

# OGC / Geoscience Gateway Final Report (Draft!)

Christopher Lynnes, Principal Investigator, GES DISC  
Ken McDonald, Principal Investigator Emeritus, NOAA  
Liping Di, Co-Investigator, George Mason University  
Ben Domenico, Co-Investigator  
Yonsook Enloe, SGT  
Dan Holloway, OPeNDAP.org  
Glenn Rutledge, Co-Investigator, NOAA/NCDC  
Wenli Yang, Co-Investigator, George Mason University  
Chengfang Hu, George Mason University  
Min Min, George Mason University

## Executive Summary

The OGC / Geoscience Gateway project developed two technologies to bridge the gap between protocols used in the Open Geospatial Consortium (OGC) community and those used within geosciences. One of those, the CEOP Satellite Data Server, provided a gateway between OPeNDAP and the OGC Web Coverage Service (WCS) that allowed scientists with OPeNDAP-enabled clients to access data served by an OGC WCS server. In addition to the technology itself, the project obtained some key findings relevant to interoperability among NASA data systems, as well as GEOSS:

1. Access to Level 2 (swath) AIRS data by GrADS clients was made transparent through the WCS server and OPeNDAP gateway.
2. The expected user access pattern has a substantial effect on the implementation and performance of both the Gateway and the WCS server. In particular, the longitudinal (over time) access desired by CEOP users demanded significant changes in the WCS server, which were reflected in the OPeNDAP gateway.
3. Maintaining data provenance from original data file through two servers to the client was demonstrated but remains a significant challenge for Service-Oriented or Resource-Oriented Architectures, particularly where chaining is involved.
4. The concept of a third-party OPeNDAP server providing OPeNDAP services for remote data (with intelligent caching) could enhance interoperability for data provided by organizations unable (or unwilling) to run OPeNDAP themselves.

The second technology was an OGC Catalog Services for the Web (CSW) server that enables searching of data in a Thematic Real-time Environmental Distributed Data Services (THREDDS) Data Server (TDS), with the following key findings:

1. The CSW server was demonstrated both with a local demonstration client and with an independent client, GI-GO, from the Earth and Space Sciences Informatics Laboratory in Florence, Italy.
2. CSW development could be improved if there were more clients supporting it, particularly jointly with WCS support.

## Background and Project Goals

The OGC / Geoscience Gateway project was formed as a combination of two selected ACCESS proposals. One of these projects was the Development and Deployment of a CEOP Satellite Data Server, which sought to develop a gateway server that would allow OPeNDAP clients to obtain data from servers providing data via Open Geospatial Consortium (OGC) Web Coverage Service (WCS). The aim was to help water and energy cycle researchers within the Coordinated Enhanced Observing Period (CEOP), many of whom use OPeNDAP clients such as the Gridded Analysis and Display System (GrADS). The other project was the Gateway for Interoperability of Atmosphere, Land, Ocean, and Modeling Science Data, which aimed to make geoscience data, such as those stored in the

THREDDS data servers, available to the OGC community. The two proposals were complementary, each attempting to bridge the gap between the geoscience research community and its family of protocols and servers such as THREDDS and OPeNDAP, with the OGC community. As a result, they were combined. The second project was also rescoped to focus on the catalog aspect of interoperability, specifically to provide OGC Catalog Services for the Web (CSW) for data stored in THREDDS Data Servers (TDS).

At a more detailed level, the first half of the project had as its chief technical goal the creation of a plug-in "handler" for the OPeNDAP server, which would be able to fetch data from a WCS server. The result should allow an OPeNDAP client such as GrADS to directly and remotely "open" a virtual OPeNDAP "file", which actually represents a response from a WCS server. This would allow, for example, Level 2 swath data (such as AIRS standard retrievals) to be reprojected on the fly, served through WCS to the OPeNDAP gateway and to a GrADS client that ordinarily could do nothing useful with the original data. Thus, the output product would be a software component, the WCS handler, to be added to the OPeNDAP released code baseline.


The CSW for THREDDS project, on the other hand, was designed to allow a client to search a database of the data available in THREDDS using the OGC CSW protocols. Ideally, the eventual return would include information about available WCS coverages, allowing the user to then request return of a particular coverage from the THREDDS data server. Thus, the end result should be a functioning server and underlying database which can ingest THREDDS metadata and reply to CSW requests from OGC-compliant clients.

