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

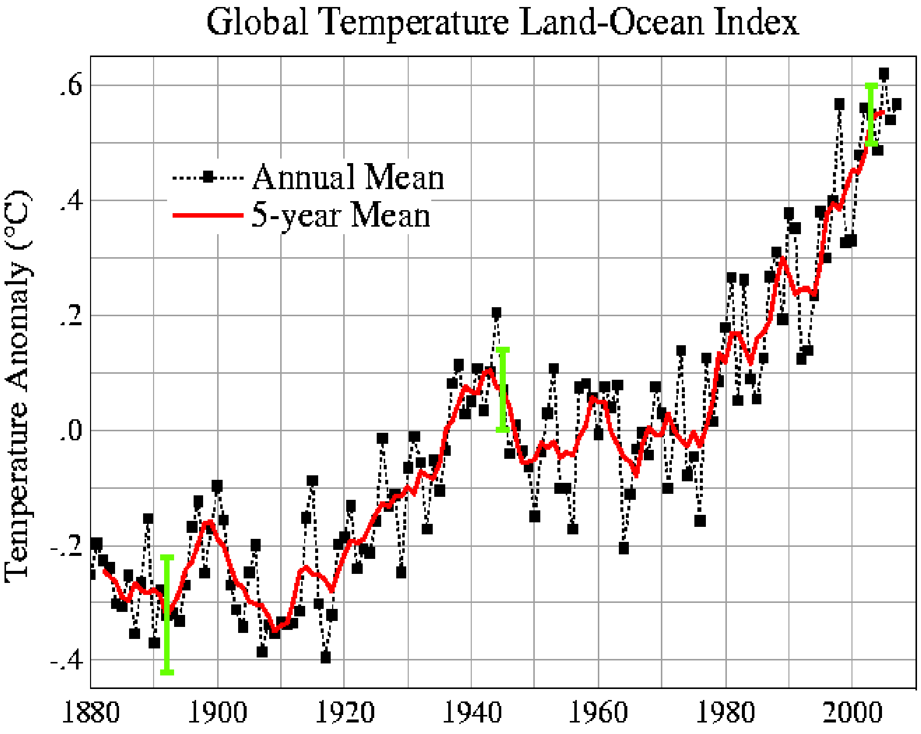
Provenance

❑ Oxford English Dictionary:

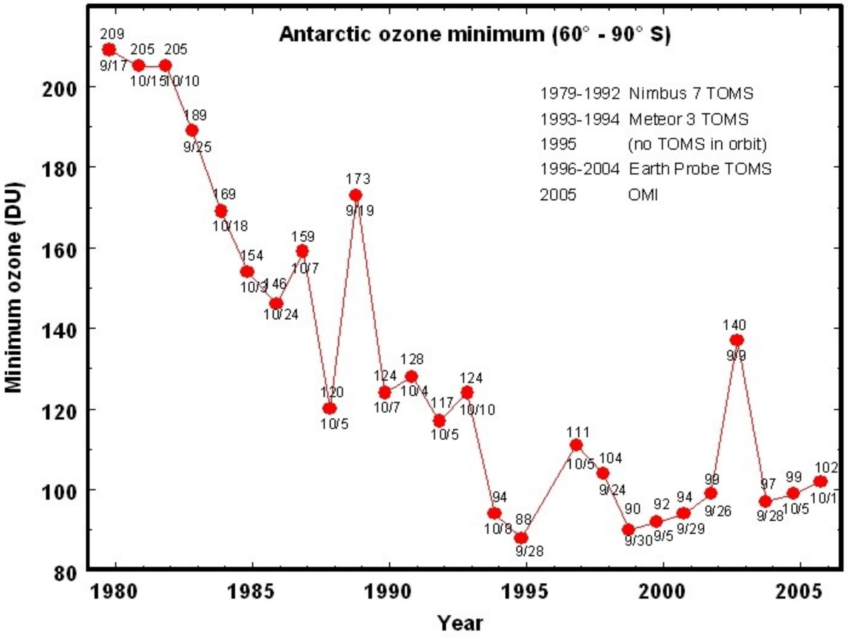
- the fact of coming from some particular source or quarter; origin, derivation
- the history or pedigree of a work of art, manuscript, rare book, etc.;
- concretely, a record of the passage of an item through its various owners.

