Provenance Services for Tracking Production of Multi-Sensor Merged Climate Data Records

Jet Propulsion Laboratory
California Institute of Technology
National Aeronautics and Space Administration

Hook Hua, Gerald Manipon, Brian Wilson, Lei Pan
Eric Fetzer, Qing Yue, Alexandre Guillaume
Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA

ACCESS-09-0022

Abstract

Multi-decadal climate data records are critical to studying climate variability and change. These often also require merging data from multiple instruments such as those from NASA's A-Train that contain measurements covering a wide range of atmospheric conditions and phenomena. However, the immense volume and inhomogeneity of data often requires an "exploratory computing" approach to product generation where data is processed in a variety of different ways with varying algorithms, parameters, and code changes until an acceptable intermediate product is generated. This process is repeated until a desirable final merged product can be generated. Typically the production legacy is often lost due to the complexity of processing steps that were tried along the way. The data product information associated with source data, processing methods, parameters used, intermediate product outputs, and associated materials are often hidden in each of the trials and scattered throughout the processing system(s).

We present a set of provenance services to better capture and interpret the production environment, the data products, metadata, and other associated materials. We also developed generic multi-platform clients to be plugged into existing code to communicate production information back to the provenance services. For any product generated, a fully traceable processing lineage is available. Utilizing this provenance information, we are able to facilitate the reproducibility of these production and analysis runs.

Objectives

- Develop generic Web Services-oriented provenance collection capability to track the production legacy of Earth science data products.
- Support first infusion case of tracking production provenance of merged and multi-decadal climate data records that are critical to studying climate variability and change.
- Provide data and processing lineage to facilitate the reproducibility of multi-sensor and multi-decadal climate data records.
- Provide a searchable, browsable, and citable data provenance of multi-sensor merged climate data records.

Approach

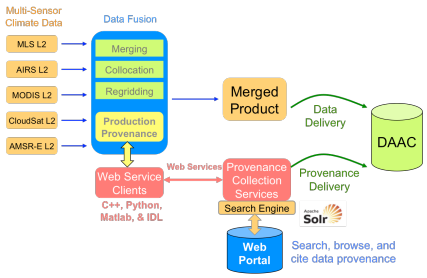
- Leverage existing provenance collection tools to enable capturing and managing of data and processing lineage.
- Develop Web Services and multi-platform service clients (Matlab, IDL, Python, and command-line) to augment provenance collection tools to be used in distributed and heterogeneous computing environments for data fusion.
- Develop web portal interface enabling the searching, browsing, and citing of data product provenance.
- Infuse provenance tracking capability into MEaSUREs data production system and provide provenance as part of climate data record delivery to DAAC.

National Aeronautics and Space Administration

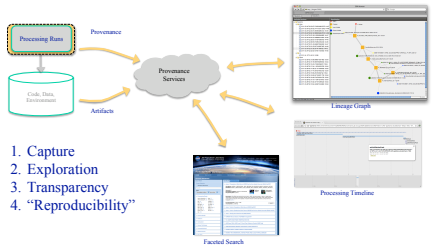
Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California
http://www.nasa.gov

Copyright 2010 California Institute of Technology.
Government sponsorship acknowledged.

Overview



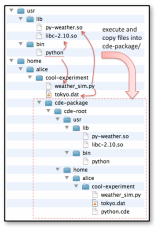
Provenance Services



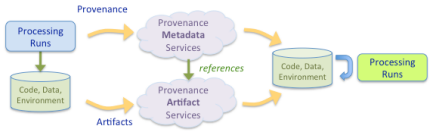
1. Capture
2. Exploration
3. Transparency
4. "Reproducibility"

Facilitate Reproducible Runs

- Leverage CDE* tool which creates a complete record what was run
- Provenance capture of Code, Data, and Environment for reproducibility.
- 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
- Collaborate with colleagues who can rerun on different machines



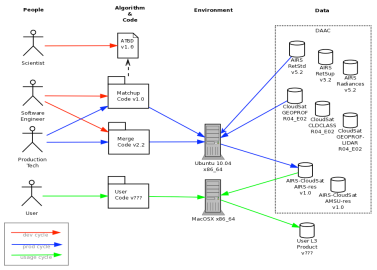
CDE package



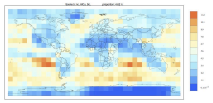
* Philip J. Guo. CDE: Run Any Linux Application On-Demand Without Installation. In Proceedings of the 2011 USENIX Large Installation System Administration Conference (LISA), December 2011.

Multi-Sensor Data Provenance Scenarios

AIRS and CloudSat Matchup Data Production

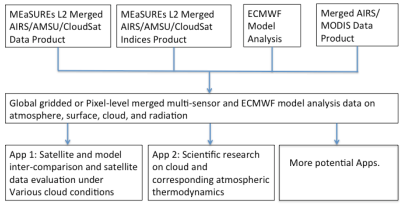


A Multi-Sensor Water Vapor Climate Data Record Using Cloud Classification



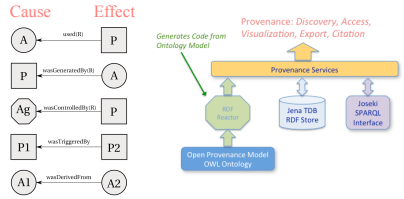
Scientists can now study climatologies that relate a small number of characteristic cloud scenes with atmospheric properties.

Level-3 Analysis using Multi-Sensor Water Vapor Climate Data Record



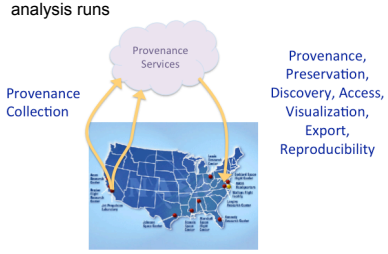
Semantic-enabled Provenance

- Leveraged Open Provenance Model OWL (OPMO)
- Following progress of W3C Provenance Drafts
- Leverage code generation from OPMO model to facilitate backend provenance storage



Long-Term Vision

- **Earth Science Collaboratory (ESC)**
  - Support generic provenance services
- **NASA ESD Preservation Content Specification**
  - Services supporting preservation
- Automated **preservation** of processing runs
- Facilitate **reproducibility** of data production and analysis runs



Persistence Plan

- MEaSUREs "Water Vapor Cloud Classification Climatology"
  - Hand-off infrastructure to MEaSUREs project
- GES DISC
  - Deliver provenance data to GES DISC
  - Provenance record collocated with MEaUSREs WVCC data products
- ESIP
  - Preservation and Stewardship Cluster
    - Continuing work on provenance and preservation model with the ESIP community
    - Effort will continue on after ACCESS task ends
  - Broader Earth Science Collaboratory (ESC) implications
- Open Source

Schedule

- 2010-05: Assess existing provenance tools.
- 2010-08: Develop Web Services for provenance collection.
- 2010-11: Develop multiple platform Web Service clients for provenance instrumentation of processing code.
- 2011-02: Infuse clients into MEaSUREs processing code.
- 2011-10: Develop web portal interface to provenance with faceted search and exploration capabilities.
- 2011-12: Enable data provenance citation.
- 2011-04: Finalize provenance tracking of latest data production.

Issues and Challenges

- Provenance data can be voluminous.
- Overhead of instrumenting code with provenance calls.
- Full reproducibility is a hard problem.
- Computation barriers when employing complex rules for inferring new information from provenance.
- Storing provenance artifacts should leverage data de-duplication