## **Results**

In brief, the technical goals of both of the subprojects were met. An OPeNDAP handler was developed and released to the community on TBD. The data used for demonstration and acquiring feedback from CEOP beta testers consisted of 28 parameters from AIRS Level 2 standard retrievals:

OPeNDAP Hyrax: WCS Coverage Offerings at NASA\_DAAAC\_WCS

http://dev1.opendap.org:8080/opendap/wcs/CEOP/BRA/ceopl2AIRTimeOffsetLineage/contents.html



## NASA DAAC Web Coverage Server

Name	Lat/Lon Envelope
<u>Surface Air Temperature</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Retrieved Atmospheric Temperature Profile</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Geopotential height, above mean sea level, for each pressure level</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Geopotential height, above mean sea level, at surface</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Surface pressure, interpolated from the NCEP GFS forecasts and local DEM</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Surface Skin Temperature</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Total Precipitable Water</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Water Vapor Mass Mixing Ratio Profile</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Saturation Water Vapor Mass Mixing Ratio Profile over equilibrium phase</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Saturation Water Vapor Mass Mixing Ratio Profile over liquid phase</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Ozone Volume Mixing Ratio Profile</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Total Ozone Burden</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Cloud top pressure for each valid cloud layer</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Cloud top temperature for each valid cloud layer</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Outgoing longwave radiation flux integrated over 2 to 2800 cm-1</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Clear sky Outgoing longwave radiation flux integrated over 2 to 2800 cm-1</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Total column CO</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Effective CO Volume Mixing Ratio Profile</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>CO effective pressure for the center of each trapezoid</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Total column CH4</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Effective CH4 Volume Mixing Ratio Profile</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>CH4 effective pressure for the center of each trapezoid</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Geopotential height, above mean sea level, for each pressure level</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Microwave surface brightness, including only emitted radiances</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Spectral microwave emissivity</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Total precipitable water vapor</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Total cloud liquid water</u>	[-158.995000 -38.035000] [150.525000 73.992000]
<u>Number of retrieved cloud layers, 0, 1, or 2</u>	[-158.995000 -38.035000] [150.525000 73.992000]

THREDDS Catalog [HTML](#) [XML](#)

Hyrax development sponsored by [NSF](#), [NASA](#), and [NOAA](#)

OPeNDAP Hyrax WCS Gateway

Figure 1. Snapshot of 28 AIRS parameters offered through OPeNDAP/OGC Gateway.

The handler was tested with several OPeNDAP clients, including two of the most popular, GrADS and Ferret. It was also used to set up a demonstration Live Access Server (LAS).

