



# **Provenance Tracking in an Earth Science Data Processing System**

Curt Tilmes

*Curt.Tilmes@nasa.gov*

**NASA GSFC  
IS&T Colloquium**

**2009-09-30**



## □ Oxford English Dictionary:

- the fact of coming from some particular **source** or quarter; **origin**, derivation
- the **history** or pedigree of a work of art, manuscript, rare book, etc.;
- concretely, **a record of the passage** of an item through its various owners.



□ Provenance can help a researcher understand:

- **Where** did a result come from? What analysis **process** led to it?
- **Who** (organization/team/scientist) produced the result?
- Are two independent analyses derived from the **same** data?
- How can I independently **reproduce** the experiment to confirm (support) or refute the finding?
- How much should I **trust** the result?

A complete provenance record increases the **credibility** of the result.



- ❑ Earth Science Data Archive volumes growing steadily
- ❑ Over time, things change:
  - Spacecraft, sensors, data processing frameworks
  - Computers, compilers, languages, libraries, formats
  - Science algorithms for transforming and analyzing data
  - (the Earth?)
- ❑ Earth Science data are being used in new ways not planned by originators
- ❑ Multiple data sets are being integrated
- ❑ Value Added Services release their own processed data from independent archives
- ❑ Tracking data provenance through processing systems and archives is a very complicated problem
  - Across organizations / agencies this just gets worse



## Identifiers for Provenance Artifacts

- ❑ Identity is a hard problem, think about the identity of a person... Do **you** have a globally unique, persistent, public identifier?
- ❑ We need to refer to all of the “artifacts” involved or related to a scientific result
  - Data/Data Sources, Sensors/Instruments/Instrument platforms
  - Test Data/Validation data/Analyses
  - Algorithms/Documentation
  - People/Teams/Projects/Organizations (reputation)
  - Published, peer reviewed scientific papers (add to credibility)
  - Computer systems/OS/Compilers/Libraries/Formats
  - Abstract things like “a data transformation event” or “a validation experiment”
  - An ephemeral execution of a web service
- ❑ Multiple versions of all of the above = Rigorous Configuration Management
- ❑ Assign everything a **unique, actionable, persistent** identifier, and maintain equivalences across system/institutional boundaries
- ❑ Identifiers enable representation of relationships and annotations



- ❑ Previous versions of data are often discarded in favor of newer ones
  - Provenance information stored as metadata along with data is usually removed along with the data itself
- ❑ Provenance information is incomplete, and often represented in non-standard forms that are difficult to follow and share
  - Imagine a phone call from a researcher “where did you get this data, and what did you do to it?”
  - Or worse, a FOIA request to NASA asking why we think the global climate is changing...
- ❑ Even if provenance is captured, some systems can't (or won't) reproduce older datasets
  - We usually rely on an error prone, manual process to attempt to reproduce data previously released



## Proprietary information

- Hardware and software designs provide a competitive advantage, why share them?

## US International Traffic in Arms Regulations (ITAR)

- Broadly applied, default is to restrict

## Cost

- Capturing/distributing provenance isn't a priority
- A project that proposes comprehensive provenance management is at a competitive disadvantage to one that doesn't.

## Competition

- Why should I share my system for reproducing my data which would give my competitor a leg up?



- ❑ Capturing complete and accurate provenance during data ingest and primary data processing.
- ❑ Assigning ***persistent identifiers*** to all related artifacts and accurately representing their relationships.
- ❑ Archiving provenance such that it can be ***easily*** retrieved and searched, even if the data are deleted.
- ❑ Supporting comprehensive ***dataset citations*** for scientific literature
- ❑ Representing provenance to human users and providing tools for navigating the graph to search and explore data provenance
- ❑ Representing provenance semantically to other systems at cooperating institutions with standard ontologies
- ❑ Allow agents to traverse inter-system provenance graphs and answer provenance questions
- ❑ Allow ***independent*** systems to mechanically reproduce data processing using the provenance information