



# ***NASA ROSES ACCESS***

**Advancing Collaborative Connections for Earth System Science**

## **Adding Semantics and OPM Ontology for the Provenance of Multi-sensor Merged Climate Data Records. Now What About Reproducibility?**

IN52A-03

AGU 2011. IN52A - Issues in Scientific Data Preservation and Stewardship I

Friday, December 9, 2011

Hook Hua, Gerald Manipon, Brian Wilson, Lei Pan,  
Eric Fetzer, Qing Yue, and Alexandre Guillaume

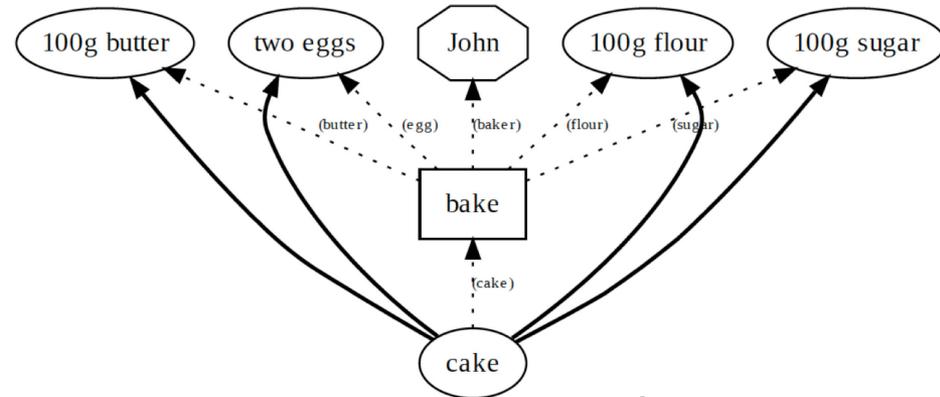
Jet Propulsion Laboratory, California Institute of Technology

Copyright 2011 California Institute of Technology.

Government sponsorship acknowledged.

# Transparency

- “We have a baked cake, but we don’t know the recipe.”

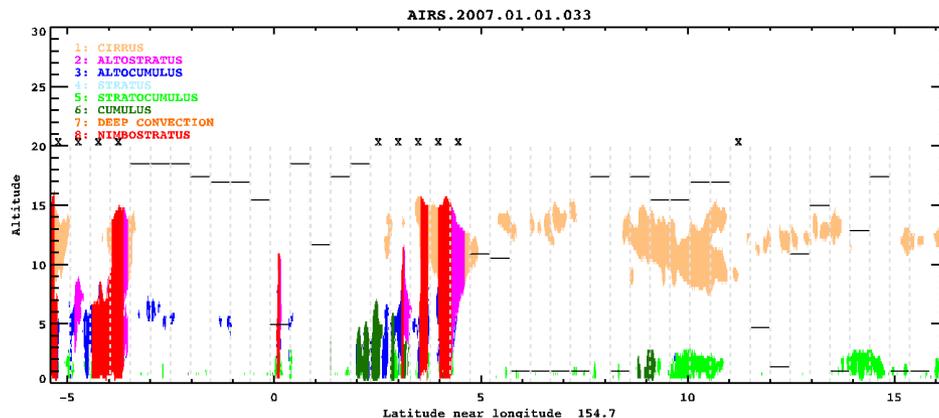


- Climategate (stolen climate emails)
  - Perception of lack of transparency.
  - Common scientific practice (ad-hoc)
  - Outcome: *there should be greater openness and information sharing*



# MEaSUREs Project: A Multi-Sensor Water Vapor Climate Data Record Using Cloud Classification

- Create a complete and consistent record of water vapor from the **A-Train satellite constellation**.
- **Merge** with water vapor observations obtained prior to the A-Train period, and with observations obtained from non-A-Train sensors.
- **Quantify** the effects of cloud variability on water vapor changes over years to decades.



