Taming Big Data Variety in the Earth Observing System Data and Information System

Christopher Lynnes, *Member, IEEE*, and Jeff Walter

***Abstract*—Although the volume of the remote sensing data managed by the Earth Observing System Data and Information System is formidable, an oft-overlooked challenge is the variety of data. The diversity in satellite instruments, science disciplines and user communities drives cost as much or more as the data volume. Several strategies are used to tame this variety: data allocation to distinct centers of expertise; a common metadata repository for discovery, data format standards and conventions; and services that further abstract the variations in data.**

***Index Terms*—data storage systems, data systems, information architecture, remote sensing, search problems.**

I. Introduction

THE National Aeronautics and Space Administration (NASA) has been launching Earth observation satellites for several decades. Since the 1990's, the data from these satellites have been archived in the Earth Observing System Data and Information System (EOSDIS). EOSDIS is a distributed system, anchored by 12 data centers across the United States and mediated by a common inventory database and a portfolio of common services.

Almost from the beginning, the volumes of science data, the first and best-known 'V' of Big Data, have presented a not-insignificant challenge to EOSDIS system architects and implementers. However, often overlooked has been the driving role of the second 'V', Variety in system architecture and operations.

The variety in NASA's Earth science data archives is rooted in the need to study the Earth as a system [1]. The Earth system components that are the subjects of remote sensing studies include the lithosphere, cryosphere, hydrosphere, atmosphere and biosphere. The union of these spheres represents a substantial diversity in the physical measurements needed to study them. Furthermore, the study of Earth as a system often requires two or more measurements to be studied together, increasing the impact of variety on the discovery and use of Earth observation data.

Due to the wide scope of Earth system science, NASA flies a wide variety of instruments on its satellites, as well as brokering data from other national space programs, such as the European Space Agency (ESA) and the Japan Aerospace Exploration Agency (JAXA). Fig. 1 shows a schematic view of the instruments orbiting in the A-Train constellation, a group of satellites that flies in formation in order to provide measurements that are roughly coincident in time and space. A-Train includes satellites operated by NASA, ESA and JAXA; EOSDIS distributes at least some products from all of the A-Train satellites.

These satellites host several different kinds of instruments, with varying horizontal and vertical resolution. Typically, horizontal coverage, vertical resolution and horizontal resolution trade off amongst each other. CALIPSO (Cloud-Aerosol Lidar and Infrared Pathfinder Satellite Observations), for example, hosts a laser-based instrument to make high-vertical-resolution measurements of atmospheric aerosols and clouds for very small areas (90 m footprint). Conversely, Aqua hosts the Moderate Resolution Imaging Spectroradiometer (MODIS), which takes pictures in the visible and infrared spectrum over very wide (>2000 km) tracks, with a horizontal resolution of 0.25 to 1 km but limited vertical resolution of the atmosphere. The Aura satellite hosts two limb scanners, the Microwave Limb Scanner and High Resolution Dynamics Limb Scanner, which look at the atmosphere at an oblique angle to produce good vertical resolution, but low (and difficult to characterize) horizontal resolution.

Satellite data are typically transmitted to the ground in raw form, and then processed to higher levels (Table 2). While the variety of data at lower processing levels (0 and 1) is driven largely by the instruments and satellites involved, the data products at level 2 often proliferate as the science team develops numerical algorithms to retrieve a multitude of geophysical measurements. Often, this is limited only by how many measurements the instrument science team thinks can be safely retrieved from the lower level data. In some cases, alternate algorithms may also be used to retrieve the same measurement. For example, ozone concentration from the Ozone Monitoring Instrument is derived using both a multi-wavelength Differential Optical Absorption Spectroscopy method [2] and a "TOMS-like" algorithm to maintain continuity with the predecessor Total Ozone Mapping Spectrometer instrument [3]. Similarly, aerosol optical depth from the MODIS instrument is computed using both a "dark target" algorithm for oceans and low-reflectance land surfaces [4] and a "deep blue" algorithm [5] for high-reflectance surfaces such as deserts. The measurements are often accompanied by ancillary variables describing data quality,

