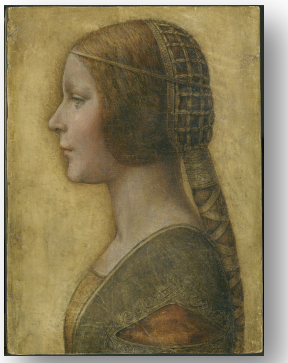
Theory & Practice of Data Cleaning

What is provenance and why is it important?

Example:The Fine Arts

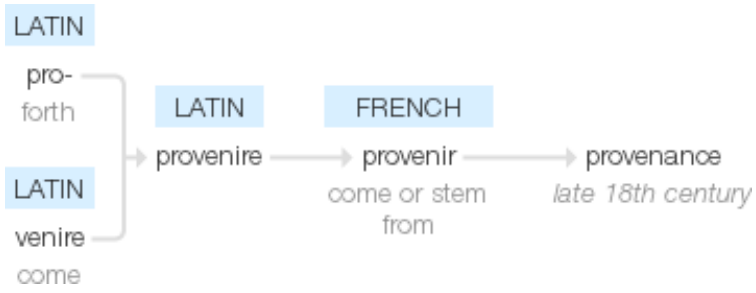


- One of these is has been sold for nearly $180 million.
- The other *could* be worth as much or more.
- Which is which?
- What is the difference?

Provenance defined …

- **Oxford English Dictionary**
  - The **place of origin** or earliest **known history** of something:
    - *an orange rug of Iranian **provenance***
  - The **beginning** of something's existence; its **origin**:
    - *they try to understand the whole universe, its **provenance** and fate*
  - A **record of ownership** of a work of art or an antique, used as a guide to **authenticity** or **quality**:
    - *the manuscript has a distinguished **provenance***

- **What is the origin (provenance!) of "provenance" ?**

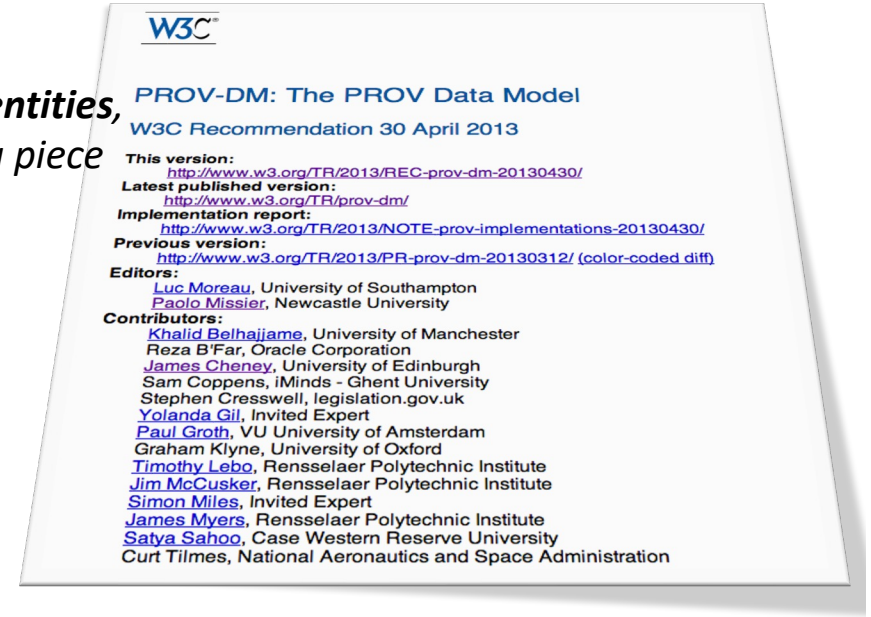


What is “provenance” (sensu W3C) ?

- *Provenance is a description of **how things came to be**, and how they came to be **in the state they are in today** (*)*

- *Provenance refers to the **sources** of information, including **entities** and **processes**, **involved in producing or delivering an artifact** (*)*

- *Provenance is a **record** that describes the **people**, **institutions**, **entities**, and **activities**, involved in producing, influencing, or delivering a piece of data or a thing in the world*

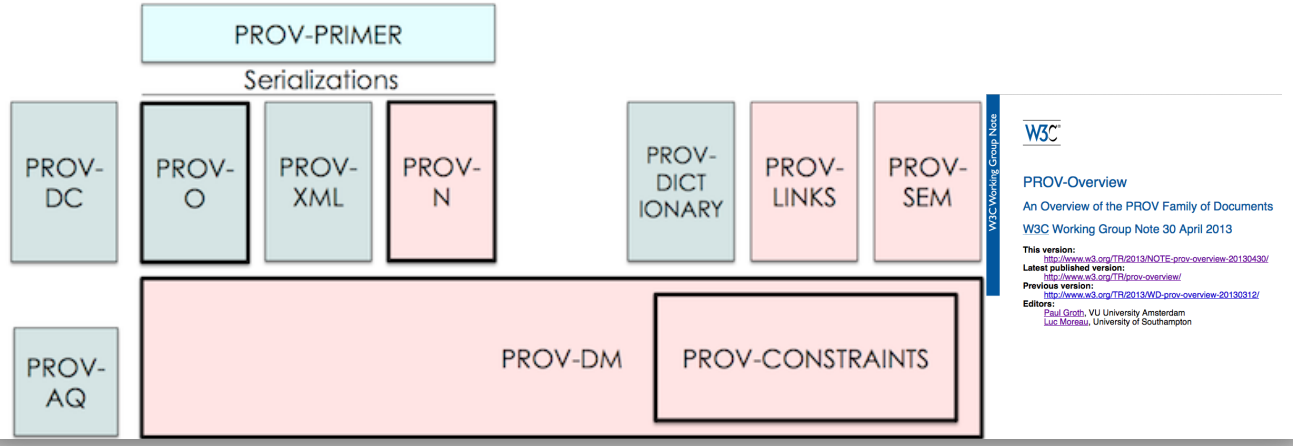

PROV-DM: The PROV Data Model

W3C Recommendation 30 April 2013

This version:
http://www.w3.org/TR/2013/REC-prov-dm-20130430/
Latest published version:
http://www.w3.org/TR/prov-dm/
Implementation report:
http://www.w3.org/TR/2013/NOTE-prov-implementations-20130430/
Previous version:
http://www.w3.org/TR/2013/PR-prov-dm-20130312/ (color-coded diff)
Editors:
Luc Moreau, University of Southampton
Paolo Missier, Newcastle University
Contributors:
Khalid Belhajjame, University of Manchester
Reza B'Far, Oracle Corporation
James Cheney, University of Edinburgh
Sam Coppens, iMinds - Ghent University
Stephen Cresswell, legislation.gov.uk
Yolanda Gil, Invited Expert
Paul Groth, VU University of Amsterdam
Graham Klyne, University of Oxford
Timothy Lebo, Rensselaer Polytechnic Institute
Jim McCusker, Rensselaer Polytechnic Institute
Simon Miles, Invited Expert
James Myers, Rensselaer Polytechnic Institute
Satya Sahoo, Case Western Reserve University
Curt Tilmes, National Aeronautics and Space Administration

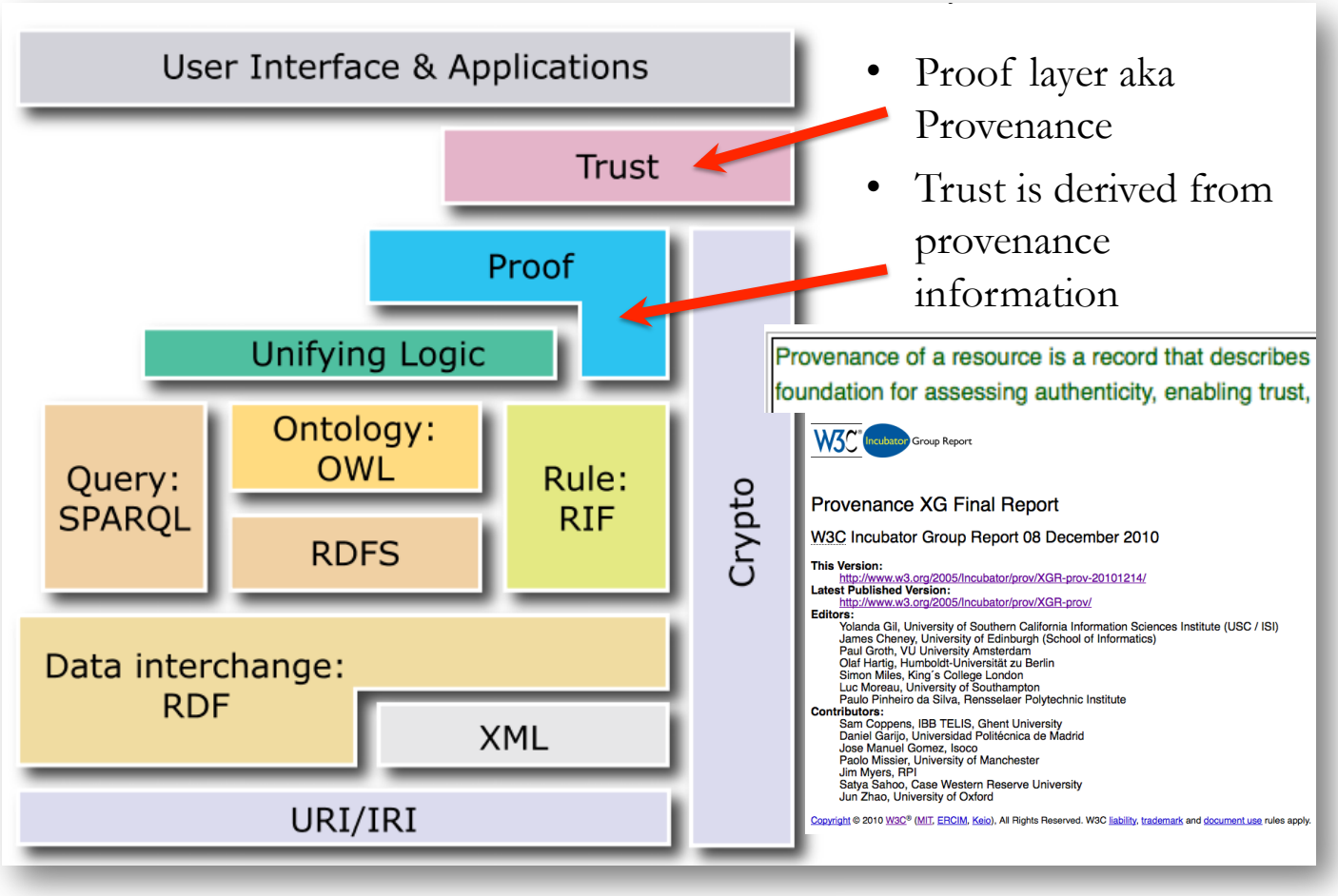
W3C PROV Family of Specifications: Modeling

- W3C Recommendations
  - PROV Data Model (PROV-DM)
  - PROV Ontology (PROV-O)
  - PROV-Constraints
  - PROV Notations (PROV-N)
- PROV Working Group Notes (selected)
  - PROV-Access and Querying (AQ)
  - PROV Dictionary
  - PROV XML
  - PROV and Dublin Core Mappings (PROV-DC)
  - PROV Semantics (using first-order logic) (PROV-SEM)



Provenance Analysis and RDF Query Processing, Satya S. Sahoo, Praveen Rao, ISWC, October, 2015.

Provenance & Semantic Web Layer Cake



Provenance Analysis and RDF Query Processing, Satya S. Sahoo, Praveen Rao, ISWC, October, 2015.

Back to the basics: Open Provenance Model (OPM) => W3C Prov

The Open Provenance Model aims to capture the causal dependencies between the artifacts, processes, and agents. Therefore, a provenance graph is defined as a directed graph, whose nodes are artifacts, processes and agents, and whose edges belong to one of the following categories depicted in Figure 1. An edge represents a causal dependency, between its source, denoting the effect, and its destination, denoting the cause.

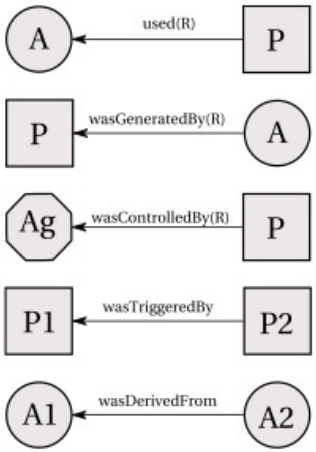


Figure 1: Edges in the Open Provenance Model: sources are effects, and destinations causes

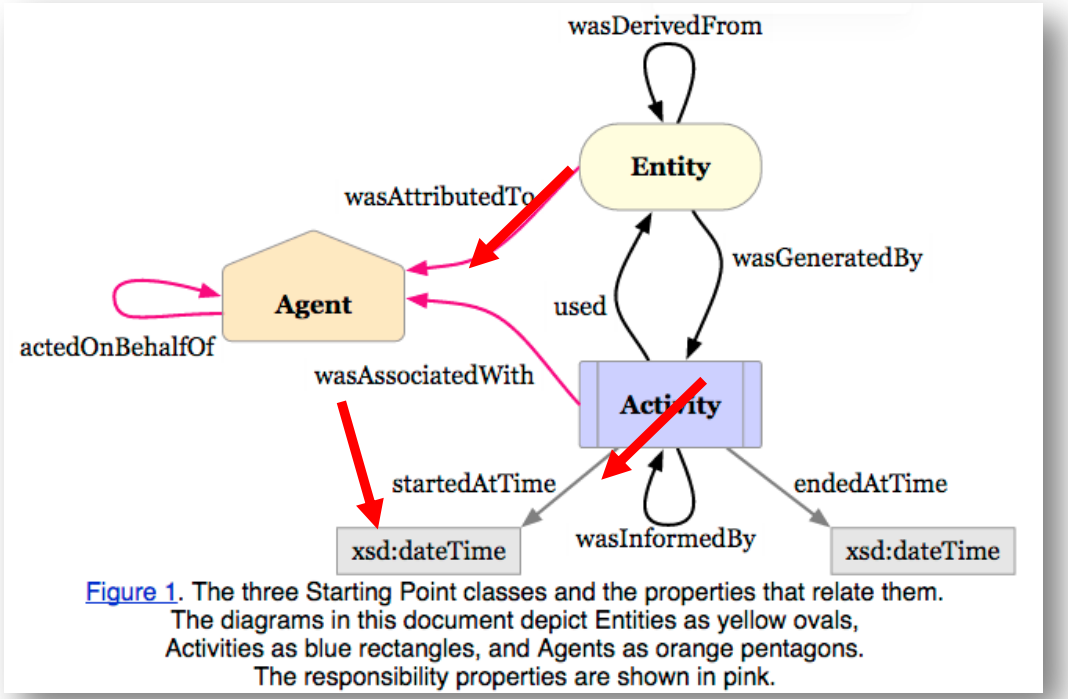


Figure 1. The three Starting Point classes and the properties that relate them. The diagrams in this document depict Entities as yellow ovals, Activities as blue rectangles, and Agents as orange pentagons. The responsibility properties are shown in pink.

Readings: Moreau, Luc, Ben Clifford, Juliana Freire, Joe Futrelle, Yolanda Gil, Paul Groth, Natalia Kwasnikowska et al. "The open provenance model core specification (v1. 1)." Future generation computer systems 27, no. 6 (2011): 743-756.