Content adapted from the EU Grid Provenance Project, sponsored by IBM UK

- ❑ Some modern scientific research is the result of lengthly computer analysis of a **very large** amount of data, building on the contributions of hundreds (thousands?) of individuals



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



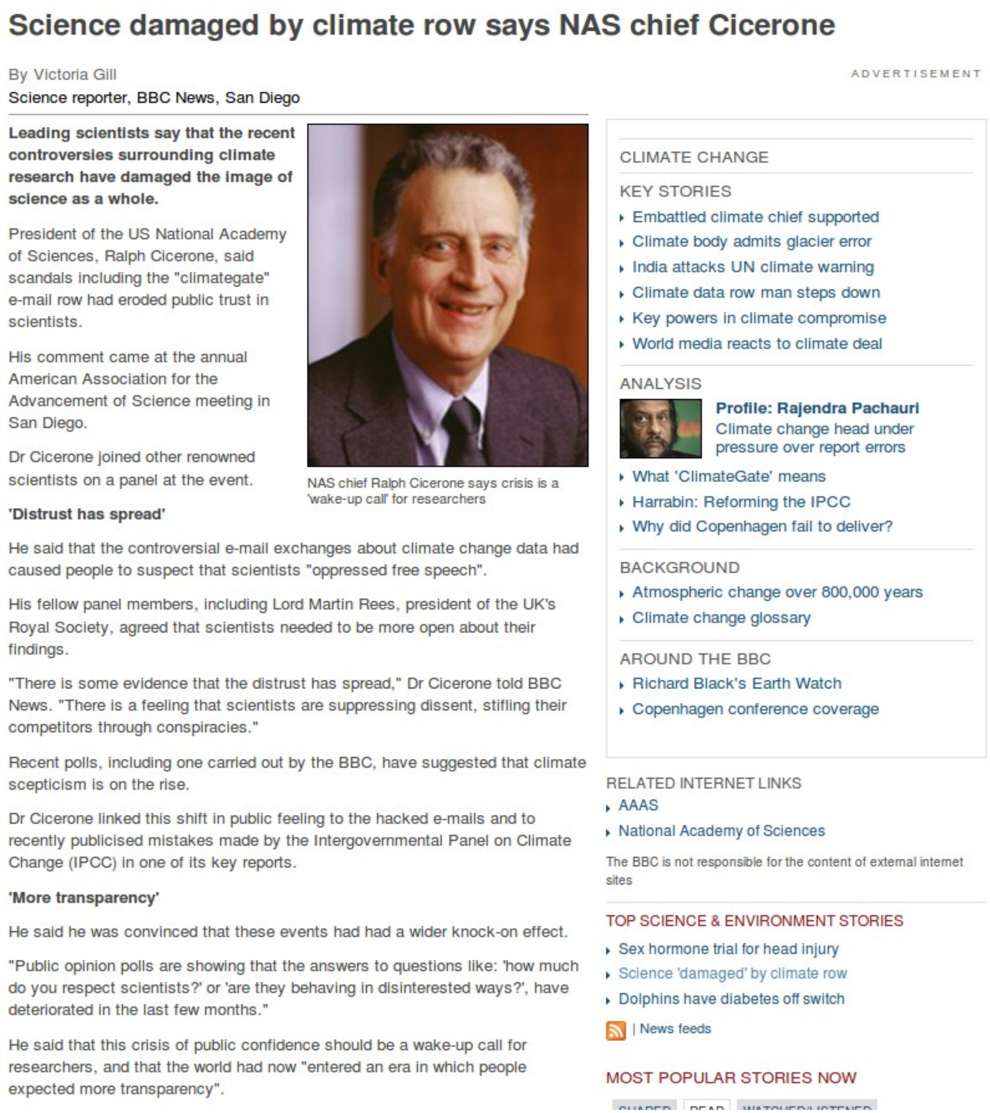
http://macuv.gsfc.nasa.gov/ozone.md

“Climategate”

“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.”



