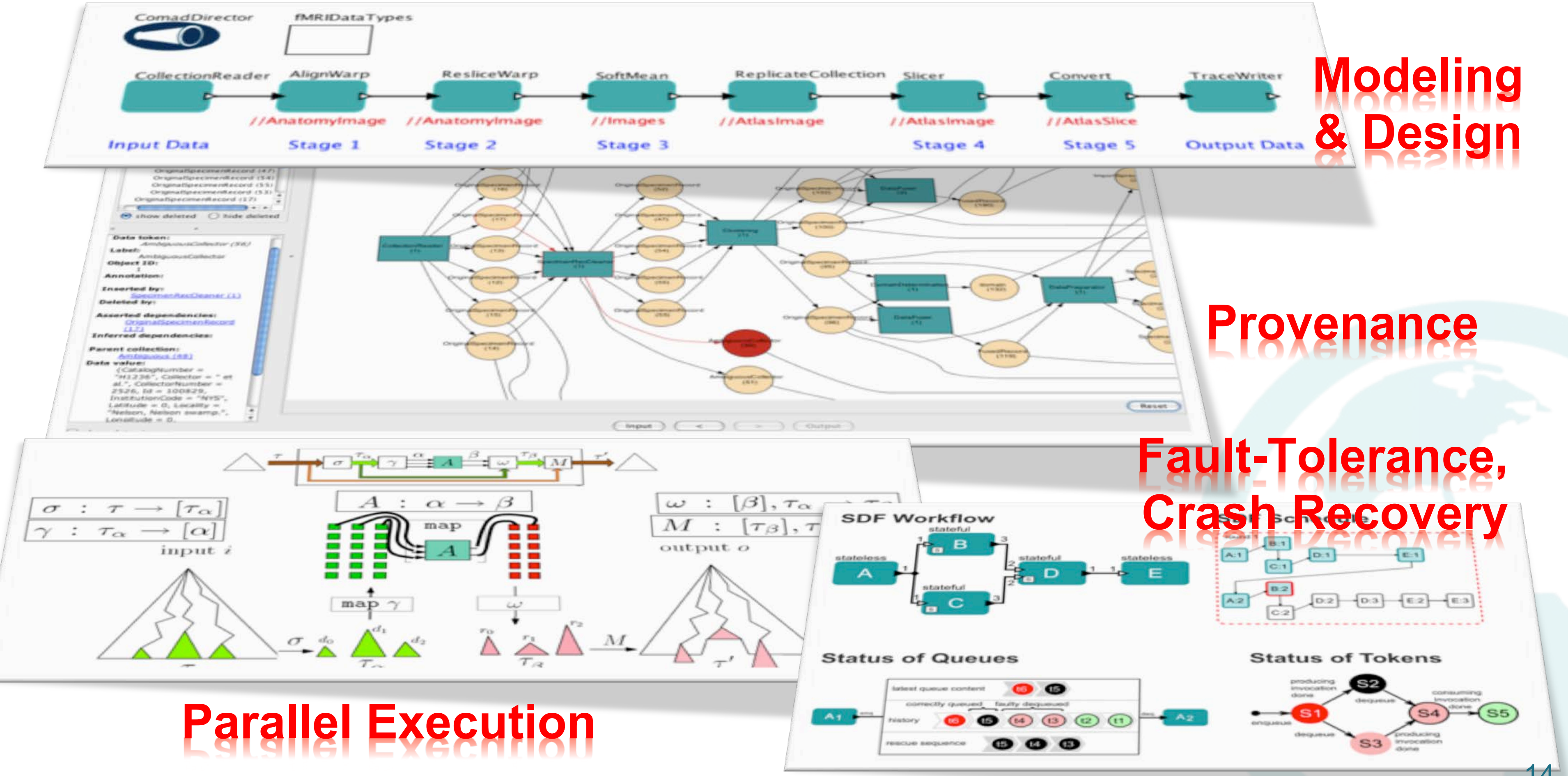
Fig. 3. Dendrogram showing general differences/similarities in how MsTMIP models formulate and parameterize (A) energy, (B) carbon, (C) vegetation, and (D) nitrogen process dynamics. Clusters are determined by Hamming distance. Models in the same "tree" share similar structural model characteristics. For example, models in the "tree" to the left (e.g., ISAM, CABLE-JPL, ORCHIDEE-JPL/LSCE) in (A) simulate ground heat flux and canopy heat storage, while models in the "tree" to the right (e.g., MC1, TEM6, VEGAS) do not. A majority of models separate live carbon into various pools (with exception of SiB-JPL), but they do so in various ways (e.g., left "tree" in (B)). Refer to the Supplement for the binary data used to create this diagram.

