Earth Science Provenance

Curt Tilmes
Curt.Tilmes@nasa.gov
Overview

- Provenance
- Earth Science Provenance
- Data Processing and Archiving
- Archiving Provenance
- Reproducibility
- Persistence
- Versions
- Identifiers
- Provenance Objectives
- Provenance Roadblocks
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.
Some modern scientific research is the result of lengthy computer analysis of a very large amount of data, building on the contributions of hundreds (thousands?) of individuals.

http://macuv.gsfc.nasa.gov/ozone.md

http://data.giss.nasa.gov/gistemp/graphs/

Earth Science
“Leading scientists say that the recent controversies surrounding climate research have damaged the image of science as a whole.”

“this crisis of public confidence should be a wake-up call for researchers”

the world had now “entered an era in which people expected more transparency.”

http://news.bbc.co.uk/2/hi/science/nature/8525879.stm
Saturday, Feb 20, 2010
How can we trust the results of such analyses?

**Internal Radiative Forcing And The Illusion Of A Sensitive Climate System By Roy Spencer**

“An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. Therefore, a condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols available in a publicly accessible database [...] or, where one does not exist, to readers promptly on request.”

- (Guide to Publication Policies of the Nature Journals, 2007)

- Science must be reproducible
  - (or it isn't science...)

- Traditionally, one could read a scientific paper, construct an identical experiment and confirm results
  - (well, most of the time...)

- **Reproducibility yields Credibility**
Earth Science Data Archive volumes growing steadily

Over time, the systems evolve:
- Spacecraft, sensors, data processing frameworks
- Science algorithms for transforming and analyzing data
- Calibration, ancillary lookups

Tracking data provenance through processing systems and archives is a very complicated problem
- Across organizations / agencies this just gets worse

Science data is being used in new ways not planned by originators

Value Added Services release their own processed data from independent archives

Remote web services can be used to transform data
Data Processing and Archiving

- 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 represented in non-standard forms that are difficult to follow
  - Imagine a phone call to a researcher “where did you get this data, and what did you do to it?”

- Even if provenance is captured, some systems can’t (or won’t) reproduce older datasets
  - Rely on an error prone, manual process to attempt to reproduce data previously released
Earth Science Provenance Artifacts

- All of the “artifacts” involved or related to the scientific result:
  - Data
  - Algorithms, Configuration Tables, Runtime Parameters
  - Documentation (ATBDs, Design Docs, Commented Source)
  - Sensors/Instruments/Instrument platforms
  - People (reputation)
  - Organizations (reputation)
  - Published scientific papers (add to credibility and understanding)
  - Computer systems, Hardware, OS, Libraries, Software
  - Abstract things like “a data transformation event,” “Software Build Event” or “a validation experiment”
  - An ephemeral execution of a web service
  - Versions from all of the above: Rigorous Configuration Management.

- Things that increase understanding or reproducibility.
Provenance “Essential” for Reproducibility

- What aspects of the provenance are “essential” for reproducibility?
- Can't record “Big Bang” provenance
  - the “butterfly effect”
- Some things are definitely “essential”
  - List of input files
- Some things are definitely “non-essential”
  - Name of processing host
- Some things aren't so clear
  - Heinrich Hertz testing Maxwell's Equations – didn't report the size of the room he worked in – turned out to be “essential”
Scientific Reproducibility

- Not necessarily a perfect match, bit-for-bit
- Different criteria depending on specific scientific meaning of the fields
- Accuracy and precision of measurements and their representation in the data structures
- Recorded provenance must be sufficient for an independent researcher to reproduce the analysis and confirm the results and conclusions
- Science software developers must develop robust code to ensure reproducibility and limit dependence on a particular computer/compiler/environment.
Capturing the provenance for every single granule of data results in a lot of data.

Most of it is very similar:
- $p_i$ uses $a_i$ and produces $b_i$

Summarize “granule” provenance into “dataset” provenance.

Answer provenance queries with “dataset” provenance where appropriate.
Versions

- Every algorithm has strict configuration management with versions mapping to revisions
- What does “version” mean to data?
- Consider Algorithm X of version 1.2 is used to produce file A
- If we revise algorithm X and reprocess with version 1.3, the produced file A is different, we note in its metadata that it was produced with version 1.3
- Now what happens if we recalibrate the instrument that produced the data that was fed to algorithm X without changing the version of the algorithm itself?
Persistence

• “It is intended that the lifetime of a [persistent identifier] be permanent. That is, the [persistent identifier] will be **globally unique forever**, and may well be used as a reference to a resource well beyond the lifetime of the resource it identifies or of any naming authority involved in the assignment of its name.”

http://www.doi.org/doi_presentations/overview_slides_4Dec2007/071205DOIOverview.ppt

- Data used to produce scientific results should be cited as rigorously and persistently as referenced papers.
- The provenance graph associated with a published component of the scientific literature should live as long as the publication is scientifically valid. (In fact, you could use a citation chain to determine which data are referenced.)
'Actionable' Identifier = *Can I click on it?*

- What happens if the resource itself is no longer around? We (NASA archive) delete old, obsolete data that takes up expensive space.

Even if the data are gone, the identifier should still be valid.

What happens if valuable data is moved from one “steward” to another? (We do this all the time...)

- An entire archive taken over by another organization
- A single dataset within the archive moved from one organization to another
- What about data served from multiple locations?
- What about data served in multiple formats?
Provenance Objectives

- Capturing complete and accurate provenance during data ingest and primary data processing
- Archiving provenance such that it can be easily retrieved and searched, even if the data are deleted
- Representing provenance to human users and providing tools for navigating graph to search and explore data provenance
- Representing provenance semantically to other systems at cooperating institutions with standard ontologies
  - Semantic Web for Earth and Environmental Terminology (SWEET)
  - Open Provenance Model (OPM)
  - Proof Markup Language (PML)
- Allow agents to traverse inter-system provenance graphs and answer provenance questions
- Allow independent systems to mechanically reproduce data processing using the provenance information
Provenance Roadblocks

- **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 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?

- **Standards for representing/sharing provenance.**
“Cool URIs don't change” –  http://www.w3.org/Provider/Style/URI

Kunze, “The ARK Identifier Scheme” –  

http://www.springerlink.com/content/edf0k68ccw3a22hu/