Provenance in Data Interoperability for Multi-Sensor Intercomparison

[1]Chris Lynnes, [1]Greg Leptoukh, [1]Steve Berrick, [1]Suhung Shen, [1]Ana Prados, [1]Peter Fox, [1]Wenli Yang, [1]Mike Min, [1]Dan Holloway, [2]Tanisha Enloe ([1]NASA GSFC, [2]RPI, [3]GMU, [4]OPeNDAP, Inc., [5]SGT)

Goddard Earth Sciences Data & Information Services Center
NASA Goddard Space Flight Center
Greenbelt, Maryland 20771 USA
chris.lynnes@nasa.gov

Abstract As our inventory of Earth science data sets grows, the ability to compare, merge and fuse multiple datasets grows in importance. This requires a deeper data interoperability than we have now. Efforts such as Open Geospatial Consortium and OPeNDAP (Open-source Project for a Network Data Access Protocol) have broken down format barriers to interoperability; the next challenge is the semantic aspects of the data.

Consider the issues when satellite data are merged, cross-calibrated, validated, inter-compared and fused. We must match up data sets that are related, yet different in significant ways: the phenomenon being measured, measurement technique, location in space-time or quality of the measurements. If subtle distinctions between similar measurements are not clear to the user, results can be meaningless or lead to an incorrect interpretation of the data. Most of these distinctions trace to how the data came to be: sensors, processing and quality assessment. For example, monthly averages of satellite-based aerosol measurements often show significant discrepancies, which might be due to differences in spatio-temporal aggregation, sampling issues, sensor biases, algorithm differences or calibration issues. Provenance information must be captured in a semantic framework that allows data inter-use tools to incorporate it and aid in the interpretation of comparison or merged products.

Semantic web technology allows us to encode our knowledge of measurement characteristics, phenomena measured, space-time representation, and data quality attributes in a well-structured, machine-readable ontology and rulesets. An analysis tool can use this knowledge to show users the provenance- related distinctions between two variables, advising on options for further data processing and analysis.

An additional problem for workflows distributed across heterogeneous systems is retrieval and transport of provenance. Provenance may be either embedded within the data payload, or transmitted from server to client in an out-of-band mechanism. The out of band mechanism is more flexible in the richness of provenance information that can be accommodated, but it relies on a persistent framework and can be difficult for legacy clients to use. We are prototyping the embedded model, incorporating provenance within metadata objects in the data payload. Thus, it always remains with the data. The downside is a limit to the size of provenance metadata that we can include, an issue that will eventually need resolution to encompass the richness of provenance information required for data intercomparison and merging.

Data Provenance: the source of data, including the execution history of the processes that produced them

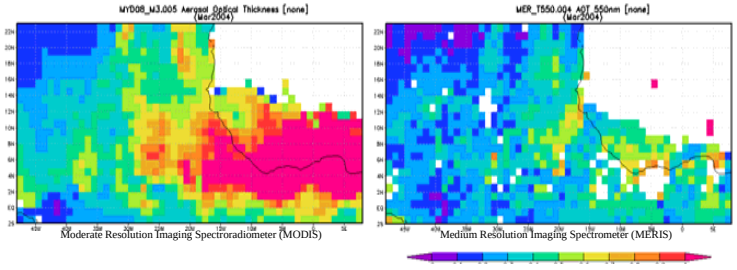
Multi-Sensor Data Matchup Use Cases

Cross-calibration
Validation
Correlation
Merging
Fusion

Provenance Semantics

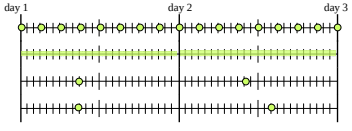
Key Matchup Characteristics in Multi-Sensor Intercomparison

- *Measured Parameter*
  - Physical Phenomena
  - Measurement Technique (e.g. ultraviolet vs. infrared ozone)
  - Processing Algorithm



MERIS underreports Aerosol Optical Thickness relative to MODIS for a high aerosol event in March 2004 west of Sahara/Sahel. This discrepancy likely arises because the MERIS aerosol product is an atmospheric correction product aimed at ocean color measurements; thus only pixels with the sky clear enough for ocean color retrievals are reported. As a result, MERIS has an effective threshold eliminating high Aerosol Optical Thickness values [Leptoukh 2007].