Christopher Lynnes and Jeff Walter are with NASA's Goddard Space Flight Center, Greenbelt, MD 20771 USA. (e-mail: christopher.s.lynnes@nasa.gov)

first-guess values, and the like. A further proliferation often takes place in Level 3 processing, which aggregates the data in space and time to a (usually regular) grid. In this case, science teams will often output a daily and a monthly set of measurements, representing different time aggregations. Some also include 5-day or 8-day aggregation for Earth observations where daily data tend to have sparse coverage, but sub-monthly variations are of interest to the science community. Finally, assimilation of the data into models, such as the Modern Era Retrospective-Analysis for Research and Applications [6], results in regularly gridded, gap-filled measurements, often corresponding to the measurements available at Level 2 or 3. In addition, first time derivative ("tendency") versions of many of the measurements are included in the model output.

II. Variety Challenges and Solutions

In many ways, the variety of data available from NASA Earth observation datasets is a boon to the science and applications communities: the chances of a suitable measurement existing for a given science application are greater. However, the variety in the types of data available also presents a challenge to the organizations responsible for providing support related to the data. The EOSDIS system has addressed this from the beginning by assigning data management duties to a network of data centers, based on science discipline (Table I). These data centers, called Distributed Active Archive Centers (DAACs) employ science experts in their assigned disciplines, provide discipline-specific documentation, and even develop tools of particular interest to the science users in that discipline, such as the Charctic Interactive Sea Ice Graph at the National Snow and Ice Data Center[1].

Of course, the variety of available data is a boon to science users only insofar as they can identify data that meets their needs. In order to mitigate the distributed nature of EOSDIS from a data discovery standpoint, a Common Metadata Repository (CMR) has been developed, which contains both dataset-level (directory) and file-level (inventory) information about the holdings at the EOSDIS data centers. The CMR is populated by metadata published by the DAACs as they produce new data or receive new data from their providers. The CMR is served by a search engine that can be accessed via an application program interface (API) or via the OpenSearch convention with Geo and Time extensions [7]. In addition, a search user interface (Earthdata Search) for the CMR is available, so that users need not know which DAAC has a particular dataset in order to discover it. As of 22 August 2015, EOSDIS presents 6055 datasets to users via Earthdata Search. In order to aid navigation of this large number of datasets, the search interface includes both a keyword search and interactive facet filtering capability.

In addition to the large number of data collections, the individual data files in most collections contain many different measurements, manifesting in an even larger number of individual data variables. For instance, gridded atmospheres products from both the MODIS and Atmospheric Infrared Sounder (AIRS) instruments contain in excess of 800 variables apiece, which span several disciplines. The AIRS data include both atmospheric chemistry (methane, carbon monoxide, ozone) and atmospheric dynamics (temperature, moisture, clouds) variables. Most end users are likely to be interested in a small subset of variables in such files. Transferring the entire file, only to have the user discard most of it is clearly an unnecessary use of network and input/output resources.

The solution to this dilemma can be found (mostly) in the Open-source Project for a Network Data Access Protocol (OPeNDAP) [8]. OPeNDAP is enabled through deployment of an HTTP-based OPeNDAP server where the data are stored and configuring it to serve the desired data over the network. In turn, an OPeNDAP client can extract the variables of interest, or even subsets ("hyperslabs") of those variables. The OPeNDAP client software is built into the popular libraries (both C and Java) for network common data form (NetCDF). The result is that any tool built with the NetCDF library (and "DAP-enabled") can read data subsets across the network as if they were read from local NetCDF files. These tools range from full-featured analysis and visualization tools like Panoply[2], the Integrated Data Viewer[3], and ArcIMS, to plug in modules for IDL and Matlab to the basic command-line utilities provided with the NetCDF library (e.g., *ncdump*). In addition, some OPeNDAP servers, particularly the ones using the Hyrax implementation[4], support a NetCDF response that allows a user to download a NetCDF file subset using web browsers or analogous command line utilities like wget or curl.

