

GHRSSST-14 DAS-TAG REPORT

Edward Armstrong ⁽¹⁾, Jean Francois Piolle ⁽²⁾

(1) Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA
Email: edward.m.armstrong@jpl.nasa.gov

(2) IFREMER, France, Email: Jean.Francois.Piolle@ifremer.fr

ABSTRACT

The DAS-TAG provides the informatics and data management expertise in emerging information technologies for the GHRSSST community. It provides expertise in data and metadata formats and standards, fosters improvements for GHRSSST data curation, experiments with new data processing paradigms, and evaluates services and tools for data usage. It provides a forum for producer and distributor data management issues and coordination.

1. Introduction

This year the DAS-TAG session had a number of presentations concerned with metadata standard reviews, new data processing capabilities and web services that allow users to apply large processing power and chained services directly to the data, GHRSSST data coordination activities and proposals to improve data curation through data lifecycle policy implementation.

2. NASA Metadata Trends

Ted Habermann from the HDF Group presented an overview metadata “dialects” including the overlap of ISO 19115 metadata with the NASA ECHO and DIF standards. XML based processing methods have been developed such that 99% of both ECHO and GCMD metadata attributes can be mapped to their ISO counterparts. Metadata description of granules, data quality and lineage, and services are very well described in the ISO 19xxx standards and the community should move to unifying within that standard.

More Overlap Than Difference

The metadata dialects currently used by ESDIS have much more overlap than difference.

The mappings are generally well understood: we are in a tweaking stage.

The translations can be implemented using well-known, standard tools that are designed for XML processing.

These are different from the programming languages generally used for scientific data processing.

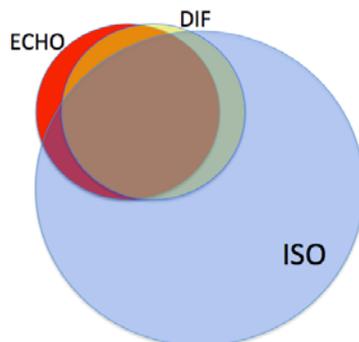


Figure 1. The overlap of metadata dialects used within NASA.

3. Web Services For Earth Science Data

Ed Armstrong from the NASA Jet Propulsion Laboratory presented a overview of emerging based web services for earth science data that will be available in the very near future from the Physical Oceanography DAAC. The RESTful nature of these services allow access from any client the can formulate a URL such as web browser or programming script. These web services allow the following capabilities and can be “chained” in sequence to provide seamless input/output from one service to another:

- Search Dataset/Granule Web Service
- Metadata for Dataset/Granule Web Service
- Extract and Subset Granule Web Service
- Image Granule Web Service

A conceptual use case was presented for ASCAT L2 ocean vector wind data that started with dataset and ISO 19115 metadata discovery, a granule search on a specific time domain, data extraction and finally visualization (Fig. 2)

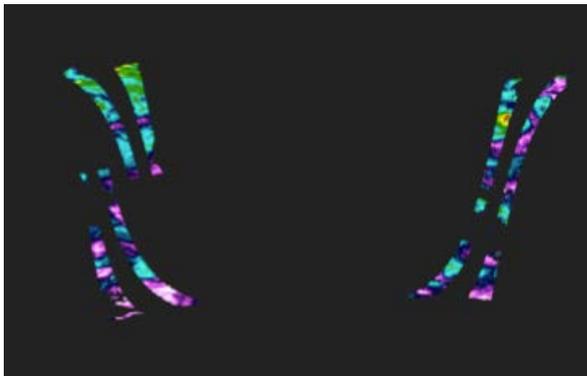


Figure 2. Visualization output after extracting an ASCAT ocean vector wind granule through web services “chaining”

4. HADOOP Usage in Medspiration

Jean Francois Piolle from Ifremer reported on the implementation of a processing framework based on Hadoop for Medspiration data. Hadoop is an open source processing paradigm based on a Map Reduce model that breaks tasks and data into smaller more modular components that can be rapidly executed independently in a distributed compute fashion. The Medspiration system is known as Nephelae and consists of 600 processing cores and 2.5 TB of memory. The computing application in this case was satellite data processing for derived products such as climatologies, anomalies and data statistics. For example, creating climatology and anomalies based on four years of regional L4 ODYSSEA SST data took 90 seconds. Processing 10 years of QuikSCAT data to retrieve daily wind speed min/max/mean took 2 minutes.

The concept here is that putting data directly in proximity to powerful computing can be leveraged by users to quickly generate results and explore new ideas.

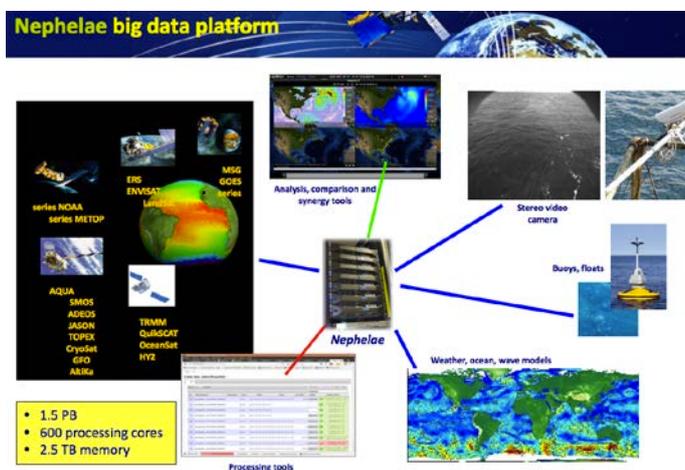


Figure 3. Concept of the Nephelae Hadoop based processing system

5. GDAC to LTSRF Data Flow

Ken Casey from the NOAA National Oceanographic Data Center (NODC) presented results from a short study on reconciling differences and inconsistencies between the GDAC and LTSRF. Some of these inconsistencies are due to missing or incomplete FGDC and DIF metadata. The LTSRF has refined its workflow to accommodate some of these issues and is now working through a backlog. In the future the transfer of GDS2 records should be easier since they do not contain external metadata records (e.g., the FR metadata records). It was also noted that the PO.DAAC metadata web service can be used to regenerate complete metadata records.

6. Dataset Lifecycle

Ed Armstrong reported on the Dataset Lifecycle Policy that the PO.DAAC has implemented for all of its new datasets including GHRSSST. This Lifecycle is designed improve data stewardship and insure that datasets that enter into the PO.DAAC distribution and archiving system met standards with regard to data formats, metadata, and even data quality and maturity. Impacts on operations, tools and distribution are assessed through the collection of various metrics including through written documentation. Of primary concern to GHRSSST data providers is a template for a "Memorandum of Understand" that includes sections for the provider to document the data uncertainty assessment and validation, and the processing lineage and algorithm history. This document is meant to be a first step to assess the dataset quality and will be eventually leveraged to improve GHRSSST ISO metadata records as well. After some discussion it was agreed that "Submission Agreement" would be a more suitable name for this template. The lifecycle concepts were agreed to be ready to be presented to a GHRSSST science team plenary session later in the week (see additional report for Thurs presentation).

7. Additional Discussion

The DAS-TAG considered the status of GDS2 production and governance of the GDS2 documentation. Some data producers have already produced GDS2 spec granules and these will be released publically in the near future. A spreadsheet was circulated for producers to enter their best estimates of the start dates of GDS2 datasets. The GDS2 is essentially frozen until the Project Office comes up with a plan for updates. An important future modification to the GDS2 is an extension for climate data records.

Copyright 2013. All rights reserved.