*AIRS Preferentially  
Samples  
Some Cloud Types*





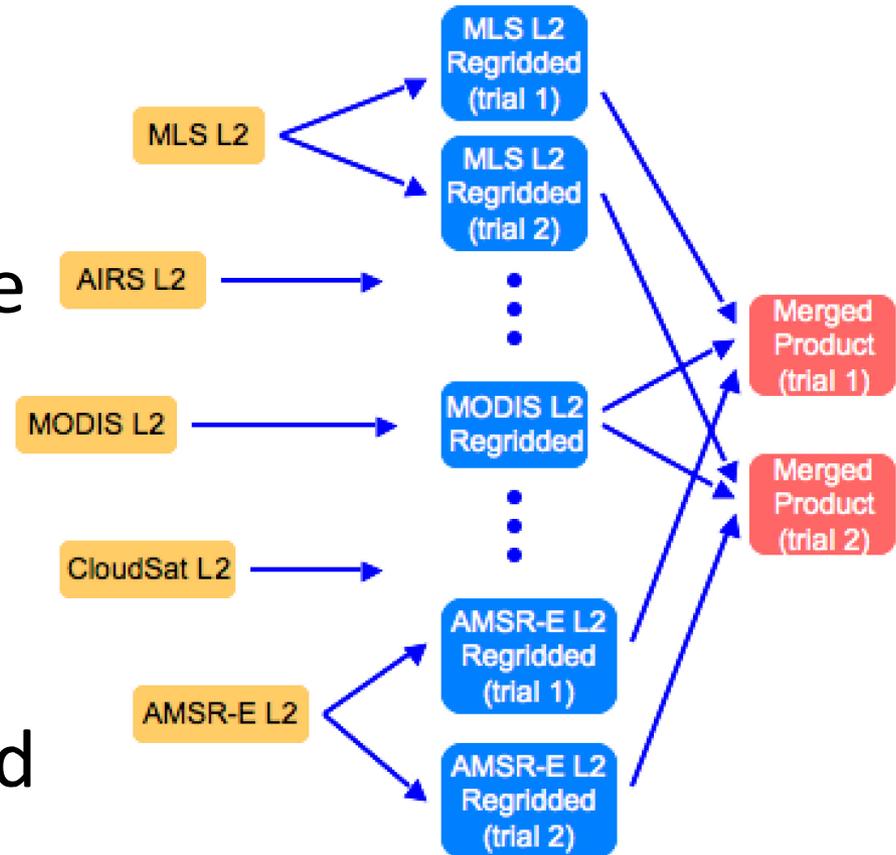
# Production Legacy Challenges

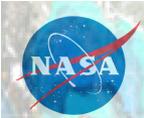
- Multi-sensor merged data products is difficult
  - **Heterogeneous** data handling
  - **Coherent** fusing of measurements
- **Reproducibility** is a basic tenet of science. However...
- Reproducibility is **difficult** with existing satellite data Level 2 data sets.
  - Processing algorithms are complex.
  - Cross-platform calibration is challenges.
- **Multi-parameter, multi-sensor** analyses are particularly difficult.
  - Sensor are rarely designed to work together.
  - Science analysis is usually very ‘free form’.



# Exploratory Analysis

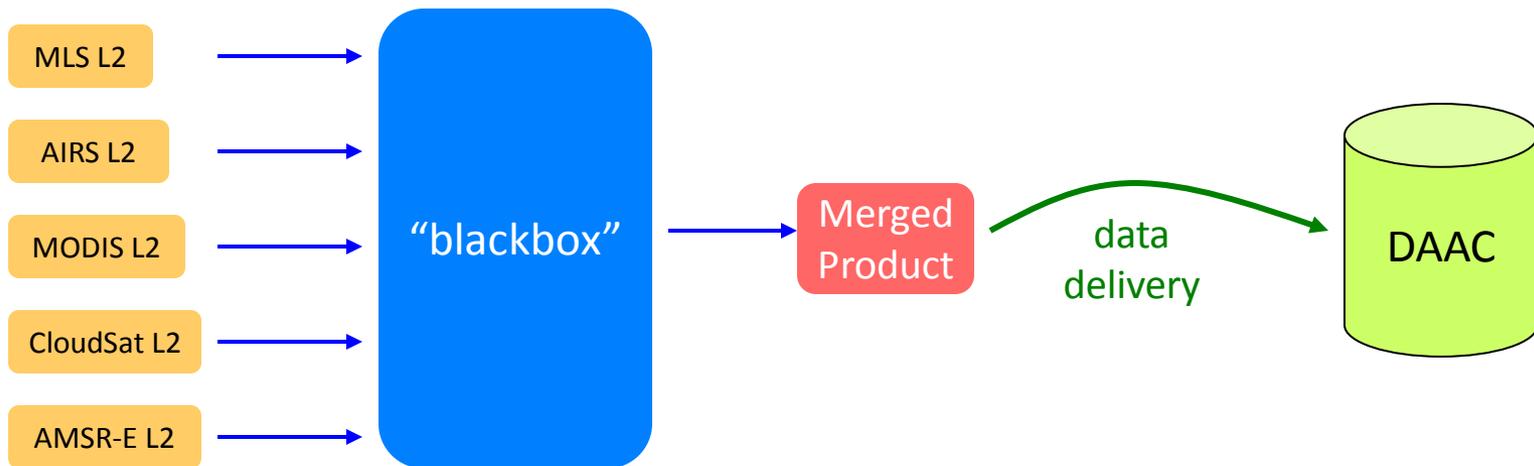
- Incredibly *informal*
- *Unstructured* process
- Automatically capture “what if” scenario runs
- How to create a *structured* data record





# Want to Avoid “Blackbox” Operations

- ATBD describes approach
- But may still want traceability and lineage for each granule

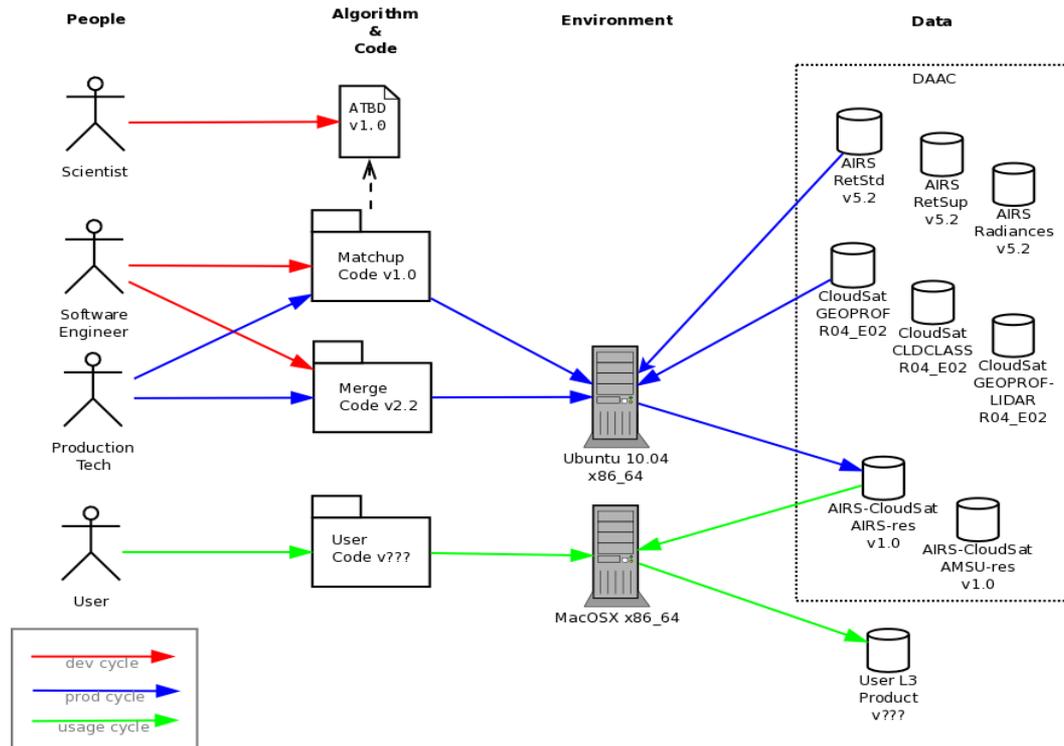




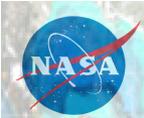
Web Services and Tools for Provenance Capture

