

## REPORT ON THE GLOBAL DATA ASSEMBLY CENTER (GDAC) TO THE 12<sup>TH</sup> GHRSSST SCIENCE TEAM MEETING

Edward M. Armstrong<sup>(1)</sup>, Andrew Bingham, Jorge Vazquez, Charles Thompson, Thomas Huang, Chris Finch

*(1) Jet Propulsion Laboratory, California Institute of Technology, 4800 Oak Grove Dr, Pasadena, CA (USA), Email : Edward.m.armstrong@jpl.nasa.gov*

### ABSTRACT

In 2010/2011 the Global Data Assembly Center (GDAC) at NASA's Physical Oceanography Distributed Active Archive Center (PO.DAAC) continued its role as the primary clearinghouse and access node for operational Group for High Resolution Sea Surface Temperature (GHRSSST) datastreams, as well as its collaborative role with the NOAA Long Term Stewardship and Reanalysis Facility (LTSRF) for archiving. Here we report on our data management activities and infrastructure improvements since the last science team meeting in June 2010. These include the implementation of all GHRSSST datastreams in the new PO.DAAC Data Management and Archive System (DMAS) for more reliable and timely data access. GHRSSST dataset metadata are now stored in a new database that has made the maintenance and quality improvement of metadata fields more straightforward. A content management system for a revised suite of PO.DAAC web pages allows dynamic access to a subset of these metadata fields for enhanced dataset description as well as discovery through a faceted search mechanism from the perspective of the user. From the discovery and metadata standpoint the GDAC has also implemented the NASA version of the

OpenSearch protocol for searching for GHRSSST granules and developed a web service to generate ISO 19115-2 compliant metadata records. Furthermore, the GDAC has continued to implement a new suite of tools and services for GHRSSST datastreams including a Level 2 subsetter known as Dataminer, a revised POET Level 3/4 subsetter and visualization tool, a Google Earth interface to selected daily global Level 2 and Level 4 data, and experimented with a THREDDS catalog of GHRSSST data collections. Finally we will summarize the expanding user and data statistics, and other metrics that we have collected over the last year demonstrating the broad user community and applications that the GHRSSST project continues to serve via the GDAC distribution mechanisms. This report also serves by extension to summarize the activities of the GHRSSST Data Assembly and Systems Technical Advisory Group (DAS-TAG).

### 1. Introduction

The GDAC serves as the key operational component for access and utility of GHRSSST data products worldwide. Its primary mission is to ensure timely and transparent access to GHRSSST datasets using a number of access protocols including FTP and OPeNDAP.

In this report we first describe key new improvements to the overall GDAC architecture. This includes implementation of the new PO.DAAC DMAS (data management and archiving system) that was reported on at the last meeting. Further sections are devoted to new products, and the development of tools and services for GHRSSST dataset subsetting and access. The metadata section documents how the GDAC has been actively improving metadata and fostering discovery of GHRSSST products, and helping to guide the development of an ISO-based metadata model. The last section details the GHRSSST data usage statistics since 2010.

## 2. GDAC integration and evolution

The original GDAC data interfaces to GHRSSST data producers, data consumers and data archive (LTSRF) were designed and implemented over 6 years ago, and as reported at the last Science Team meeting, a new PO.DAAC data management architecture, DMAS has now been implemented for all GHRSSST data streams. This new architecture has several improvements including scalability to handle increasing volumes of data ingest and dissemination.

In addition to aforementioned ingest and dissemination capabilities, further DMAS functions included metadata registry into an upgraded Master Metadata Repository (MMR) in an Oracle database in conjunction with its web-based search and discover interface, FGDC metadata generation and implementation of the NODC interfaces for GHRSSST data transfer for archiving, ingest data latency tracking and distribution metric capturing, and other enhanced operator functions. DMAS also assumes data management roles of the MODIS L2P RDAC including L2P ancillary filling. As shown in Figure 1, DMAS is a multi-mission data system that

offers data ingestion, validation, catalog, archive, and distribution capabilities. All GHRSSST data have been handled operationally by DMAS since June 2009.