Provenance

- *Externally facing*
- *Internally facing*


input_effect_variable
fetch_effect_variable
effect_variable_1
convert_effect_variable_units
effect_variable_2
create_land_water_mask
detrend_deseasonalize_effect_variable
land_water_mask
effect_variable_3
input_drough_variable
init_data_variables
define_droughts
fetch_drought_variable
predrought_effect_variable_1
drought_value_variable_1
recovery_time_variable_1
drought_number_variable_1
sigma_dv_event
month_dv_length
drought_variable_1
calculate_data_variables
recovery_time_variable_2
drought_value_variable_2
predrought_effect_variable_2
drought_number_variable_2
export_recovery_time_figure
export_drought_value_variable_figure
export_predrought_effect_variable_figure
export_drought_number_variable_figure
output_recovery_time_figure
output_drought_value_variable_figure
output_predrought_effect_variable_figure
output_drought_number_figure

Scientific Workflow Research



Modeling & Design

Provenance

Fault-Tolerance, Crash Recovery

Parallel Execution

Wait, there is more …

https://goo.gl/53OCdW

- **Fine-grained** vs **coarse-grained** provenance
- **Black-box** vs **white-box** provenance

- **Standards**:
- OPM ➔ **PROV**
- D-OPM ➔ **ProvONE**

- **Database Community**:
- why-, where-, how-, why-not provenance
- links to causality
- … logical derivations, proofs, …

Live Demonstration

Facilitate reproducible science

- Creating and managing provenance information
- Communicating script and model workflows
- Storing and sharing
- Using provenance information for search

Creating and managing provenance information

Investigator tools (ITK)

Matlab DataONE Toolbox

Recordr R Library

Java YesWorkflow Tool

Functions being added

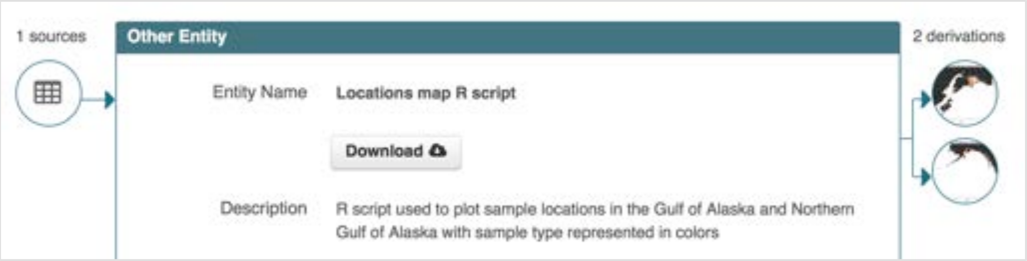
`record()`
`startRecord()`
`endRecord()`
`listRuns()`
`deleteRuns()`
`viewRun()`
`publish()`

`set()`
`get()`
`saveConfig()`
`loadConfig()`
`listConfig()`

See: Run Manager API document

Example: R programming

```
# Generate map of locations by type
library(recordr)
recordr <- new(“Recordr”)
pkg <- record(recordr, “./hcdbSites.R”, “locations-by-type-png”)
```



Other Entity

Entity Name: Locations map R script

Description: R script used to plot sample locations in the Gulf of Alaska and Northern Gulf of Alaska with sample type represented in colors

R: managing script runs

```
>   listRuns(recordr)

    Script        StartTime             EndTime               PublishedTime  Tag                     RunID

    hcdbSites.R   2015-05-07T18:53:09Z  2015-05-07T18:53:09Z  unpublished    locations-by-type-png
    C85A ...
>   deleteRuns(recordr, “locations-by-type-png”)

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

```
>   viewRun(recordr, “locations-by-type-png”)

