COL/NW/03

# Integrating Web Services:
# Bi-directional Coupling of OPeNDAP (DODS) and LAS

FY 2003 Proposal to the NOAA HPCC Program

August 19, 2002

| Title Page | Proposed Project | Budget Page |


| | |
|---|---|
| Principal Investigator: | **Steve Hankin** |
| Line Organization: | OAR |
| Routing Code: | R/E/PM |
| Address: | NOAA/PMEL<br>7600 Sand Point Way NE<br>Seattle, WA 98115 |
| Phone: | (206) 526-6080 |
| Fax: | (206) 526-6744 |
| E-mail Address: | mailto:hankin@pmel.noaa.gov |

Jonathan Callahan
callahan@pmel.noaa.gov


| | |
|---|---|
| Proposal Theme: | **Collaboration, Visualization and Analysis** |
| Funding Summary: | FY 2003 $ 75,000 |


Steven C. Hankin
PI/Program Manager
PMEL

Cynthia Loitsch
Program Support Officer
PMEL

Dennis More
Division Leader
PMEL/OCRD

Eddie N. Bernard
Director
PMEL

# Integrating Web Services:
# Bi-directional Coupling of OPeNDAP (DODS) and LAS

Proposal for FY 2003 HPCC Funding

Prepared by: Steve Hankin

## Executive Summary:

**Objective: Improve web services integration (machine-to-machine interoperability) by providing an OPeNDAP protocol interface to the Live Access Server (LAS).**

Web services (machine-to-machine interoperability) are a major theme in modern, Internet software development. Coupling functionality available through one service with different functionality available through other, potentially remote services creates a larger system that is distributed, flexible, scalable and robust. This proposal will focus on completing machine-to-machine interoperablity between two Web services that are of significance to NOAA: OPeNDAP (a.k.a. DODS) for binary data access and LAS for access to data products. The result will be improved access to distributed NOAA data including: 1) direct (OPeNDAP) access to metadata enhanced data served by LAS; 2) a new and very general framework for making local data sets available in the OPeNDAP framework; and 3) much more robust 'data fusion' in LAS.

It is acknowledged that NOAA's data holdings need to be made available to the widest possible audience – from scientists to policy makers to educators. However, much NOAA data, although technically accessible, is still unusable outside of specific scientific communities because of the tremendous variety in data formats and the inscrutable nature of the attendant metadata: unintelligible variable names, missing units, lack of geo-referencing, etc.

Since 1998, the HPCC funded LAS has been putting a friendlier face on NOAA data with web based user interfaces that allow data visualization and download. A typical LAS installation will contain a great deal of metadata beyond that provided with the raw data. Additionally, LAS can harness multiple back end software packages (Ferret, GRADS, IDL, etc.) to provide access to data from a wide variety of formats as well as derived (calculated) variables.

This proposal will augment existing LAS capabilities by providing OPeNDAP access to LAS data holdings. The resulting blended web service will give OPeNDAP clients (IDL, Matlab, Ferret, GRADS , ncbrowse etc.) access to: 1) richer and more standardized metadata utilizing the LAS metadata database; 2) many existing (and future) types of data – Matlab files via Matlab, GRIB files via GrADS, RDBMS data via perl code, netCDF, flat binary, and ASCII via Ferret, etc.; and 3) derived variables provided by on-the-fly analysis and regridding in the back end packages.

As LAS is itself an OPeNDAP client, the resulting system will also produce a more robust 'data fusion' environment. The ability to perform on-the-fly data fusion requires machine-to-machine binary data access that hides messy details taken care of in LAS by its back-end packages. This proposal uses OPeNDAP networking to create an abstraction layer which hides this machinery. Data fusion thus becomes possible between any two LAS datasets, regardless of the details of their generation.

## Problem Statement:

Over the past few years, the OPeNDAP framework has proved exceptionally useful as a system for accessing remote data. Users with OPeNDAP enabled client software (Matlab, IDL, Ferret, GRADS, ncbrowse, etc.) can use Gigabytes of remote data just as if it were on a local disk. The OPeNDAP approach has been from the bottom up in the sense that the strictest requirements are placed upon the lowest levels of data transport – the structured binary data (arrays, grids, sequences, …) To minimize entry barriers, OPeNDAP deliberately minimizes semantic metadata requirements such as variable or axis names, units or geo-referencing

Other systems (NOAAServer, GCMD, etc.) have tried to approach data access from the top down, focusing on 'search' metadata as is typically found in metadata documents adhering to the FGDC, GILS or DIF standards. These systems contain information about the 'who', 'what', 'why', 'when' and 'where' of data. However, through the top down emphasis these systems have omitted the functionality needed to ensure inter-operable transport of binary data.

A consequence of the OPeNDAP approach is that the end user is often in the position of successfully accessing raw data but needing to find further metadata to make full scientific sense of it. The intrepid user will end up with usable data access but only after some sleuthing, guesswork and trial and error. For data fusion systems that provide comparisons between data at two different sites the problem is further complicated. Data differencing in these systems will be done not by the end user, but by a machine implying much stricter metadata standards than are needed by the end user.

**The fundamental goal of this proposal is to provide self-describing, binary access to data that is stored in a variety of formats that are already addressed by LAS.** The HPCC funded Live Access Server (LAS) has been a popular tool for providing graphical browse and download of formatted subsets of data. Through the LAS framework installers are able to augment or modify metadata associated with a dataset, initialize non-standard data and define derived variables. The improved data is available for visualization and download but, in the current form of LAS, not for binary access. This proposal will address this deficit, enabling 'direct' access to any dataset provided through LAS augmented by the metadata enhancements that are configured into LAS.

