



National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

Distributed Oceanographic Webification Service

Thomas Huang, Zhangfan Xing, Edward M. Armstrong

Jet Propulsion Laboratory, California Institute of Technology,
4800 Oak Grove Drive,
Pasadena, CA 91109-8099, United States of America.



PO.DAAC

- The **NASA Physical Oceanographic Distributed Active Archive Center (PO.DAAC)** at Jet Propulsion Laboratory is an element of the **Earth Observing System Data and Information System (EOSDIS)**. The EOSDIS provides science data to a wide communities of user for NASA's Science Mission Directorate.
- Archives and distributes data relevant to the physical state of the ocean
- The mission of the PO.DAAC is to preserve NASA's ocean and climate data and make these universally accessible and meaningful.



<http://podaac.jpl.nasa.gov>

The screenshot shows the PO.DAAC website interface. At the top, there are navigation links for NASA Earth Data, Data Discovery, Data Centers, Community, and Science Disciplines. Below this is the Jet Propulsion Laboratory logo and the PO.DAAC title: "PHYSICAL OCEANOGRAPHY DISTRIBUTED ACTIVE ARCHIVE". A search bar is present with the text "SEARCH FOR DATASETS: Enter Dataset Keyword". The main content area is divided into several sections: "Parameter" (with sub-links for Collections, Platform, Sensor, Spatial Coverage, Latency), "OCEAN STORY" (highlighting the AQUARIUS satellite detecting river flooding), "DATA ACCESS TOOLS & SERVICES" (listing protocols like FTP, OPeNDAP, THREDDS, LAS, HITIDE), "ANIMATION & IMAGES", "EVENTS" (listing meetings from 2013-2014), "ANNOUNCEMENTS" (listing hardware maintenance and outages), "LEARN ABOUT" (listing oceanographic parameters), "DATASET HIGHLIGHT" (featuring the Reconstructed Sea Level Dataset from June 17, 2013), "OCEAN STORIES", "PO.DAAC SERVICES & TEAM", and "EOSDIS". The footer includes the USA.gov logo and links for Privacy, FAQ, and Feedback.





PO.DAAC Labs

- Explore New Ideas
- Prototypes
- Next Generation of Tools and Services
- Coming Soon, Data Intensive Science Data Services

http://podaac.jpl.nasa.gov/podaac_labs

The screenshot shows the PO.DAAC Labs website interface. At the top, there is a navigation bar with links for NASA Earth Data, Data Discovery, Data Centers, Community, Science Disciplines, and a search bar for EOSDIS. Below this is the Jet Propulsion Laboratory logo and the text 'BRING THE UNIVERSE TO YOU' with social media icons. The main header features the PO.DAAC logo and 'PHYSICAL OCEANOGRAPHY DISTRIBUTED ACTIVE ARCHIVE CENTER'. A secondary navigation bar includes links for HOME, DATASET DISCOVERY, DATA ACCESS, MEASUREMENTS, MISSIONS, MULTIMEDIA, USER COMMUNITY, HELP, and FORUM. Below the navigation, there is a section titled 'PO.DAAC LABS' with a sub-header 'Explore New Ideas, Prototypes and Tools. Offer feedback directly to the engineers who developed them.' and an email contact: podaac@podaac.jpl.nasa.gov. The main content area lists three services: 'State of the Ocean 2D (SOTO 2D)', 'Webification', and 'Coastal Marine Discovery Services', each with a small thumbnail image and a brief description.

This block shows the continuation of the PO.DAAC Labs website content. It features three more service entries, each with a thumbnail image and a description. The text is partially obscured by a vertical line on the right side of the page, but the layout and structure are consistent with the previous block.