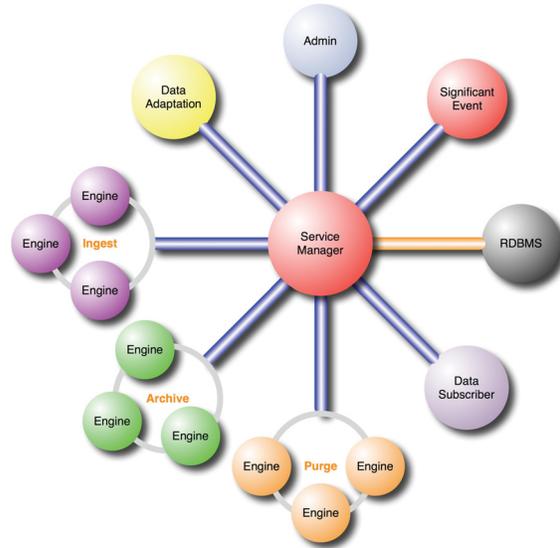


Figure 1. The top-level DMAS system architecture view.

## 3. New Products

The GDAC has continued to support the ingestion of new L2P and L4 products. These include:

- A global version of the Level 4 MEaSURES Multisensor (MUR) 1 km dataset
- GOES-13 L2P
- MTSAT2 L2P

Forthcoming products:

- Windsat L2P
- Global Level 4 DMI\_OI

## 4. Tools and Services

The PO.DAAC has made operational the existing beta version of the Dataminer Level 2 subsetter (Figure 2, <http://podaac-tools.jpl.nasa.gov/dataminer/aegina/src/da>)

taminer.php). This technology is an adaptation of the Ifremer NAIAD (Enhanced Satellite Archive Dataminer) tool described in detail in last years report. The core of the NAIAD system is the “virtual tile” database whereby each swath data granule is tiled, or divided into regions (typically representing 500km x 500km). The spatial and temporal metadata associated with each data region is stored inside of a tile, as well as that region’s statistical properties. Using this approach a user can quickly perform a space/time query and download only the granules meeting the search criteria via OPeNDAP connections to either the GDAC or LTSRF. Currently AMSRE, and Aqua and Terra MODIS L2P granules are accessible via Dataminer for subsetting.

**Dataminer Swath Subsetting Tool**



- seamless access to L2 products across data centers
- minimal set-up requirements for product inclusion
  - web services can be used directly without GUI
  - no dependency on orbital parameters for search
- up to 4 constraints: space, time, product, data summaries
  - comprised of a collection of technologies

**SOAP web services**

query  imaging  
extraction

**Tile catalog**



mysql database

**Javascript interface**



powered by web services

**Data access**

**OPeNDAP**  
standardized protocol

**Product tilters**



summarize data at sub-file resolution

Figure 2. The features and components of the NAIAD/Dataminer system.

A Google Earth-based interface called State Of The Ocean (SOTO) has been implemented as a core visualization tool for the physical disciplines supported by the PO.DAAC including sea surface temperature. As shown in Figure 3, a user can globally visualize SST fields from the previous five days using GHRSSST MODIS L2P, AMSRE L2P, or G1SST L4 or some combination thereof including SST anomaly data. Other parameters including wind, SSH and ocean chlorophyll are also available. No specialized software other than a web browser and the Google Earth plug-in is required to run this system. SOTO can be accessed from: <http://podaac-tools.jpl.nasa.gov/soto/>

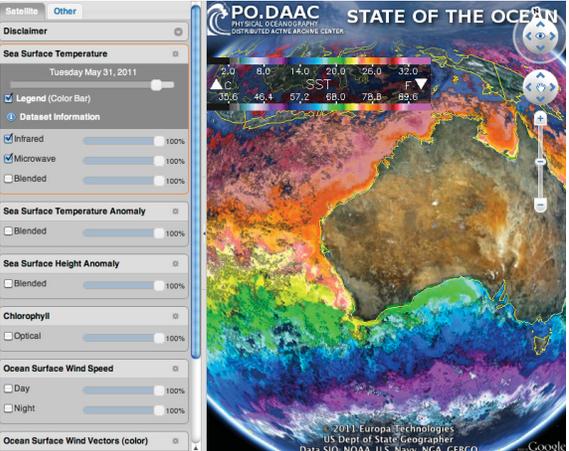


Figure 3. The SOTO Google Earth interface for GHRSSST SST (and others).

The POET Level 3/4 subsetter and visualization tool has also received an upgrade to the user interface and has improved speed and functionality. It can be found at <http://newpoet.jpl.nasa.gov> . A number of GHRSSST global Level 4 products are accessible via POET.