Science damaged by climate row says NAS chief Cicerone

By Victoria Gill
Science reporter, BBC News, San Diego

ADVERTISEMENT

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

President of the US National Academy of Sciences, Ralph Cicerone, said scandals including the "climategate" e-mail row had eroded public trust in scientists.

His comment came at the annual American Association for the Advancement of Science meeting in San Diego.

Dr Cicerone joined other renowned scientists on a panel at the event.

NAS chief Ralph Cicerone says crisis is a 'wake-up call' for researchers

'Distrust has spread'

He said that the controversial e-mail exchanges about climate change data had caused people to suspect that scientists "oppressed free speech".

His fellow panel members, including Lord Martin Rees, president of the UK's Royal Society, agreed that scientists needed to be more open about their findings.

"There is some evidence that the distrust has spread," Dr Cicerone told BBC News. "There is a feeling that scientists are suppressing dissent, stifling their competitors through conspiracies."

Recent polls, including one carried out by the BBC, have suggested that climate scepticism is on the rise.

Dr Cicerone linked this shift in public feeling to the hacked e-mails and to recently publicised mistakes made by the Intergovernmental Panel on Climate Change (IPCC) in one of its key reports.

'More transparency'

He said he was convinced that these events had had a wider knock-on effect.

"Public opinion polls are showing that the answers to questions like: 'how much do you respect scientists?' or 'are they behaving in disinterested ways?', have deteriorated in the last few months."

He said that this crisis of public confidence should be a wake-up call for researchers, and that the world had now "entered an era in which people expected more transparency".

CLIMATE CHANGE

KEY STORIES
- Embattled climate chief supported
- Climate body admits glacier error
- India attacks UN climate warning
- Climate data row man steps down
- Key powers in climate compromise
- World media reacts to climate deal

ANALYSIS

Profile: Rajendra Pachauri
Climate change head under pressure over report errors

- What 'ClimateGate' means
- Harrabin: Reforming the IPCC
- Why did Copenhagen fail to deliver?

BACKGROUND
- Atmospheric change over 800,000 years
- Climate change glossary

AROUND THE BBC
- Richard Black's Earth Watch
- Copenhagen conference coverage

RELATED INTERNET LINKS
- AAAS
- National Academy of Sciences

The BBC is not responsible for the content of external internet sites

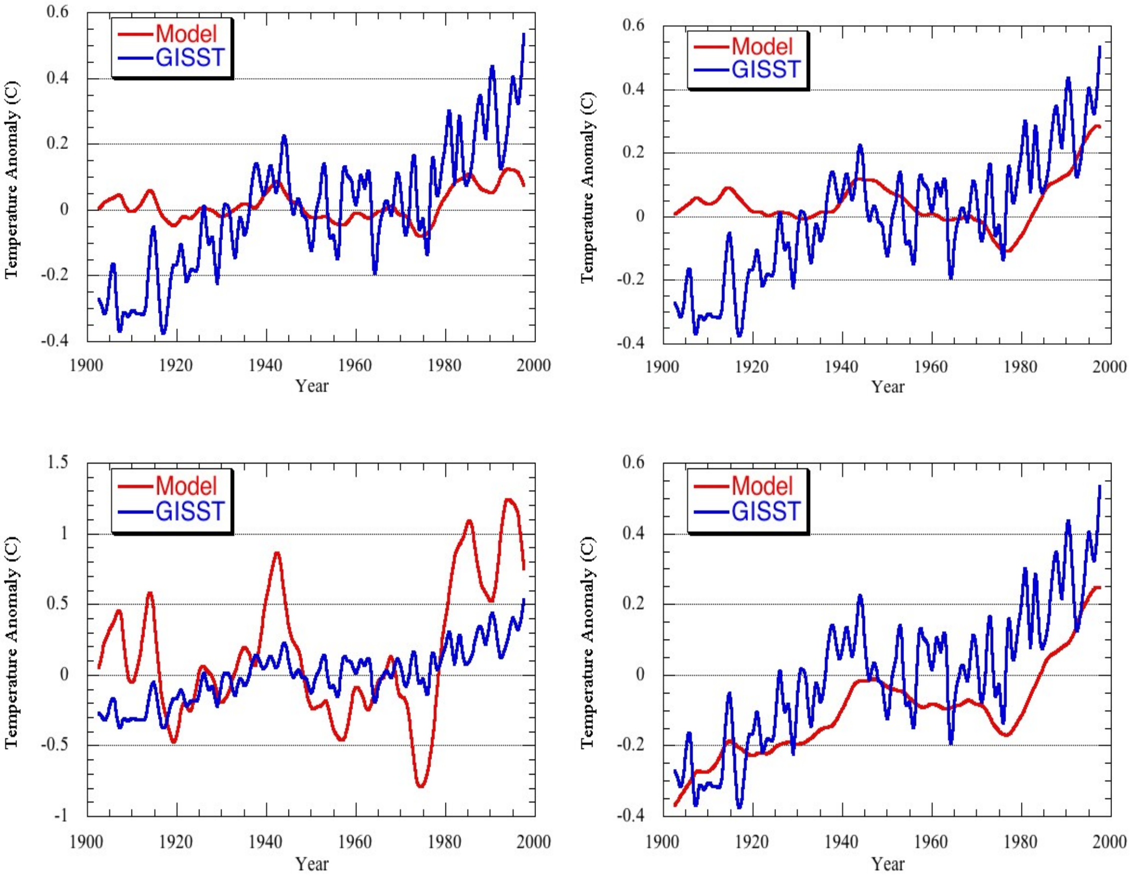
TOP SCIENCE & ENVIRONMENT STORIES
- Sex hormone trial for head injury
- Science 'damaged' by climate row
- Dolphins have diabetes off switch

