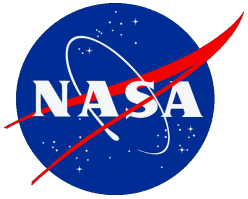


UMBC
AN HONORS UNIVERSITY IN MARYLAND

Provenance Artifact Identification in the Atmospheric Composition Processing System (ACPS)

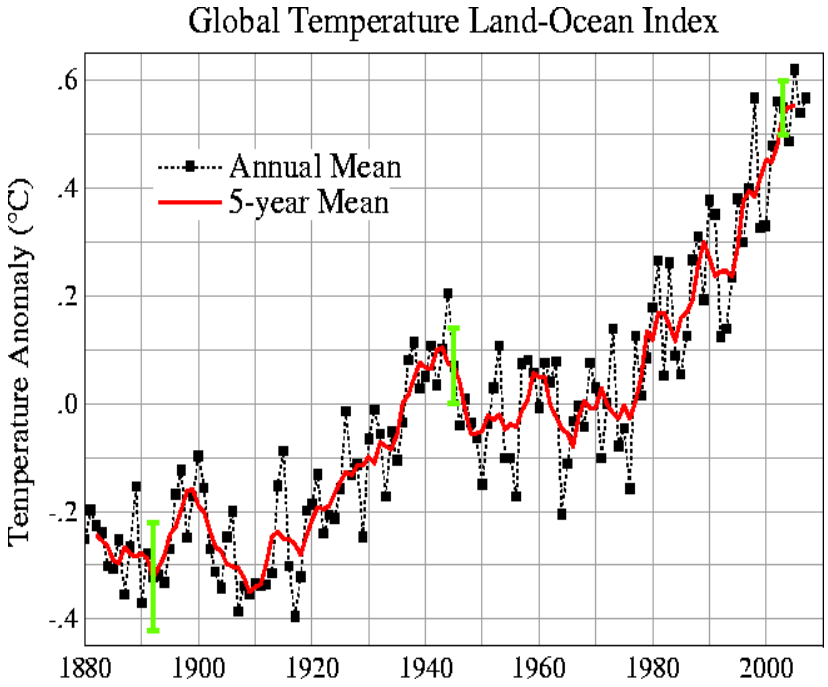
Curt Tilmes
NASA/UMBC

Yelena Yesha
UMBC

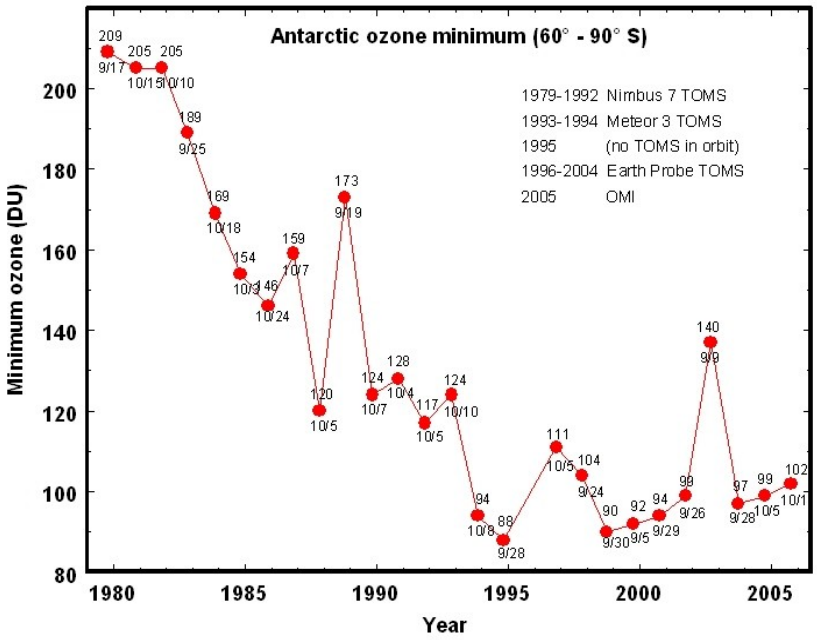
Milton Halem
UMBC

Overview

- Background
- Earth Science Processing Artifacts
- Persistence
- Actionable Identifiers
- Earth Science Data Versions
- Granularity
- ArchiveSets
- Persistent URLs
- Artifact Web Server
- Semantic Web and Linked Data



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



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

"Climategate"

"scandals including the `climategate' e-mail row had eroded public trust in scientists"

"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

Science damaged by climate row says NAS chief Cicerone

By Victoria Gill
Science reporter, BBC News, San Diego

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

Background

- Modern research in earth science often involves sifting through mounds of data from a variety of sources (field sensors, satellite data, etc.) and applying various algorithms to reduce/transform/massage that data in various ways
- The data are likely the result of the work of hundreds of individuals from multiple organizations over decades.
- They are stored in multiple long term archives (which often change over time as well).
- This science relies on representing the provenance of such scientific results in a manner conducive to exploration, understanding and reproducibility.
- We need persistent identifiers to represent the artifacts of processing and their relationships.