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

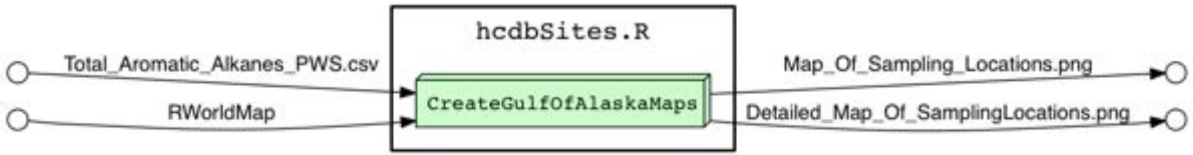
```
>   publishRun(recordr, “locations-by-type-png”)

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

Communicating workflow provenance

YesWorkflow tool

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



hcdbSites.R

Total_Aromatic_Alkanes_PWS.csv

RWorldMap

CreateGulfOfAlaskaMaps

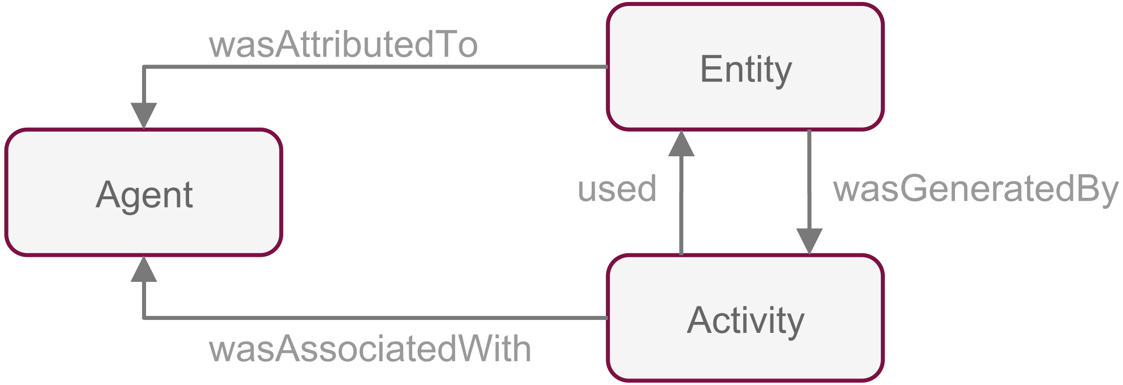
Map_Of_Sampling_Locations.png

Detailed_Map_Of_SamplingLocations.png

Storing and sharing provenance information

Using a common model

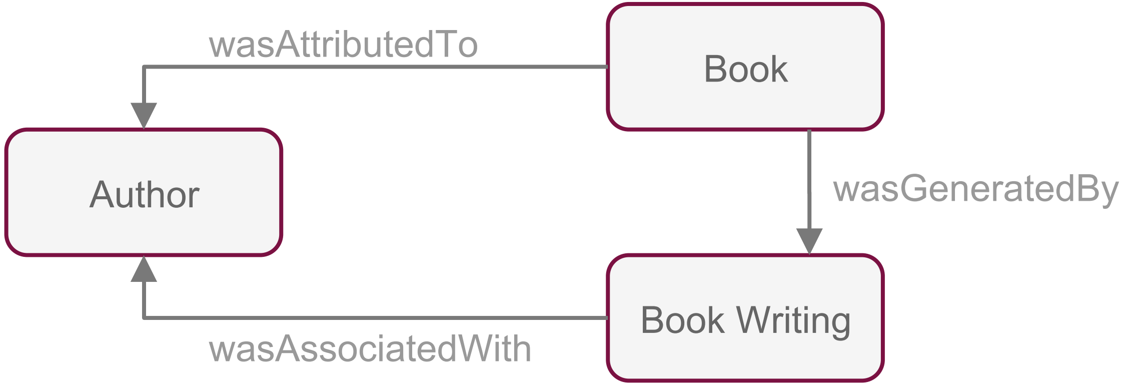