The serving of data through OPeNDAP also mitigates another aspect of Variety, the heterogeneity of data formats. EOSDIS has long promoted use of Hierarchical Data Format as a standard within the system. Nonetheless, a significant number of data collections are stored in other formats, including NetCDF, ASCII and unstructured binary formats. However, serving such data through OPeNDAP provides a standard, uniform interface to data for all of these formats.

Until recently, OPeNDAP was deployed for a minority of eligible datasets within EOSDIS. However, a recent U.S. federal initiative named the Big Earth Data Initiative (BEDI) [9] has enabled the proliferation of OPeNDAP-served datasets within EOSDIS. Note that not only does this help address the Variety aspect of multi-variable data files, but the spread of subsetting capabilities also helps the Volume challenge faced by many end users, who typically have more limited computer and data management resources available than institutional data centers.

In addition to the subsetting capabilities, the Hyrax implementation of OPeNDAP includes a plug-in data handler based on NetCDF Markup Language (NcML) with three capabilities that help to further smooth over the data variety

[1] http://nsidc.org/arcticseaicenews/charctic-interactive-sea-ice-graph/

[2] Available at http://www.giss.nasa.gov/tools/panoply/

[3] Available at http://www.unidata.ucar.edu/software/idv/

[4] See http://docs.opendap.org/index.php/Hyrax

problem (Fig. 2). The NetCDF-based tools that support map visualization (such IDV, Panoply, and ArcIMS) often depend on a standard representation of coordinate variables in the data files that follows the Climate-Forecast (CF) convention. The NcML handler allows the data center deploying the server to amend or even substitute for the internal file metadata to satisfy the CF convention, thus smoothing over some of the irregular variations among data products and making them more usable in the analysis tools.

Another key capability of the NcML handler is the ability to present just a subset of the data file's variables to the end user or client. Thus, the data center serving AIRS data might deploy one NcML handler configuration that presents only the methane-related variables, another that exposes only the carbon-monoxide variables, another that shows only temperature, and so forth. This capability has the salutary effect of allowing us to quickly create new data products tailored to specific communities without duplicating the actual data. This also allows us to form "single-parameter" data products, which are easier to align with layer-oriented data tools such as visualizers or map servers such as the EOSDIS Global Imagery Browse Services, which are also being populated as part of the BEDI effort.

On the other hand, DAACs are typically oriented toward managing actual data files, whereas these NcML based single-parameter subsets exist largely in a virtual, on-demand sense. Rather than force DAACs to instantiate a full set of file-level metadata for the single-parameter versions of data served through OPeNDAP, the Common Metadata Repository has been enhanced to support "virtual data products." These are derivative products whose metadata are cloned from their parent "real" data products and modified slightly to describe the virtual product. Then, as file-level URLs for the parent data product are published to the CMR, the derivative URLs are computed automatically, using configurable generation rules, and inserted into the file-level CMR database. This capability has proven to have additional applications, particularly with respect to on-demand standard products.

A third benefit of serving data through OPeNDAP is the ability to represent a data collection of individual files as a single virtual "granule" of data through aggregation. For both the Hyrax and THREDDS Data Server varieties of OPeNDAP, this is done via an NcML file, while "data descriptor files" serve the same purpose with the GrADS Data Server. The BEDI effort is improving the performance of the Hyrax implementation (the most common deployment type in EOSDIS) with respect to aggregation performance. Aggregation reduces the complexity of dealing with variations in temporal resolution data collections, as a user can simply request a time range of data from a single virtual "granule".

III. Conclusion

The increase in data available through OPeNDAP that is enabled by BEDI is mitigating part of the Variety problem of EOSDIS data collections. However, in order to realize the full benefit, the user community needs simple instructions on how to access, and particularly to subset and reformat, data through OPeNDAP. Since many prospective users are scientists or applications users, i.e.., not software engineers, a how-to guide is envisioned that offers several options for accessing subsets via OPeNDAP. These will include command-line versions of access recipes in addition to popular scripting languages such as Python, which has modules that support NetCDF (and thus OPeNDAP) access. The NetCDF Common Operators (NCO) is especially useful in this context, as the package supports specification of spatial subsets using latitude and longitude. The package converts these specifications into the grid indices required in a typical OPeNDAP constraint expression, for datasets that support CF coordinates.