W3C Prov: some finer points

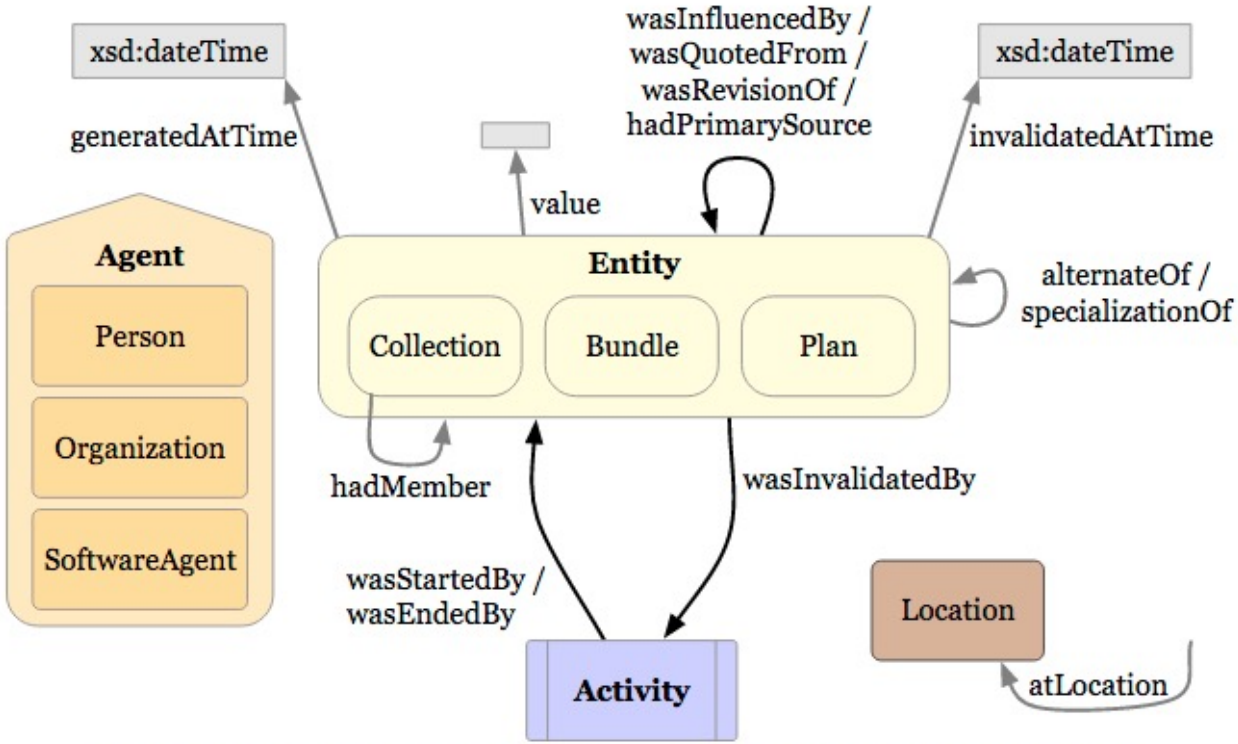
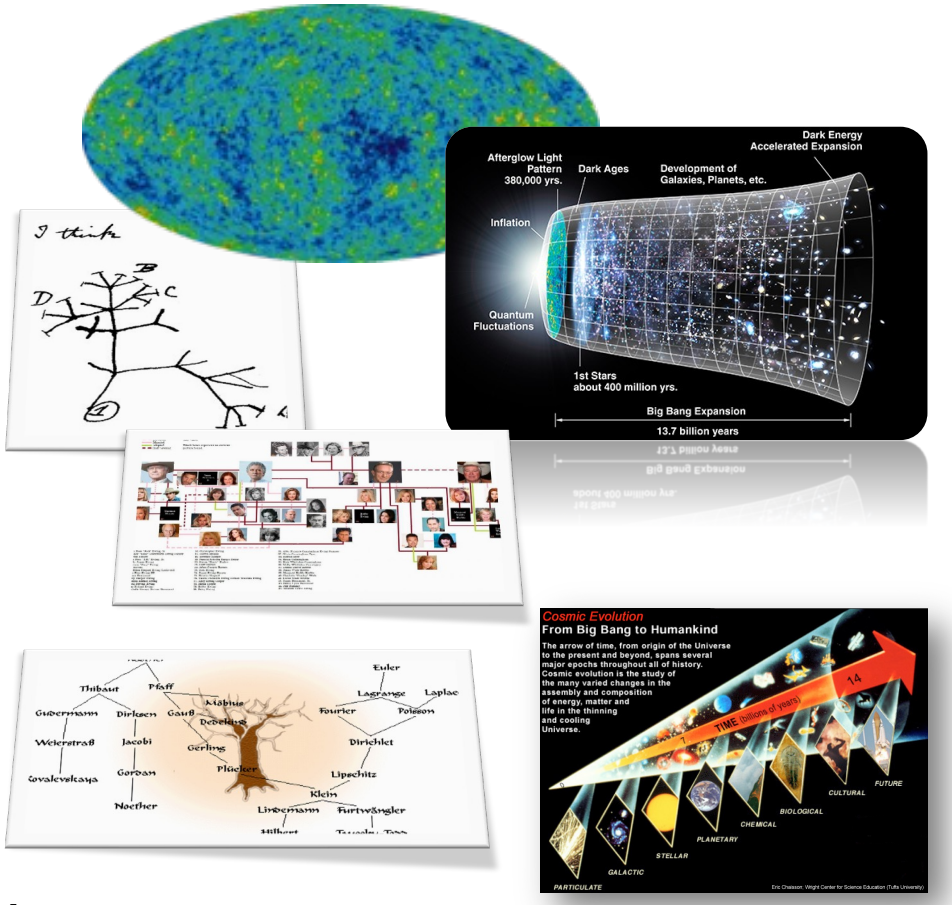


Figure 3. The expanded terms build upon those in the Starting Points section.

Provenance-in-Science Palooza

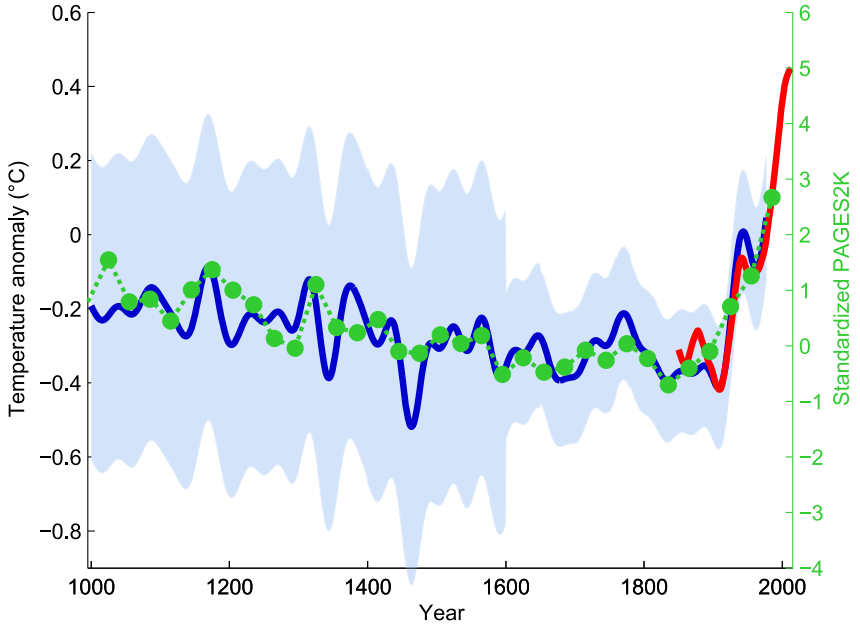
- What are those?
- *Cosmology*
- *Geology, Stratigraphy*
- *Phylogeny*
  - the *Tree of Life*
- *Genealogy*
  - your family: literally
- *Academic Pedigree*
  - *"Doktorvater" (oder Doktor-Mutter)*
- *Etymology*
- *Chain of custody*
  - of art(ifacts)
- Yes: all about **origins** and **history**

  …



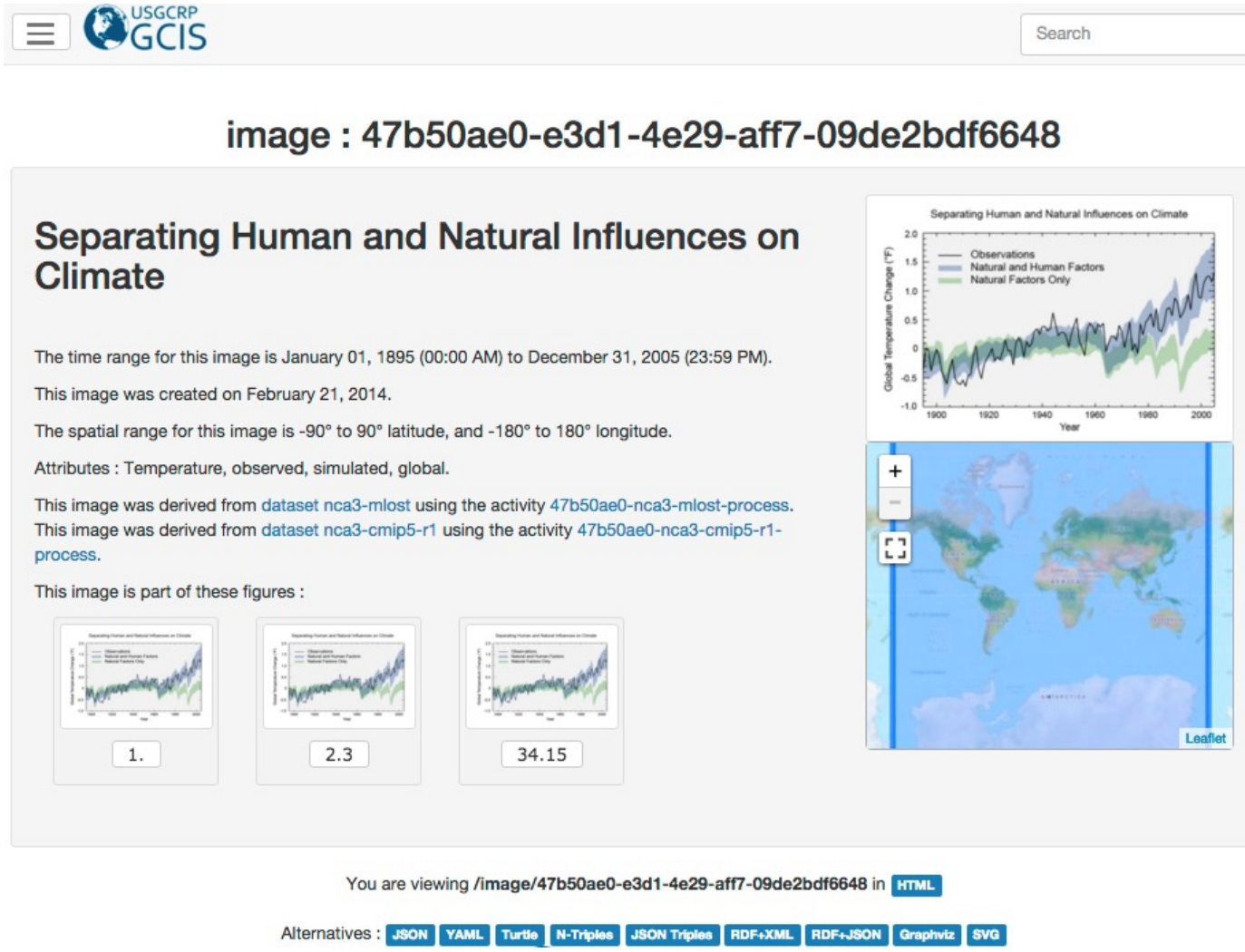
Using Provenance for *Transparency, Reproducibility*

- What ***input data*** went into this study?
- What ***methods*** were used?
  - … with what ***parameter*** settings, ***calibrations***, …?
- *Can we **trust** the data and methods?*

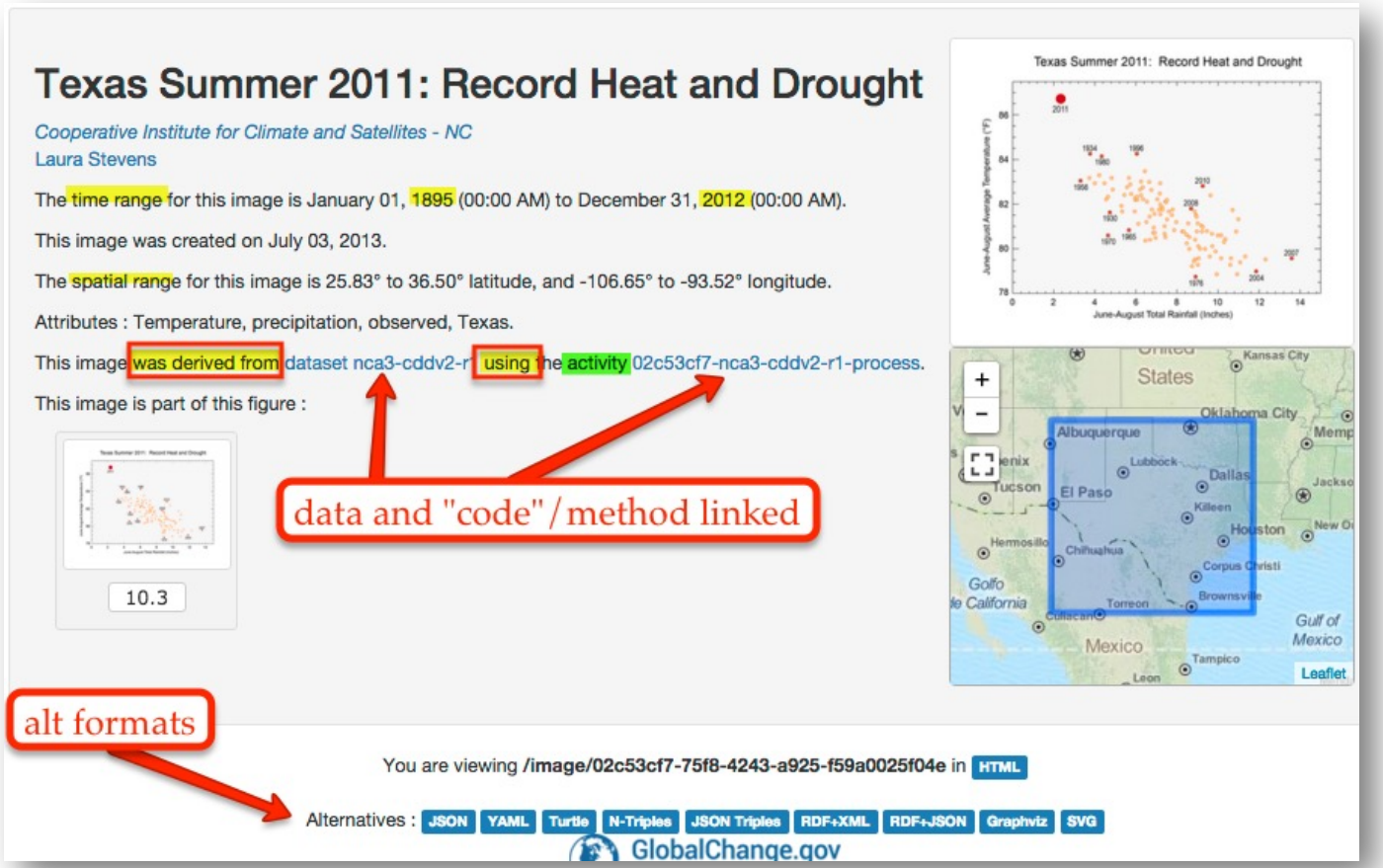


- **Provenance** (*lineage*): track **origin** and **processing history** of data ➔ trust, data quality ~ audit trail for attribution, credit
- **Discovery** of data, methodologies, experiments

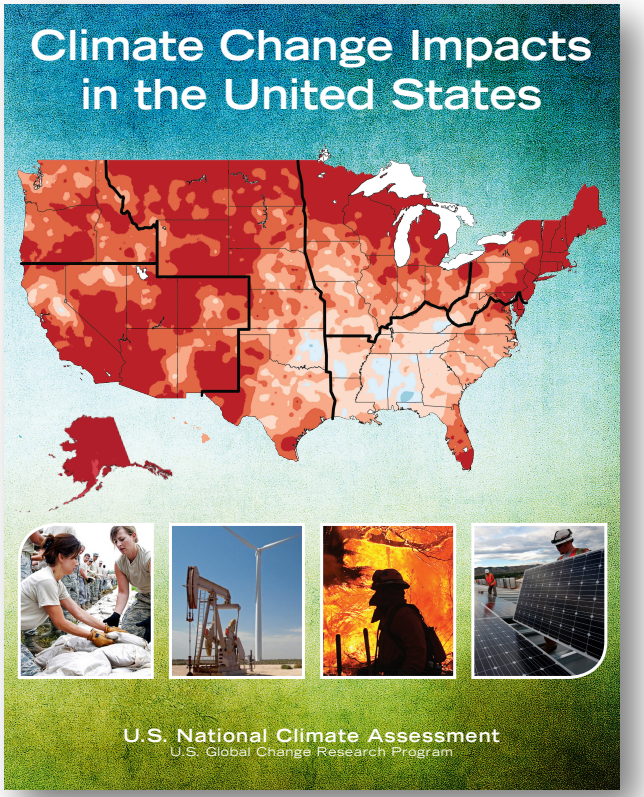
Climate Change: Whodunnit?



Tracing the sources (data, code)



Provenance today: Important but *hard*

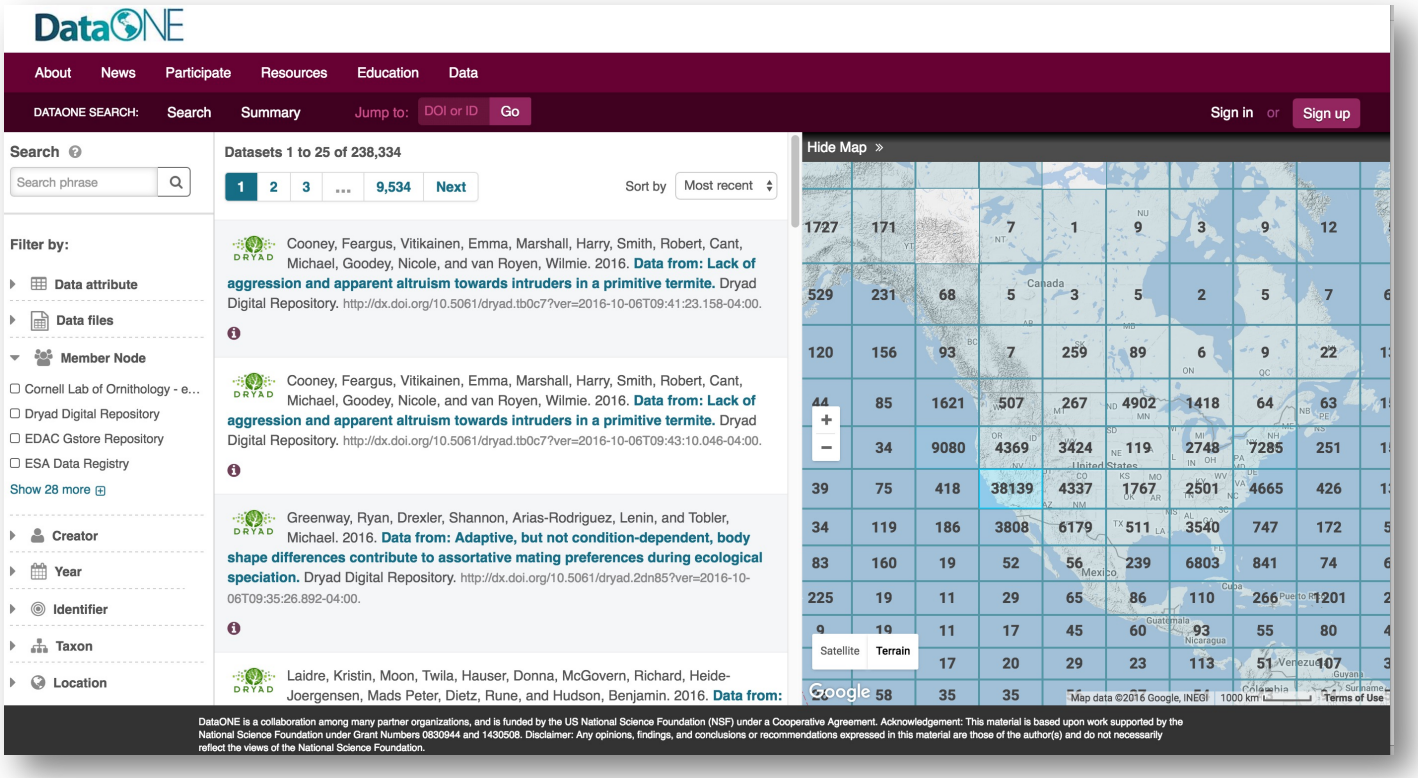


"This report is the result of a **three-year** analytical effort by a team of **over 300 experts**, overseen by a broadly constituted Federal Advisory Committee of **60 members**. It was developed from information and analyses gathered in over 70 workshops and listening sessions held across the country."

