Provenance Tracking in an Earth Science Data Processing System

Curt Tilmes[1] and Albert J. Fleig[2]

[1] NASA Goddard Space Flight Center, Greenbelt, MD 20771,USA,
Curt.Tilmes@nasa.gov
[2] PITA Analytic Sciences, 8705 Burning Tree Rd., Bethesda, MD 20817, USA,
Albert.J.Fleig@nasa.gov

Abstract. NASA and other organizations involved with climate research have captured huge archives of earth observations. The sensors, spacecraft, and science algorithms for transforming and analyzing the data and the processing frameworks are evolving over time. Science Data Processing Systems (SDPSes) should capture, archive, and distribute provenance information of all externally received data and algorithms, as well as describing all internal processes used for data transformation. This will make the data sets produced by the systems easier to understand, enable independent scientific reproducability, and ultimately, increase the credibility of the scientific research that makes use of those data sets.

1 Introduction

Earth science data have been captured from remote sensing satellites for several decades now, and numerous national data centers hold vast quantities of such data. In addition to the initial raw data received directly from sensors, the data include calibration processes and geolocation determination. The data are used with a variety of scientific retrieval algorithms to produce derived geophysical products, and they undergo transformations to reformat, regrid, subset, etc. the data to massage it into forms useful for scientists to perform research. Over time, the systems that perform this long series of data transformations from observation through product generation evolve. New technologies are developed, later generations of spacecraft, sensors, and data processing frameworks have different characteristics. The science algorithms for transforming and analyzing the data also improve over time with our growing understanding of earth science and the overall climate.

Tracking the provenance of earth science data throughout this process is a difficult problem. Research that makes use of multiple data sets from multiple data sources housed in multiple archives distributed among multiple organizations or agencies with different standards and policies simply exacerbates the problem. Science data is being used in new ways not planned by the originators of a given data set. We now find value added services (such as SOAR[1]) are building new archives that have transformed data from other sources, and

re-distributed the data in a new form. Some of these systems even provide the capability to automatically retrieve data from a data archive on-demand and perform dynamic alterations, distributing requested data to an end user without retaining a copy of the data. Maintaining complete provenance information through a processing chain that includes ephemeral data from such a "virtual archive" can be even more complicated.

This paper will discuss some of the general concerns of science data processing, and provenance in the context of two specific science data processing systems in operation at NASA's Goddard Space Flight Center: the MODIS Adaptive Data Processing System (MODAPS) [3] and the OMI Data Processing System (OMIDAPS) [2]. MODIS, the Moderate Resolution Imaging Spectroradiometer, is an instrument on the NASA Terra spacecraft launched in 1999, and on the Aqua spacecraft launched in 2002. OMI, the Ozone Monitoring Instrument, is a Dutch instrument launched on the NASA Aura spacecraft in 2004. These systems will provide examples and serve as case studies.

2 Science Data Processing

2.1 Data Archiving

There are two parts of every data file in the data processing system, the actual **data** ("bunch of bits") and the **metadata** with information that describes or relates to the data.

The data files are assigned a unique identifier and stored in an archive system where they can be retrieved by that identifier. We refer to the smallest "chunk" of individually identified data as a *granule* of data. A granule could refer to a year, a month or a day of data.

For both MODIS and OMI, the level 0, or raw, data are provided to the processing systems in 2 hour granules. MODIS data is quite voluminous, so the Level 1/2 data are stored in 5 minute granules. These are canonicalized on even 5 minute boundaries, e.g. 00:00:00 - 00:05:00, 00:05:00 - 00:10:00, etc. MODIS Level 3 data are organized somewhat differently for each of three climate categories, Land, Oceans and Atmospheres. The MODIS Land Discipline organizes its data with a integerized sinusoidal projection on a latitude longitude grid. [4] There are 326 land tiles, identified by their horizontal and vertical tile coordinates. The Level 3 gridded data are stored on various temporal resolutions as well, typically including daily, 8 day, 16 day and 32 days of data.

OMI's purpose is to monitor atmospheric constituents (Ozone of course, but also several others), which it retrieves from measures of backscattered solar radiation. It also has a lower resolution and lower data rate than MODIS. For these and historical reasons, the data are organized into contiguous data on an orbit by orbit basis.

