UMBC

AN HONORS UNIVERSITY IN MARYLAND

Data Provenance Management for Earth Science Reproducibility

Curt Tilmes
NASA/UMBC

Yelena Yesha
UMBC

Milton Halem
UMBC

MC^2

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

Science

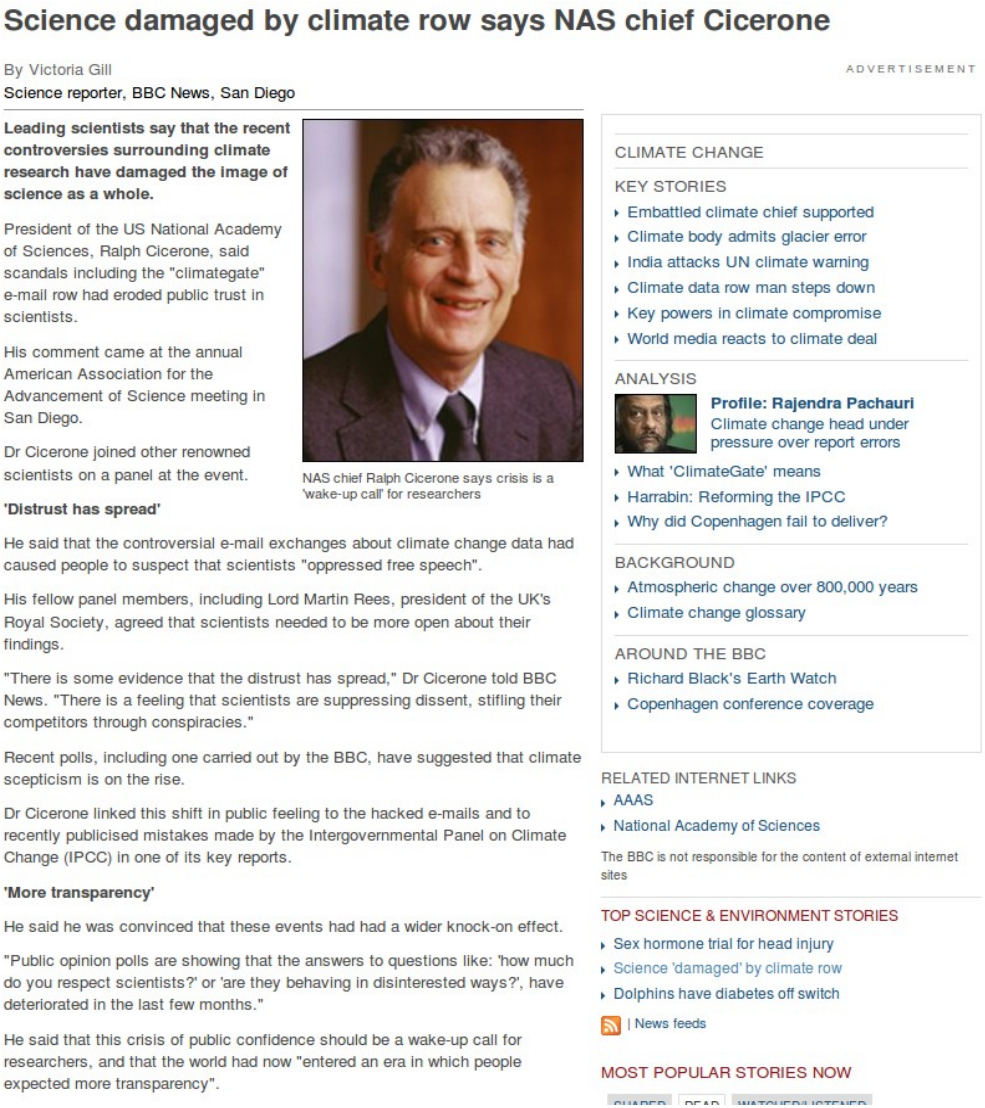
- “An inherent principle of publication is that others should be able to replicate and build upon the authors' published claims. Therefore, a condition of publication in a Nature journal is that authors are required to make materials, data and associated protocols available in a publicly accessible database [...] 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*