➔many research projects, groups conduct R&D on provenance methods, tools, …

Example: DataONE

A scientific data federation: DataONE
Data Observation Network for Earth

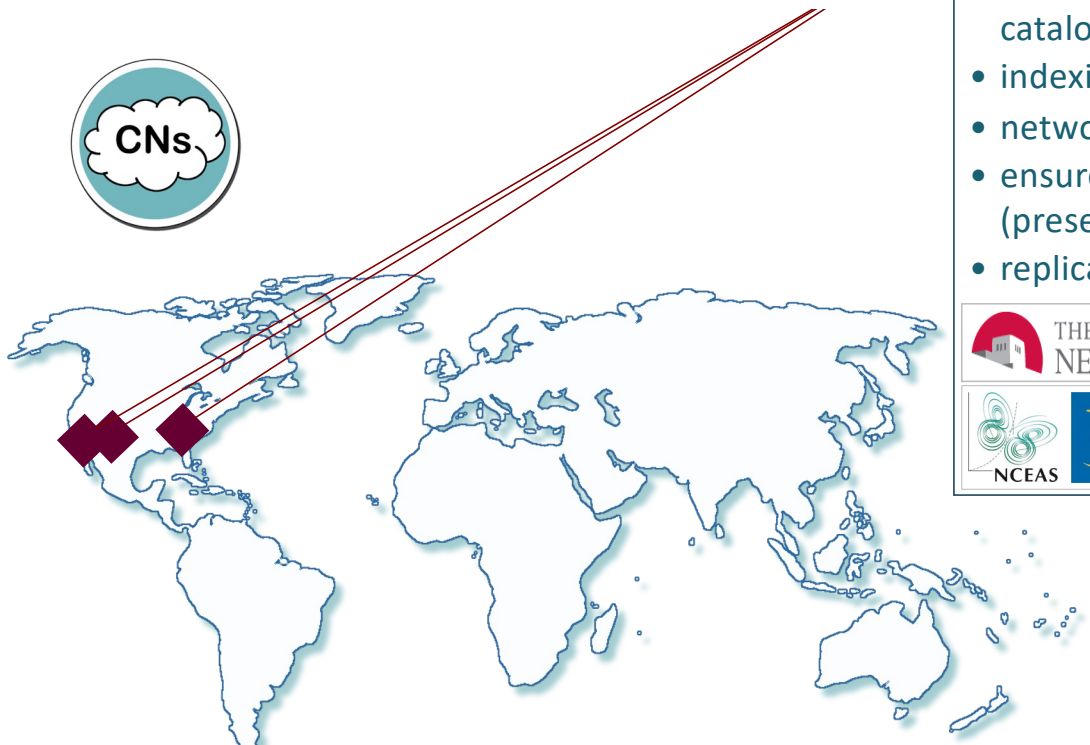


DataONESearch

search.dataone.org

DataONE Cyberinfrastucture:
Coordinating Nodes

Components for a flexible, scalable, sustainable network



Coordinating Nodes

- retain complete metadata catalog
- indexing for search
- network-wide services
- ensure content availability (preservation)
- replication services

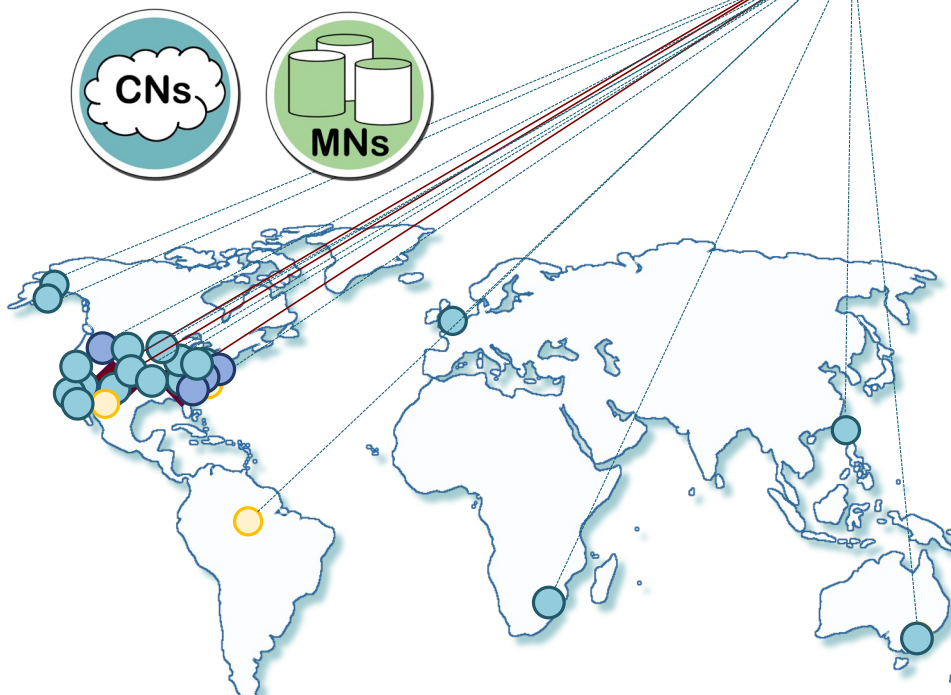
THE UNIVERSITY of NEW MEXICO
OAK RIDGE National Laboratory
NCEAS
UCSB

DataONE

www.dataone.org/coordinating-nodes

DataONE Cyberinfrastructure:
Member Nodes

Components for a flexible, scalable, sustainable network



Coordinating Nodes

Member Nodes

- diverse institutions
- serve local community
- provide resources for managing their data
- retain copies of data

DataONE

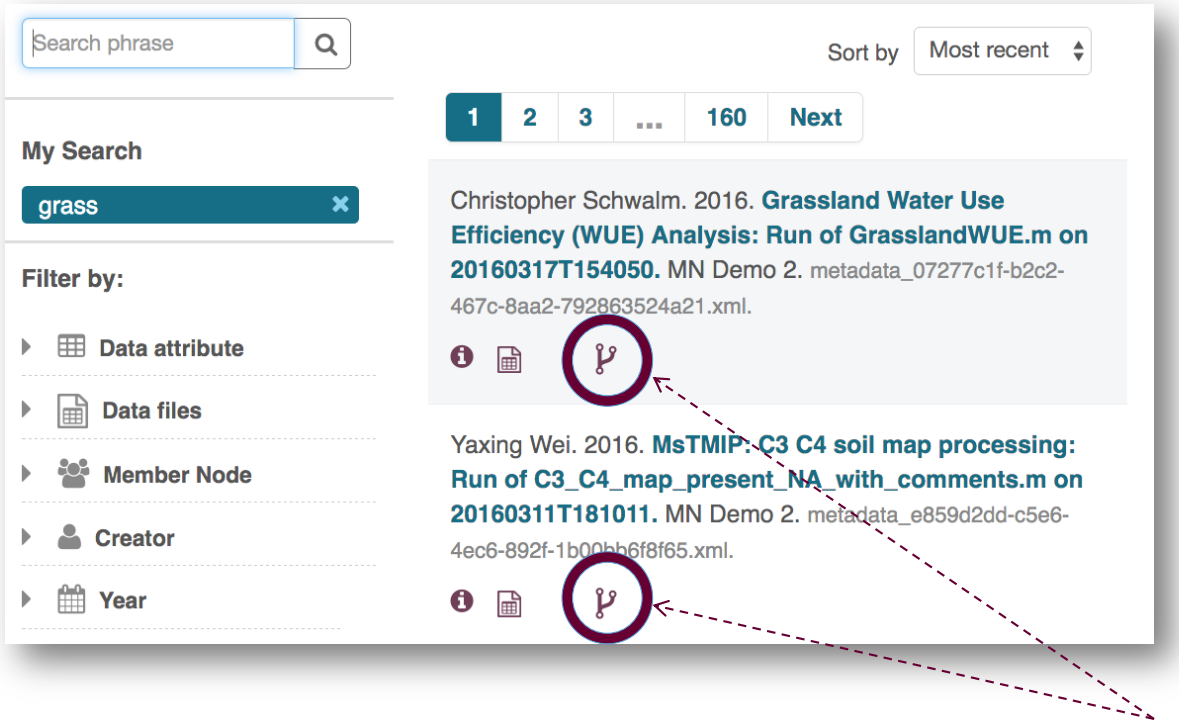
www.dataone.org/member-nodes

DataONE Cyberinfrastructure:
Investigator Toolkit

Components for a flexible, scalable, sustainable network

www.dataone.org/investigator-toolkit

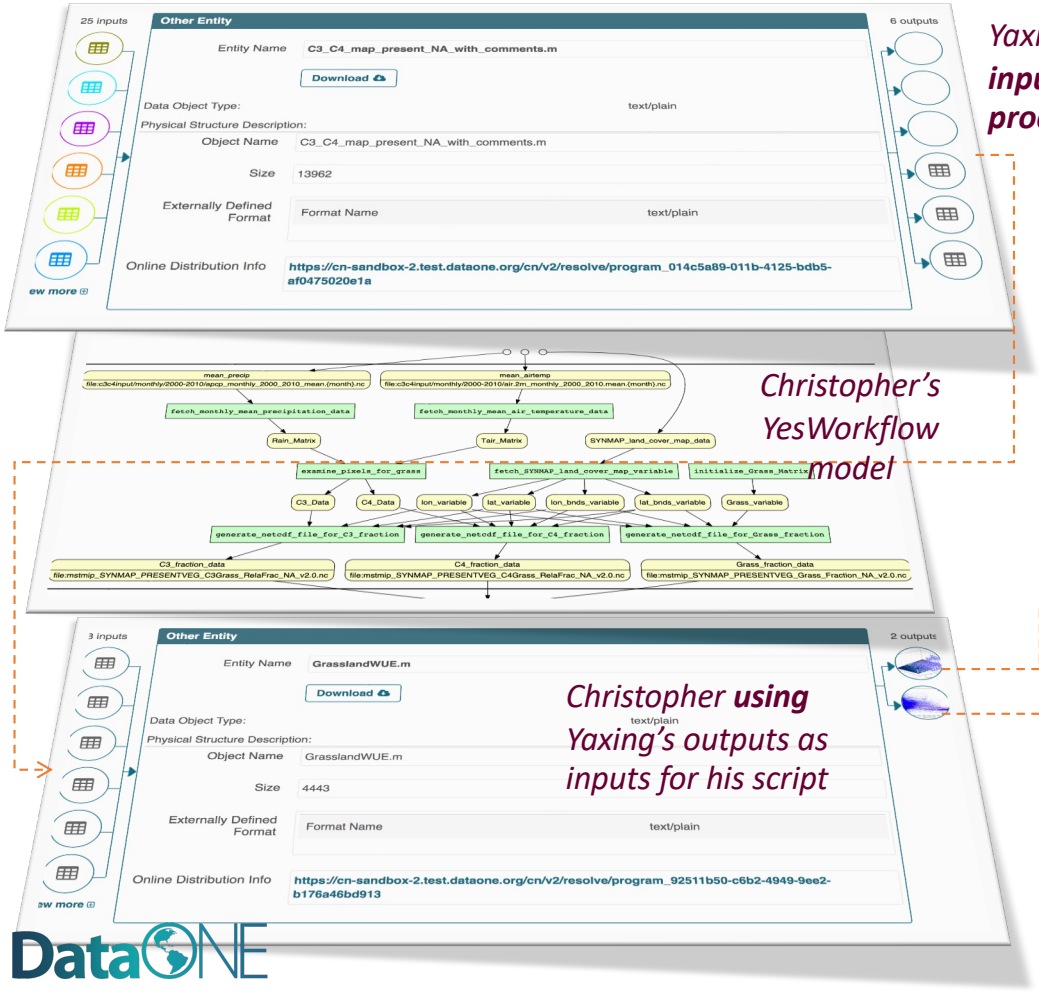
Provenance in Action: Benefits & Impact



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

DataONE

DataONE: Support for Provenance

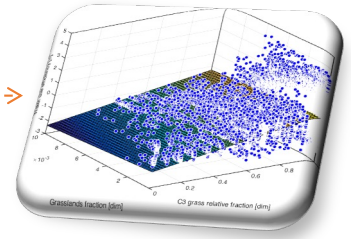


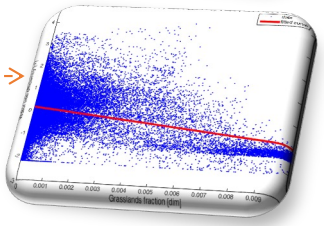
*Yaxing's **script** with **inputs** & **output products***

Christopher's YesWorkflow model

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

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





Theory & Practice of Data Cleaning

YesWorkflow: Provenance for script-based workflows

From Workflows & Provenance to Provenance for Script-based Workflows …

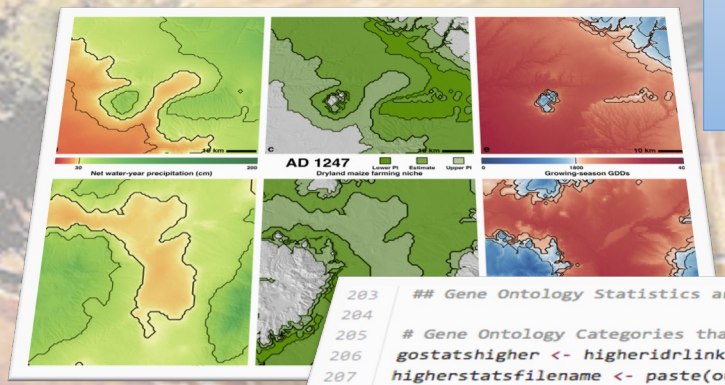
- What workflow tools are (most) scientists using?
  - Workflow systems
  - … vs scripts (Python, R, MATLAB, …)

- What provenance tools are their?
  - Workflow system support
  - Tools for “workflow” scripts!?

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

```
## 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)))
```

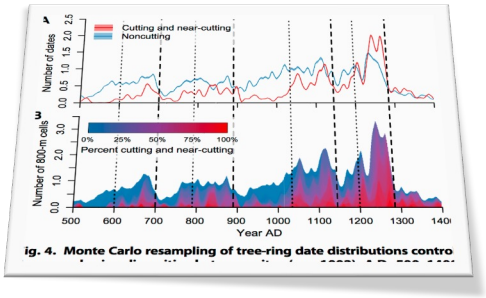
… implemented as an R Script …

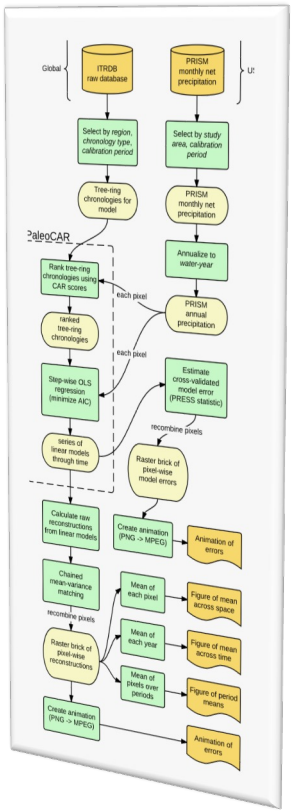
Painting by Roy Andersen

Provenance Support for Reproducible Science Example: Paleoclimate Reconstruction

Science paper (OA) uses:

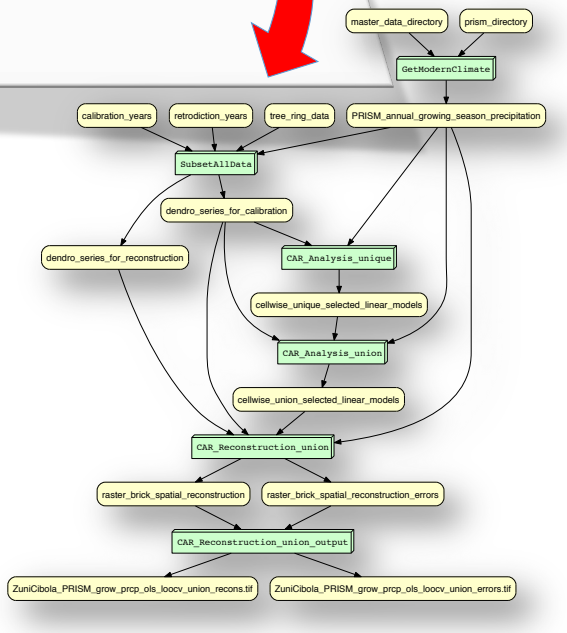
- open source code:
  - R, PaleoCAR, …
- Is that all we need?
- What was the “workflow”?
- Is there prospective and/or retrospective provenance?





