DAS-TAG Report

Chair : Edward Armstrong
Co Chair: Jean Francois Piolle

14th GHRSST Science Team Meeting

Woods Hole, MA, USA

21 June 2013

National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

What do we do for GHRSST?

* Provide technology infrastructure and informatics expertise
* Foster data usability, stewardship and understanding for present and future generations

Agenda

Emerging trends in metadata, *Ted Habermann (remote)*

PO.DAAC integrated web services, *Ed Armstrong*

Reconciling GHRSST archive integrity and data flows, *Ken Casey*

A Hadoop framework for data mining and analyses of large datasets, *Jean-Francois Piolle*

Proposals for new GHRSST dataset policies, *Ed Armstrong and Gary Corlett*

Summary

* Metadata evolution
  * Proven ability to translate among metadata formats at the dataset level
  * ISO 19115 provides powerful ways to capture dataset lineage and quality that are important to GHRSST
* Data accountability and Lifecycle
  * Need improved reconciliation of data flows between GHRSST data centers
  * Even users should be given tools to run “reports”

Summary

* Dataset Lifecycle …
  * A process to improve the stewardship of GHRSST data within the GDAC
  * Proposal for a "Submission Agreement" for data provider
    * Document data lineage, quality and uncertainty
* Technology improvement development
  * Bring processing and services of data directly to the user !
    * Integrated web services for dataset and granule discovery, metadata, data subsetting, extraction and imaging
    * Input of one service to call another
    * GDAC will rollout these in 6 months

Summary

* Technology improvement development….
  * Bring processing and services of data directly to the user !
    * Hadoop framework for distributed computing -- Ifremer proof of concept
    * ***Nephelae*** system with 600 cores
      * Example: process 4 years of L4 data to produce anomalies from a climatology in 90 seconds
      * 10 years of QuikSCAT L2 data to produce daily wind statistics in 2 minutes

Issues and concerns update

* Way forward on implementing on dataset lifecycle policy
* GDS governance has been resolved ?
* Implementing schedule for new GDS2 datasets
* Non standard distribution of L2P data still outstanding