"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

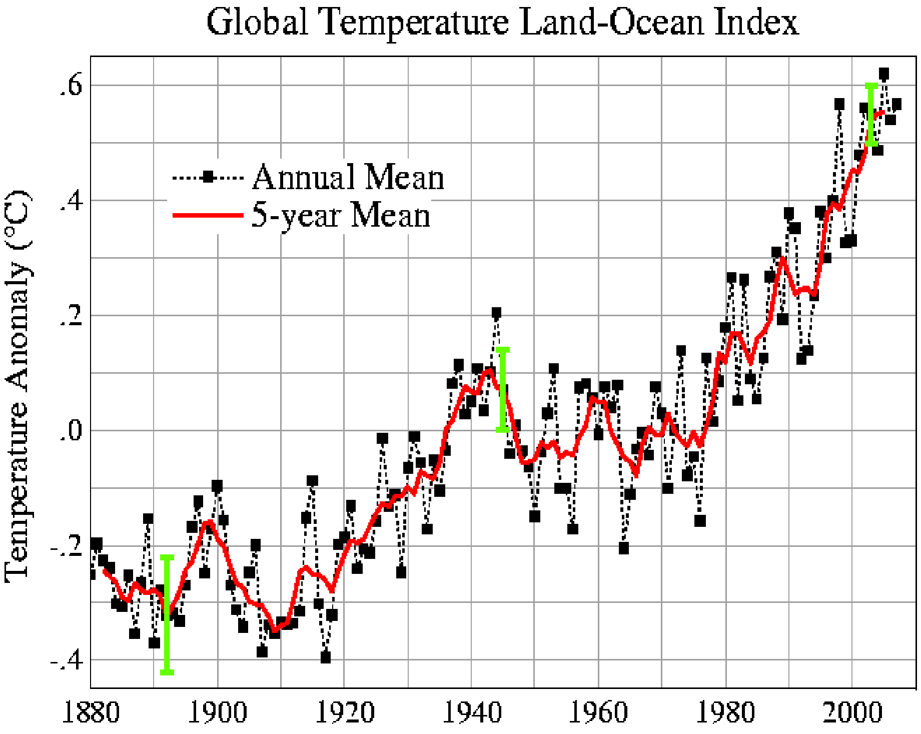
TOP SCIENCE & ENVIRONMENT STORIES
- Sex hormone trial for head injury
- Science 'damaged' by climate row
- Dolphins have diabetes off switch