YesWorkflow:
Yes, scripts are workflows, too!

```
203  ## Gene Ontology Statistics are Calculated Here.
204
205   # Gene Ontology Categories that were shown to be relatively Higher (more expressed) in the Experimental Condition.
206   gostatshigher <- higheridrlinkedtogenes[1]
207   higherstatsfilename <- paste(outputDirectory, "/", runName, "_", conditions[1],"_GOStatsHigher_",mytestcond[1], "_v
208   write.table(gostatshigher,file=higherstatsfilename, row.names=FALSE, col.names=FALSE, quote=FALSE, sep="\t")
209   geneListHigherCHR <- gostatshigher$SYMBOL
210   geneListHigherLinkedtoEntrezIds <- select(hgu133plus2.db, keys= geneListHigherCHR, "ENTREZID", "SYMBOL")
211   GOstatsGenesH <- geneListHigherLinkedtoEntrezIds[,2]
212
213  x <- org.Hs.egACCNUM
214  mapped_genes <- mappedkeys(x)
215  xx <- as.list(x[mapped_genes])
216 geneUniverse <- (unique(names(xx)))
```



- **Script** *vs* Workflows/**ASAP**:
  - **A**utomation: *****
  - **S**caling: **
  - **A**bstraction: *
  - **P**rovenance: **

YW annotations: Model your Workflow!

```
1  # @BEGIN collect_data_set
2  # @PARAM cassette_id @PARAM accepted_sample @PARAM num_images @PARAM energies
3  # @OUT sample_id @OUT energy @OUT frame_number
4  # @OUT raw_image_path @AS raw_image
5  # ... @URI file:run/raw/{cassette_id}/{sample_id}/e{energy}/image_{frame_number}.raw
6  run_log.write("Collecting data set for sample {0}".format(accepted_sample))
7  sample_id = accepted_sample
8  for energy, frame_number, intensity, raw_image_path in collect_next_image(
9                          cassette_id, sample_id, num_images, energies,
10                         "run/raw/{cassette_id}/{sample_id}/e{energy}/image_{frame_number:03d}.raw"):
11     run_log.write("Collecting image {0}".format(raw_image_path))
12 # @END collect_data_set
13
14 # @BEGIN transform_images
15 # @PARAM sample_id @PARAM energy @PARAM frame_number
16 # @IN raw_image_path @AS raw_image
17 # @IN calibration_image @URI file:calibration.img
18 # @OUT corrected_image @URI file:run/data/{sample_id}/{sample_id}_{energy}eV_{frame_number}.img
19 # @OUT corrected_image_path @OUT total_intensity @OUT pixel_count
20     corrected_image_path = "run/data/{0}/{0}_{1}eV_{2:03d}.img".format(sample_id, energy, frame_number)
21     (total_intensity, pixel_count) = transform_image(raw_image_path, corrected_image_path, "calibration.img")
22     run_log.write("Wrote transformed image {0}".format(corrected_image_path))
23 # @END transform_images
```

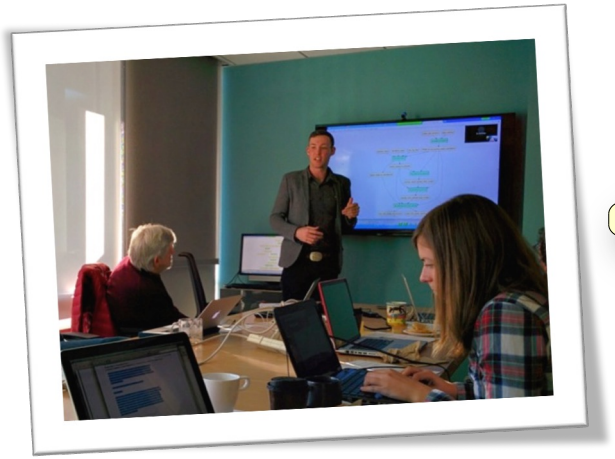

Figure 1. YW-annotated fragment of a Python script for data collection from protein crystal samples. YW-annotations `@BEGIN` and `@END` delimit code blocks; `@IN` and `@OUT` tags model relevant input and output data elements of a block; `@PARAM` identifies a block's parameters. `@URI` templates for raw images (line 5) and corrected images (line 18) link conceptual-level data elements such as `raw_image` with runtime resources (data files and their file paths). Executable script code is greyed out to emphasize YW-annotations. A program variable (`raw_image_path`) is highlighted in the code (lines 8, 11, 21): aliases (lines 4, 16) are used to link such program-level objects to the scientist's concepts (here: `raw_image`). Full example available from [MBL15].

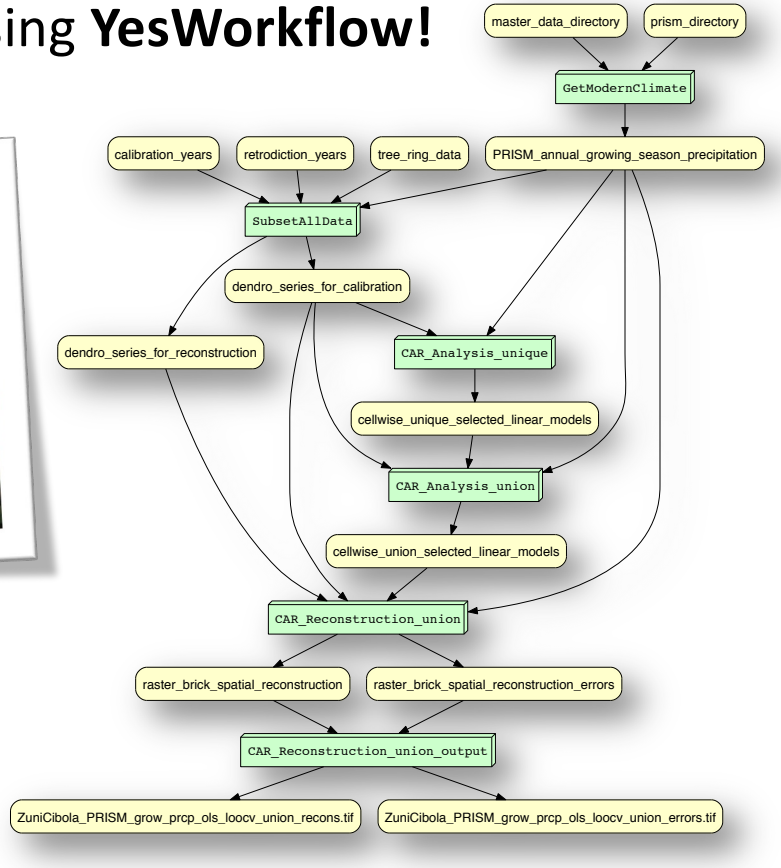
Paleoclimate Reconstruction (EnviRecon.org) …

… explained using **YesWorkflow!**

Kyle B., (computational) archaeologist:

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

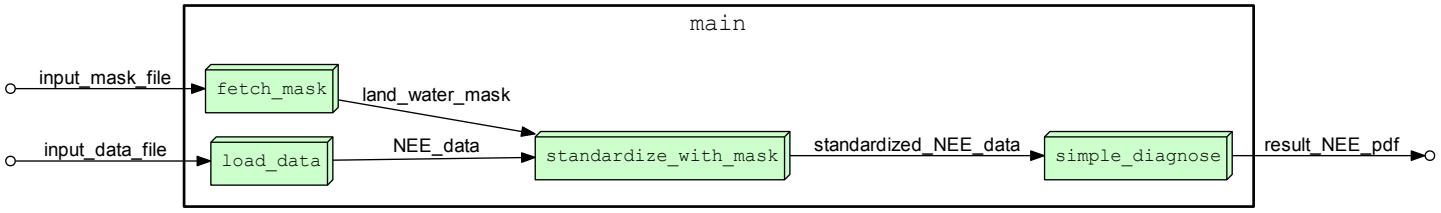




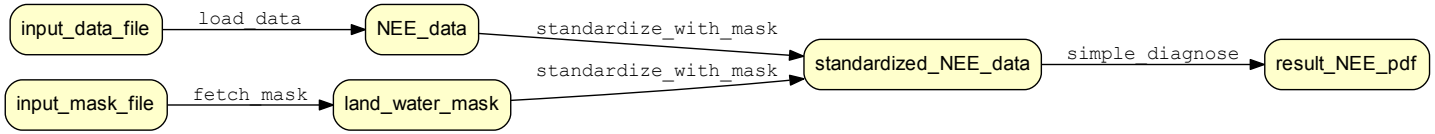
SKOPE + Kurator + DataONE (Data Observation Network for Earth)

=> YesWorkflow.org

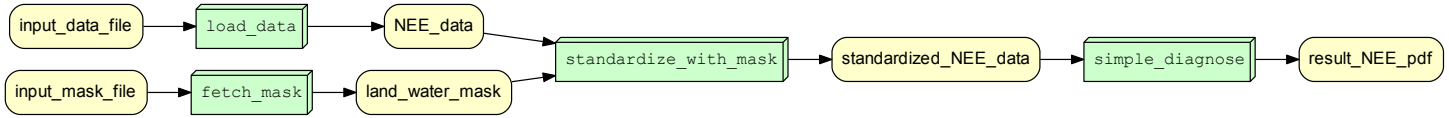
Get 3 views for the price of 1!



Process view



Data view

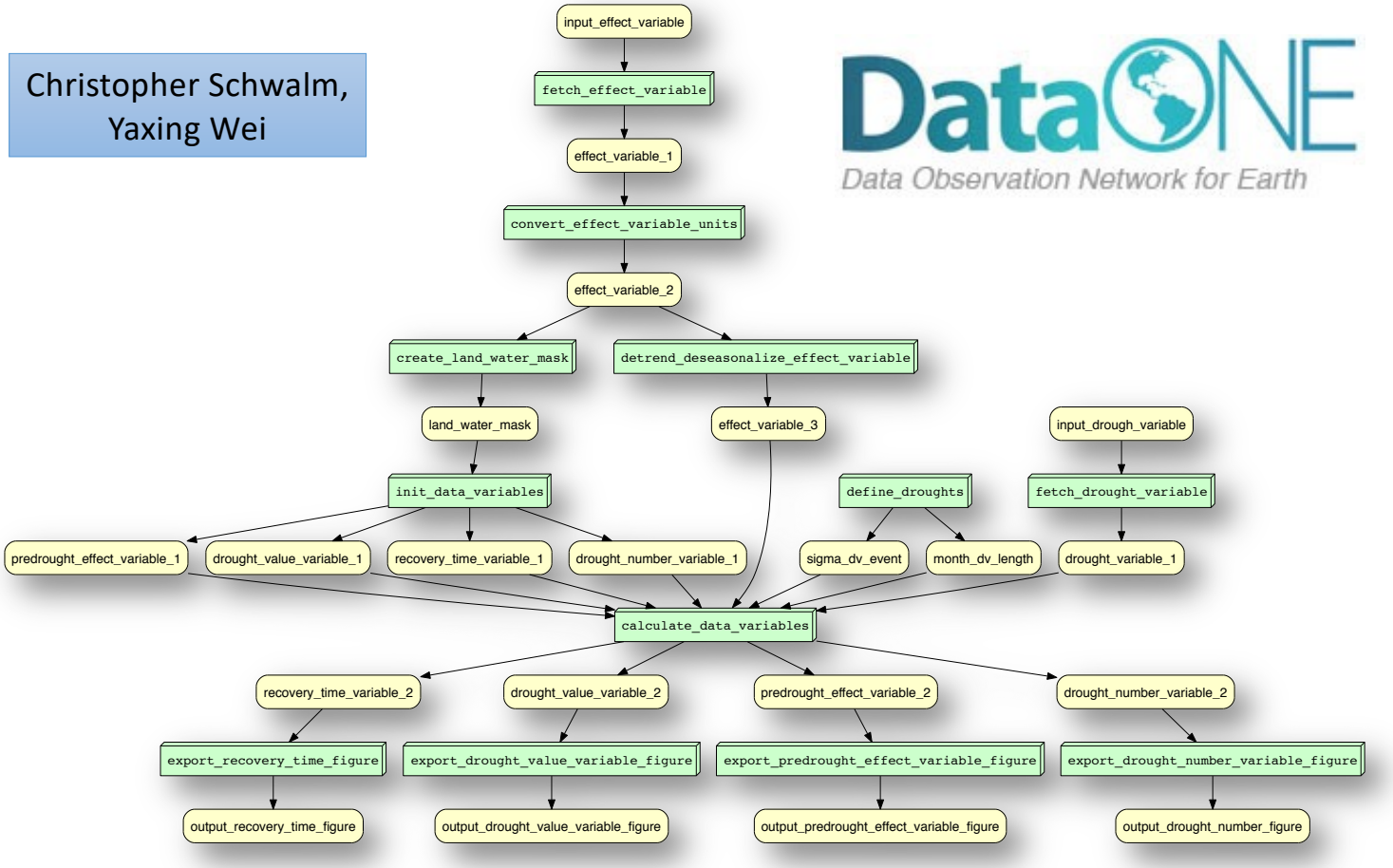


Combined view

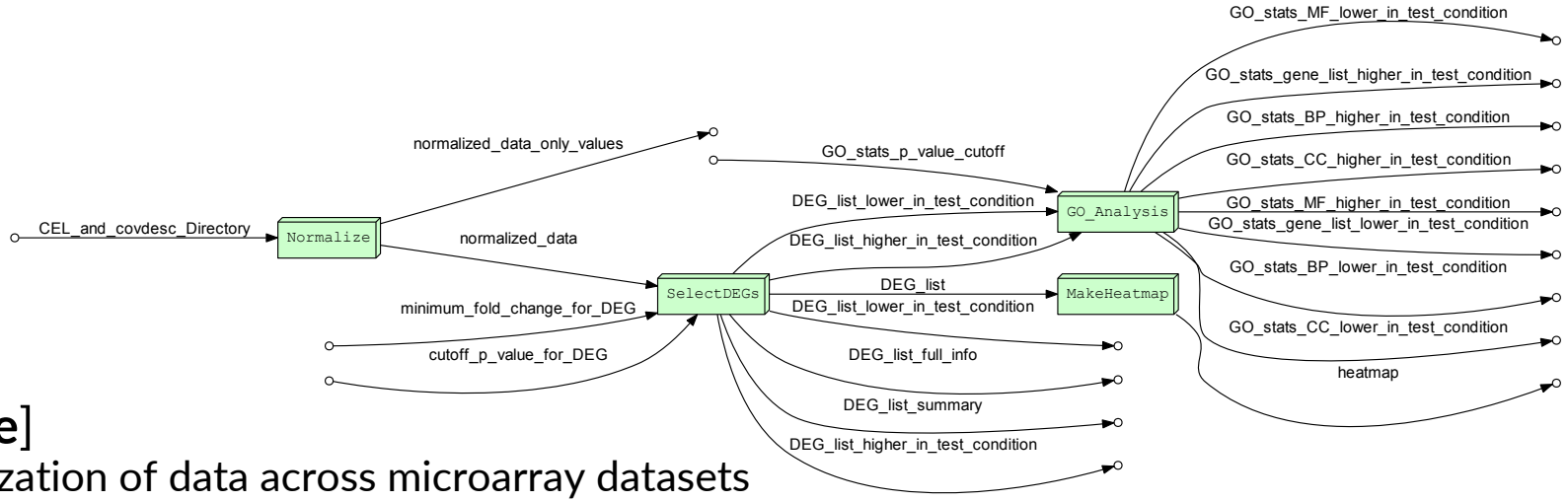
Multi-Scale Synthesis and Terrestrial Model Intercomparison Project (MsTMIP)