W3C has published the 'PROV' family of recommendations



See w3.org/TR/prov-o/

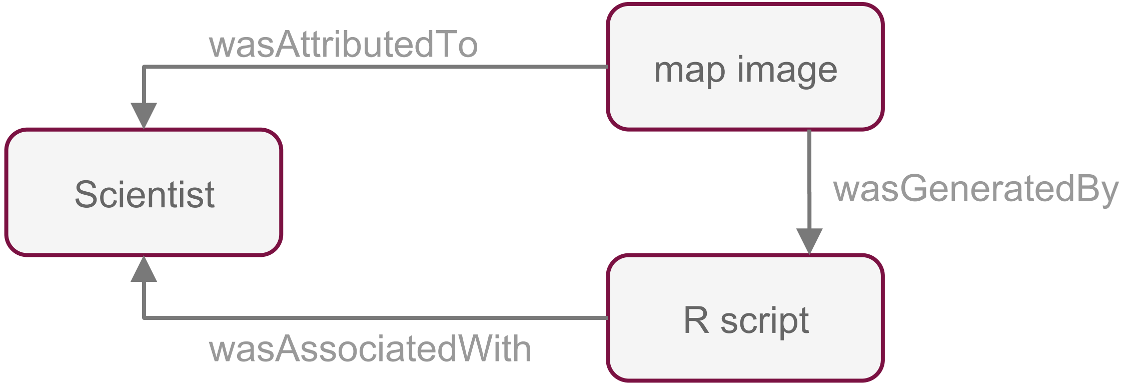
Using a common model

Example: Book writing activity



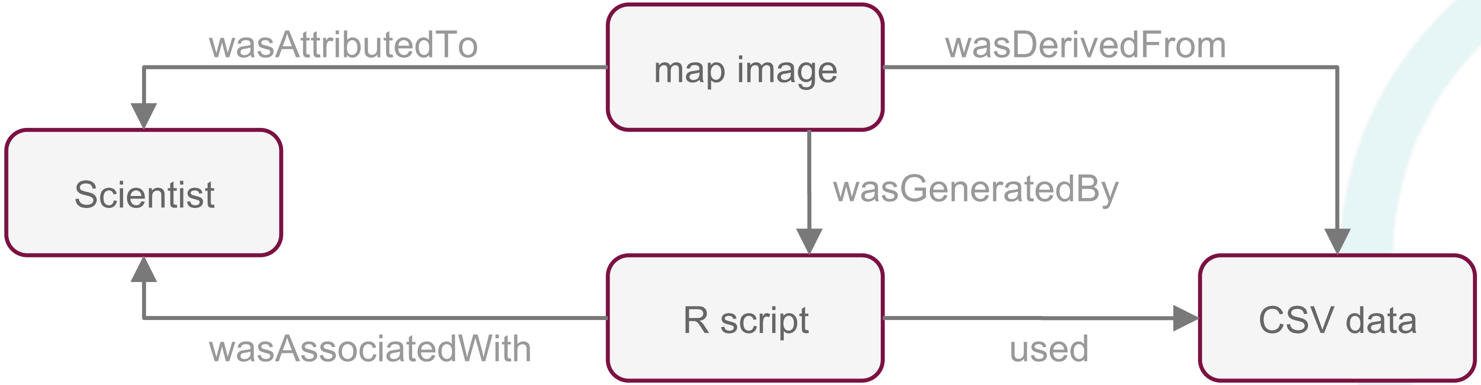
Using a common model

Example: Scientific workflow



Using a common model

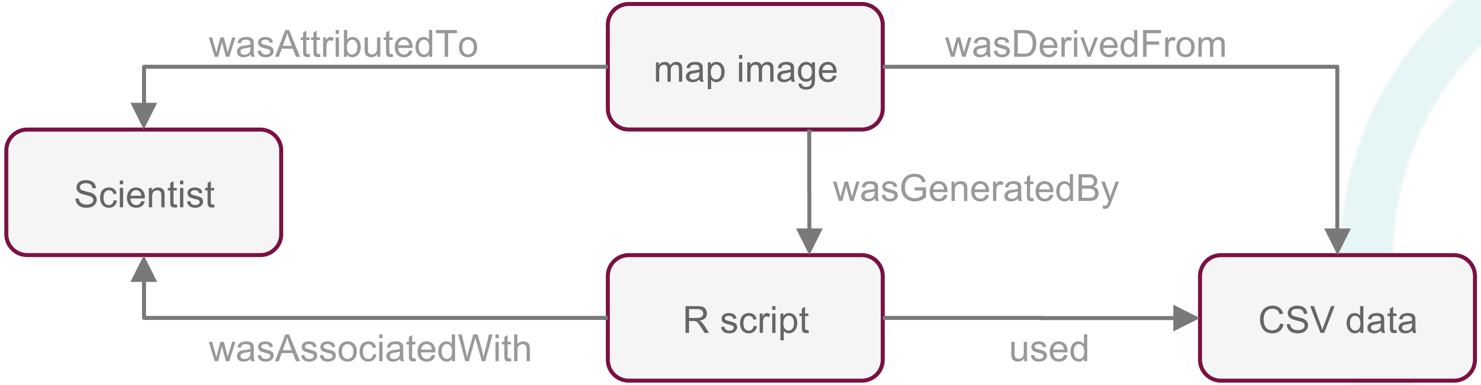
Example: Scientific workflow



Using a common model

Example: Scientific workflow

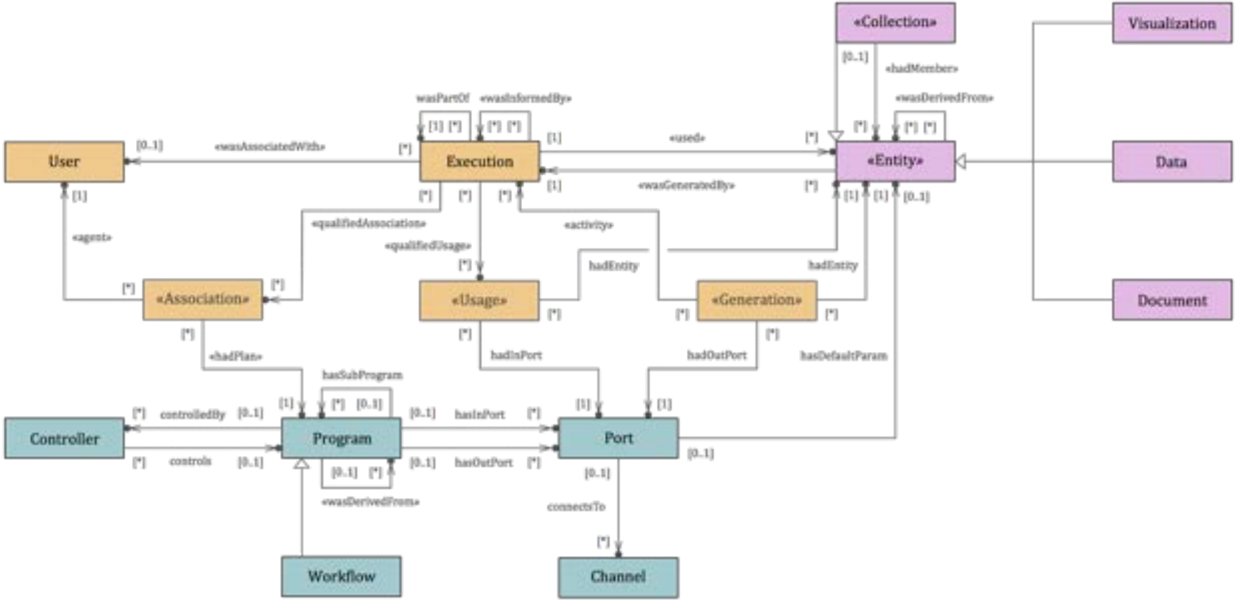
< “map image” wasDerivedFrom “CSV data” >



Using a common model

- W3C PROV model : very generic, universal