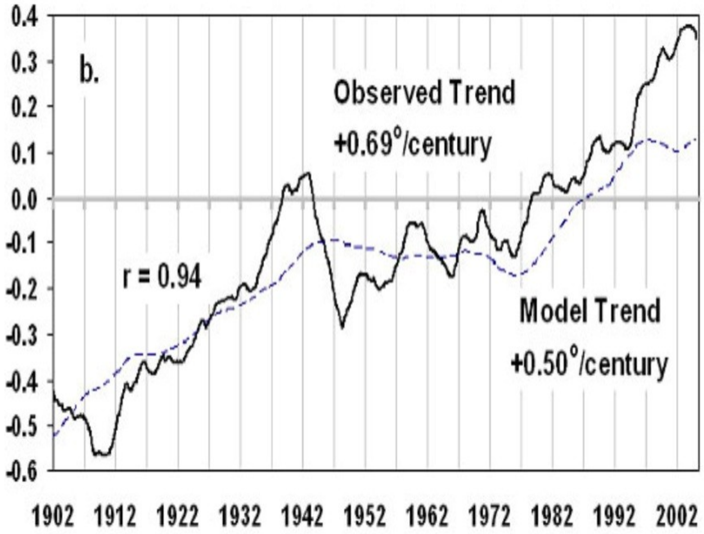
| News feeds

MOST POPULAR STORIES NOW

http://news.bbc.co.uk/2/hi/science/nature/8525879.stm
Saturday, Feb 20, 2010

How can we trust the results of such analyses?