# PROVENANCE SERVICES

- Track data pedigree
  - Upstream data, executables, algorithms, parameters, instrument sources, environment, and users

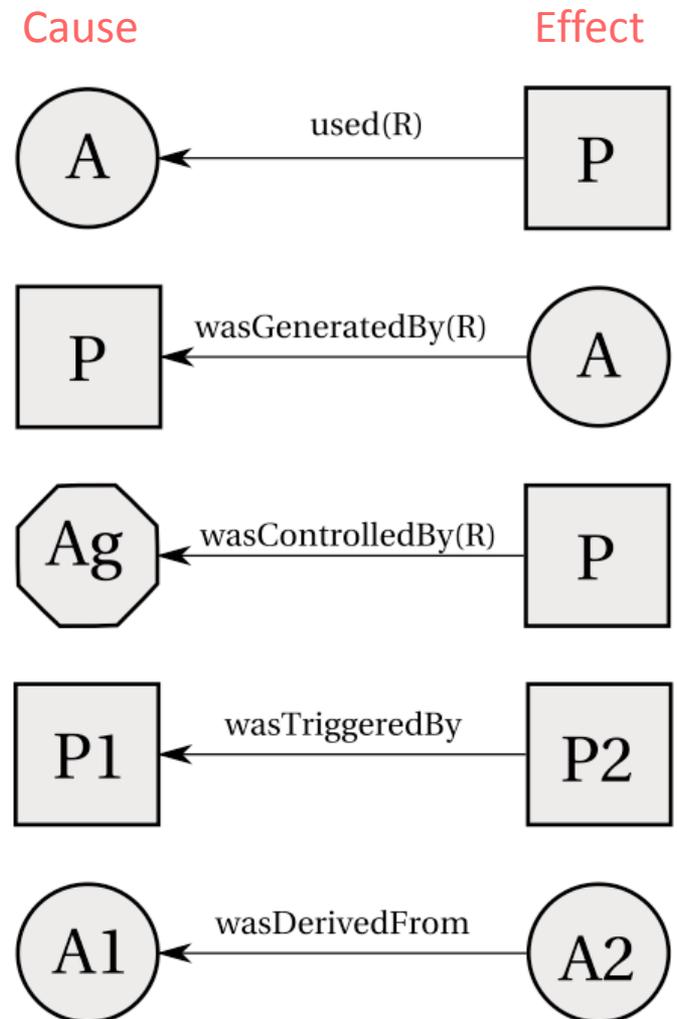


- Provenance can be used to enable reproducibility
- Context allows the use of the information for something else.



# Open Provenance Model: Abstract Model

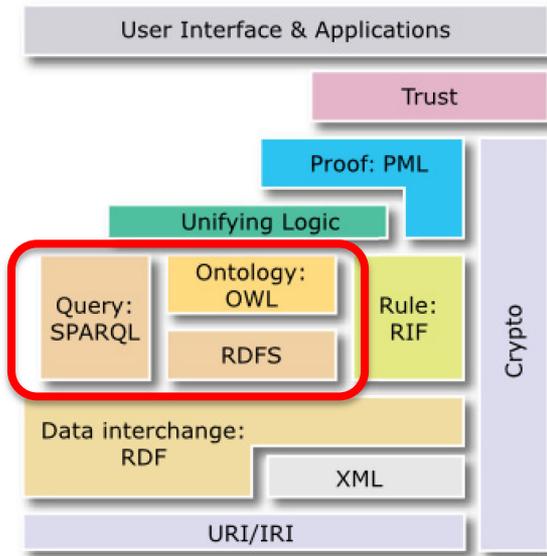
- Allows for process-oriented and dataflow oriented views
- Nodes
  - **A**rtifacts, **P**rocesses, **A**gents
- Edges (Dependencies)
  - Causal dependencies between the artifacts, processes, and agents
- A model of artifacts in the past, explaining how they were derived
  - Invert edge direction for present tense
- *Looking into mappings to ISO Lineage (LI and LE)...*