BIG DATA

Capture

Curation

Storage

Search

Sharing

Transfer

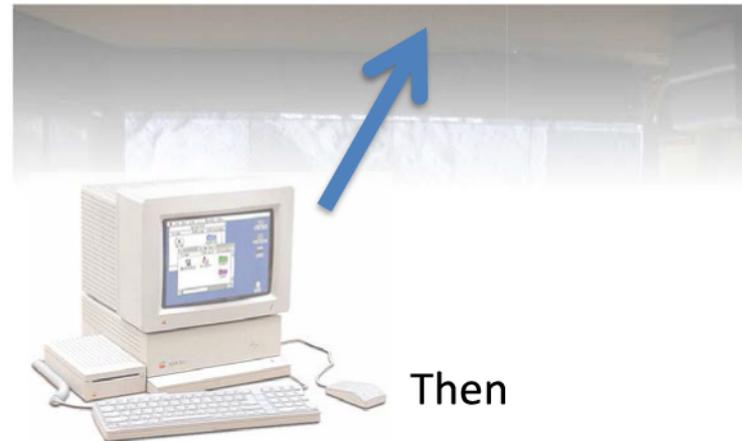
Analysis

Visualization



Big Data

- **Handles**
 - More Requests
- **Processes**
 - More Transactions
- **Gathers**
 - More Metrics
- **Utilizes**
 - More Science Data Artifacts
- **Delivers**
 - More Results With Less Time





~~BIG DATA~~

DATA-INTENSIVE SCIENCE

“The goal is to have a world in which all of the science literature is online, all of the science data is online, and they interoperate with each other.”

– Jim Gray, *The Fourth Paradigm: Data-Intensive Scientific Discovery*



Scientific Data Management in the Coming Decade – Jim Gray

“The scientific file-formats of HDF, NetCDF, and FITS can represent tabular data but they provide minimal tools for searching and analyzing tabular data. Their main focus is getting the tables and sub-arrays into your Fortran/C/Java/Python address space where you can manipulate the data using the programming language.”

“Performing this *filter-then-analyze*, data analysis on large datasets with conventional procedural tools runs slower and slower as data volumes increase. “

“**Set-oriented file processing** will make file names increasingly irrelevant – analysis will be applied to “all data with these attributes” rather than working on a list of file/directory names or name patterns.”



The Fundamental Physics

- Moving/copying science data (and managing copies) is more expensive than computation
- Hardware & software do not yet make science data analysis easy at terabyte scales
- Current analytics are mostly I/O bound.
- Next generation - “advanced” analytics will be compute bound (simulations, distributed linear algebra). Efficiency matters.
- **Bad data architecture** will generally not cause catastrophic failures
- Instead, it will erode our ability to compete

“It’s hard to know when you are sucking.”

– Peter Wang, Founder and President of Continuum Analytics



The real challenge is how to scale the **architecture** so it can respond to an exponential increase in the amount of data we are serving.

Science Data System Architecture is an orchestration of

People

Process

Policies

Technologies



Webification (w10n)

- Webification (w10n) is an enabling technology
- Developed as an interface specification with two implementations
 - **Pomegranate** (<http://pomegranate.nasa.gov>): Python WSGI application for Apache HTTPD. A RESTful web-service that webifies various file formats (NetCDF, HDF 4/5, GRIB, FITS, etc.)
 - **Juneberry** (<http://juneberry.jpl.nasa.gov>): Java servlet-based implementations to webifies some of our planetary image data formats (VICAR/PDS, FITS, TIFF, JPEG, GIF, etc.)
- **Goal:** make data easy to use in the “web” way
- **Idea:** inner components of an arbitrary data store, such as attributes, labels, image bands, and data arrays, are directly addressable and accessible by well-defined and meaningful URLs.
 - Allows contents of remote HDF and NetCDF files to be extracted via RESTful URLs



Extensibility

Webification is a specification that can be apply to any structured data store/container/data model. Already supported under the PO.DAAC's w10n instance, Pomegranate

- File system directories and files
 - With wildcard support
 - <http://example.com/test/data/>*
 - [http://example.com/test/data/\[a-zA-Z\]*\[0-9\]*/](http://example.com/test/data/[a-zA-Z]*[0-9]*/)
 - [http://example.com/test/data/\[a-zA-Z\]*\[0-9\]/*.nc/](http://example.com/test/data/[a-zA-Z]*[0-9]/*.nc/)
- Data formats (NetCDF, HDF, GRIB, FITS, TEXT)
- Remote service (OPeNDAP)
 - Use w10n to proxy to any local and remote OPeNDAP service
- Work in progress
 - Data Model (DAP)
 - Databases (relational, non-sql)