| News feeds

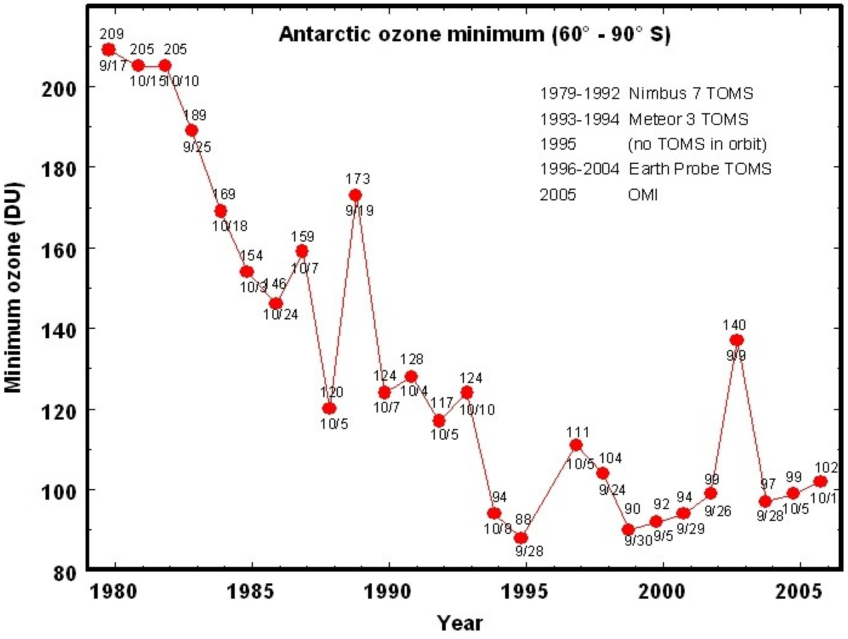
MOST POPULAR STORIES NOW

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

- ❑ 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

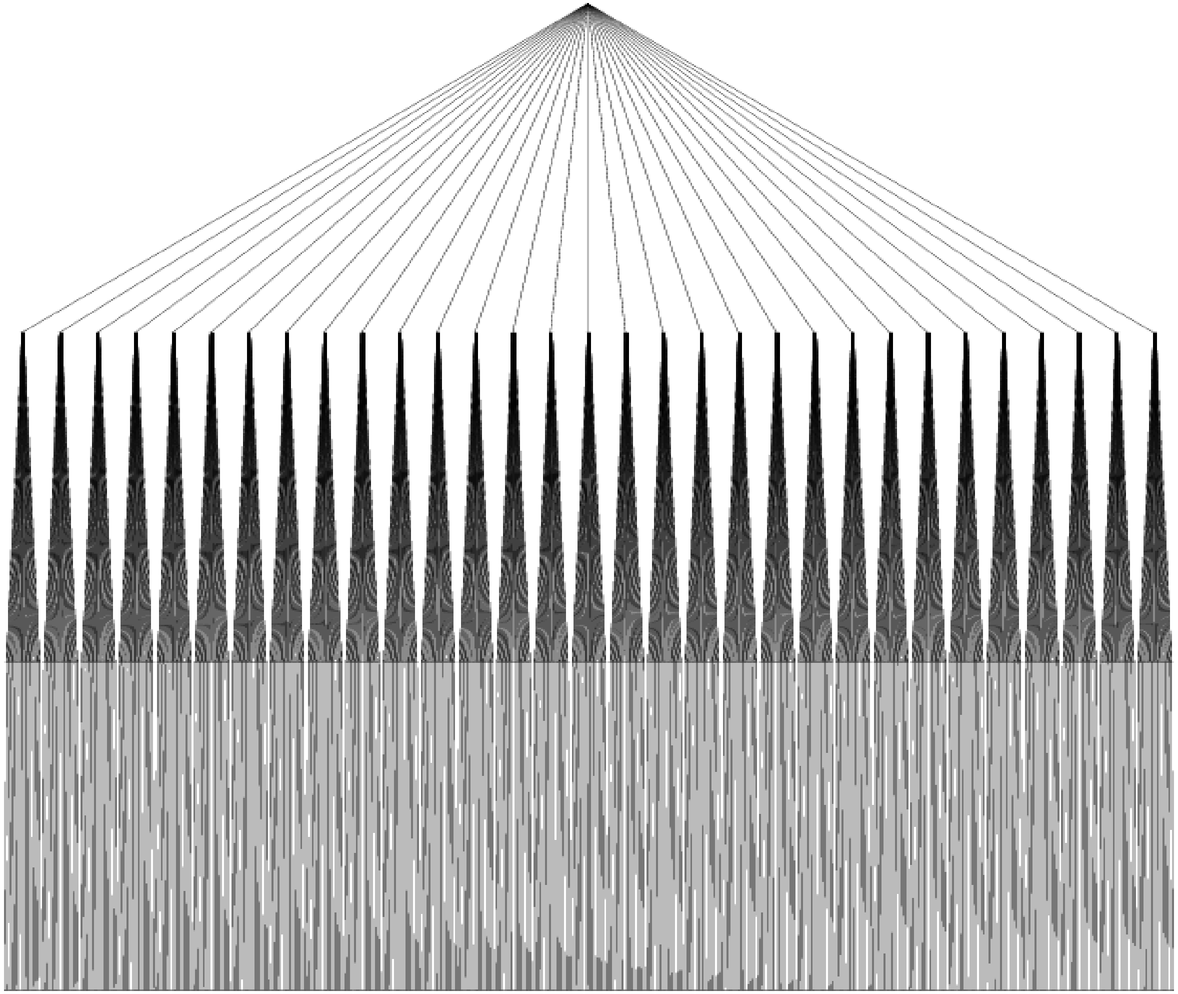


http://data.giss.nasa.gov/gistemp/graphs/



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

Ozone Processing



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

- ❑ 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 Provenance Artifacts

- ❑ All of the “artifacts” involved or related to the scientific result:
  - Data
  - Algorithms, Processes, Configuration Tables, Runtime Parameters (“Workflow Provenance”)
  - Documentation (ATBDs, Design Docs, Commented Source)
  - Sensors/Instruments/Instrument platforms
  - People/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.
  - Specific relationships between all the artifacts.
- ❑ Things that increase *understanding* and enable *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"
  - Workflow artifacts
- ❑ Some things are definitely "non-essential"
  - Name of processing host
  - These are useful for auditing and increase credibility of provenance.
- ❑ 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* in diverse, heterogeneous environments and *limit dependence* on a particular computer/compiler/environment.