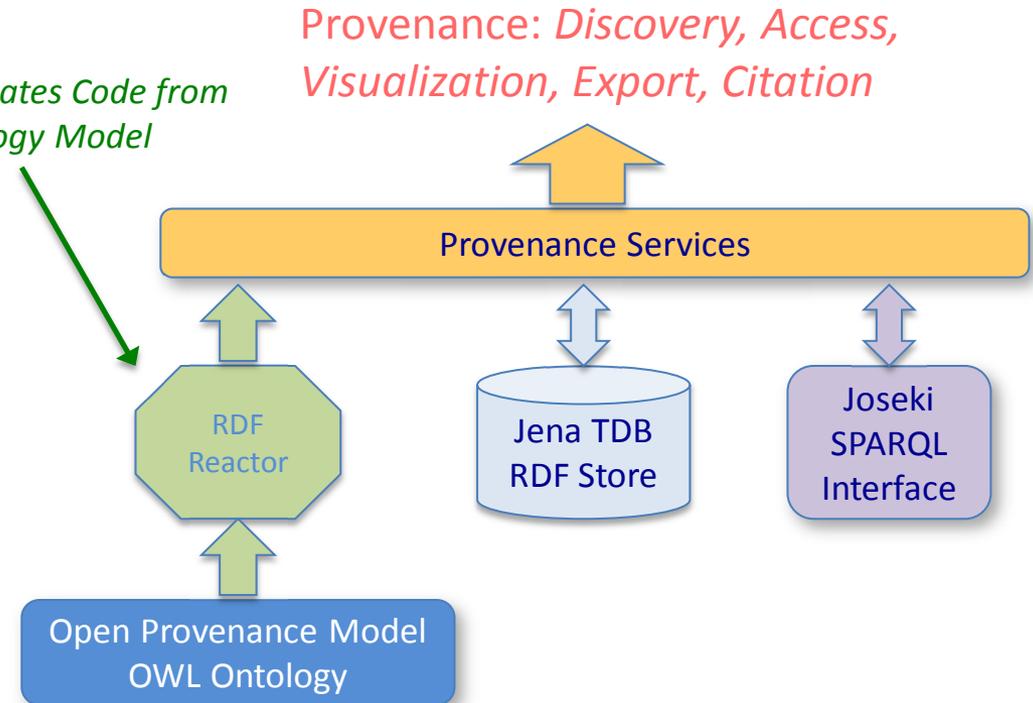


# Semantic Technologies

- Reuse open source semantic web tools

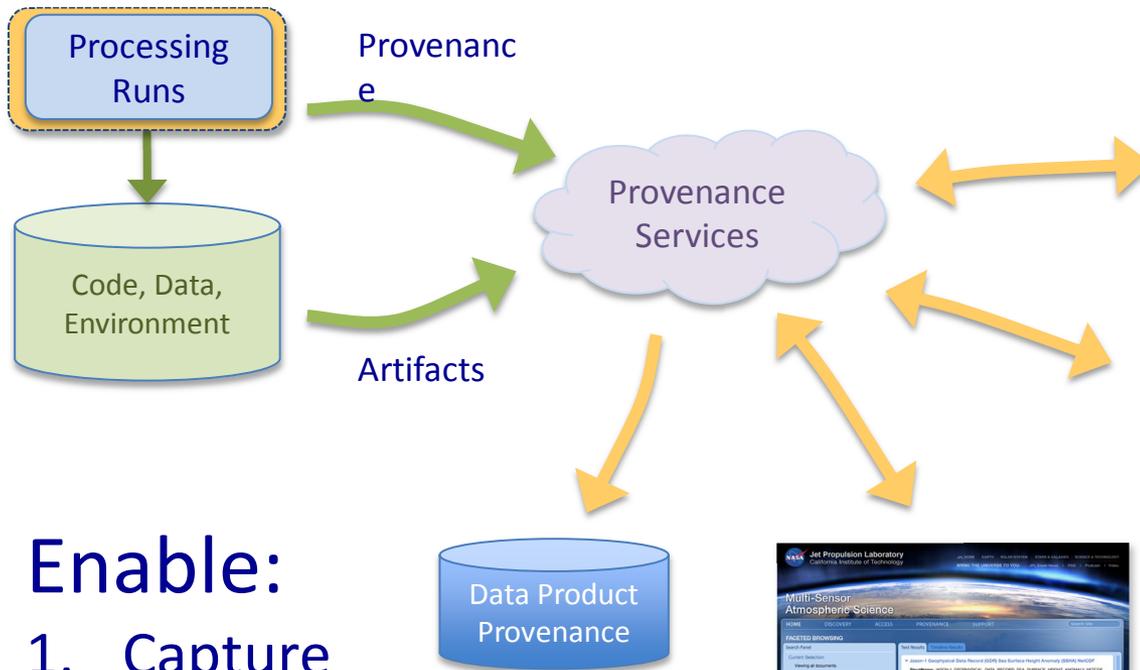


Generates Code from  
Ontology Model



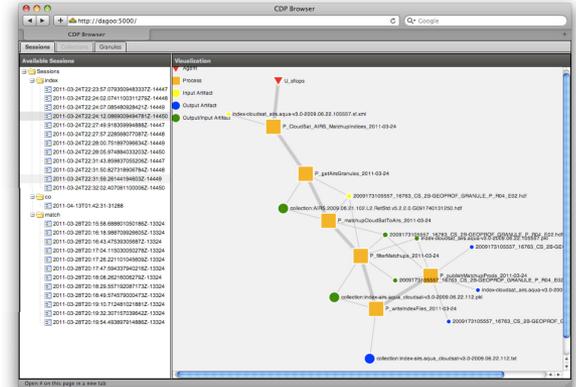


# Provenance Services

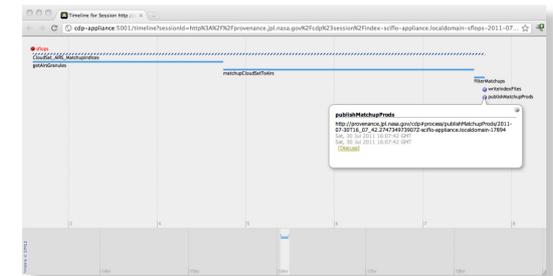


## Enable:

1. Capture
2. Exploration
3. Transparency
4. "Reproducibility"



Lineage Graph



Processing Timeline



Faceted Search



# Provenance Capture Approaches

- *“One size does not fit all”*
- Support multiple client-side approaches
  1. Simple REST web service calls to provide real-time provenance
  2. Instrument processing code. Buffering provenance to log file for deferred single upload
  3. Provenance from workflow engine
  4. Automatic capture of runs



# Approach #1: REST Web Service Invocations

- Simple REST web services for “real-time” provenance capture
- Example calls using “*one-liner commands*”:
  - Process controlled by agent

```
curl -f "http://provenance.jpl.nasa.gov:8099/services/process/was_controlled_by_agent?process=process/generateMatchupIndices/2010-12-16T22_03_59.842924-node1-9813&agent=agent/gmanipon/node1"
```