One of the most nettlesome of the remaining Variety problems arises from the plethora of data collections in EOSDIS. For certain measurements, such as Ozone and Aerosol Optical Depth, the number of available data collections is several hundred, so that the winnowing process is still problematic, even with keyword-based filtering and faceted browse. In order to tackle this, we are working on improving relevancy ranking. Rather than treating the data collection metadata records primarily as documents for text-based search (which produces a number of false positives), the Common Metadata Repository search engine will be enhanced to employ a relevancy ranking algorithm that is more in line with how science users make decisions about data collections. This will include a heightened emphasis on terms that describe the measurements themselves, heuristics on which data sets are more likely to be usable to the widest community, and metrics on how many users download data for a given data collection. In addition, temporal relevance and spatial relevance will be computed, with preference given to data collections that cover the whole area or time period specified by the user.

References

[1] G. Asrar, J. A. Kaye, and P. Morel, "NASA research strategy for Earth system science: Climate Component," *Bull. Amer. Meteor. Soc.*, vol. 82, no. 7, pp. 1309-1329, 2001.

[2] J. P. Veefkind, J. F. de Haan, E. J. Brinksma, M. Kroon, and P. F. Levelt, "Total ozone from the Ozone Monitoring Instrument (OMI) using the DOAS technique," *IEEE Trans. Geosci. Remote Sens.*, vol. 44, no. 5, May 2006, doi: 10.1109/TGRS.2006.871204.

[3] P. K. Bhartia and C. W. Wellemeyer, "TOMS-V8 total ozone algorithm," in *OMI Algorithm Theoretical Basis Document, OMI Ozone Products,* P. K. Bhartia, Ed. Greenbelt, MD: NASA/Goddard Space Flight Center, 2002, vol. 2.

[4] R. C. Levy, S. Mattoo, L. A. Munchak, L. A. Remer, A. M. Sayer, and N. C. Hsu, "The Collection 6 MODIS aerosol products over land and ocean," *Atmos. Meas. Tech. Discuss*., vol. 6, pp. 159-259, 2013.

[5] N. C. Hsu, S. C. Tsay, M. D. King, and J. R. and Herman, J. R.: Deep blue retrievals of Asian aerosol properties during ACE-Asia, *IEEE T. Geosci. Remote*., vol. 44, pp. 3180–3195, 2006.

[6] M. M. Rienecker, M. J. Suarez, R. Gelaro, R. Todling, J. Bacmeister, E. Liu, M. G. Bosilovich, S. D. Schubert, L. Takacs, G.-K. Kim, S. Bloom, J. Chen, D. Collins, A. Conaty, A. da Silva, W. Gu, J. Joiner, R. D. Koster, R. Lucchesi, A. Molod, T. Owens, S. Pawson, P. Pegion, C. R. Redder, R. Reichle, F. R. Robertson, A. G. Ruddick, M. Sienkiewicz, and J. Woollen. "MERRA: NASA's Modern-Era Retrospective Analysis for Research and Applications," *J. Climate*, vol. 24, 3624–3648. 2011, doi: http://dx.doi.org/10.1175/JCLI-D-11-00015.1

[7] *OGC OpenSearch Geo and Time Extensions*, Open Geospatial Consortium standard 10-032r8, http://www.opengis.net/doc/IS/opensearchgeo/1.0, 2014.

[8] P. Cornillon, J. Gallagher, and T. Sgouros, "OPeNDAP: accessing data in a distributed heterogeneous environment," Data Science Journal, vol. 2, pp. 164-174, 5 November 2003.

[9] J. Holdren, National Plan for Civil Earth Observations, 62 pp., Natl. Sci. and Technol. Counc., Washington, D. C., 2014. Available: http://www.whitehouse.gov/sites/default/files/microsites/ostp/NSTC/national_plan_for_civil_earth _observations_-_july_2014.pdf .]