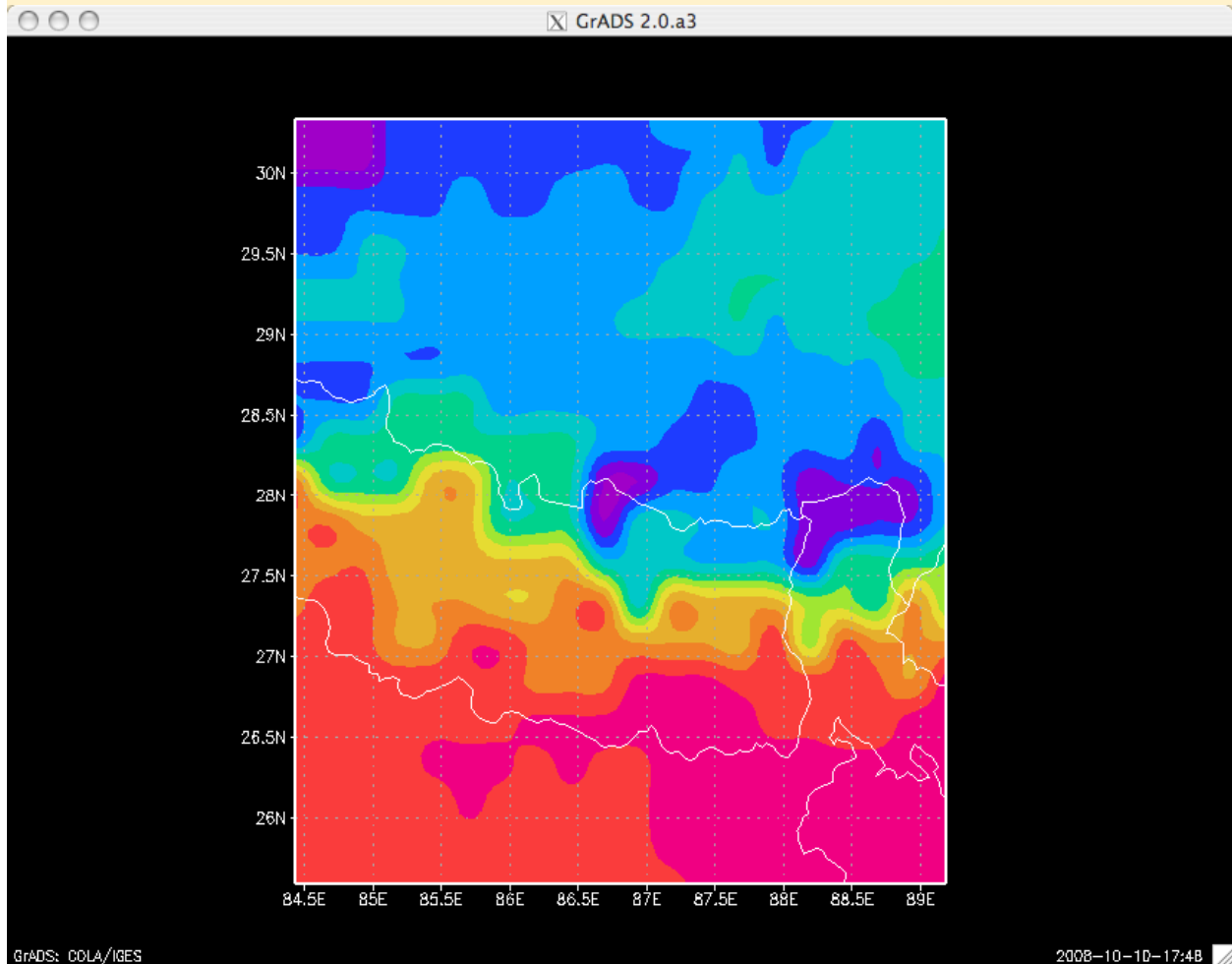


Figure 2. Surface Air Temperature from AIRS for the Himalayan CEOP Reference Site on 2002-12-01. The image was produced in GrADS by acquiring data via the OPeNDAP / OGC Gateway.

## Lessons Learned

### The Importance of Time

The most significant lesson arose from feedback from test users: it became clear that the users were primarily interested in looking at the variation in the data over time, particularly for the small CEOP reference site areas. The GrADS tools is often used for this kind of time series analysis. This exposed a key difficulty in serving data to this community. The WCS protocol is primarily geographically oriented, though it does have some (little-used) support for the time dimension. Furthermore, the original AIRS and MODIS data are stored in files representing individual time slices. In theory, a WCS server can be constructed to rearrange the data into a three-dimensional space-time cube. However, for long time series (such as the two years representing the CEOP-3 and CEOP-4 observing period), this is not practical due to the performance impact, and indeed most clients will time out before anything can be returned. An alternative is to use an "aggregation server" on the external side of the OPeNDAP server, resulting in multiple requests to the OPeNDAP Gateway and WCS server. This avoids timeouts at the WCS server, but essentially moves them to the aggregation server.

A second significant impact of the time-oriented user view was the need to make the data evenly spaced in the time dimension. In reality, satellite scene and swath data for a specific geographic area have varying collection times, depending on the particular orbital geometry and orbit repeat times. To solve this problem, we artificially aligned the data on pre-set times (0100 and 1300). To maintain the time information in the response, we added a `time_offset` variable which, when added to the main time dimension, would yield the actual collection time.

## WCS Implementation Lessons

In addition to the time oriented view, the target clients (particularly GrADS) and the organization of the source data imposed further constraints on the WCS implementation. Many of these are highly technical, but important details; as a result they have been described in a separate document "**CEOS WGISS: Lessons Learned From WCS Server Design And Implementation**". These lessons include:

- Limiting the length of the URLs to stay within client limits
- Making the access URLs deterministic (i.e., predictable) and informative so that user scripts can be generated
- Quality screening the data before reprojection in the WCS server
- Including ancillary information in the WCS response
- Reprojecting and mosaicking Level 1 and Level 2 data

## OPeNDAP Lessons

The work identified several areas where interoperability between OPeNDAP's Data Access Protocol (DAP) and the OGC/WCS interface presents challenges.