Christopher Schwalm,
Yaxing Wei

DataONE
Data Observation Network for Earth



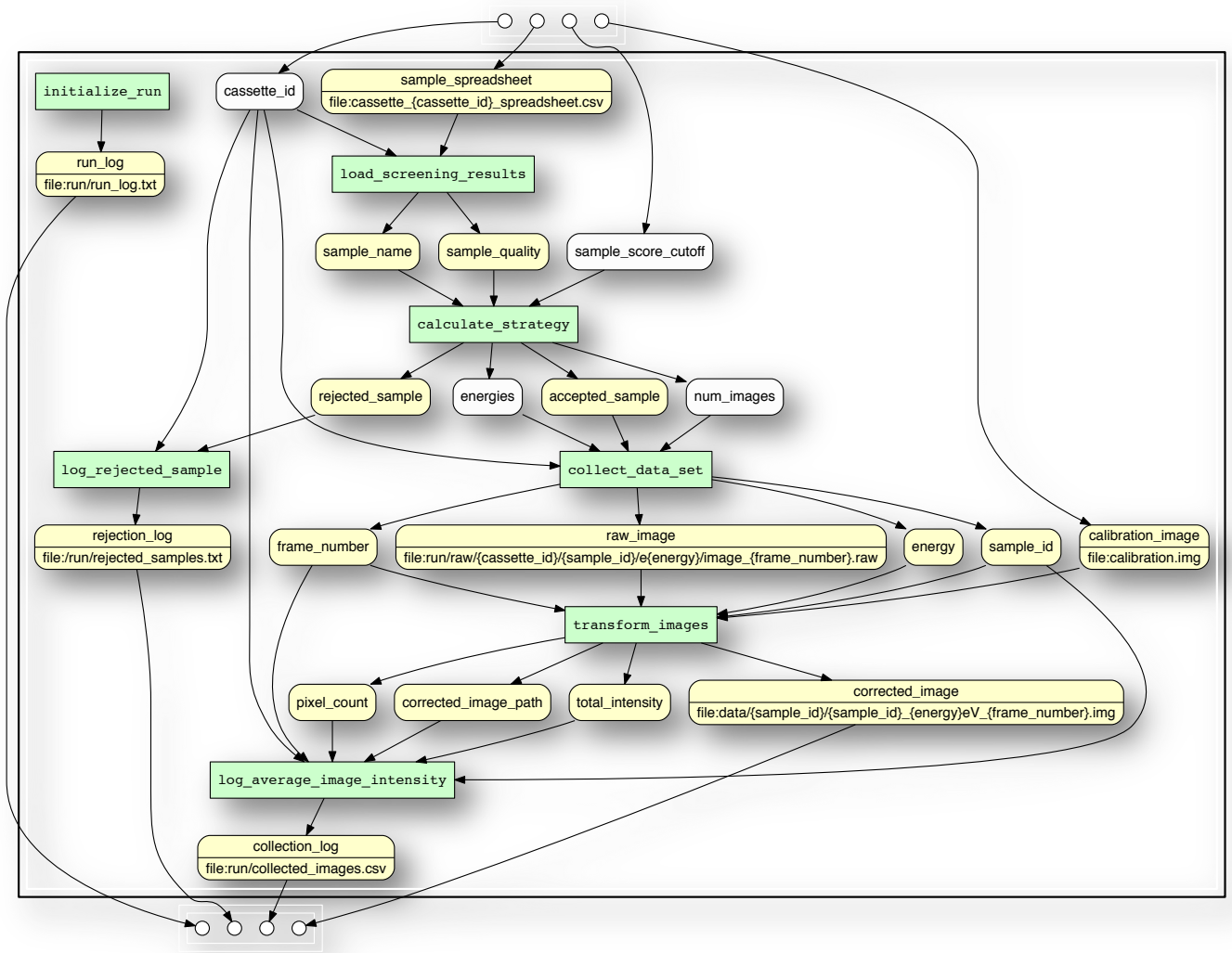
Gene Expression Microarray Data Analysis



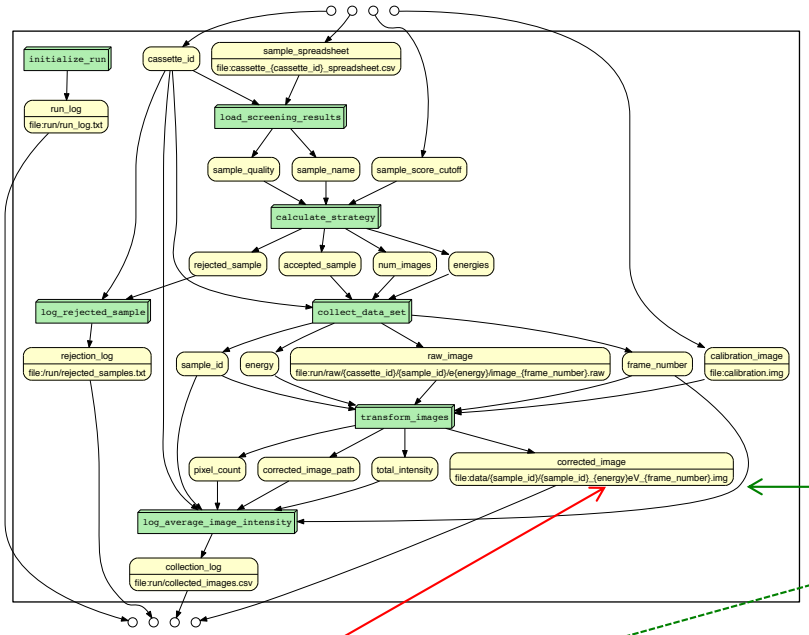
- [**Normalize**]
  - Normalization of data across microarray datasets
- [**SelectDEGs**]
  - Selection of differentially expressed genes between conditions
- [**GO Analysis**]
  - determination of gene ontology statistics for the resulting datasets
- [**MakeHeatmap**]
  - creation of a heatmap of the differentially expressed genes

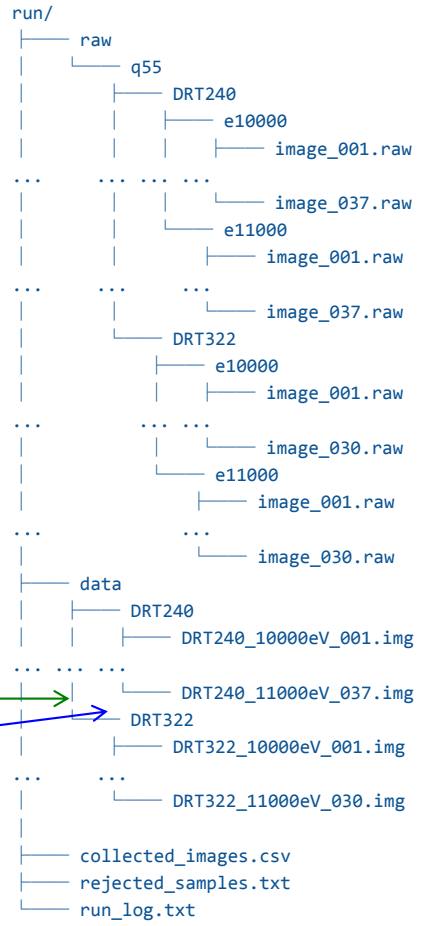
Tyler Kolisnik, Mark Bieda

Data collection workflow (X-ray diffraction)



YW-RECON: Prospective & Retrospective Provenance … (almost) for free!



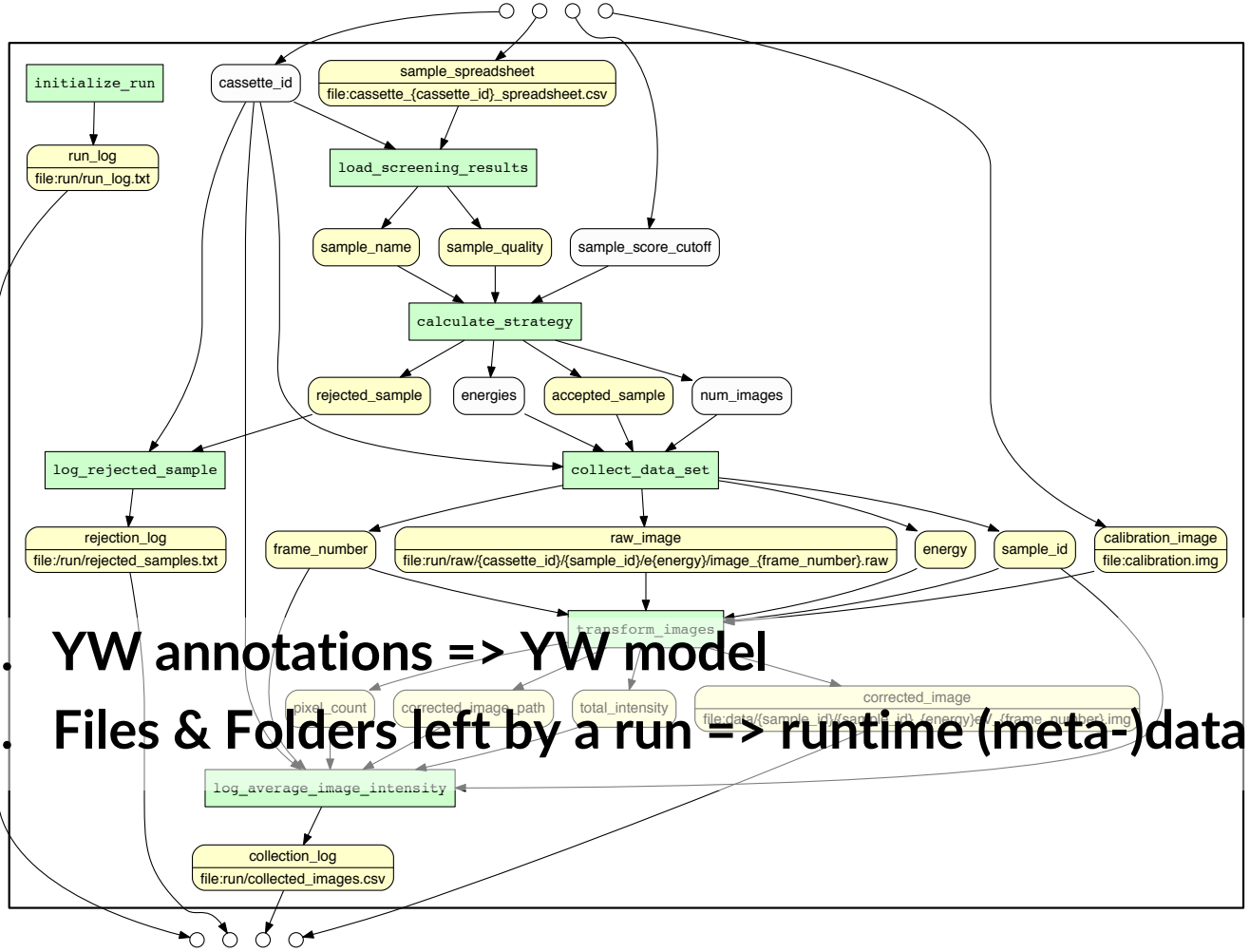


- **URI-templates *link*** conceptual entities to **runtime provenance** "left behind" by the script author …
- … facilitating provenance **reconstruction**

YW (*prospective*) and YW-Recon (*retrospective*) Provenance

- **1. YW: Annotate Script => YW Model**
  - Annotate **@BEGIN..@END, @IN, @OUT**
  - Visualize, share, be happy ☺
- **2. Run script**
  - Files are read and written
  - Folder- & Filenames have metadata
- **3. YW-Recon**
  - Use **@URI** tags **that link YW Model ⇔ Persisted Data**
  - Run URI-template queries
    - cf. “ls -R” & RegEx matching
- **4. YW-Query**
  - Answer the user’s provenance queries

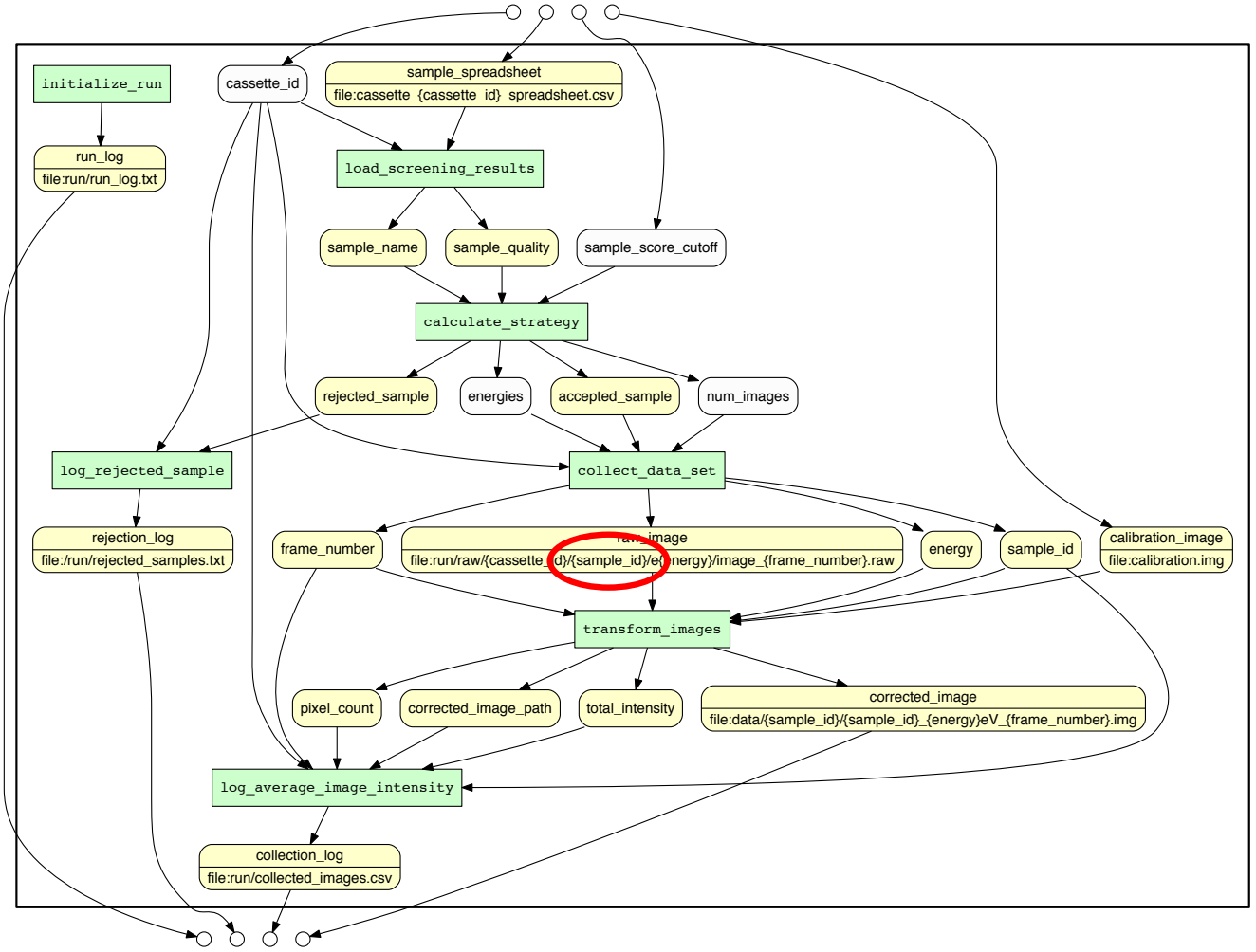
Data collection workflow: runtime data

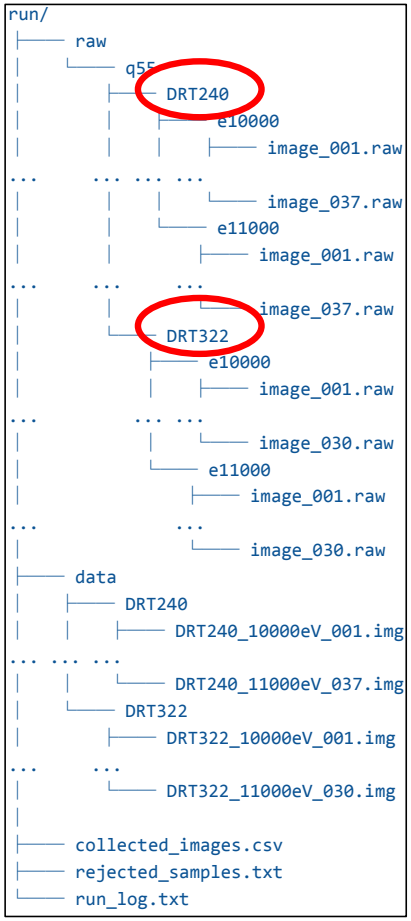


1. YW annotations => YW model
2. Files & Folders left by a run => runtime (meta-)data

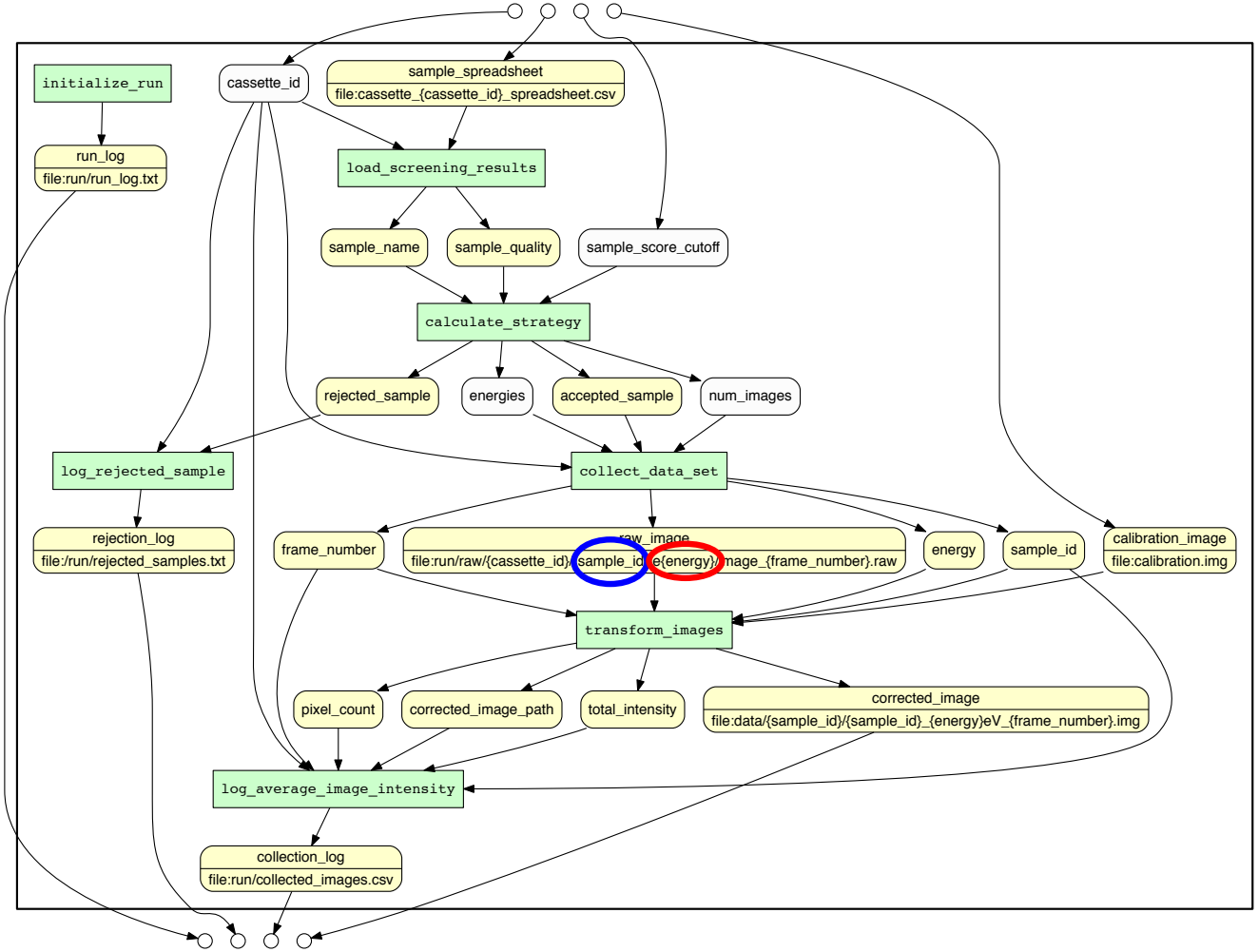
```
run/
├── raw
│   └── q55
│       ├── DRT240
│       │   ├── e10000
│       │   │   ├── image_001.raw
...    ...  ... ...  ...
│       │   │   └── image_037.raw
│       │   └── e11000
│       │       ├── image_001.raw
...    ...  ...  ...  ...
│       │       └── image_037.raw
│       └── DRT322
│           ├── e10000
│           │   ├── image_001.raw
...        ... ...
│           │   └── image_030.raw
│           └── e11000
│               ├── image_001.raw
...            ...
│               └── image_030.raw
├── data
│   ├── DRT240
│   │   ├── DRT240_10000eV_001.img
... ... ...
│   │   └── DRT240_11000eV_037.img
│   └── DRT322
│       ├── DRT322_10000eV_001.img
...    ...
│       └── DRT322_11000eV_030.img
│
├── collected_images.csv
├── rejected_samples.txt
└── run_log.txt
```

Q_1: What samples did the script run collect images from?