- *Space-Time Alignment*
  - Synoptic: Geostationary satellite, models
  - Time-averaged: Gridded satellite data
  - Sun-synchronous: Polar-orbiting satellites
  - Diurnally varying: Low-inclination orbits
  - Climatological



- *Data Quality*
  - Pixel-level: e.g. cloud-contaminated
  - Geographically varying: e.g. latitude-dependent, esp. models
  - Physiographically varying: e.g. aerosols over land and water

Matchup Reasoning Use Cases

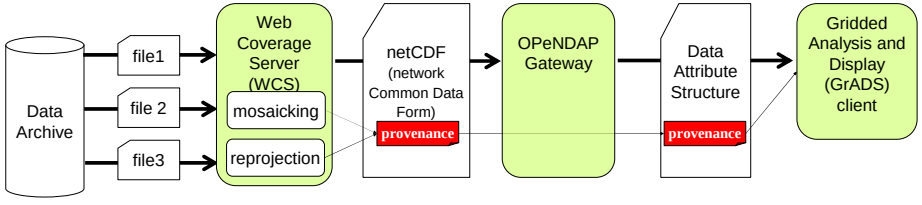
- *Parameter Matchup*
  - Given two parameters, can they be safely compared, based on their provenance? Merged? Fused??
- *Quality Screening and Weighting*
  - Apply congruent quality screening to compare two parameters
  - Apply quality-based weighting to merge or fuse two parameters

References: Leptoukh, G., S. Cox, J. Farley, A. Gopalan, J. Mao, and S. Berrick, S. (2007). Exploring NASA and ESA atmospheric data using GIOVANNI, the online visualization and analysis tool. *ENVISAT Symp.*, Montreux, Switzerland, 2007 Apr 23-27.

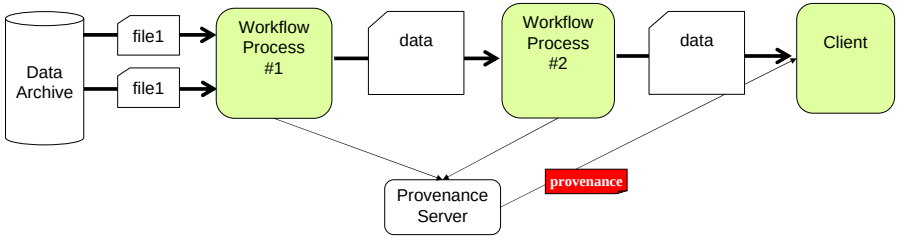
Provenance Mechanics

As data and services become more interoperable and remotely accessible, multi-sensor intercomparisons are more likely to result from multiple chained workflows. The mechanics of distributed provenance now become more important.

Embedded Approach: Provenance is embedded as a metadata attribute in the data payload
Example shown from the OPeNDAP/WCS Gateway Project



Out-of-Band Approach: Provenance is available via separate request



Embedded Approach: Works with legacy clients, facilitates persistence, BUT limited by data payload constraints
Out-of-Band Approach: Flexible and scalable, but difficult to use with legacy clients. Heterogeneous workflows (multiple different workflow systems) require provenance interoperability among the systems.

Conclusion: The task of automated or semi-automated intercomparison of two apparently comparable parameters exposes many challenges, one of them the proper consideration of the data provenance. Foremost amongst these is describing the provenance with enough semantic richness to assess and eventually assure the scientific validity of an intercomparison operation. Ongoing ontology efforts are beginning to tackle this but there is much to be done. For example, see the Earth Science Information Partners (ESIP) Data / Services Ontology at http://wiki.esipfed.org/index.php/Semantic_Web

Complicating this task is the dispersion of data and services to multiple sources, to be accessed via heterogeneous workflows. Persisting and transmitting the rich provenance needed for intercomparison will require provenance interoperability in addition to data interoperability.