- Process was triggered by another process

```
curl -f "http://provenance.jpl.nasa.gov:8099/services/process/was_triggered_by_process?triggered_process=process/getMatchupIndices/2010-12-16T22_03_59.843623-gerald-manipon.jpl.nasa.gov-9813&process=process/generateMatchupIndices/2010-12-16T22_03_59.842924-node1-9813"
```

- Process used an artifact

```
curl -f "http://provenance.jpl.nasa.gov:8099/services/process/used_artifact?process=process/getMatchupIndices/2010-12-16T22_03_59.843623-node1-9813&artifact=file://node1/data/sensors/atrain/airs.aqua/L2/v5/2006/07/15/airx2ret/AIRS.2006.07.15.091.L2.RetStd.v5.0-14.0.G07294185430.hdf/2010-08-04T15_06_41.273418"
```

- Artifact was generated by process

```
curl -f "http://provenance.jpl.nasa.gov:8099/services/process/was_generated_by_process?artifact=file://node1/data/measures/pickles/airs.aqua_cloudsat/v2.1/2006/07/15/index-cloudsat_airs.aqua-v2.1-2006.07.15.082803.pkl/2010-12-16T22_13_02.403408&process=process/getMatchupIndices/2010-12-16T22_03_59.843623-node1-9813"
```

- Artifact was derived from another artifact

```
curl -f "http://provenance.jpl.nasa.gov:8099/services/process/was_derived_from_artifact?derived_artifact=file://node1/data/measures/pickles/airs.aqua_cloudsat/v2.1/2006/07/15/index-cloudsat_airs.aqua-v2.1-2006.07.15.082803.pkl/2010-12-16T22_13_02.403408&artifact=file://node1/data/sensors/atrain/airs.aqua/L2/v5/2006/07/15/airx2ret/AIRS.2006.07.15.091.L2.RetStd.v5.0-14.0.G07294185430.hdf/2010-08-04T15_06_41.273418"
```



# Approach #2: Instrumentation Functions

Python

```
processStart('generateMatchupIndices', '/home/sflops/runMatchups.py', 'v0.1')
processUsedArtifacts('data:AIRS:RetStd',
['/home/sflops/AIRS.2009.06.23.031.L2.RetStd.v5.2.2.0.G09175065058.hdf'], 'input')
processGeneratedArtifacts('indexFiles', ['/home/sflops/index-001.txt'], 'output')
processEnd('generateMatchupIndices')
```

Matlab

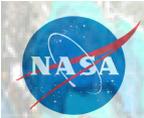
```
processStart('generateMatchupIndices', EXECUTABLE_PATH, VERSION, PROJECT_NAME, SESSION_NAME);
processUsedArtifacts(AIRS_COL, AIRS_GRANULES, 'input');
processGeneratedArtifacts(INDEX_COL, INDEX_FILES, 'output');
processEnd('generateMatchupIndices');
```

IDL

```
LOG_PROCESS_START, 'generateMatchupIndices', '/home/sflops/runMatchups.pro', 'v0.1'
LOG_PROCESS_USED_ARTIFACTS, 'data:AIRS:RetStd',
'/home/sflops/AIRS.2009.06.23.031.L2.RetStd.v5.2.2.0.G09175065058.hdf', 'input'
LOG_PROCESS_GENERATED_ARTIFACTS, 'data:indexFiles', '/home/sflops/index-001.txt', 'output'
LOG_PROCESS_END, 'generateMatchupIndices'
```

Command  
Line

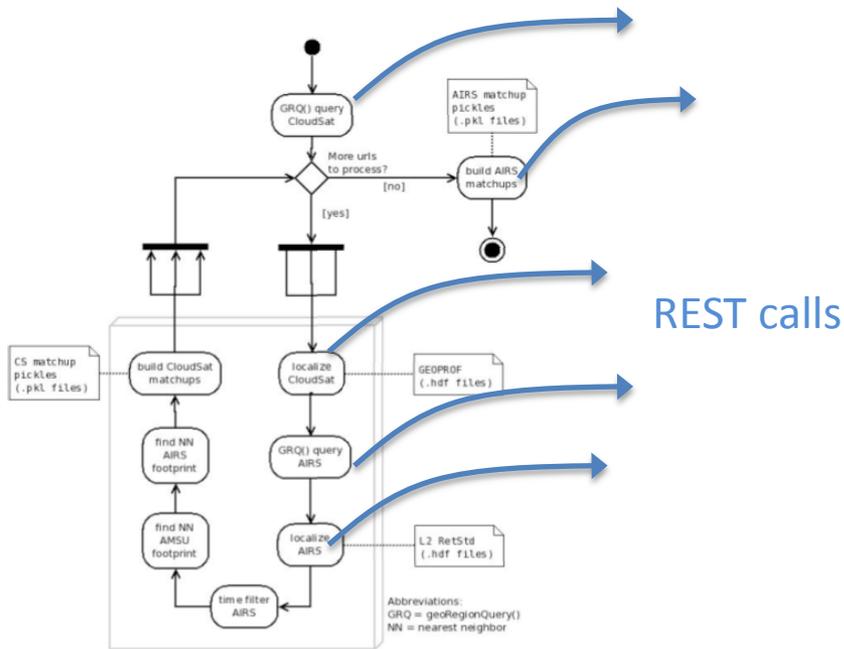
```
curl -f -v -X POST --data-binary @generateMatchupIndices_provenance.log
http://provenance.jpl.nasa.gov:8099/services/logfile/upload
```