Each different type of data is assigned an "Earth Science Data Type" (ESDT) that identifies a set of data files. For example, `OML1B` for OMI Level 1B, or `OMTO3` for OMI Total Ozone. The ESDT encodes multiple pieces of metadata, including

the instrument, the level, the spacecraft (in the case of MODIS which has two instances currently flying), and the type of data.

2.2 Primary and Secondary Metadata

Depending on the data level, and the metadata associated with a particular granule, a unique identifier is constructed from a minimal set of metadata. For example, there is one OMI Level 1B granule for each orbit of data. For orbit number 123, the particular granule could be described with the tuple {`OML1B`, `123`}. The Level 1B data from MODIS on Terra captured between 10:50 and 10:55 on Feb. 17, 2008 could be described with the tuple {`MODL1B`, `2008-02-17`, `1050`}. A MODIS level 3 land tile at tile coordinates (12, 17) of type MODVI (vegetation index) from data captured on Jan. 13, 2008 could be described with the tuple {`MODVI`, `2008-01-13`, `(12,17)`}.

This *primary metadata* is a minimal set of metadata that can be used to find a particular granule of interest by searching an indexed database, resulting in a pointer to the data file of interest.

Secondary metadata is a much larger set. It can include any other information useful to the user of the data. This can include a large variety of information:

- Geographic information that can be used to limit a spatial search,
- Quality information ("Data is bad for some reason," "Granule is cloud obscured"),
- Instrument configuration information ("Instrument in spectral zoom mode," "Spacecraft maneuver in progress"),
- Extra information about the data files themselves: file size, checksum for data integrity verification
- Provenance information (Where did I get this file? How did it come to exist?)

Secondary metadata can also include data annotations added after processing, or by another organization. For example, after the data are processed, the science data quality can be assessed by independent QA group and the granules annotated with that assessment.

2.3 Reprocessing

Both MODAPS and OMIDAPS are operational systems that currently receive data from active satellites and run the various science algorithms continuously on newly acquired data. Science keeps marching forward however, and new research and analysis of the data yield new versions of the algorithms. The change could resolve a bug that introduces an artifact into the data, or simply improve the quality of the data. It can be complicated to assess the effect of the change on the data. Sometimes the algorithm is run in parallel on a significant quantity of data that are then compared to the older version. If the new version is substituted into the operational system, a discontinuity in trends can occur, affecting research that might depend on such a trend. Sometimes it is better to keep producing a

dataset consistent with known problems than to produce an inconsistent data set. For example, consider monitoring a long term trend. If a particular measurement is sufficiently *precise*, even in the absence of perfect *accuracy*, the trend may still be useful. If in the middle of such a dataset the accuracy suddenly improves, introducing a jump in the trend, the dataset may be less useful for monitoring the long term trend. The approach that MODAPS and OMIDAPS typically take is to periodically go back to the beginning of the mission and reprocess all the data with the best known set of algorithms, thus producing an improved and consistent data set. We refer to these periodic large scale reprocessing campaigns as a *collection*. MODAPS is currently completing production of collection 5.

All the science algorithms are carefully configuration controlled and versioned throughout the processing system. The metadata for every product always includes the version number of the algorithm that produced it within the system.

3 Provenance

Provenance refers to the source of information and the historical process that led to its existence. Provenance information is critical to end users trying to understand where a particular data file came from. To this end, the system records all aspects of the data production flow. This includes:

- The source of all externally supplied data files. This could include a reference to the specific file in another archive responsible for the stewardship [5] of that data file.
- The source of the algorithms used to transform the data within the system. "Source" here refers to the origin of the algorithm, but also important in understanding an algorithm is its source code. Where possible and legal, we store the source code in a controlled configuration management (CM) repository that tracks changes across multiple versions of the same algorithm. When used properly, the CM system can also store comments, bug report numbers, references to other papers, and other information that can help a researcher understand the reasons behind changes. possible and legal
- Algorithm Design Documents. While the source code is the most up to date form of an algorithm, it is seldom the best way to understand the scientific functioning of the algorithm. Where possible, we also store or reference any design information which describe the mathematical basis and physical science behind the algorithms in the form of formulas, text, diagrams, tables, and graphs. These can also reference peer-reviewed science journal articles or other information about the algorithm. Our goal is to store or reference anything that can help someone understand the algorithm better.
- A complete description of the processing environment. This includes things as basic as what particular computer ran the program and what hardware resources it had. It could easily be the case in the future that the exact same hardware might be found only in a museum, but listing the particular

hardware could be useful to someone trying to analyze the data. More important than the hardware is the software in the environment. This includes the operating system and software library versions.
- A complete description of the processing framework. Just as we CM the science algorithms, every module that is part of the processing system is stored in a CM repository.
- A record of each job's execution. This is a list of all of the outputs of the production rule execution process, including runtime parameters (things like "Orbit Number," "Data Date," "Debug Mode," "Algorithm Control Flags") and a list of all input files. We also store extra information about the execution including the clock time it started and finished, CPU and disk resource utilization, etc. These can help in the analysis of data processing performance and optimization.