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.

□ A data citation should include a persistent identifier for the specific data used in the research.

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 are 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?

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?
- Versions that change too often aren't useful. (You used version 1,714 of the data, I used 1,759. What's the difference?)

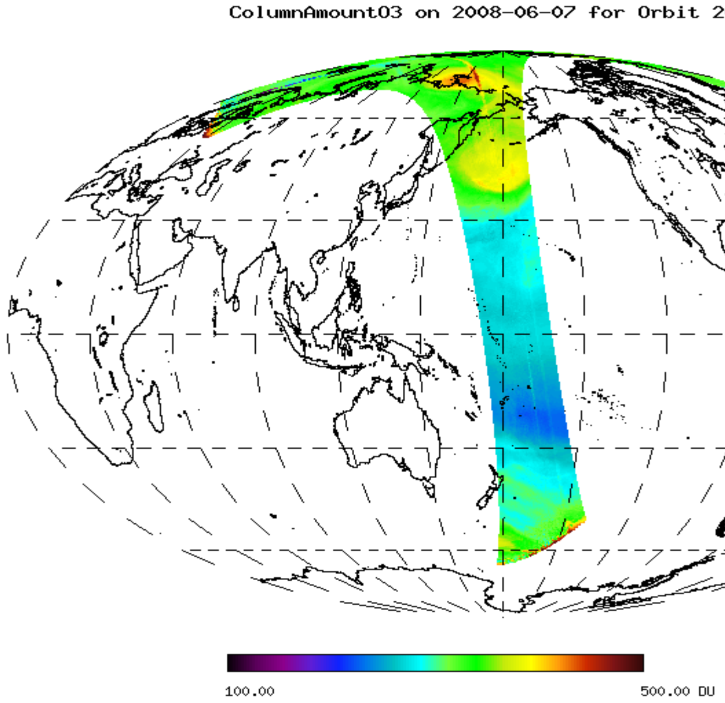
Granularity

- ❑ 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 unique, well-defined **Granularity.**
  - 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

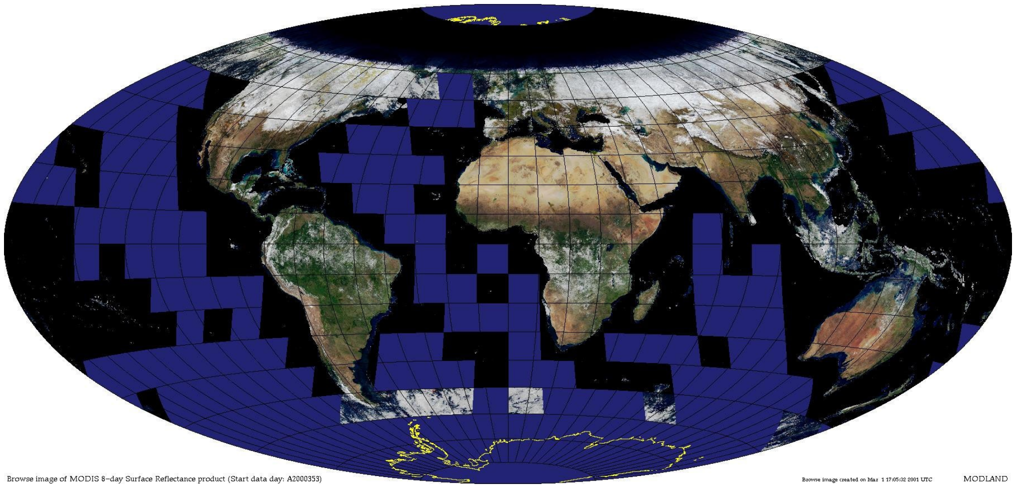
- Earth Science Data Type (**ESDT**) defines a short key for each standard data product:
  - A specific algorithm (with published Algorithm Theoretical Basis Document 'ATBD')
  - A specific data format
  - A specific data **Granularity**

Granularity Example: OMTO3



ESDT = OMTO3
Granularity = Orbital
Granule Key = 20718

Granularity Example: MODIS 8day LSR



Browse image of MODIS 8–day Surface Reflectance product (Start data day: A2000353)

Browse image created on Mar 1 17:05:32 2001 UTC

MODLAND

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

Granularity Edge Cases

- ❑ Those examples are the easy cases.
- ❑ For weird things, I resort to Key is “something unique” and the Iterator is simply “list of the Keys already used”.
- ❑ The concepts still hold.