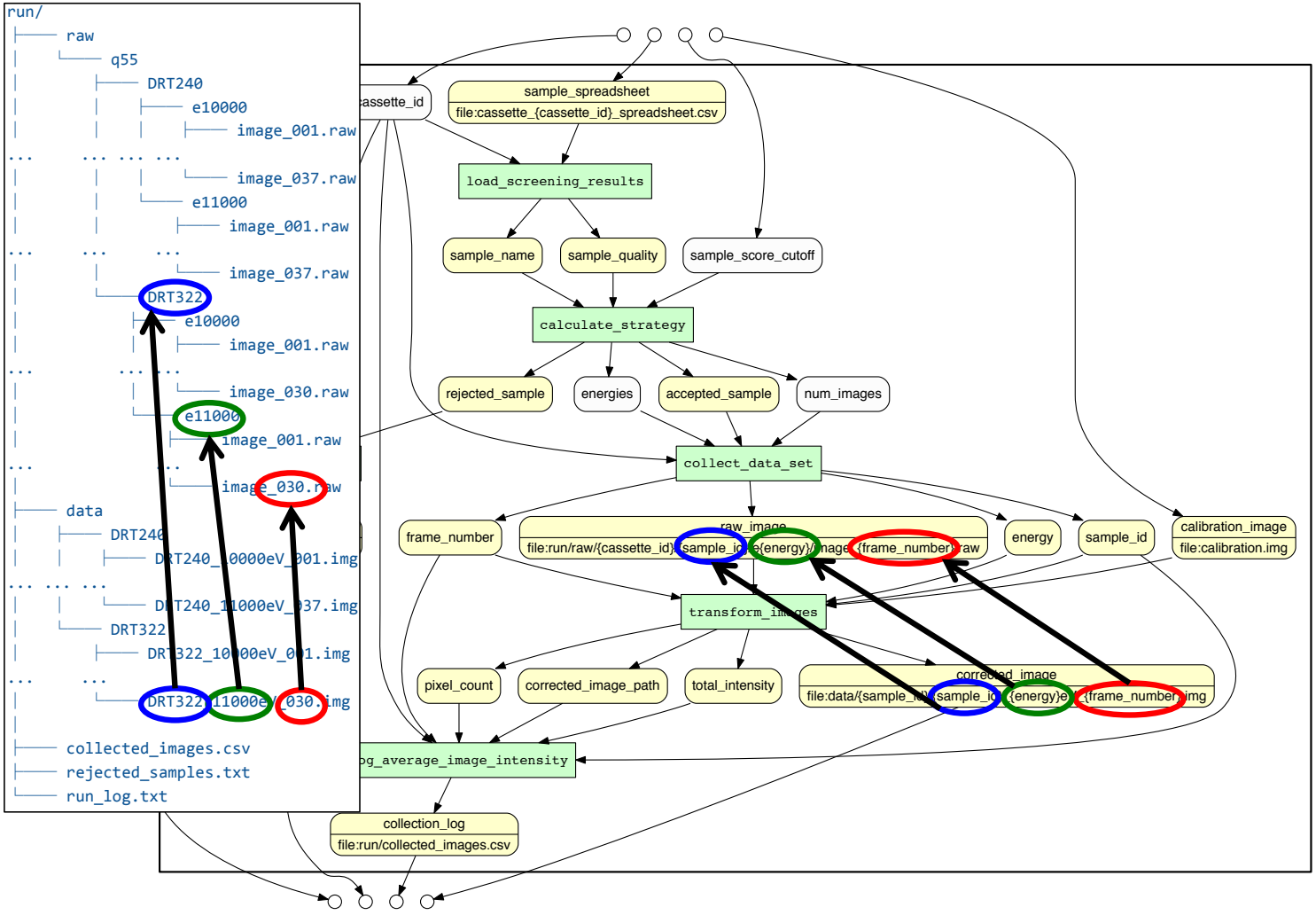


Q_2: What energies were used for image collection from sample DRT322?

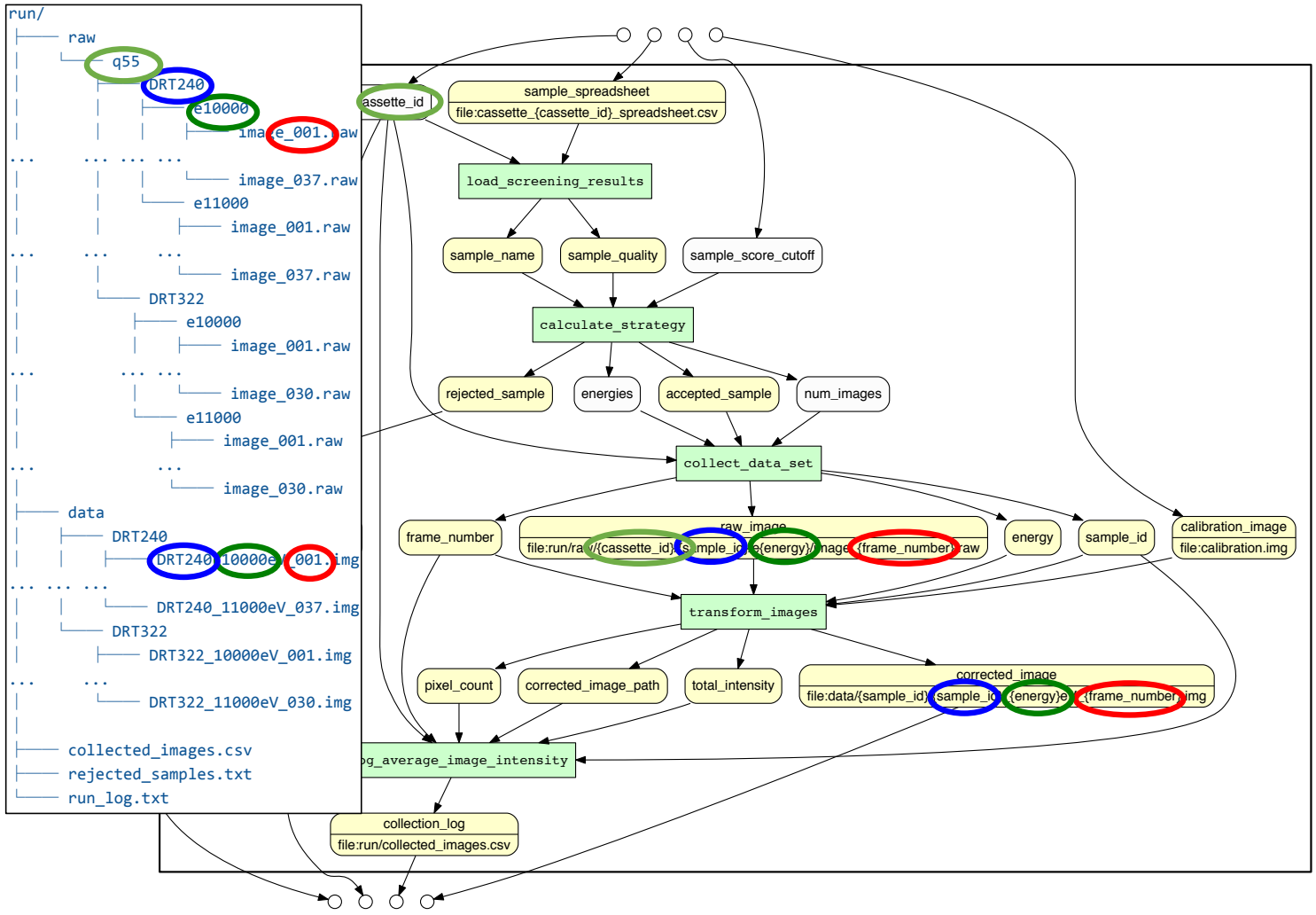



```
run/
├── raw
│   └── q55
│       ├── DRT240
│       │   ├── e10000
│       │   │   ├── image_001.raw
...    ...  ... ...
│       │   │   └── image_037.raw
│       │   └── e11000
│       │       ├── image_001.raw
...    ...    ...
│       │       └── image_037.raw
│       └── DRT322
│           ├── e10000
│           │   ├── image_001.raw
...        ... ...
│           │   └── image_030.raw
│           └── e11000
│               ├── image_001.raw
...            ...
│               └── image_030.raw
├── data
│   ├── DRT240
│   │   ├── DRT240_10000eV_001.img
... ... ...
│   │   └── DRT240_11000eV_037.img
│   └── DRT322
│       ├── DRT322_10000eV_001.img
...    ...
│       └── DRT322_11000eV_030.img
│
├── collected_images.csv
├── rejected_samples.txt
└── run_log.txt
```

Q_3: Where is the raw image of corrected image DRT322_11000ev_030.img?



Q_5: What cassette-id had the sample leading to DRT240_10000ev_001.img?



Yin & Yang: Demonstrating

complementary provenance from noWorkflow & YesWorkflow

João F. Pimentel, Saumen Dey, Timothy McPhillips,
Khalid Belhajjame, David Koop, Leonardo Murta,
Vanessa Braganholo, **Bertram Ludäscher**

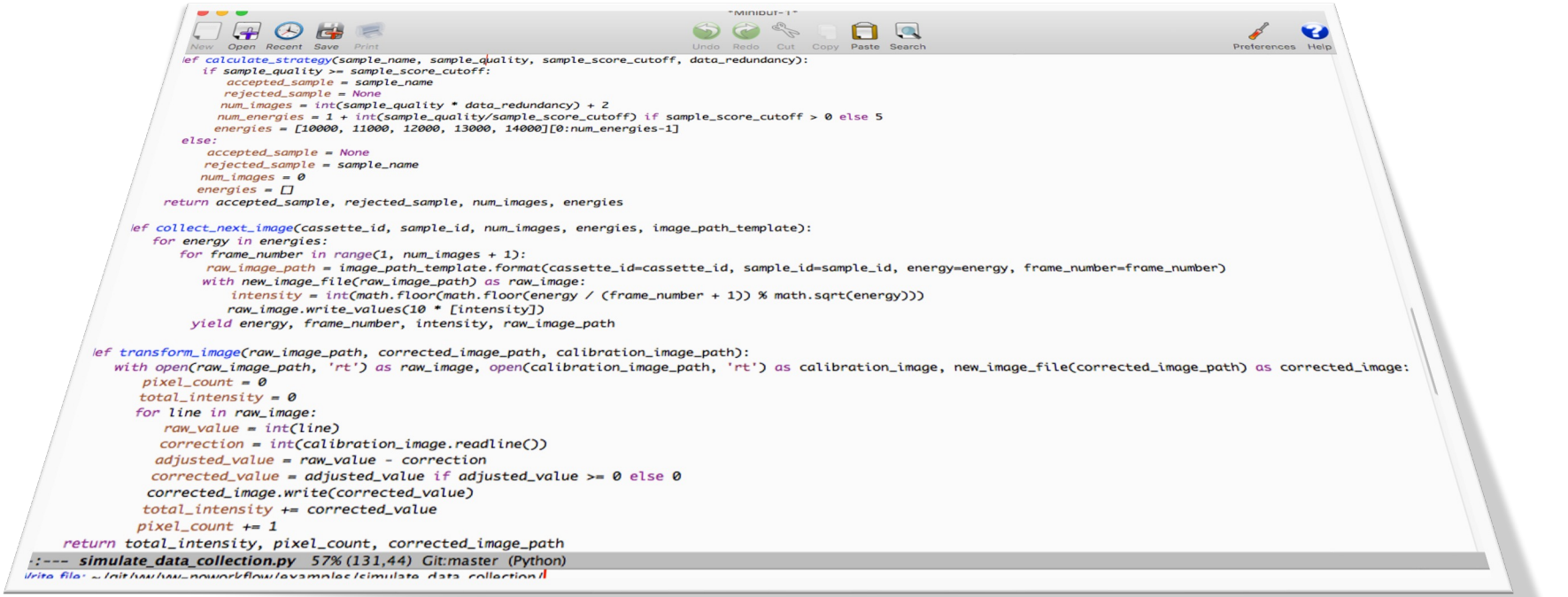
Instituto de Computação

ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

DAUPHINE
UNIVERSITÉ PARIS

UNIVERSITY OF MASSACHUSETTS · DARTMOUTH 1895 ·

Using Provenance from Script Runs



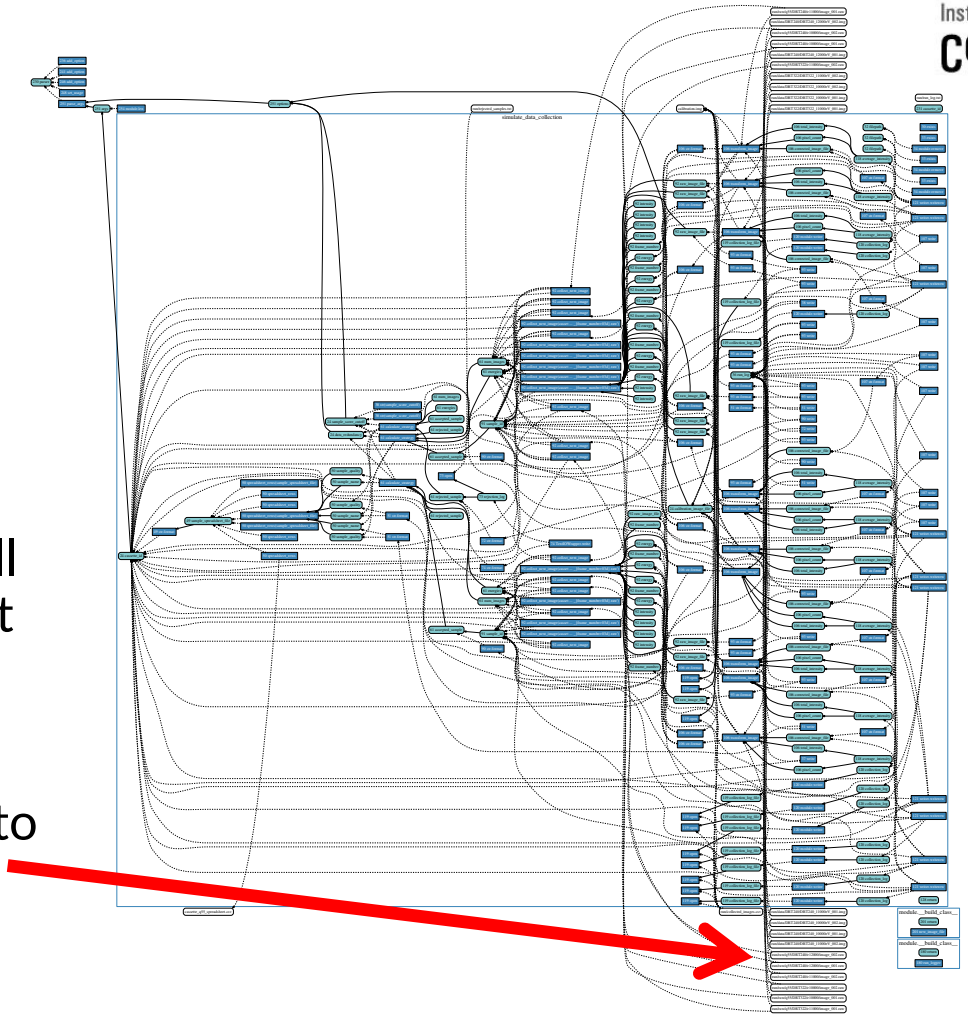
Example from the log-file:

```
2016-06-07 20:32:36 Wrote run/data/DRT240/DRT240_11000eV_002.img
```