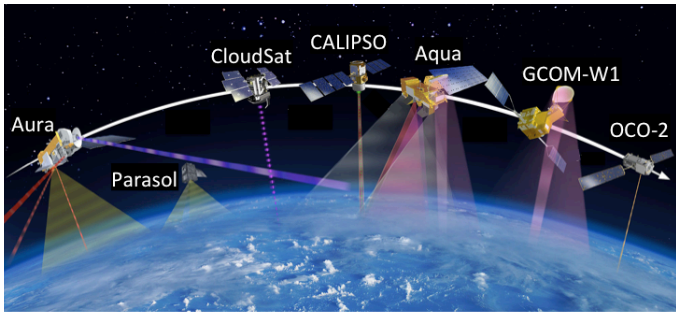

Fig. 1. Satellites in the A-Train constellation. These satellites host a number of instruments with very different scanning geometries (indicated schematically by the colored cones emanating from the satellites.) This diagram (not to scale) is adapted from the original at http://atrain.nasa.gov/images.php.

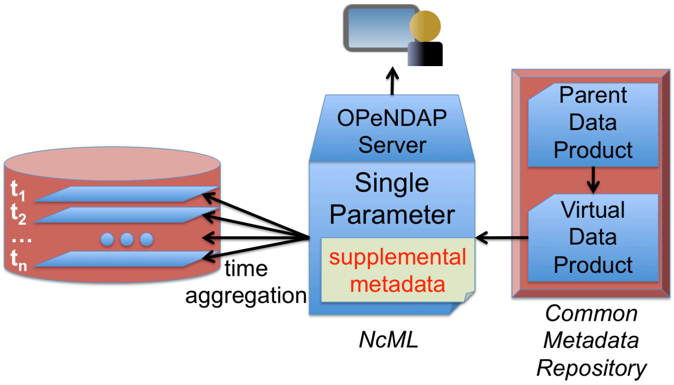

Fig. 2. NcML (NetCDF Markup Language) helps mitigate the Variety problem in Earth observation data by allowing a provider to present a single parameter, by allowing supplemental metadata to non-CF-compliant data, and by presenting a time-aggregated view of the dataset. EOSDIS's Common Metadata Repository further abstracts the single-parameter view by supporting virtual data products derived from actual parent data products.

TABLE 1
ALLOCATION OF SCIENCE DISCIPLINES TO DAACS

DAAC	Disciplines / Subject Areas
Alaska Satellite Facility DAAC	Synthetic Aperture Radar, sea ice, polar processes, geophysics
Atmospheric Science Data Center	Aerosols, clouds, radiation budget, tropospheric chemistry
Crustal Dynamics Data Information System	Solid earth, space geodesy
Global Hydrology Resource Center	Hydrologic cycle, sever weather interactions, lightning, atmospheric convection
Goddard Earth Sciences Data and Information Services Center	Atmospheric composition and dynamics, global modeling, global precipitation, solar irradiance, water and energy cycle
Land Processes DAAC	Ecosystem variables, land cover, radiation budget, surface reflectance / radiance, surface temperature, topography, vegetation indices
Level 1 and Atmosphere Archive and Distribution System	Atmosphere, MODIS radiance
National Snow and Ice Data Center DAAC	Cryosphere, frozen ground, glaciers, ice sheets, sea ice, snow, soil moisture
Oak Ridge National Laboratory DAAC	Biogeochemical dynamics, ecological data, environmental processes
Ocean Biology DAAC	Ocean biology, sea surface temperature
Physical Oceanography DAAC	Gravity, ocean currents and circulation, ocean surface topography, ocean winds, sea surface salinity, sea surface temperature
Socioeconomic data and applications data center	Environmental stability, geospatial data, human interactions, land use

TABLE 2
SATELLITE DATA PROCESSING LEVELS

Processing Level	Definition
0	Reconstructed, unprocessed instrument/payload data at full resolution (with communications artifacts removed)
1A	Level 0 data that have been time-referenced and annotated with ancillary information (calibration and georeferencing parameters) computed and appended but not applied
1B	Level 1A data processed to sensor units, usually georeferenced and with calibration applied
2	Derived geophysical variables at the same resolution and locations s the Level 1 source data
3	Variables mapped on uniform space-time grids, involving temporal and/or spatial aggregation
4	Model output or results from analyses of lower level data (e.g., variables derived from multiple measurements)