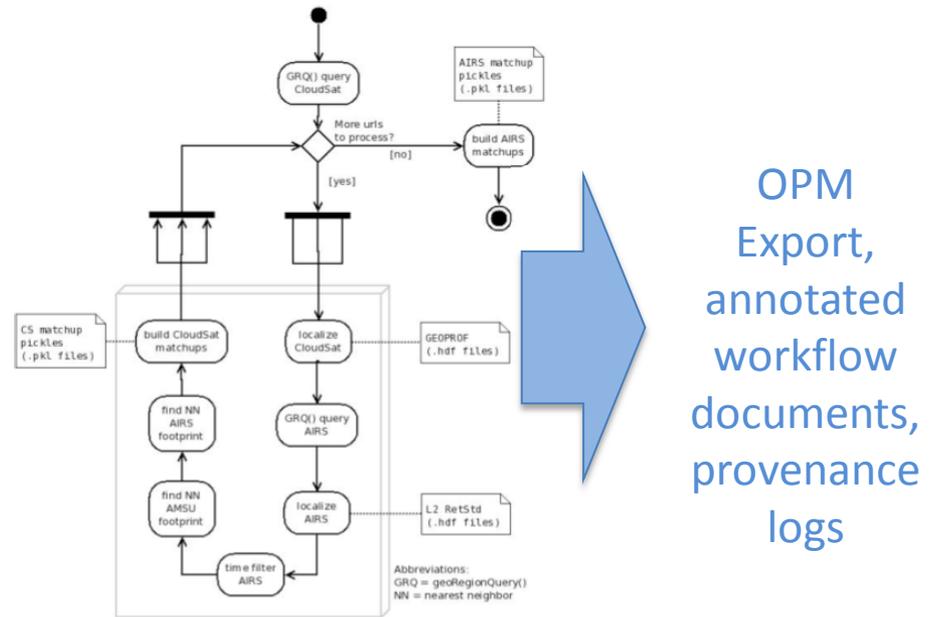
# Approach #3: Workflows

- Two ways to capture provenance from workflows

(a) Instrument workflow tasks to make REST service calls

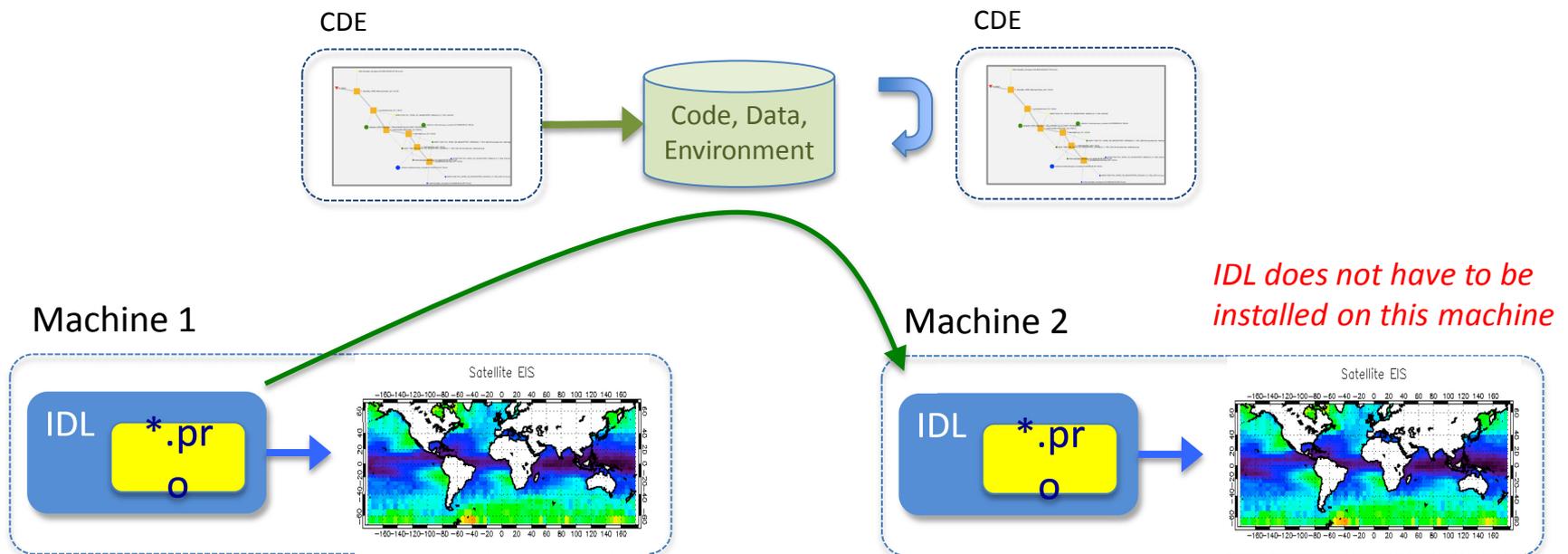


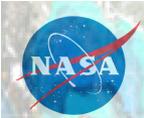
(b) Workflow to produce provenance records