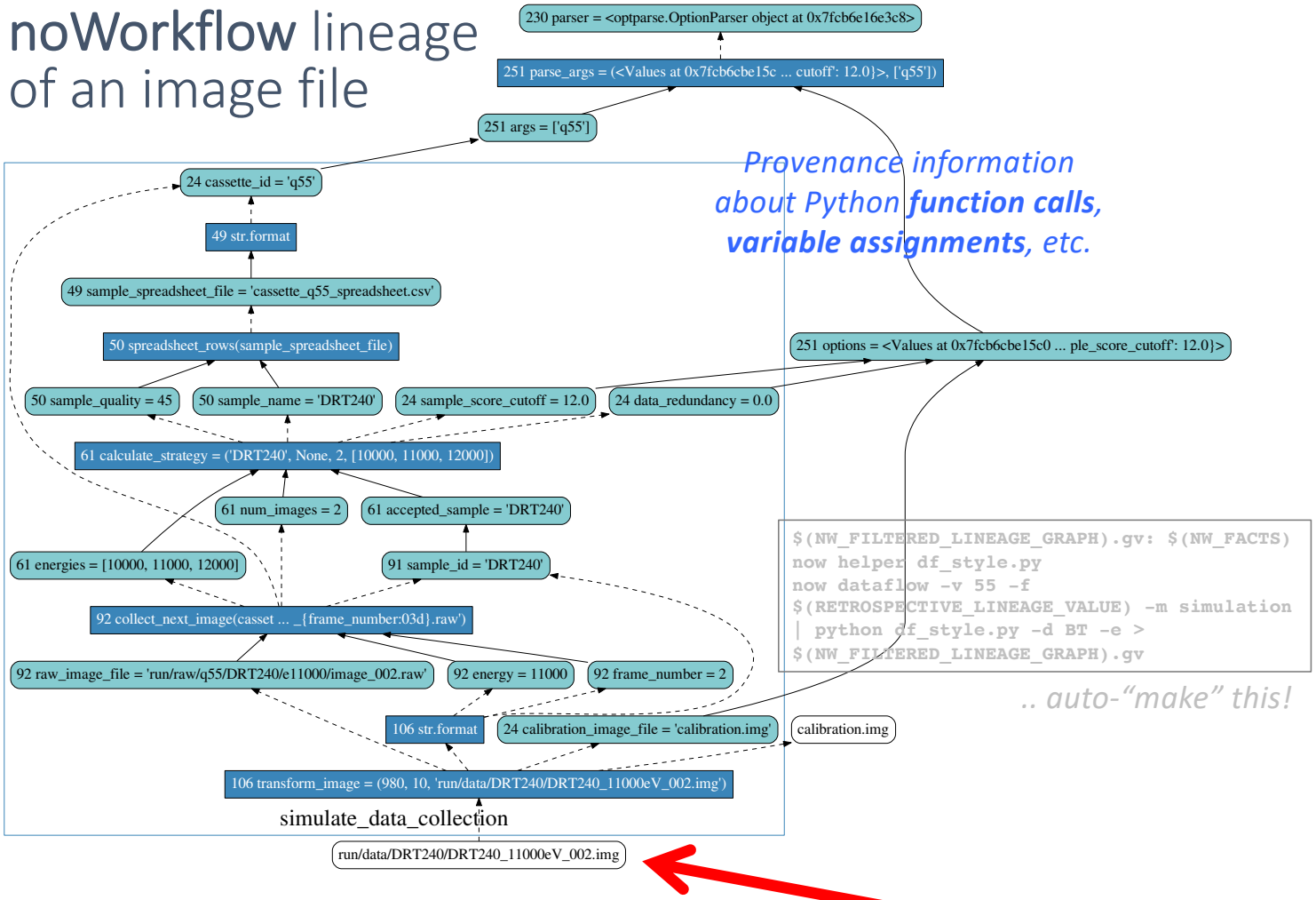
But how was that image derived?? *("Provenance for Self!")*

noWorkflow: not only Workflow!

- **Scripts have provenance, too!**
- **Transparently** capture some/all provenance from Python script **runs**.
- Use filter **queries** to “zoom” into relevant parts ..



noWorkflow lineage of an image file



*Provenance information about Python **function calls**, **variable assignments**, etc.*

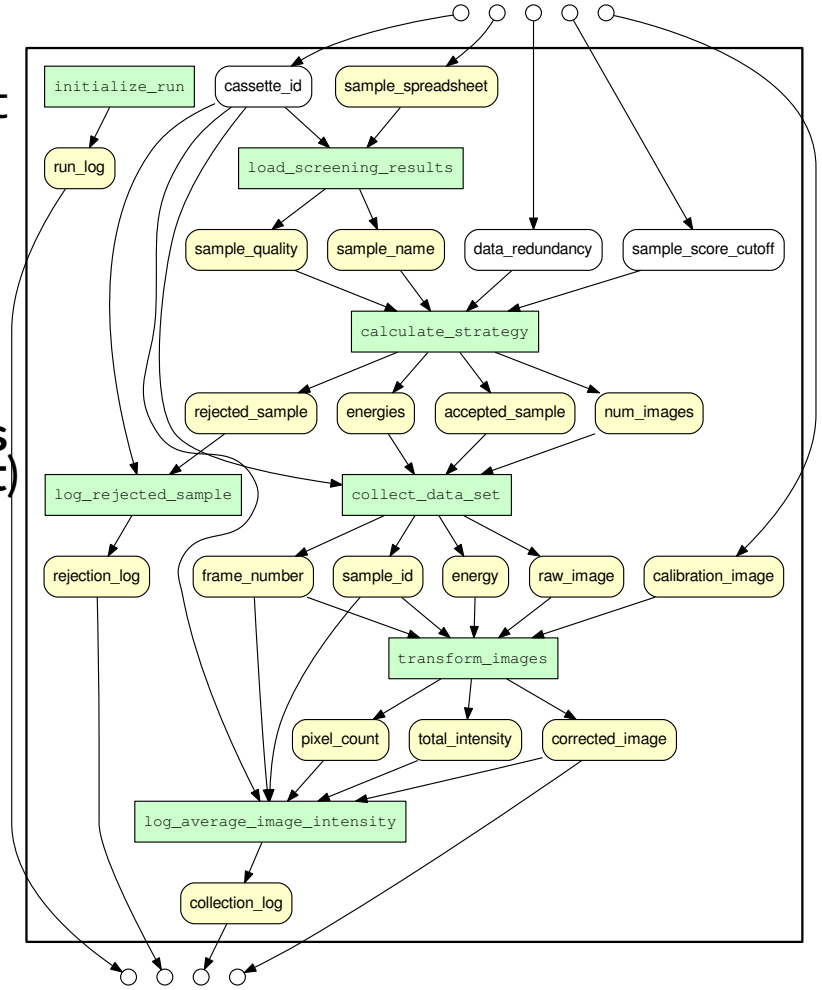
```
$(NW_FILTERED_LINEAGE_GRAPH).gv: $(NW_FACTS)
now helper df_style.py
now dataflow -v 55 -f
$(RETROSPECTIVE_LINEAGE_VALUE) -m simulation
| python df_style.py -d BT -e >
$(NW_FILTERED_LINEAGE_GRAPH).gv
```

.. auto-"make" this!

$ now dataflow -f "run/data/DRT240/DRT240_11000eV_002.img"

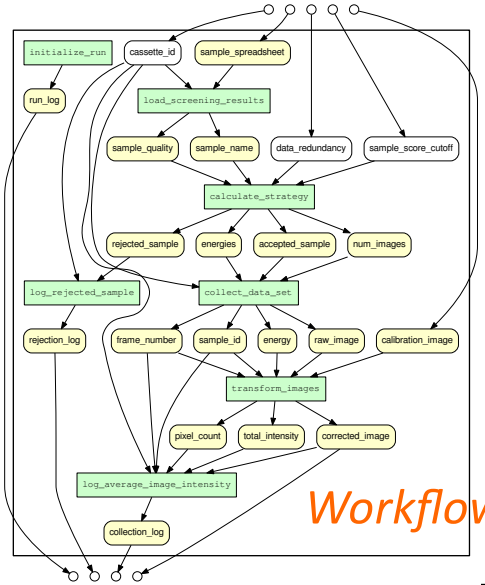
YesWorkflow: Yes, scripts are Workflows, too!

- Use **YW** annotations `@begin...@end, @in, @out` to **reveal hidden conceptual workflow** (**prospective** provenance)
- Script isn't changed:
  - annotations via **comments** (=> **language independent**)
- For understanding and sharing the "big picture"
- **Query** and visualize!

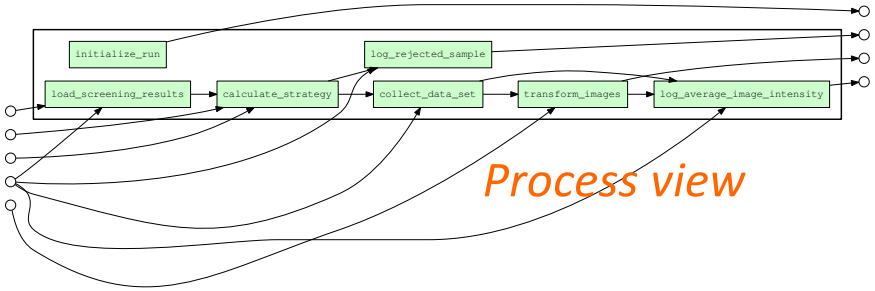


ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Alternate YW Views



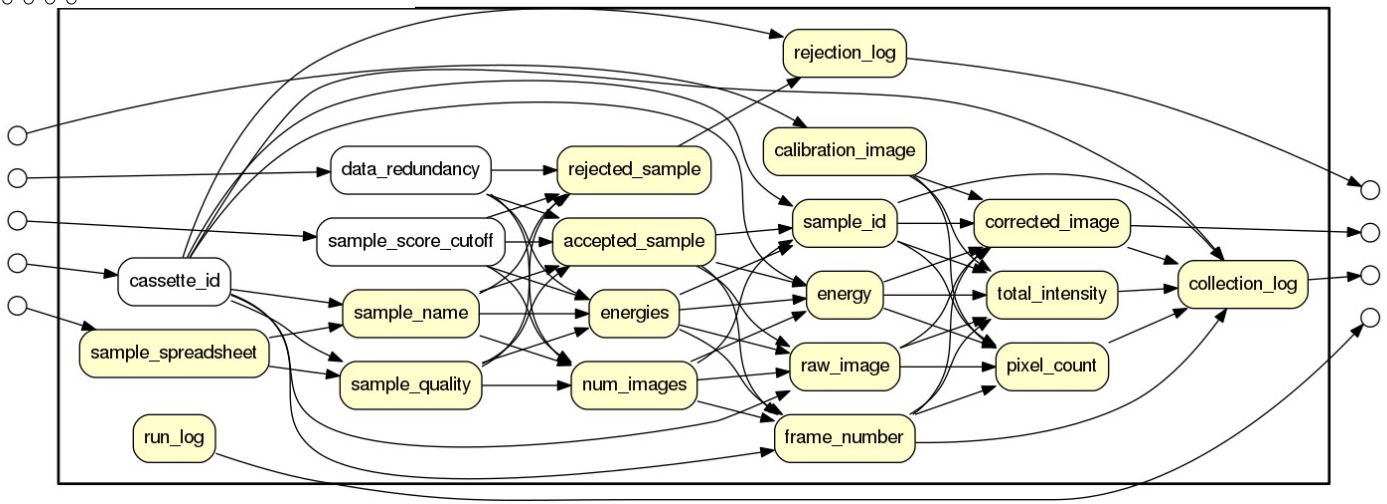
Workflow view



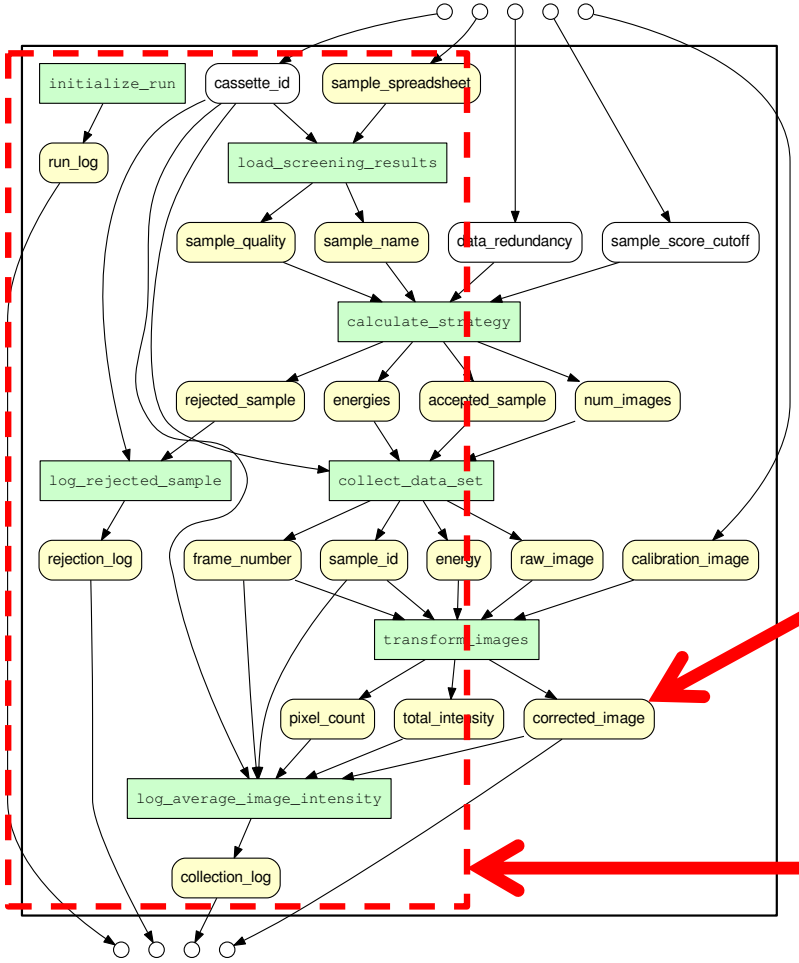
Process view

Data view

simulate_data_collection



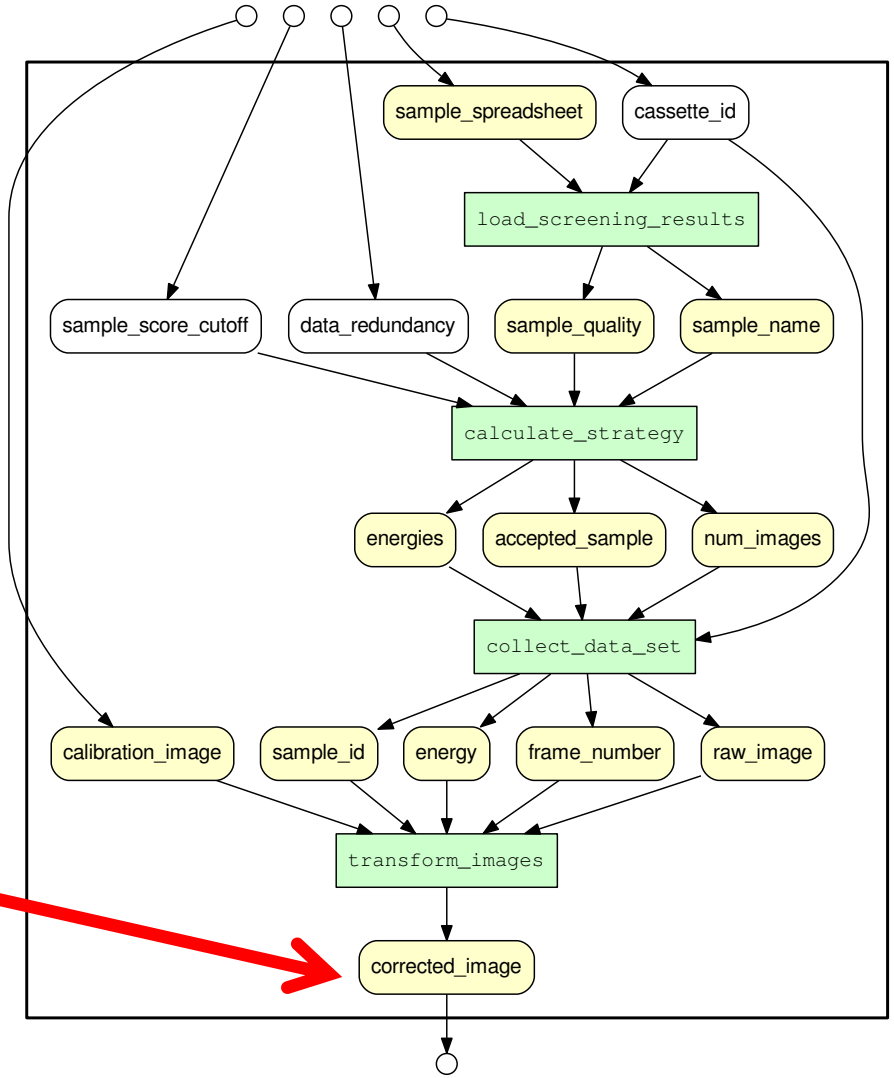
What is the lineage of "`corrected_image`"?



From here on "upwards":
What led (leads) to this?