PO.DAAC has also experimented with an implementation of a THREDDS (Thematic Realtime Environmental Distributed Data Services) server, and will eventually deploy a version to aggregate GHRSSST datasets into yearly and annual catalogs.

A GHRSSST specific user forum has been established and will be managed at the PO.DAAC

(<http://podaac.jpl.nasa.gov/forum/forum/4>).

This forum will allow a single location for future collaborations among science team members, technical advisory groups and eventually even users.

## 5. Metadata and Discovery

In the area of data discovery and metadata the GDAC has made significant progress as part of new strategies adopted in the PO.DAAC DMAS infrastructure. First, GHRSSST datasets are now discoverable via a faceted search mechanism directly from the PO.DAAC website. An example of this interface is shown in Figure 4, where one option is to browse the entire “collection” of GHRSSST datasets. Users are presented with a metadata summary page for each GHRSSST dataset including an interface to access individual granules. Examples of other facets that can be browsed are “sensor”, “platform” and “resolution.” This interface is extendable and can be modified to suit the requirements and preferences of the user community.



Figure 4. Faceted search capability for all PO.DAAC datasets including GHRSSST. In this example the entire “collection” of metadata of all 57 GHRSSST datasets will be browsed.

For dataset and granule discovery web services, the PO.DAAC has implemented a new data discover system: the Oceanographic Common Search Interface (OCSI) (Figure 5). In addition to serving as the backend infrastructure of PO.DAAC’s faceted web search capability, OCSI is designed to support discovery of PO.DAAC data according to various metadata standards. Currently the system supports the ESIP Discovery Specification, the NASA extensions of the standard OpenSearch protocol (<http://www.opensearch.org>). This protocol specifies a way to discover data and return XML (atom/rss) structured search results based on a pre-defined user query. Initial search constraints are limited to keyword, space/time queries but can be eventually extended to other attributes one richer metadata are indexed at the granule level. OCSI has also implemented prototype support of ISO 19115-2 metadata records for GHRSSST datasets following the metadata model specifications in GDS 2 (Figure 6).

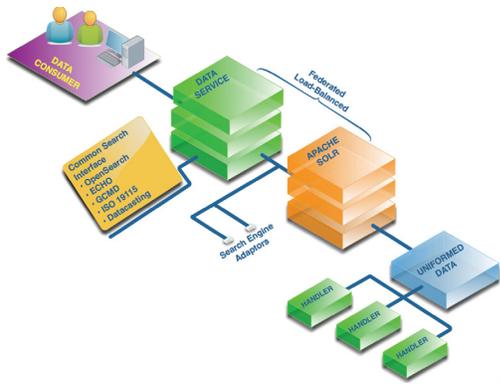


Figure 5. The web services of the Oceanographic Common Search Interface (OCSI) in yellow that interact with the PO.DAAC dataset inventory.

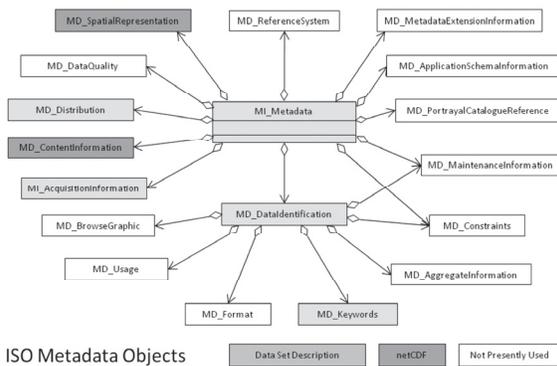


Figure 6. ISO 19115-2 objects for the GDS 2.0 metadata specification

From the perspective of the dataset metadata, the PO.DAAC has embarked on a major metadata quality improvement project. This effort is focused on improving the accuracy and completeness of the metadata attributes including the quality of the dataset description or

abstract. A database interface and maintenance tool developed by the PO.DAAC makes this task much more straightforward and ensures consistency across all GHRSSST datasets. Once this quality effort is complete by Fall 2011 all GHRSSST datasets will be exported in the Directory Interchange Format (using OCSI) to the NASA Global Change Master Directory. This should significantly enhance the exposure of GHRSSST products to a broader community of potential users.

The PO.DAAC continues to be active in providing GHRSSST metadata to the NASA's Earth Observing System Clearinghouse (ECHO), a metadata search interface to all NASA earth science data holdings at the granule level. Currently, ECHO contains 32 GHRSSST data sets with more than 270,000 granules. These data sets and granules are available for search through the new ECHO Reverb interface.