Webification (w10n)

- PO.DAAC Labs is now hosting **Pomegranate (w10n)** under <http://podaac-w10n.jpl.nasa.gov>) to webify most our data holdings
- **General w10n URL Format:**
`http://host:port/some_path/webifiable_store/identifier?queryString`
- **Example:**
`http://podaac-w10n.jpl.nasa.gov/w10n/allData/ghrsst/data/L2P/MODIS_A/JPL/2013/168/20130617-MODIS_A-JPL-L2P-A2013168000000.L2_LAC_GHRSSST_D-v01.nc.bz2/sea_surface_temperature[0:1,0:20:2,30:40:4]?output=json`



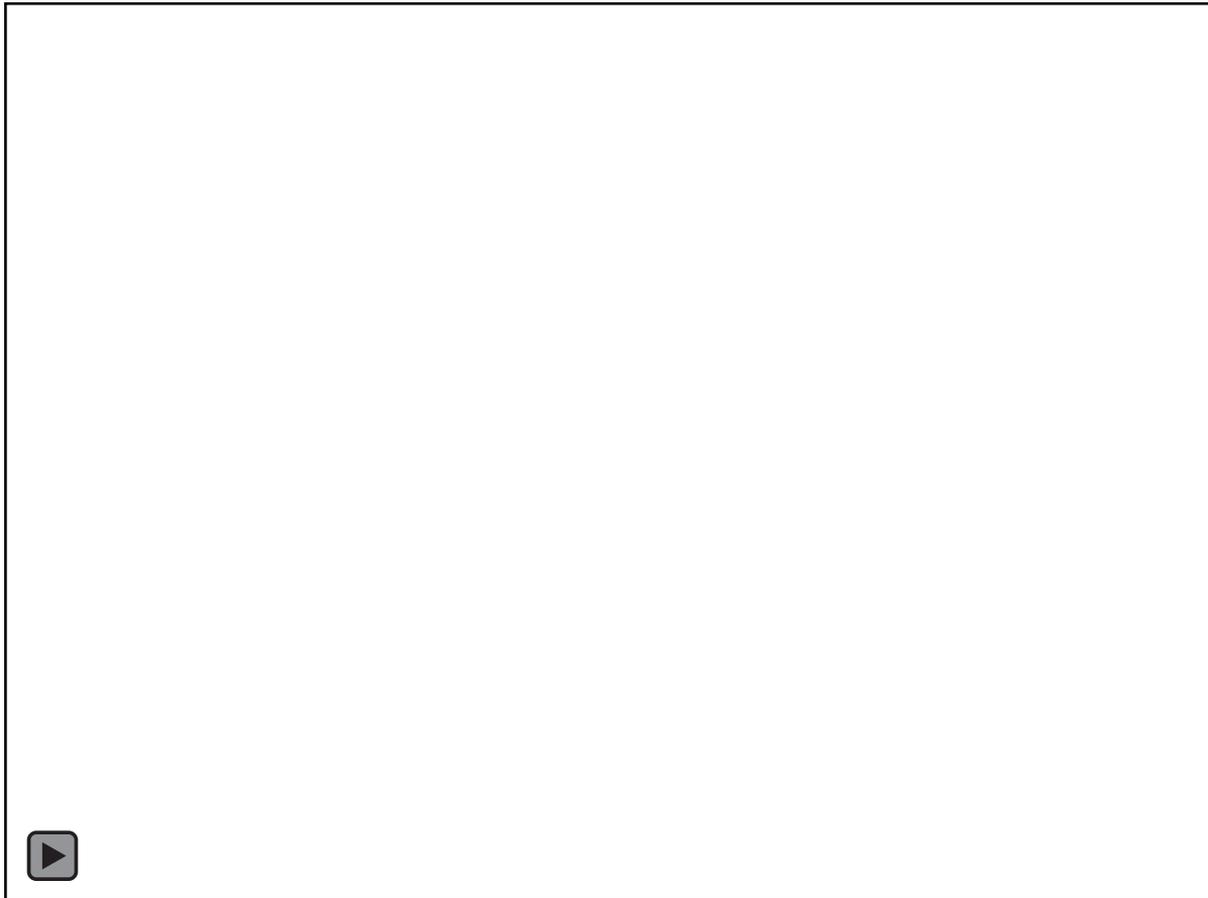
Webification and OPeNDAP

Characteristics	Webification	OPeNDAP
Data store structure	Hierarchical tree, consisting of nodes and leaves with associated meta and data information	Internal flat data model to support common numerical array types
URL Syntax	Predictable URL. Extends the "path" portion of classic URL syntax by introducing the forward-slash '/' notation for meta info, and square-bracket '['] notation for data info.	Use pre-defined keywords and "queryString" to denote data info.
RESTfulness	Fully RESTful with support for both read and write	Partially RESTful. Read-only
Array Response	JSON, NetCDF, big/little-endian binaries, CSV, etc	Less structure ASCII text and data info (array) can be in ASCII, NetCDF, DAP binary
Compression	HTTP compression	Own internal compression