It is expected that other archives and suppliers of all artifacts (data, algorithms, documents, etc.) will capture, archive, and distribute their own provenance information in a well-defined manner. Ideally, this should provide a complete distributed provenance graph even back to information describing the spacecraft and instrument that captured the original observations. This provenance helps to put scientific results derived from the data into context and allows future researchers to understand the entire data flow. Currently, questions like "Was the ozone input to the weather model derived from back-scattered ultraviolet or microwave radiation measurements?" require a human to read natural language data descriptions, visit various web sites, and/or call up scientists personally to manually determine the provenance of the dataset.

3.1 Scientific Reproducibility

While provenance information is nice to have for a researcher trying to understand a data set and algorithm, especially for climate research using remote sensing data, it is critical for scientific reproducibility. Many systems recognize this ideal and strive to store sufficient information that a dedicated (sometimes very dedicated) researcher who is able to expend sufficient effort could theoretically construct a system capable of reproducing the data. Other systems can reproduce the latest version of the data, but do not support obtaining older data.

Our goal is to make it not just possible, but *easy* to reproduce any data file that gets distributed from our system. To that end, we archive all versions of fully integrated algorithms. As a next step, we plan to distribute a processing framework that can access the integrated algorithms directly and interact with our system to download the information needed to replicate the environment and re-run the algorithms. Additionally, since the integrated algorithms are available and encapsulated provenance information is available, so that remote users can use their local execution framework to reproduce any of our files within their own systems. This provides complete scientific reproducibility and allows an independent verification and validation of all data provided by our main system. Providing this capability can increase the *credibility* of the science results that use the data.

3.2 Process on Demand and Virtual Archives

As algorithms improve and are inevitably changed over time, older data sets become obsolete and the expense of storing all data physically on disk outweighs their historical value. Typical archives keep previous versions of data around long enough to analyze its differences with current data, then remove it in favor of the new data. Those archives also store the metadata (including provenance information) colocated with the data, and deletion of the data often causes deletion of metadata and provenance information as well.

Our system also removes old data files, but, as described above, we retain sufficient provenance information to reproduce deleted data sets if needed. This functions as an extreme form of compression where the provenance information suffices to re-create a file. The provenance is a proxy for the physical contents of the file.

The next logical step, already implemented on MODAPS, is *Process on Demand.* Some MODIS products are very large, and less widely used (Level 1B), while others are much smaller and more widely used (Level 2 and above). The Level 1B products are created in normal processing and used as inputs to Level 2. After keeping them around for 30-60 days for the most interested users to retrieve, they are removed from the archive. Since the system retains the ability to re-create them as needed, users can order the older files from the archive whereupon the files are scheduled for reprocessing. Depending on the level of requests, the system can use a small amount of processing capability as a stand-in for a very large amount of disk.

Since archive space has historically been a very limiting factor, science teams make very considered, deliberate and often limiting choices when deciding which official data products to produce and archive. *Process on Demand* allows a "virtual" archive of many more products thereby relaxing some of the self-imposed limitations. This approach has led to the development of "services" that can transform data dynamically to very specific forms requested by users [1]. It is important that such services don't overlook the intensive verification and validation functions performed by the science teams, and that complete provenance information is captured, even for dynamically created, ephemeral data products served from a virtual archive.

3.3 Provenance Problems

As noted above, systems often store provenance information in the metadata along with the data files, and when the data are removed, so is the metadata. Someone later researching a science paper with results citing a specific data set may find that not only are the data no longer available, but also there is no information about how that data set was produced.

When data files used in production come from external providers, our provenance information can refer to that source, but it must also refer to the specific file so that it can be retrieved from that provider. If upstream providers don't

archive or distribute sufficient provenance information for significant inputs, they can become a "dead end" in the graph.