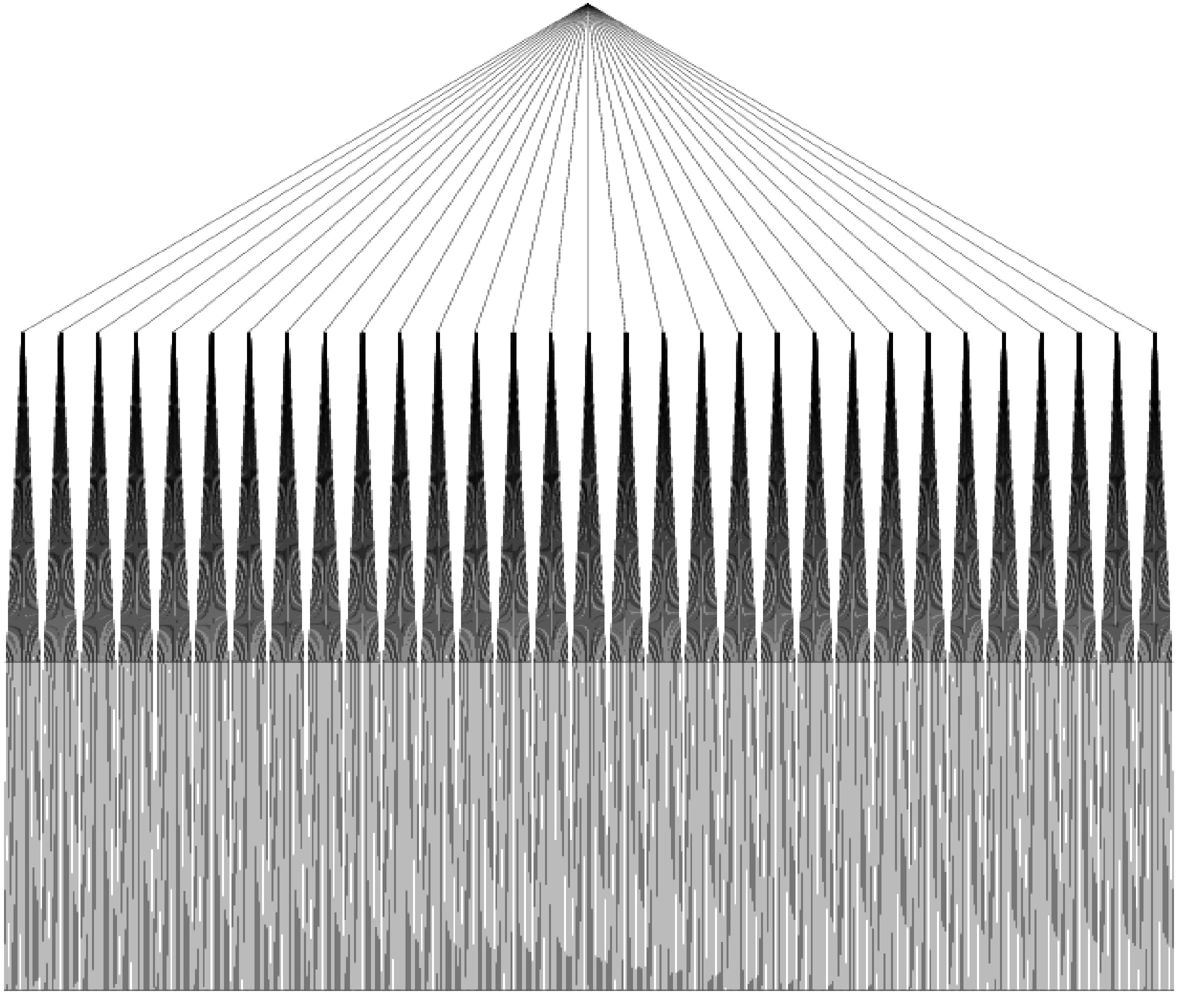




Internal Radiative Forcing And The Illusion Of A Sensitive Climate System By Roy Spencer
http://climatesci.org/2008/04/22/internal-radiative-forcing-and-the-illusion-of-a-sensitive-climate-system-by-roy-spencer/

“How to cook a graph in three easy lessons,” raypierre, 2007-05-21.
http://www.realclimate.org/index.php/archives/2008/05/how-to-cook-a-graph-in-three-easy-lessons/

Ozone Processing



Science

- □ “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*

Data Processing and Archiving

- ❑ 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.

Dataset vs. Granule Provenance

- ❑ 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

Earth Science Data Versions

❑ 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 Identifiers

- ❑ '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.

References

- ❑ “Cool URIs don't change” – `http://www.w3.org/Provider/Style/URI`
- ❑ Kunze, “The ARK Identifier Scheme” – `http://tools.ietf.org/html/draft-kunze-ark-15`
- ❑ Buneman, “Why and Where: A Characterization of Data Provenance” – `http://www.springerlink.com/content/edf0k68ccw3a22hu/`
- ❑ Wolfe, Linda, “MODIS Science Software and Product Versioning White Paper” – `http://modis.gsfc.nasa.gov/sci_team/meetings/199812/presentations/wolfe_paper.pdf`