## 6. GDAC data metrics

The following figures (Figure 7 and 8) are representative summaries for the data volume (compressed) and number of users of GHRSSST data from the GDAC since early 2006. One metric to note is that the GDAC has distributed 30 million granules representing about 50 TB of compressed volume since June 2010 (around 4.0 TB/month). More enhanced statistics will be reported at the June 2011 Science Team meeting.

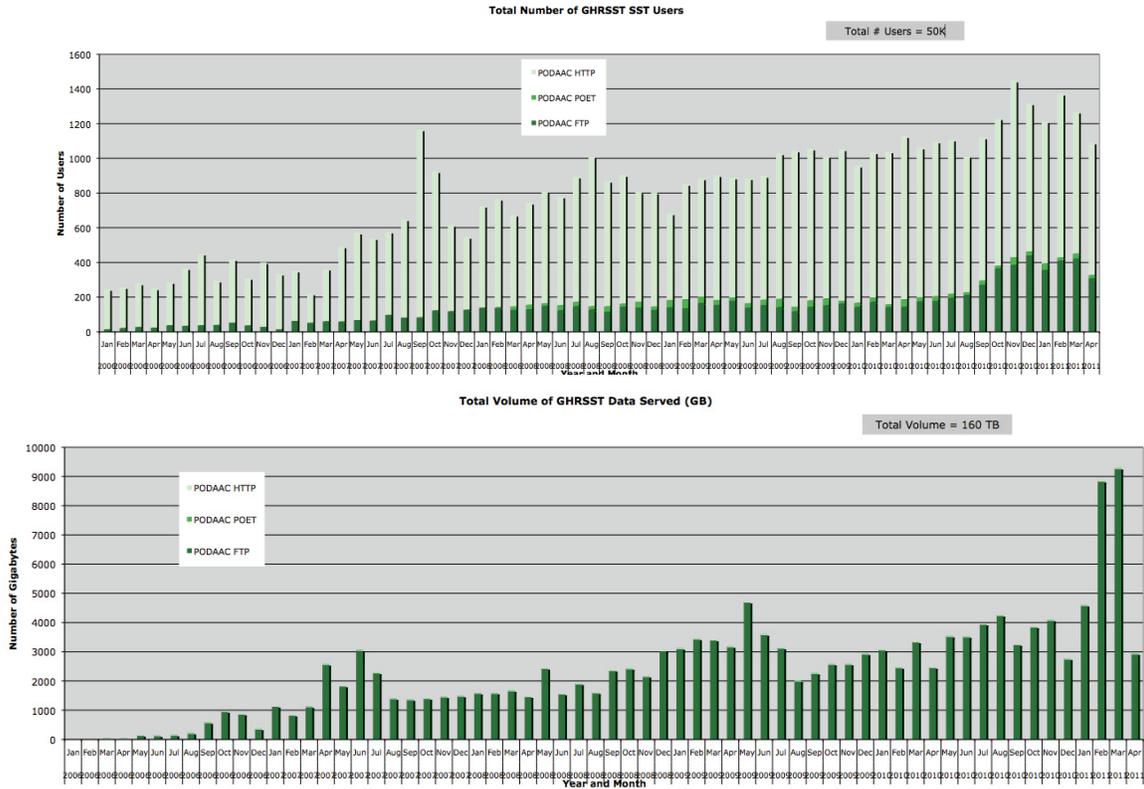


Figure 7 and 8. GDAC user and data distribution (compressed volume) summaries.

## 7. Conclusion

The Global Data Assembly Center (GDAC) continues to meet its requirements to distribute increasing numbers of GHRSSST datasets and volumes, foster data discovery, maintain meaningful metadata records, implement robust data stewardship practices, and build new data utilization tools and services. GHRSSST datastreams can now leverage off an improved and scalable data management system that has recently be put into place at the PO.DAAC as well as new subsetting, visualization, data discovery, web services and

metadata tools. NASA has recognized the importance of GHRSSST data (with the recent 2011 NASA Physical Oceanography proposal call emphasizing these products) while supporting the concept that leading edge research cannot be fostered without strong data management principles and infrastructure. The GDAC is committed to maintaining GHRSSST data for all users in conjunction with the NOAA Longterm Stewardship and Reanalysis Facility well into the future.

*This work was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration.*

© 2011 California Institute of Technology. Government sponsorship acknowledged.