The example above described a (very simplified) scenario where science leads to an algorithm, which is coded into software, which is used to process data. Sadly, this ideal seldom matches reality. We often find software evolving in new directions that simply aren't retroactively captured in design documents and published papers. Keeping the entire provenance chain up to date requires dedication and discipline.

Sometimes provenance information is captured, but the information is restricted. Hardware and software designs provide a competitive advantage, so some organizations are reluctant to release proprietary information in the processing chain. In particular, due to past problems with distribution of satellite and rocket technology the U.S. International Traffic in Arms Regulations (ITAR) is particularly restrictive of certain types of information. Even where the information isn't particularly sensitive, the default ITAR position is to restrict data, and sometimes it is simply easier to avoid the procedural burden to get permission to release information.

Most systems attempt to capture provenance information, but we have found that it is often incomplete, and represented in non-standard forms that can be difficult to follow. Often it is reduced to a phone call to the scientist asking "Where did you get this data, and what did you do to it?" Based on personal discussions, we have found that capturing and distributing good, usable provenance often simply isn't a priority for scientists. They are more than willing to talk about provenance and explain their methodologies with colleagues, but sometimes don't see the usefulness of incorporating provenance into the production system.

Even if provenance is captured, archived, and distributed, some systems can't (or won't) reproduce older datasets. They can rely on an error prone, manual process to attempt to reproduce data previously released.

4 Conclusion and Future Work

Access to complete provenance information is essential for many aspects of the use of Earth science data. It is possible to build science data processing systems that automatically capture the provenance information with little impact on resources, operations or the participating scientists involved in creating the data. This will make it possible for users to know how a data set was made, reproduce the results of the initial processing, and understand differences over time periods even after the original producers are no longer available for consultation.

With complete provenance a user will know what input data was used for any product including details of where it came from and what version it was. The user will be able to know what exact algorithms were used to make a product, what exact input data was used, what exact system the data was produced with and what processing system it was made on. This will improve the credibility of the data set and make it possible to determine whether differences over time of

a remotely sensed data set come from true geophysical changes or are artifacts of the production system.

We are working on development of methods to distribute the processing framework used to make a product in such a way that remote scientists can access the algorithms that were used, interact with our system to download the information needed to replicate the environment, and run time parameters and reproduce the results or modify any component and asses the impact of the change.

Complete provenance requires that input data obtained from external sources also comes with its own provenance. We are working on identifying tools, content, and standards for this and on encouraging other data sources to provide this information. We note with concern that there is not a current commitment in the science community to require adequate stewardship to maintain and support complete provenance even if it is available. We also note that the requirement for the scientists providing processing algorithms to also provide complete documentation of the final version of their process does not receive adequate support.

It is our hope that by showing that all of the needed information can be easily captured if available and that it can assure data reproducibility we can encourage further development in these areas.

Acknowledgment

The authors thank the many individuals comprising MODIS and OMI teams for making development of MODAPS and OMIDAPS successful. Some of the information presented here was taken from a number of documents and web sites throughout the two projects.

References

1. Halem, M., Yesha, Ye., Tilmes, C., Goldberg, M., Shen, S., Zhou, L.: Service Oriented Atmospheric Radiances (SOAR): A Community Research Tool for the Synthesis of Multi-Sensor Satellite Radiance Data for Weather and Climate Studies. Proc. 3rd Intl. Conf. on Web Information Systems and Technology. (2007)
2. Tilmes, C., Linda, M., Fleig, A.: Development of two Science Investigator-led Processing Systems (SIPS) for NASA's Earth Observation System (EOS). Proc. IEEE Geoscience and Remote Sensing Symposium. (2004) 2190–2195
3. Masuoka, E., Tilmes, C., Ye, G., Devine, N.: Producing Global Science Products for the Moderate Resolution Imaging Spectroradiometer (MODIS) in the EOSDIS and MODAPS. Proc. IEEE Geoscience and Remote Sensing Society (2000)
4. Wolfe, R., Roy, D., Vermote, E.: The MODIS land data storage, gridding and compositing methodology: LEVEL 2 Grid. IEEE Trans. on Geoscience and Remote Sensing **36** (1998) 1324–1338
5. Diamond, H., Bates, J., Clark, D., Mairs, R. Archive management: the missing component. Proc. 20th IEEE/11th NASA Goddard Conf. on Mass Storage Systems and Tech. (2003) 40–48
6. W3C Semantic Web Activity. `http://www.w3.org/2001/sw/`.