Earth Science Processing Artifacts

- All of the "artifacts" involved in the provenance of a scientific result:
  - Data
  - Algorithms
  - Documentation
  - Sensors/Instruments/Instrument platforms
  - People (reputation)
  - Organizations (reputation)
  - Published scientific papers (add to credibility)
  - 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

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

- 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?

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?

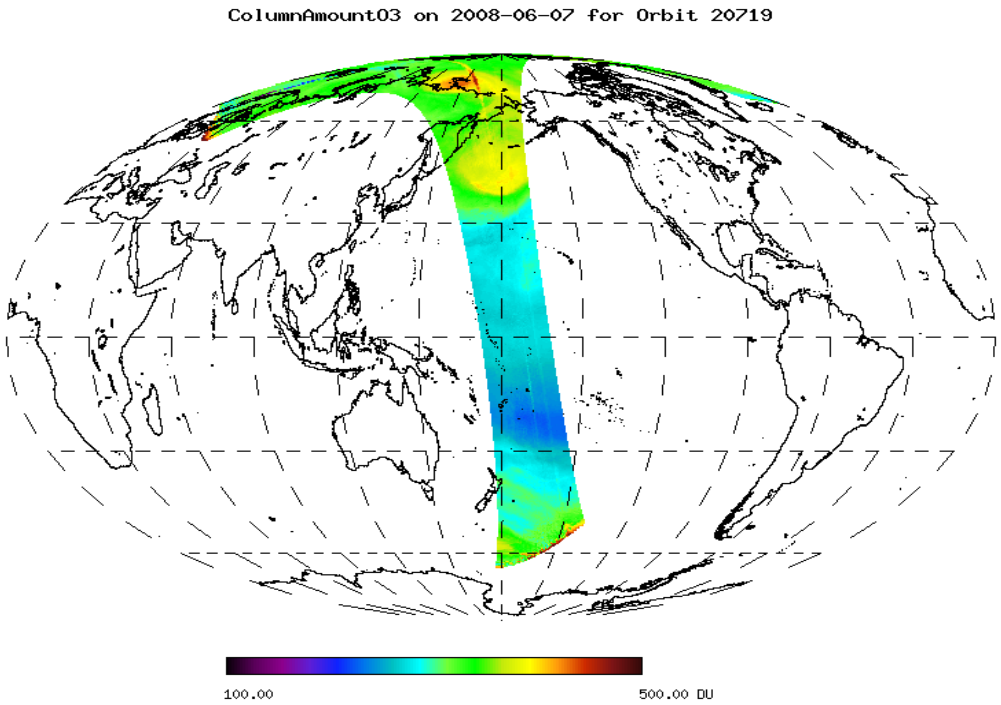
Granularity

- Dealing with data at the extremes of granularity is awkward:
  - All data from all places for all times
  - A single measurement of some property for a single place at a single instant in time.
- Convention breaks down data into “granules” where neither the size of a single granule nor the total number of granules in a dataset are overwhelming.
- For a large amount of very consistent data, we can define:
  - A consistent granule definition (spatial/temporal/other)
  - A **Granule Key** that can uniquely identify a granule in a dataset.
  - A well-defined mechanism for iterating through the granules in a dataset.