*.. and what is **irrelevant***
*and should be **pruned**??*

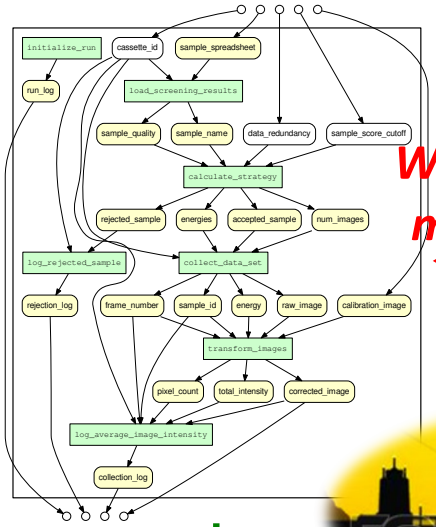
Subgraph resulting from lineage query on YW workflow model

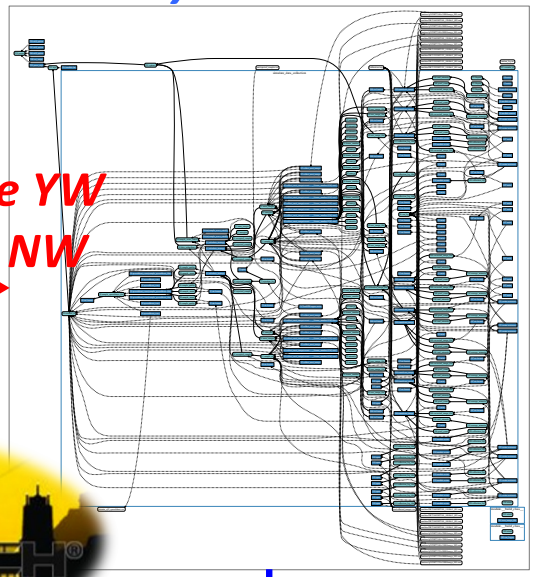


What is the lineage of `corrected_image`***?***

YesWorkflow:
Conceptual *workflow model*

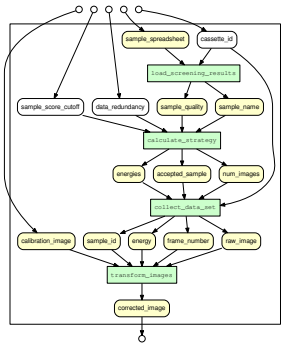
noWorkflow:
Python *trace model*





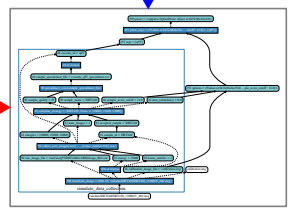
Would like to use YW model to query NW data!

lineage query



But how do we bridge this gap???

lineage query



"Workflow-Land" (*prospective* provenance) ⇔ *"Trace-Land"* (*retrospective* provenance)

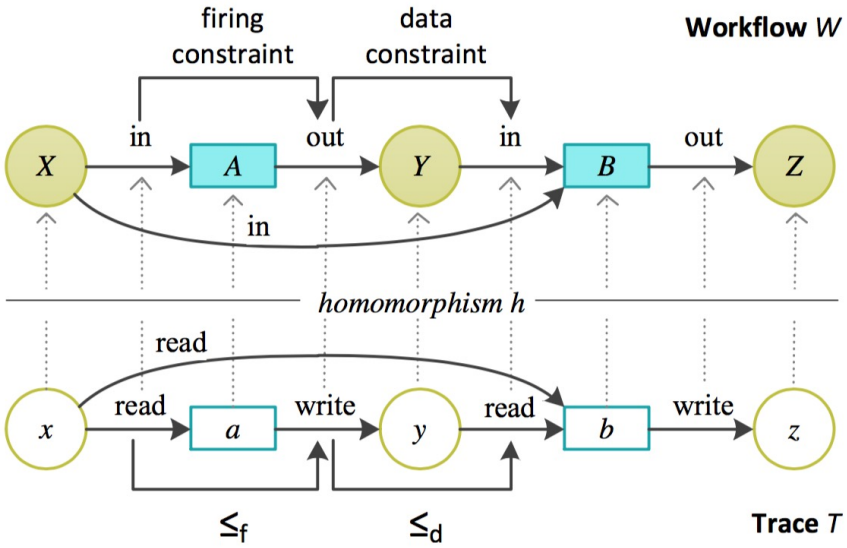
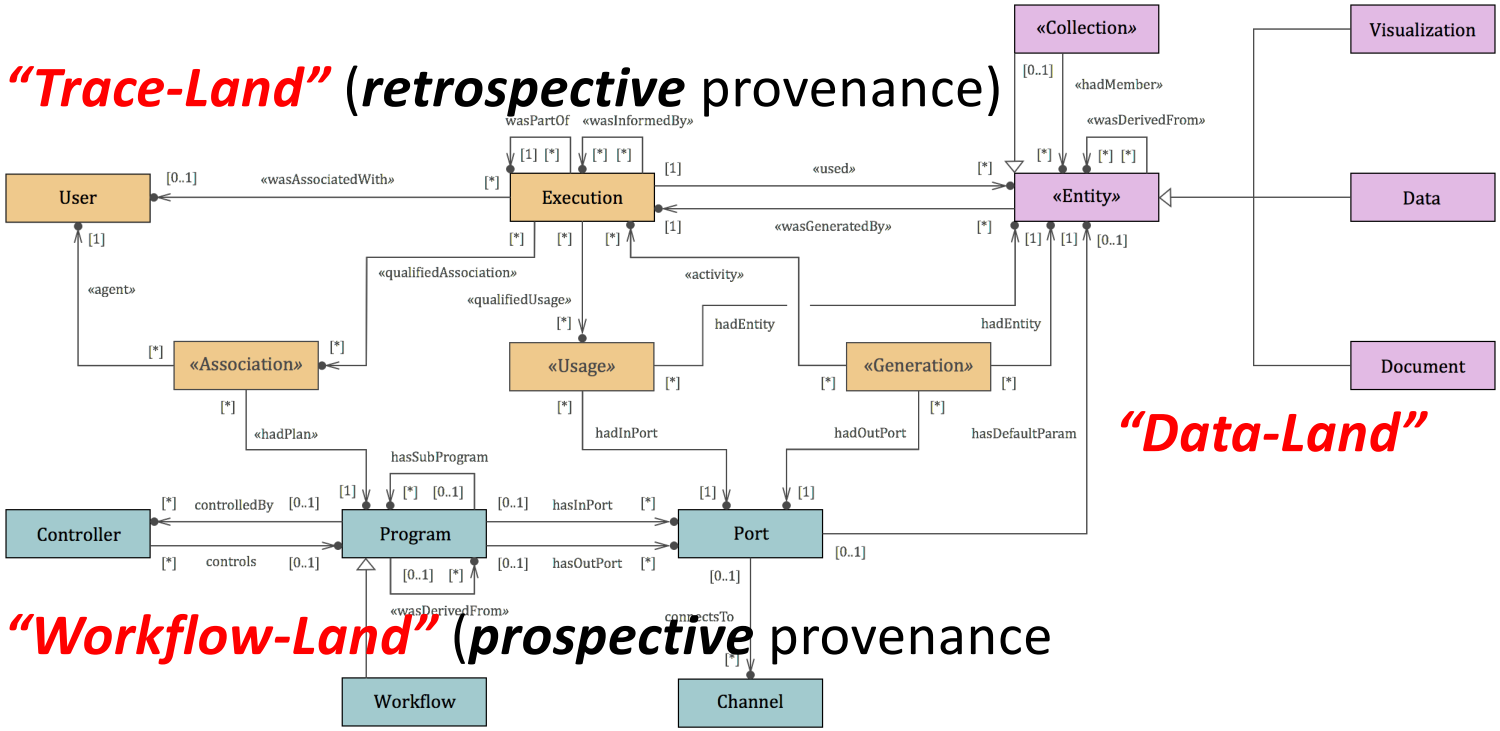


Figure 6: A homomorphism h from trace T to workflow W guarantees structural validity. Workflow-level constraints induce temporal constraints $\leq_f$ and $\leq_d$ on traces [DKBL12].

ProvONE: PROV for scientific workflows

"Trace-Land" (***retrospective*** provenance)



"Data-Land"

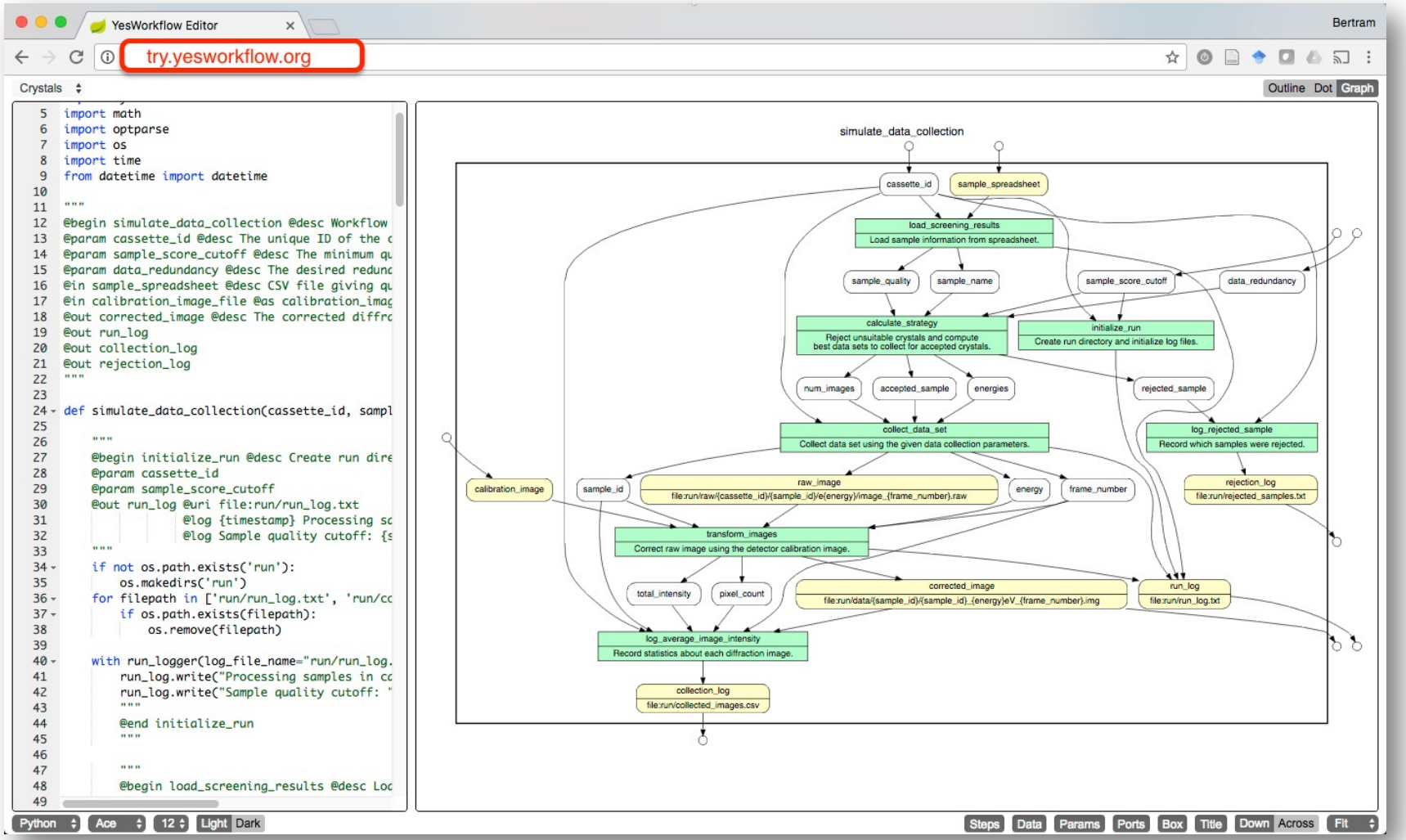
"Workflow-Land" (***prospective*** provenance

DataONE

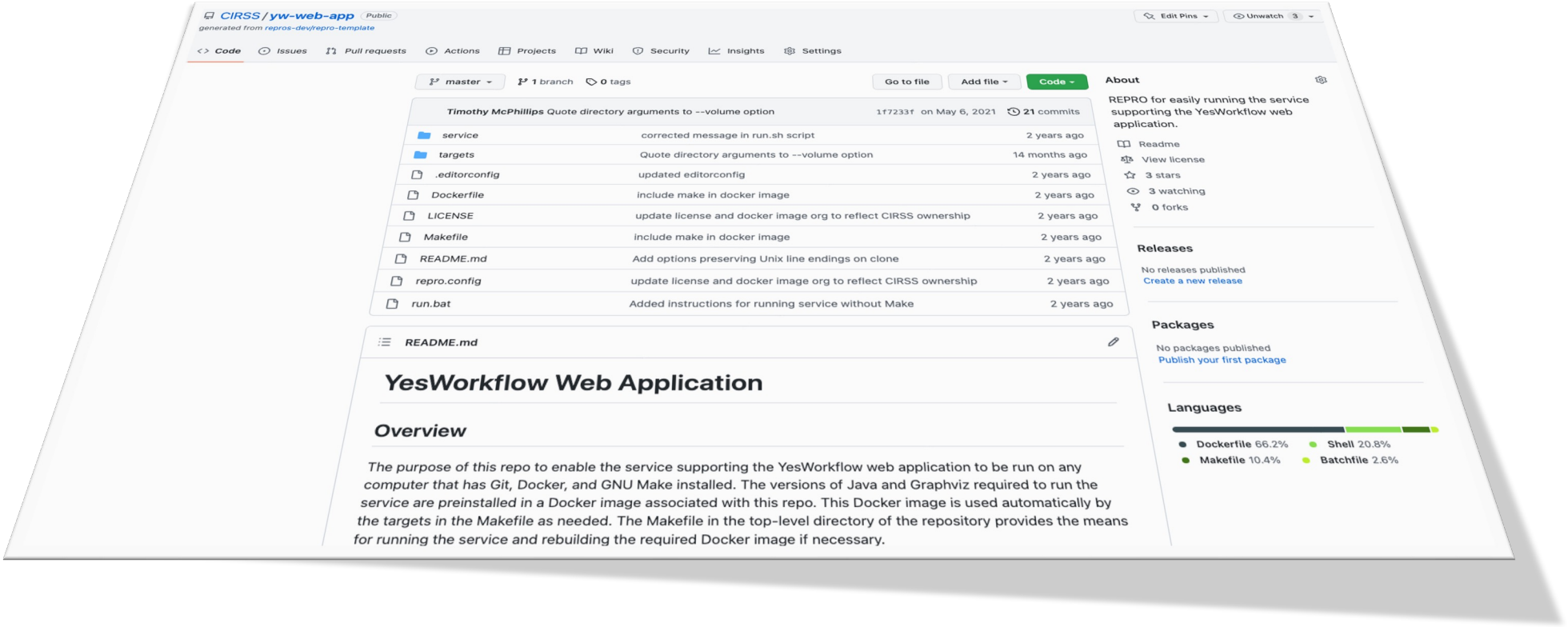
https://purl.dataone.org/provone-v1-dev

Theory & Practice of Data Cleaning

YesWorkflow Demo


YesWorkflow Editor
Bertram
try.yesworkflow.org
Crystals
Outline Dot Graph
import math
import optparse
import os
import time
from datetime import datetime
"""
@begin simulate_data_collection @desc Workflow
@param cassette_id @desc The unique ID of the c
@param sample_score_cutoff @desc The minimum qu
@param data_redundancy @desc The desired reduno
@in sample_spreadsheet @desc CSV file giving qu
@in calibration_image_file @as calibration_imag
@out corrected_image @desc The corrected diffra
@out run_log
@out collection_log
@out rejection_log
"""
def simulate_data_collection(cassette_id, sampl
"""
@begin initialize_run @desc Create run dire
@param cassette_id
@param sample_score_cutoff
@out run_log @uri file:run/run_log.txt
@log {timestamp} Processing sc
@log Sample quality cutoff: {s
"""
if not os.path.exists('run'):
os.makedirs('run')
for filepath in ['run/run_log.txt', 'run/cc
if os.path.exists(filepath):
os.remove(filepath)
with run_logger(log_file_name="run/run_log.
run_log.write("Processing samples in cc
run_log.write("Sample quality cutoff: "
"""
@end initialize_run
"""
"""
@begin load_screening_results @desc Loc
simulate_data_collection
cassette_id
sample_spreadsheet
load_screening_results
Load sample information from spreadsheet.
sample_quality
sample_name
sample_score_cutoff
data_redundancy
calculate_strategy
Reject unsuitable crystals and compute best data sets to collect for accepted crystals.
initialize_run
Create run directory and initialize log files.
num_images
accepted_sample
energies
rejected_sample
collect_data_set
Collect data set using the given data collection parameters.
log_rejected_sample
Record which samples were rejected.
calibration_image
sample_id
raw_image
file:run/raw/{cassette_id}/{sample_id}/e{energy}/image_{frame_number}.raw
energy
frame_number
rejection_log
file:run/rejected_samples.txt
transform_images
Correct raw image using the detector calibration image.
total_intensity
pixel_count
corrected_image
file:run/data/{sample_id}/{sample_id}_{energy}eV_{frame_number}.img
run_log
file:run/run_log.txt
log_average_image_intensity
Record statistics about each diffraction image.
collection_log
file:run/collected_images.csv
Python
Ace
12
Light Dark
Steps Data Params Ports Box Title Down Across Fit

New: Container-based YW Web-app



https://github.com/CIRSS/yw-web-app