CAPTURING, HARMONIZING AND DELIVERING DATA AND QUALITY PROVENANCE

Gregory Leptoukh, Christopher Lynnes

NASA Goddard Space Flight Center

1. INTRODUCTION

Satellite remote sensing data have proven to be vital for various scientific and applications needs. However, the usability of these data depends not only on the data values but also on the ability of data users to assess and understand the quality of these data for various applications and for comparison or inter-usage of data from different sensors and models. In this paper, we describe some aspects of capturing, harmonizing and delivering this information to users in the framework of distributed web-based data tools.

2. WHY PROVENANCE AND WHY NOW?

For the purposes of this paper, we define provenance as a structured history of the data from the very moment of a measurement through all the processing steps (with description of inputs/outputs and the actual algorithm steps) along with the necessary information about the instrument, platform, measurement conditions, and quality of the data. Knowing this history, data users can make educated assessments of the scientific results derived from the data. With advances in information technology, it has become much easier to find and get various data. The web-based access, visualization and analysis tools have significantly increased productivity of users of all levels of experience and have provided a much lower entry-barrier for inexperienced users. In addition, with more attention attracted to climate data, the results derived from remote-sensing and model data undergo increased scrutiny. Trust and transparency go together, and therefore provenance (which underpins transparency) and quality (which underpins trust) have become important considerations.

3. QUALITY ASPECT OF PROVENANCE

Even if detailed information about all the inputs, algorithms and processing steps is captured and delivered to a user, it is not sufficient for proper data usage. Important additional information about the corresponding data quality needs to be captured and delivered too. The complication with data quality is that it has several different facets. One of these facets, quality control, accompanies the data values as quality flags or quality confidence flags associated with each individual measurement. In some cases, additional flags are also provided for each

value, describing environmental conditions such as surface type, observing geometry, and/or the retrieval algorithm behavior (success, limited success, etc.) In most cases, these quality or environment flags represent the retrieval algorithm's "satisfaction" with its ability (or inability) to converge to a reasonable solution. This "satisfaction" may only indirectly indicate the actual quality of the data value, and rarely translates directly to uncertainty or bias. Moreover, it can be very difficult to use these flags to infer fitness of the data for a specific use. The latter, the "fitness-for-purpose" quality information is seldom provided and is left for users of the data to guess based on limited validation efforts documented in scientific papers.

4. CAPTURING PROVENANCE

We distinguish knowledge provenance from the processing provenance. The former contains information about the instrument, satellite, retrieval algorithm – basically, the information embedded in the data before processing begins; it typically not change from one data file to another. Some of this information can be encapsulated in the so-called metadata while the more detailed but vital information is dispersed in various Algorithm Theoretical Basis Documents and validation papers, and then systematized and assembled into the data-specific documentation. The processing provenance contains both general and file-specific information about all the processing steps occurred from the raw satellite data through Levels 0 to 3, plus whatever processing steps happen during post-processing by data access and analysis tools.

5. PROVENANCE HARMONIZATION

When comparing data from different sensors, it is important to know and understand the differences in their histories. However, in reality, both knowledge and processing provenance are captured and presented differently for different products. To harmonize provenance is to systematize provenance items or steps for each sensor and system in such a way that provenance of different data products can be compared. The challenge is to identify common classes. An ontological approach helps here to identify and capture relations between various aspects of data and processing. The ontology also helps to map similar but differently named items from different data products. The ontology, together with rulesets, allow provenance capture in a form that a computer can recognize and infer similarities and differences. In Giovanni [1, 2] the processing provenance harmonization is achieved by the very fact of having all the data being harmonized into a single HDF4 format upon entering the Giovanni system. The Giovanni workflow then moves the already harmonized data through the common processing steps.

6. DELIVERING PROVENANCE

Once provenance is captured and harmonized (if needed), it should be delivered together with data. There are at least two distinct methods: (1) create/append and propagate provenance together with the data through workflow

steps; (2) generate provenance on request using a separate workflow. The former imposes more requirements on each processing step to carry through and log additional information (thus impeding performance) but ensures consistency between the data and the corresponding quality/provenance. The latter relies on the ability of a data system to be queried for provenance that can be then aggregated and visualized – here consistency may suffer if some of the processing aspects cannot be reproduced.

7. VISUALIZING PROVENANCE

The streamlining of science analysis processes provided by web-based tools highlights an increased need for transparency. Presenting provenance in overly technical terms is not useful to the intended users of the system, namely, scientists. Therefore the challenge is to convert knowledge and processing provenance that has been captured and internally expressed in computer and information science terms (cryptic for most users) to a clear description and explanation of the science concepts and processing history of the data. By doing this, we can increase user trust, understanding, and reduce misinterpretation or generation of inconsistent results.

8. EXAMPLES

Web-based science analysis and processing tools allow users to access, analyze, and generate visualizations for vast amounts of data without requiring the user to directly manage the data or the data analysis processes or understand the limits on the underlying data.

8.1. Giovanni

NASA Giovanni [1, 2] is a tool that displays Earth science data from NASA satellites directly on the Internet, without the difficulties of traditional data acquisition and analysis methods. With a few clicks, data from various instruments on NASA satellites can be displayed in a variety of formats, including area plots, time series, meridional averages, zonal averages and vertical profiles, among others. Animations and numerical outputs are also available. Users can analyze phenomena ranging from the environment surrounding a Saharan dust storm to the impact Hurricane Katrina had on ocean surface chlorophyll concentrations. Single and multiple parameters can be plotted for specified ranges and time periods. Data is accessed and displayed via a collection of Giovanni portals, each one targeting specific projects or communities. For each generated plot, users have options to see the corresponding Giovanni processing provenance (called lineage in Giovanni) where all the steps leading to this particular steps are described along with input and output data pointers for each step.

8.2. Multi-sensor Data Synergy Advisor

The goals of the Multi-sensor Data Synergy Advisor (MDSA), the NASA-ESTO funded project [3], are to provide users of remotely sensed aerosol data with clear, cogent information on salient differences between data candidates for intercomparison, merging and fusion to enable scientifically and statistically valid conclusions. MDSA project focused on the value-added capabilities of NASA Giovanni online tool for data access, visualization and analysis to improve usage of NASA's remote-sensing data, and specifically, addressing aerosol data. Augmenting Giovanni with semantic web technologies and ontologies to support data inter-comparisons from different sensors or models, encoding dataset variable characteristics and related quality to derive inter-comparison rules, and adding data provenance (essential parameter details, quality and production caveats) can greatly enhance scientists' ability to perform valid comparisons, draw quantitative conclusions, and then mergeor fuse data from multiple sensors.

9. ACKNOWLEDGEMENTS

This work was supported in part by the NASA ESTO AIST and ACCESS programs.

10. REFERENCES

[1] J. Acker, G. Leptoukh, S. Shen, T. Zhu, and S. Kempler, "Remotely-sensed chlorophyll a observations of the northern Red Sea indicate seasonal variability and influence of coastal reefs," Journal of Marine Systems, vol. 69, 2008, pp. 191-204.

[2] S.W. Berrick, G. Leptoukh, J.D. Farley, and H. Rui, "Giovanni: A Web Service Workflow-Based Data Visualization and Analysis System," IEEE Transactions on Geoscience and Remote Sensing, vol. 47, Jan. 2009, pp. 106-113.

[3] S. Zednik, P. Fox, D. L. McGuinness, "System Transparency, Or How I Learned to Worry about Meaning and Love Provenance!," IPAW 2010, to appear in Springer conference proceedings.