Earth Science Data Type

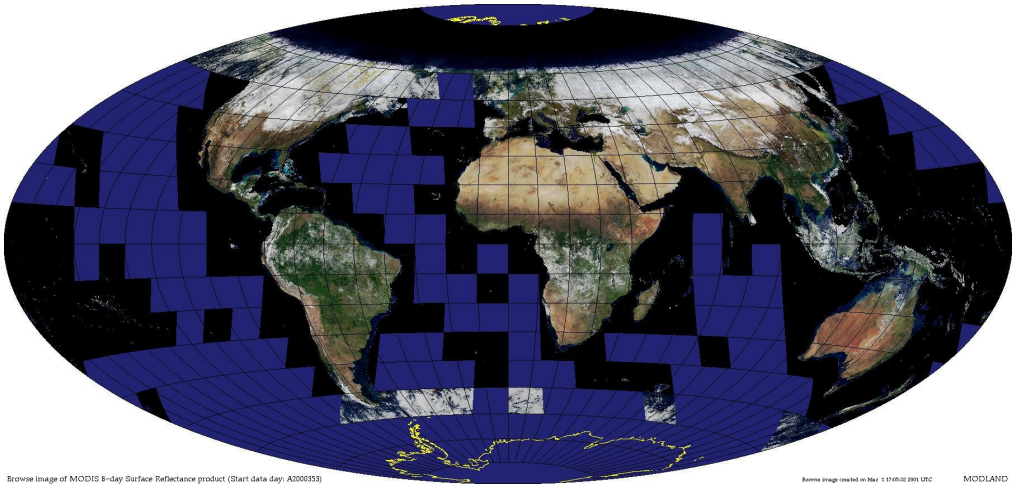
- Earth Science Data Type (**ESDT**) defines a short key for each standard data product:
  - A specific algorithm (with published Algorithm Theoretical Basis Document 'ATBD')
  - A specific data format
  - A specific data **Granularity**

Granularity Example: OMTO3



ESDT=OMTO3
Granularity = Orbital
Granule Key = 20718

Granularity Example: MODIS 8day LSR



ESDT=MOD09A1
Granularity = 8DayTiled
Granule Key = 2000353,12,17 (year/doy,Hor., Ver.)

ArchiveSets

- The ACPS uses **ArchiveSet**s to differentiate processing runs, experiments, etc.
- The key concept is that {ArchiveSet,ESDT,Granule Key} is always unique at a point in time.
- If a newly created file matches one already in the ArchiveSet, the old one is automatically removed from the 'current' ArchiveSet.
- We call {ArchiveSet,ESDT} a **DataSet.**
- A Granularity Iterator can be used to enumerate all the Granule Keys in a DataSet.
- Timestamps are used to precisely maintain the granule membership at any historic point in time, so {DataSet,Timestamp} refers uniquely to a set of files, none of which have the same Granule Key.

PURL: Persistent URL

- Very simple indirect mapping that redirects from a PURL to a URL with standard HTTP redirect
- Includes “partial redirects” to relocate whole hierarchies

<scheme>://<PURL resolver>/<name>

http://purl.org/*mypath/mylocalid*

http://purl.org/NET/ACPS/<ArtifactType>/<ArtifactIdentifier>

PURL Examples

```
http://purl.org/NET/ACPS/Granularity/Orbital

http://purl.org/NET/ACPS/ESDT/OMTO3

http://purl.org/NET/ACPS/APP/OMTO3/v1.2.5

http://purl.org/NET/ACPS/DataEvent/52782

http://purl.org/NET/ACPS/BuildEvent/125526

http://purl.org/NET/ACPS/Granule/17/OMTO3/28794

http://purl.org/NET/ACPS/Granule/17/OMTO3/28794/2009-12-01T17:15:28

http://purl.org/NET/ACPS/Dataset/17/OMTO3/2009-12-01T17:15:28
```

Data Citations can include the 'DataSet' identifier, fully qualified with a timestamp to refer to a specific set of granules.

Artifact Web Server

- Each identifier is 'actionable' and will return the metadata (or data) associated with that artifact, including the relationships with other artifacts.
- Maintain the metadata and relationship graph even if the data themselves are deleted.
- Multiple fomats returned based on HTTP Content-Type/Accept headers:
  - YAML – A human friendly format useful for debugging and testing.
  - XML – The modern standard for data interchange, easy to parse and transform
  - JSON – A lightweight data-interchange language that is particularly easy to incorporate into dynamic web sites.
  - RDF/OWL – Suitable for ingest into triple stores supporting complex queries, reasoning and data mining.

- The RDF/OWL representation allows our provenance graphs to be easily traversed and handled by standard Semantic Web software.
- We can also establish equivalences and relationships with other entities following the principles of Linked Data, linking to scientific literature publications, standard instrument identifiers, scientist identifiers, etc.
- We plan to be compatible with OPM RDF/OWL representations, and are also experimenting with Proof Markup Language (PML).