Extensibility	Supports various data format including OPeNDAP data model	Flat data model = difficult to support anything that is hierarchical in nature



L2 Swath Data: Using WebGL to scale a lot of numbers

- Using Level 2 near real-time granule from MODIS that is directly above *Typhoon Haiyan*
- The netCDF file we visualize contains over 2.8 million data-points. We are not only able to visualize the data in the browser, but we can dynamically model the data to flat or spherical modes, with added terrain mode to visualize deltas.





Visualization as a Service

- Using w10n to remotely get L3 Grace data from a NetCDF file located somewhere within PO.DAAC.





Virtualized Quality Screening Service (VirtualQSS)

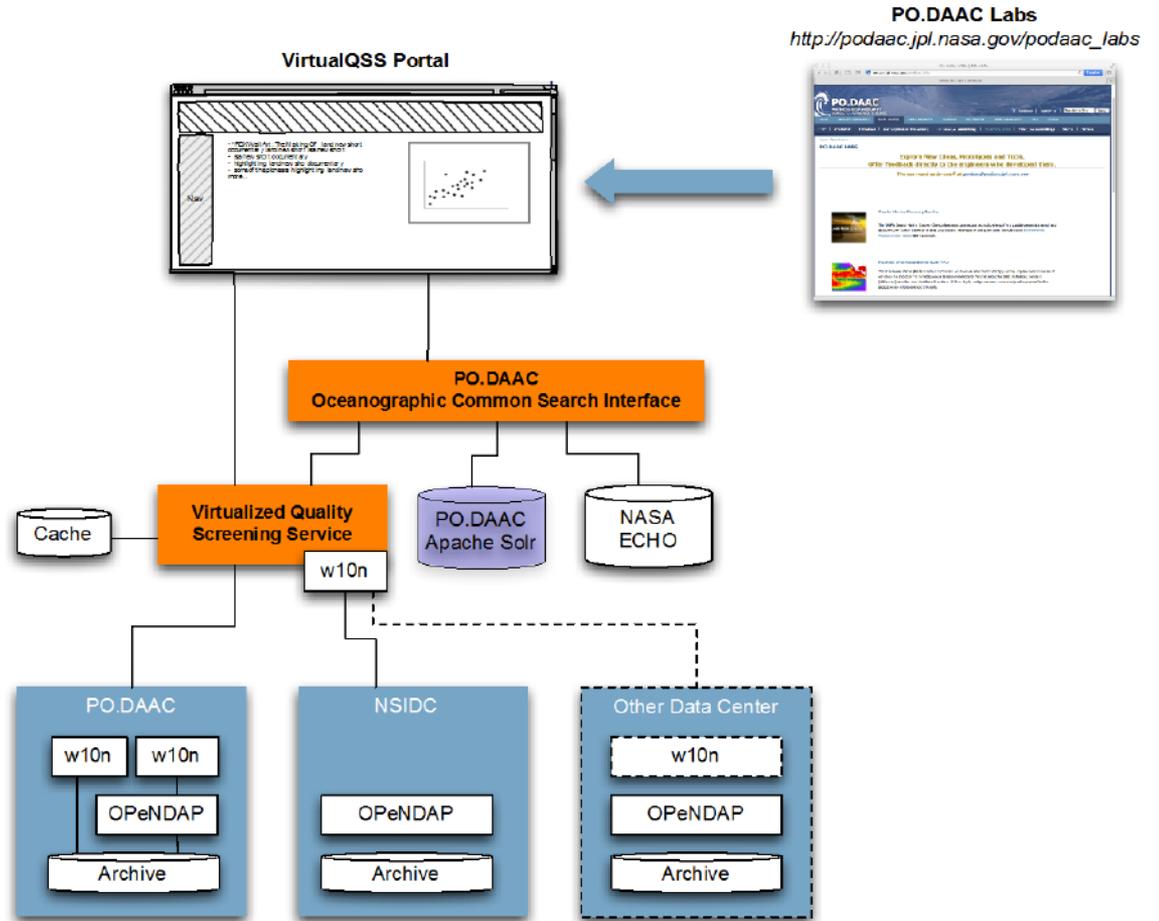
- PI: E. Armstrong; CO-Is: T. Huang, Z. Xing, and T. Chin
- Respond to ROSES 2013 - NRA NNH13ZDA001N-ACCESS
- Data of interest frequently contains commensurate quality information in the form of a flag or other metric that a user must apply to actual geophysical data to make it meaningful and understandable
 - This is a three-step process using OPeNDAP
 1. Request geophysical data
 2. Request the quality data
 3. Apply quality information (e.g. a flag) to the geophysical data in order to make the result usable for further analysis or even visualization