# Approach #4: Automatic Capture of Runs

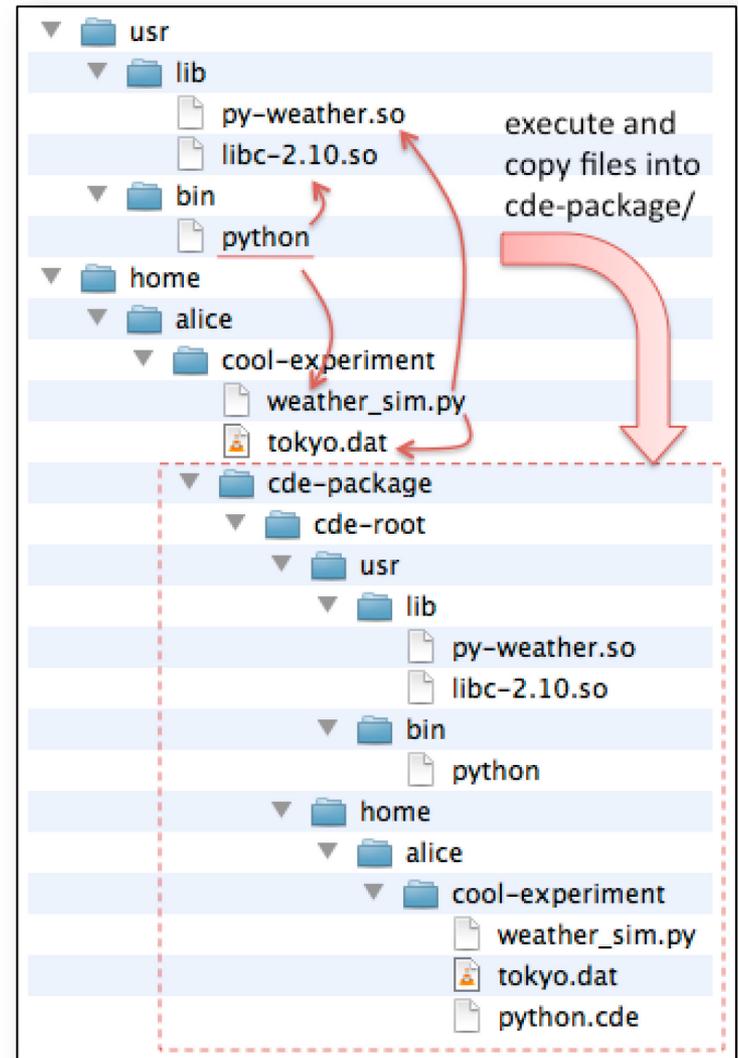
- Leveraged *open source* tool **CDE**: Automatic packaging of Code, Data, and Environment
  - Philip Guo [pg@cs.stanford.edu](mailto:pg@cs.stanford.edu)
- **No modifications needed to processing code.**
  - High-level interaction recorded. Built upon *strace* tool.





# Some Reproducible Research

- CDE\* creates a complete record of what was run
- Works well for “*incredibly informally*” research approaches
- Perform reproducible research
  - **Records** high-level process spawning, all input files used, and all output files generated.
  - **Packages** up all files and information needed to reproduce a run (*with some limitations*)
- Collaborate with colleagues who can **rerun on different machines**



\* Philip J. Guo. CDE: Run Any Linux Application On-Demand Without Installation. In Proceedings of the 2011 USENIX Large Installation System Administration Conference (LISA), December 2011.



# THE BIG PICTURE

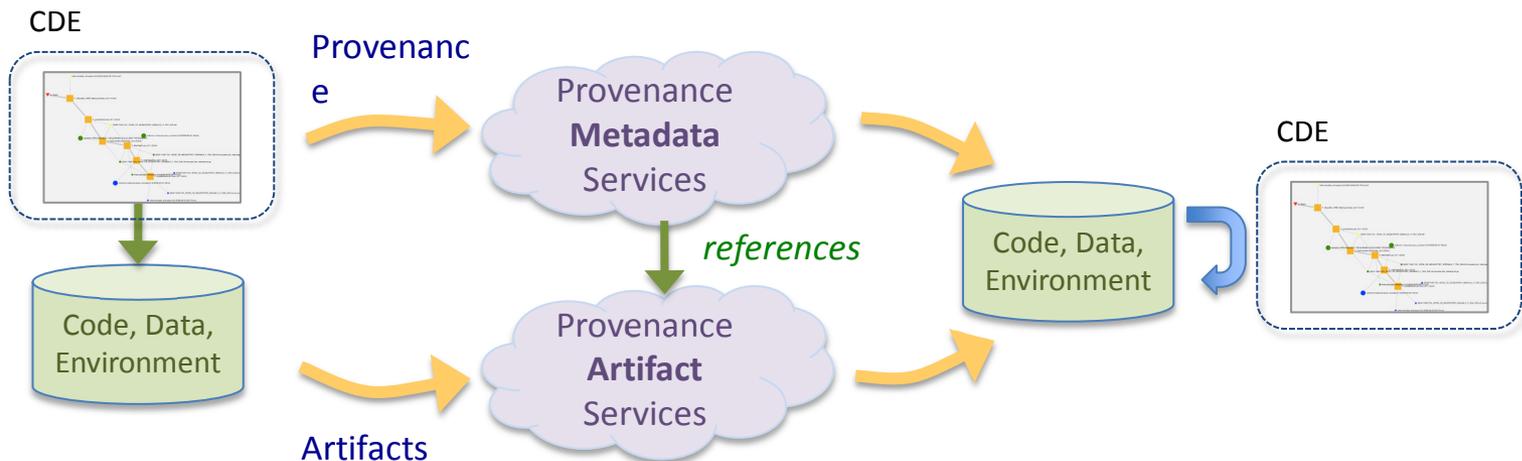


# NASA Earth Science Data Preservation Content Specification

- *“...identify all the content items that need to be preserved to ensure their availability to support future investigations in long-term global change research.”*
- These provenance services can support preservation of:
  - Data
  - Product software used in processing
  - Algorithm inputs (e.g. ancillary data)
  - Software tools used in processing and analysis