1: CEOP scientists frequently use DAP-enabled NetCDF applications for data analysis, and by design these applications typically iterate read operations against the data they're operating upon. On the WCS side, the WCS service generates the coverage response, (i.e., the coverage is a virtual dataset) from a larger data source, potentially using compute-intensive operations. The problem we're presented is a client application operating in a stateless environment making iterative read requests against a virtual dataset. Insuring timely response required implementing a caching solution. Caching could be implemented in several places, at the WCS, within the DAP `wcs_gateway_handler`, at the DAP server, or in the client. We chose to implement caching at the DAP server (i.e., hyrax). After initial work to implement caching into the `wcs_gateway_handler` itself, we chose to employ a Squid Web Proxy to provide high-performance caching in front of the DAP server. This solves the performance issue with regard to repetitive access, but creates the problem for tuning the cache to insure consistency with the WCS on the back-end.

2: An integral aspect of the gateway is to provide a large number of parameters each corresponding to the CEOP reference sites, covering the EOP time periods. For AIRS alone this translated into close to one million discrete virtual datasets. Both the DAP and WCS interfaces provide techniques to facilitate representing these large collections as singular objects, both provide data models that support multiple independent dimensions that can be used to provide an aggregate view for a large data collection. While both the DAP and WCS support this usage not all implementations, in either community, employ them. Using the DAP the approach would be to provide an aggregated view for each parameter by reference site, with the aggregation represented using the time dimension since the spatial extent per reference site is the same. This approach would still require hundreds of aggregated datasets, but several orders of magnitude smaller than a single dataset per parameter, per site, per time step. However, providing a single view over this large data collection would require the capacity to support extremely compute-intensive operations, and it was decided that providing an aggregated view could not be supported currently.

3: The DAP server (i.e., hyrax) by default can provide a filesystem-like directory response, listing the data sources available. To support the gateway we developed a filesystem-like representation to describe these virtual

collections, where the directory hierarchy followed from 'reference site' to 'parameter', and at each of the lowest levels the datasets were delineated by the 'date' that the dataset presented of the parameter observation. For AIRS, this extrapolated to approximately one-million discrete virtual files, or datasets. Initially, the WCS provided a THREDDS catalog representation of this virtual collection. We implemented a new THREDDS handler within Hyrax (.i.e., OLFS dispatch\_handler), using the THREDDS API provided by Unidata as the mechanism to provide this filesystem-like directory view. We experienced poor performance using this approach, and as the overall collection grew in size we completely outgrew the memory capacity of the THREDDS API to support this. Our subsequent approach was to query the WCS via the GetCapabilities and DescribeCoverage request/responses, which required implementing another OLFS dispatch\_handler that interacts with WCS. Using the WCS responses for GetCapabilities and DescribeCoverage this dispatch\_handler will generate the filesystem-like interface, on demand, as the user traverses the directory hierarchy. This handler queries the remote WCS at server startup and caches the responses from the remote WCS, using the cached information when performing on-demand directory responses. The initial development was performed using the WCS implemented at GDAAC for the AIRS collection. This WCS provides both spatial and temporal domain extents, consistent with the hierarchical representation we desired. The dispatch\_handler we implemented is general in the sense that it can make these requests to any WCS instance, but we found that how WCS express their coverages can vary widely. So widely, that we were unable to implement a general solution in the dispatch\_handler that we thought would be useful for the end-user. Additionally, we found that many of the XML responses returned would not validate against OGC schemas, and that was active discussion within the WCS developer community concerning the different revisions of the interface and schemas. At the following URL: [http://docs.opendap.org/index.php/IOOS\\_Gateway](http://docs.opendap.org/index.php/IOOS_Gateway) is a discussion of issues encountered, as well as solutions proposed for various aspects of our work on this project.

## Data Provenance

In the third year of the project, inclusion of data provenance information was prototyped. This was identified as an issue when the team demonstrated OPeNDAP access to a DataFed WCS server, which was serving GOCART aerosol model output obtained from Goddard Space Flight Center. Because the data had passed through so many servers, its origin was effectively obscured. In addition, it was difficult to tell if the data at the client end had been modified along the way. Accordingly, we developed a prototype that used included information about the original data files and the processing done to them in the WCS response from the server. These were encoded in a NetCDF ProcessingHistory attribute, using a structure that mimicked the ISO 19115 metadata standard. In addition, we developed demonstration scripts to be run from GrADS that would access the attribute and parse it to obtain the URLs of the original data files. A detailed document was written describing this: **CEOS WGISS: Provenance within Data Interoperability Standards**, including the demonstration scripts.

## Interoperability between OGC/CSW and WCS

The second half of the project dealt with providing search capabilities for data in THREDDS Data Servers, using the OGC Catalog Services for the Web. The ideal use case is a user with a Web Coverage Service client, searching for specific Web Coverage Service parameters. The client should be able to both search via CSW and use the responses to acquire WCS coverages for display or analysis in the client.