Manual data discovery and movement as well as local computation



VirtualQSS – The Cloud-based Service

- High performance indexed, temporal spatial engine to generate predictable w10n array requests
- Leverage elastic nature of Cloud Computing to retrieve geophysical data and quality information
- Leverage Map-Reduce architecture to apply quality information to create quality screened products

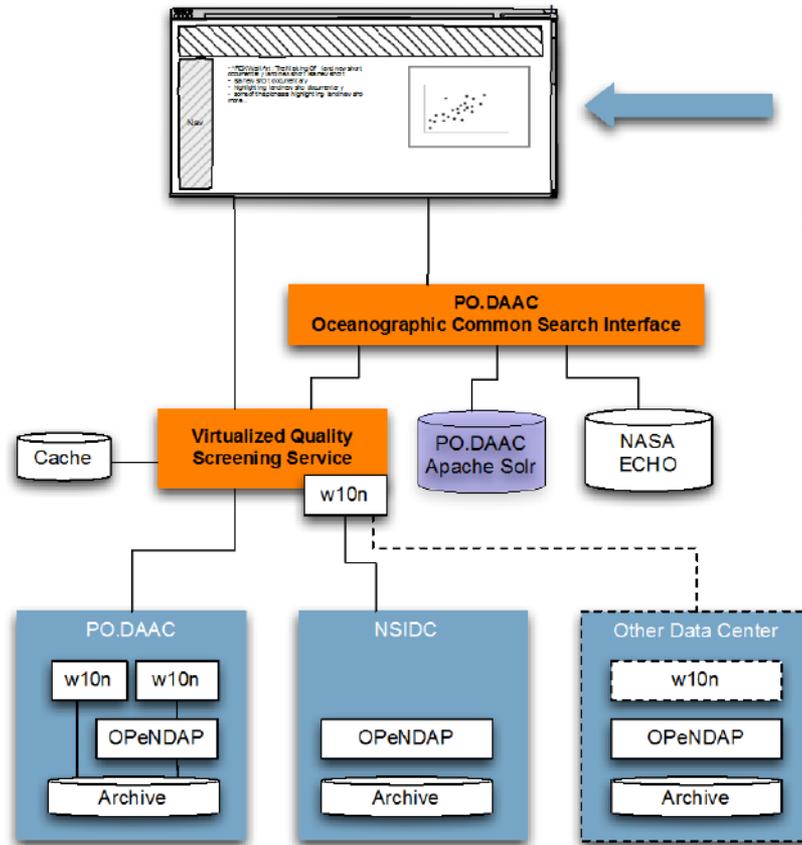


PO.DAAC Labs

http://podaac.jpl.nasa.gov/podaac_labs



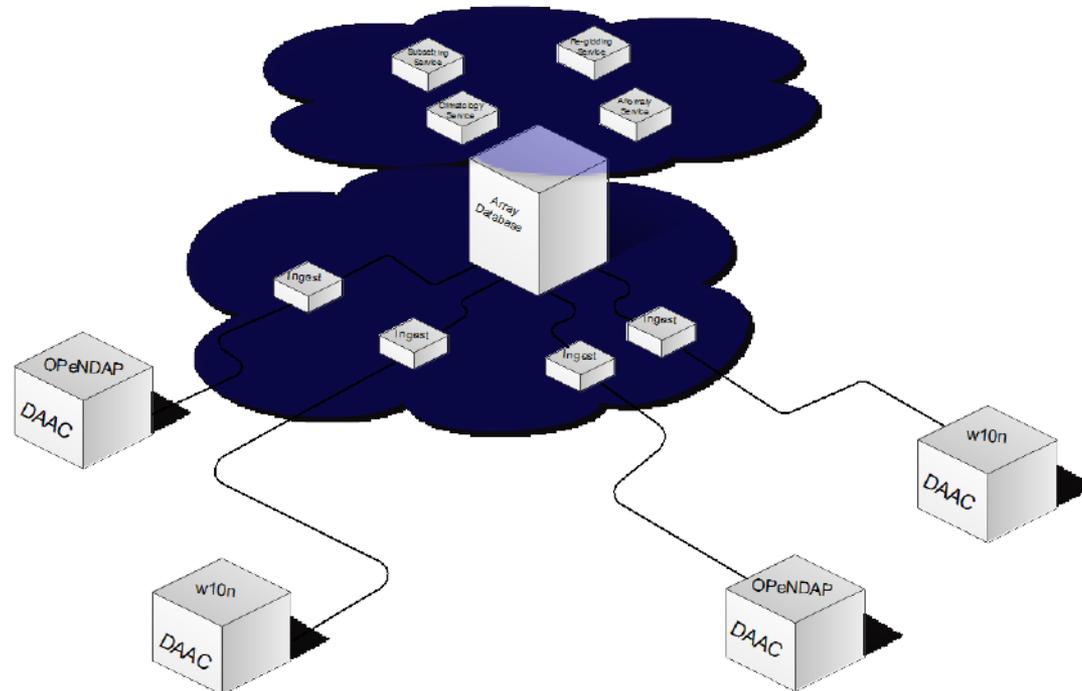
VirtualQSS Portal





Cloud Computing at Earth Science Data System Working Group (ESDSWG)

- The Earth Science Data System Working Groups (ESDSWG) is a NASA organization established under the auspices of NASA Headquarters in 2004
- Funded activity: Science Data Service Platform
 - Prototype a file-free solution for the next generation of science tools and services
 - Retain provenance information and lineage to the data center for granule retrieval
 - Using w10n as the enabling technology for harvesting array data into the Cloud environment





SOME “BIG DATA” GRAND CHALLENGES

How do we handle 700 TB/sec of data coming off the wire when we actually have to keep it around?

- Required by the Square Kilometre Array

I want to draw a box anywhere on the globe, just get me all the data in that box and plot the temporal change/trend.

- Required by PO.DAAC User Working Group

We need a better way to help our users find the right dataset(s) and related dataset(s).

- Required by PO.DAAC User Working Group

How do we compare petabytes of climate model output data in a variety of formats (HDF, NetCDF, Grib, etc.) with petabytes of remote sensing data to improve climate models for the next IPCC assessment?

- Required by the 5th IPCC assessment and the Earth System Grid and NASA

How do we catalog all of NASA’s current planetary science data?

- Required by the NASA Planetary Data System



National Aeronautics and
Space Administration

Jet Propulsion Laboratory
California Institute of Technology
Pasadena, California

THANKS

Questions, and more information

Thomas.Huang@jpl.nasa.gov