# A Portable Runnable Preservation Record

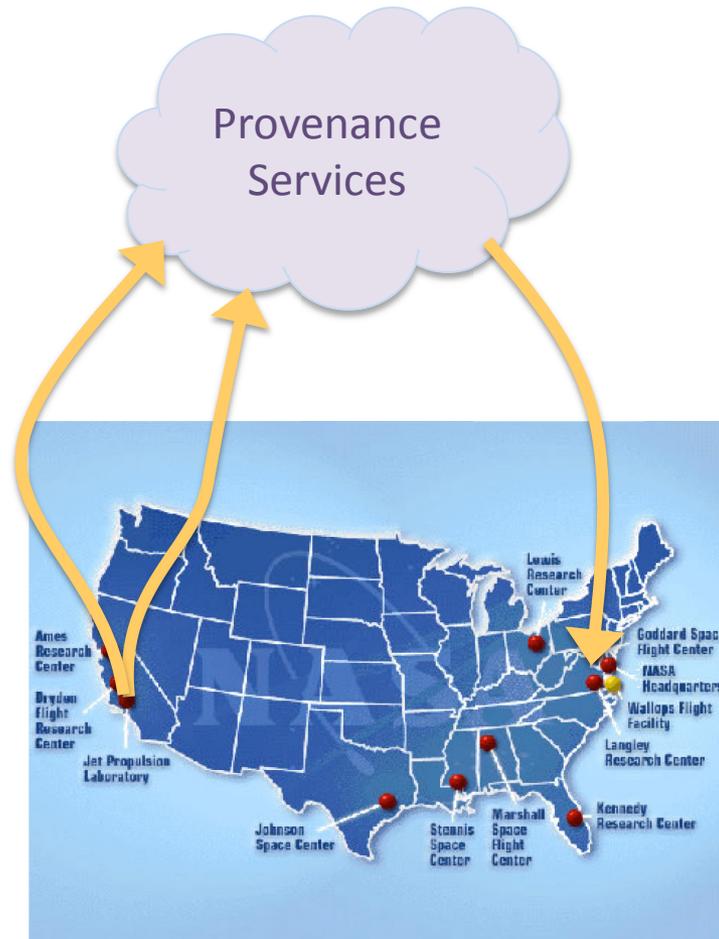
- Provenance services addressing **preservation** of data product generation and analysis
- Provenance services enabling *some* **reproducibility** of results
- Provenance services facilitating **collaboration**



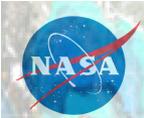


# Provenance Collection for an “Earth Science Collaboratory” (ESC)

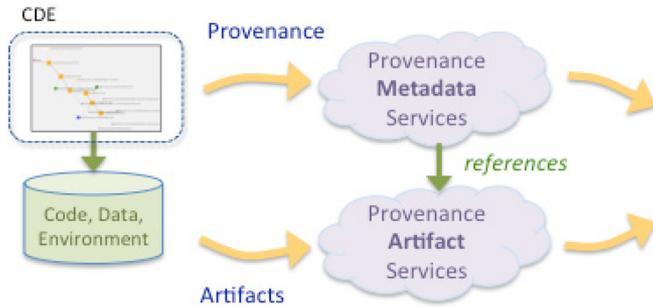
Provenance  
Collection



Provenance,  
Preservation,  
Discovery, Access,  
Visualization,  
Export,  
Reproducibility



# Improving the Understanding of Production Legacy



- Improve understanding of data and processing lineage
- Near instant
  - Faceted search
  - Lineage graph
  - Lineage Timeline

The screenshot shows the NASA Jet Propulsion Laboratory website for Multi-Sensor Atmospheric Science. The navigation bar includes links for HOME, DISCOVERY, ACCESS, PROVENANCE, and SUPPORT. The main content area is titled 'FACETED BROWSING' and features a search panel on the left and text results on the right.

**Search Panel:**

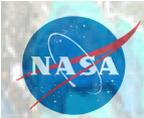
- Current Selection: Viewing all documents!
- Product Types: artifact (15), process (6)
- Instruments: AIRS (7), CloudSat (7)
- Datasets: 2B-CLDCLASS (3), 2B-GEOPROF (1), AIRSM\_CPR\_MAT (1), AIRS\_CPR\_MAT (1), L1B.AIRS\_Rad (1), L1B.VIS\_Rad (1), L2.RetStd (1), L2.RetSup (1)
- Versions: R04\_E02 (6), v3.1 (4), v5 (4)
- Agents: U\_sflops (21)
- Sessions: S.CloudSat\_AIRS\_MergedData\_2011-11-02T05:42:59 (15), S.colocate\_ceres\_2011-11-02T06:03:05 (6)

**Text Results:**

- P\_11261\_2011-11-02T06:03:08**
  - Last Updated: 2011-11-02T04:27:29.946Z
  - Type: process
  - Session: S\_colocate\_ceres\_2011-11-02T06:03:05
  - Agent: U\_sflops
  - GeneratedBy: ceres\_2008061004519\_09796\_CS\_2B-CLDCLASS\_GRANULE\_P\_R04\_E02.nc
  - Used: 2008061004519\_09796\_CS\_2B-CLDCLASS\_GRANULE\_P\_R04\_E02.hdf
  - Buttons: Lineage, Timeline, Reproduce
- P\_11260\_2011-11-02T06:03:05**
  - Last Updated: 2011-11-02T04:27:28.833Z
  - Type: process
  - Session: S\_colocate\_ceres\_2011-11-02T06:03:05
  - Agent: U\_sflops
  - Buttons: Lineage, Timeline, Reproduce
- 2008061004519\_09796\_CS\_2B-CLDCLASS\_GRANULE\_P\_R04\_E02.hdf**
  - Last Updated: 2011-11-02T04:27:26.835Z
  - Type: artifact
  - Session: S\_colocate\_ceres\_2011-11-02T06:03:05
  - Artifact Timestamp: 2011-10-30T21:24:54Z
  - Agent: U\_sflops
  - Instrument: CloudSat
  - Dataset: 2B-CLDCLASS
  - Version: R04\_E02
  - Used: P\_11261\_2011-11-02T06:03:08
  - Buttons: Lineage, Timeline, Reproduce



**BACKUP SLIDES**



# NASA ESD Preservation Content Specification

- Level 1 Satellite Mission Data System Requirements states
  - *“The <<project name>> shall transfer to the <<designated NASA Earth Science Division-assigned Data Center>> all the information and documentation required for long-term preservation of knowledge about the products resulting from <<project name>>, as defined in the NASA Earth Science Data Content Specification.”*
- defines the contents of data, metadata and associated documentation to be preserved beyond the life of missions funded by NASA’s Earth Science Division.
- identify all the content items that need to be preserved to ensure their availability to support future investigations in long-term global change research.
- content items are divided into eight categories: Preflight/Pre-Operations, Products (Data), Product Documentation, Mission Calibration, Product Software, Algorithm Input, Validation and Software Tools.