- Tracking provenance in scientific workflows requires specialization of PROV
- The ProvONE model extends PROV to provide this

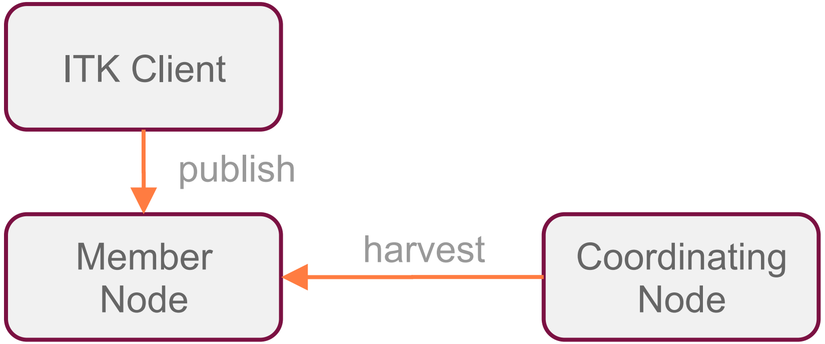
ProvONE builds on W3C PROV

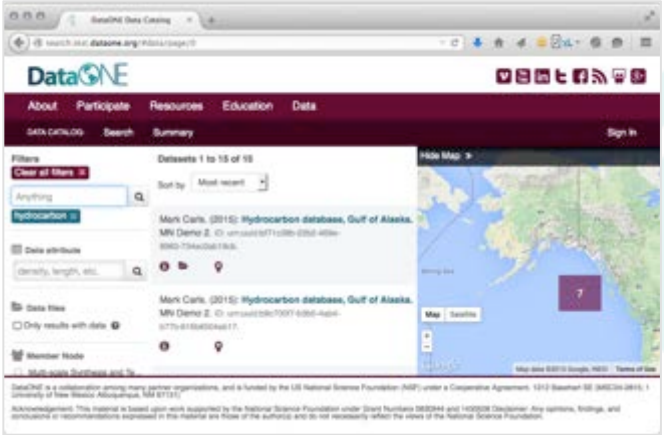


Using provenance information for search

Facilitating search

DataONE harvests provenance information and indexes it





Datasets 1 to 15 of 15

□ Only results with data

Mark Carls. (2015): Hydrocarbon database, Gulf of Alaska. MN Demo 2. ID: [illegible]

Mark Carls. (2015): Hydrocarbon database, Gulf of Alaska. MN Demo 2. ID: [illegible]

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.