The team successfully developed a system that ingests metadata from a THREDDS Data Server catalog into a relational database. That database is served by a server that responds to OGC CSW requests following the ISO 19115 profile. (CSW servers can support a number of different profiles, such as EBrim.) The CSW server supports a hierarchical style of drill down to specific data servers and files, with automatic on-demand ingest of metadata at the lower levels. A test web client was also developed to demonstrate the server capabilities.

However, locating a client supporting both CSW (ISO 19115) and WCS proved more difficult. Discussions were conducted with both ESRI and the University of Florence (developers of GI-GO). At the close of the project, it appears that GI-GO is close to achieving this: it can display Web Coverages and it can search via CSW. The connection has not yet been made between the two protocols as of this writing, but we are hopeful that it will be made in the near future. Even in this case, however, a subtle mismatch between the CSW and the WCS protocols

was exposed in this case. Because the data are inventoried at a file level, they can correspond to a number of coverages, some of which may not match the CSW search criteria. As a result, the client has to sift through the results (essentially a second search) in order to present only the coverages "of interest". The lessons from this half of the project have been described in more detail in the document **CEOS WGISS: Interoperability between OGC CS/W and WCS Protocols**.

## Where To From Here?

The project was originally conceived as a technology development project. In each case, the proposed technologies were developed successfully, though the implementations exposed further work needed in interoperability to achieve results that would be scientifically usable and sustainable. The resulting lessons learned were documented for interoperability groups such as the CEOS WGISS, with the hope that they will be taken onboard for further development in the field of interoperability. In addition, an enduring product of the project was the WCS handler for OPeNDAP, which has been released as part of the OPeNDAP baseline.

The long-term prospects of the respective servers is more problematic. They will be kept running on a best effort basis, or as they are rolled into related work. On the WCS side, the GES DISC is looking at the prospect of keeping the CEOP WCS server running in order to support further experimentation with provenance, particularly through inclusion into an instance of Giovanni, which also is addressing data provenance issues.

An unexpected insight was the potential for gateway servers to provide value-added service to data stored at a remote location. This expedient could allow a site with significant OPeNDAP expertise to enable OPeNDAP access to data from providers (such as science P.I.s) that are unable or reluctant to deploy their own OPeNDAP server. This could make data sets available to a wider variety of clients at fairly modest cost.

## Output Products

### New Technology

- CEOP Satellite Data Server (WCS Handler for OPeNDAP), submitted as a New Technology Report GSC-15475-1

### Technical Notes

- CEOS WGISS: Interoperability between OGC CS/W and WCS Protocols
- CEOS WGISS: Provenance within Data Interoperability Standards
- CEOS WGISS: Lessons Learned From WCS Server Design And Implementation

### Abstracts / Talks

Yang, W., Min, M., Bai, Y., Lynnes, C., Holloway, D., Enloe, Y. and L. Di, 2008. Operational Interoperable Web Coverage Service for Earth Observing Satellite Data: Issues and Lessons Learned, AGU Fall Meeting, San Francisco, 2008.

Hu, C., Di, L., Yang, W., Lynnes, C., Domenico, B., Rutledge, G., and Y. Enloe, 2007. Interoperability Between Geoscience And Geospatial Catalog Protocols, AGU Fall Meeting, San Francisco, 2007.

Holloway, D., Yang, W., Enloe, Y., and C. Lynnes, 2007. Bridging OGC and Earth Science-based Data Access Protocols, AGU Fall Meeting, San Francisco, 2007.

Min, M., McDonald, K., Yang, W., Di, L., Enloe, Y., and D. Holloway, 2007. Extending OGC data services for the CEOP science community, Geoscience and Remote Sensing Symposium, 2007. IGARSS 2007.

McDonald, K., Enloe, Y., Di, L., Holloway, D., 2006. A gateway to support interoperability of OPeNDAP and OGC protocols, Geoscience and Remote Sensing Symposium, 2006. IGARSS 2006.

## **Acknowledgements**

We gratefully acknowledge the NASA ACCESS program for the funding that made this work possible. We thank Michael Bosilovich of NASA/GSFC for his crucial insights on user access needs for the CEOP Satellite Data Server, as well as for extensive tirekicking and feedback. We also thank Kenji Taniguchi of U. Tokyo and Beate Geyer of GKSS for testing the server and providing valuable feedback. For the CSW THREDDS work, we thank Stefano Nativi, Enrico Boldrini and Lorenzo Bigagli of the Earth and Space Science Informatics Laboratory for working with us to make their GI-GO client work with our CSW server.