ArchiveSets

- ❑ **ArchiveSets** differentiate processing runs, experiments, etc.
- ❑ The key concept is that **{ArchiveSet,ESDT,Granule Key}** is always unique at a point in time, or more generally, **{ArchiveSet,ESDT,Granule Key,TimeStamp}** maps to a single unique granule within a system.
- ❑ If a newly created file has a granule key that matches one already in the ArchiveSet, the old one is automatically removed from the 'current' ArchiveSet.
- ❑ For each ESDT, maintain a list of the 'best' Archivesets
- ❑ Multiple ESDTs within an ArchiveSet generally have similar characteristics (processed from the same lower level data, calibrated in a common way).

DataSets

- ❑ We call **{ArchiveSet,ESDT}** a **DataSet.**
- ❑ A Granularity Iterator can be used to enumerate all the possible Granule Keys in a DataSet.
- ❑ Timestamps are used to precisely maintain the granule membership set at any historic point in time, so **{DataSet,Timestamp}** refers uniquely to a set of files, none of which have the same Granule Key.
- ❑ Note: Granules in a DataSet could be generated in a different way from one another (e.g. start processing with version 1.2, upgrade to 1.3 and process some more).

Granules and Datasets

- Each Granule and Dataset has a persistent identifier.
- Dereferencing that identifier will lead to all the provenance artifacts associated with that granule/dataset – Any identifier can be an entrance into the overall provenance graph.
- Dereferencing a granule identifier could lead to the specific file(s) (bunch of bits) holding that data associated with that granule.
- It could also lead to things like "produced this granule, but later deleted" or even "never made for reason *X*"
- A Granule or DataSet can be copied to another system.
- A Granule or DataSet can be reproduced mechanically by repeating the essential events that led to its creation.

Dataset vs. Granule Provenance

- ❑ Capturing the provenance for every single granule of data results in a lot of data, this can be difficult for people to work with.
- ❑ Most of it is very similar
  - p_i uses a_i and produces b_i
- ❑ Summarize “granule” provenance into “dataset” provenance
- ❑ Coalesce commonalities (1-1000 were made with version 1.2), but maintain differences (1-1000 were made with version 1.2, 1001-2000 were made with version 1.3)
- ❑ Answer provenance queries with “dataset” provenance where appropriate
- ❑ In particular, I'm interested in comparing dataset provenance – What is the difference between dataset *A* and dataset *B*?
- ❑ Use the “essential” property to differentiate between things I care about (version of the algorithm used to produce granule *X*) from things I probably don't (granule *X* was produced on host *fred* in dataset *A* and host *barney* in dataset *B*)

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 to another system/archive.

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

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

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

PURL Example URIs

```
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
```

A DOI (Digital Object Identifier) could map to the DataSet without timestamp. A data citation could include the DOI + a timestamp. That would refer to the specific set of Granules that were part of that DataSet at that time.

Timestamp is ISO 8601, and hierarchical, so for most purposes “Year-Month-Day” would be sufficient.

Artifact Web Server

- ❑ Each identifier is 'actionable' and will return the metadata (or data) associated with that artifact, including the relationships with other artifacts.
- ❑ Can redirect hierarchy subsets to other compatible servers.
- ❑ Maintain the metadata and relationship graph even if the data themselves are deleted.
- ❑ Multiple formats 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 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 (CiteSeer et al), standard instrument identifiers, scientist identifiers, etc.
- ❑ Plan to be compatible with OPM RDF/OWL representations, and are also experimenting with Proof Markup Language (PML) ontologies.

Reproducibility

- The data processing system is complicated, it has complex scheduling rules, production rules with database queries to manage processing dependencies and determine the best possible inputs for each run.
- For reproducibility, we don't need all that. The answer is just “run the same way you did before.”
- A light-weight processor can read the provenance graph and reproduce a single granule, or iterate arbitrarily within a DataSet to reproduce any subset.
- With virtual processing environments, we can archive not just the program, but also the right (and minimal) environment to run it in.
- Cloud processing systems could be harnessed to extend reprocessing to anyone who wants to pay a cloud provider some money.

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.