Relationship to HPCC objectives:

This proposal directly addresses the primary goal of NOAA HPCC: "to provide greater access to its vast holdings of real-time and historical information to users in a more complete, more usable form".

## Proposed Solution:

We propose to create a LAS DODS Server (LDS) that will combine the functionality provided by LAS with the structured binary access to data provided by OPeNDAP. This is a strategy that has already been successfully demonstrated by the software developers at COLA (Center for Ocean-Atmosphere Land Studies) who developed the GrADS DODS Server (GDS). The LDS will have a larger scope as it handles multiple back-end packages, rather than GrADS, alone.

Typically, the Live Access Server acts as an OPeNDAP client, requesting subsets of remote binary data from OPeNDAP data servers from which it produces graphics and formatted files. The communication between the client and server uses the highly abstracted and generalized OPeNDAP data model, which allows a very wide variety of data to be served. Various OPeNDAP servers have been developed to serve particular data formats, such as netCDF, HDF, JGOFS, and others.

To recast LAS as an OPeNDAP server we will need to write custom code to translate between the OPeNDAP protocol for structured binary data and the LAS protocol for data products. Along with the protocol translation, LAS will add its internal 'use' metadata to the OPeNDAP representation of the data and will translate from the data models understood by the LAS "back-ends" to the data model expected by OPeNDAP. The resulting LAS-OPeNDAP Server will appear to OPeNDAP clients as a standards-conforming OPeNDAP server.

Technology

The OPeNDAP libraries are available as Java code and a large portion of LAS is also written in Java. The resulting LAS OPeNDAP server will be implemented as a Java servlet (Java code running within the context of a Web server) connecting these two pieces of pre-existing software. The OPeNDAP libraries will provide the protocol by which information is exchanged with the client across the internet while LAS will handle the actual data operations including metadata enhancement, subsetting and on-the-fly analysis. The Jakarta Project's Tomcat package will be used to provide an execution environment for the servlet.

## Analysis:

As systems that provide web based data and metadata access mature, many are being converted to web services – machine accessible repositories of functionality rather than just data. LAS-relevant examples include NASA's Global Change Master Directory (GCMD) which provides search and metadata content services; the Open GIS Web Coverage Service (WCS); the Unidata/THREDDS real-time cataloging services; the GrADS-DODS Server which provides reformatting and analysis services; and LAS which acts as a front end for a variety of services provided by Ferret and other back end packages. To bind these various data and metadata services more tightly together a Network Data Access Protocol is needed. This is precisely what is provided by OPeNDAP.

Both OPeNDAP and LAS are already well established projects – many projects both inside and outside of NOAA depend upon them for data dissemination. Currently, LAS is a client for remote data provided by OPeNDAP servers, thus allowing LAS sites to put their local data in the context of remote data ranging from huge reference dataset to data provided by investigators at other sites. Still missing is the converse relationship: letting OPeNDAP clients request data from LAS servers.

We would like to note that our solution does not require the development of a new data model – something that takes years to develop and see through to maturity. Both OPeNDAP and LAS already have well defined and proven data models. Instead, the protocol translation essentially creates an abstraction layer which hides any machinery involved in connecting OPeNDAP clients to LAS servers.

Scope

The need for accurate data access with 'consistent and complete' metadata is universal and LAS is among the most popular packages for providing this to users of NOAA data. The LAS-OPeNDAP Server will improve access to 'use' metadata and also enable binary access to any of the derived variables created on-the-fly at a particular LAS installation. A growing network of LAS servers will also be able to perform data fusion by making OPeNDAP requests of each other. The final deliverable from this proposal will thus have a very wide scope both within NOAA and beyond.

Leverage

This proposal uses preexisting software that represents many years of development and funding from many different agencies. The Live Access Server and Ferret have been funded at various times by HPCC, ESDIM, OGP, NASA, NSF, and ONR while OPeNDAP/DODS has received funding from NOAA, NASA, NSF, and most recently NOPP. Each of these projects is receiving increasing attention from providers of data as a solution to the needs of data users. HPCC funding of this proposal will benefit both from the work that has gone before but also from the established base of LAS sites.

Cost/Benefit

The cost/benefit ratio achieved by adding functionality to established, popular software systems is extremely favorable. The problems being tackled by this proposal include creating cross-linkages between Web services protocols and improving metadata associated with remote data access. The metadata problem has been considered intractable by many. We believe the technical solutions presented here will efficiently harness the human energy already being directed at this problem at individual institutions. Importantly, our solution will not require any additional work on the part of any data provider; it will merely enable better use of work that is already being done by them.

Alternatives

We are not aware of any remote data access framework that provides the functionality, simplicity of use, and support available with OPeNDAP.

## Performance Measures:

### Milestones

Month 02 – milestone 1: architecture document describing implementation strategy
Month 08 – milestone 2: operational prototype (beginning of testing period)
Month 12 – milestone 3: document and add to standard release of LAS

### Deliverables

1 – publicly available version of LAS which functions as an OPeNDAP protocol server (Alternatively, this can be thought of as an OPeNDAP server with the full flexibility and functionality of LAS behind it – a novel application of this networking protocol.)