COL/NW/02

Discovering and Fusing Distributed LAS Data Sources using Unidata/THREDDS and Web Services

FY 2004 Proposal to the NOAA HPCC Program

August 19, 2003

| Title Page | Proposed Project | Budget Page |

Principal Investigator: **Steve Hankin**

Line Organization: OAR
Routing Code: R/E/PM
Address:

NOAA/PMEL
7600 Sand Point Way NE
Seattle, WA 98115

Phone: (206) 526-6080
Fax: (206) 526-6744
E-mail Address: Steven.C.Hankin@noaa.gov

Jonathan Callahan
Jonathan.S.Callahan@noaa.gov

Proposal Theme: **Collaboration, Visualization and Analysis**

Steven C. Hankin PI/Program Manager PMEL	Cynthia L. Loitsch Program Support Officer PMEL
Dennis W. Moore Division Leader	Eddie N. Bernard Director

PMEL/OCRD PMEL

Discovering and Fusing Distributed LAS Data Sources using Unidata/THREDDS and Web Services

Proposal for FY 2004 HPCC Funding

Prepared by: Steve Hankin and Jon Callahan

Executive Summary:

Objective: This proposal utilizes XML-based Unidata/THREDDS catalogs to allow independent LAS sites to "fuse" with one another – creating a powerful data intercomparison framework.

Approximately 50 independent Live Access Server (LAS) sites representing on the order of a terabyte of data are in operation at the time of this writing. We propose to utilize Thematic Realtime Environmental Data Distributed Services (THREDDS) catalogs to tie independent LAS servers into a scalable network that provides web services for data discovery, data browse, data access and data fusion. Individual LAS sites will "publish" automatically-generated THREDDS catalogs to advertise the data contents and web services available at each site. End users (including LAS sysadmins) will be able to discover these 'preconfigured' data resources and related services. LAS will be enhanced so data sources located in this manner will be immediately available for fusion into other LAS sites. The labor-intensive manual configuration step typically required before data comparison becomes possible is achieved in a scalable manner through the aggregate efforts of the LAS sysadmins at each individual site.

The threefold impacts of this proposal include: 1) greatly simplifying the effort to tightly integrate ("fuse") data from the rapidly expanding list of LAS servers installed within NOAA, NASA, DOE, Navy and the academic and international research communities; 2) demonstrating the power of the XML-based Unidata/THREDDS catalog (see associated **letter of support** from Unidata's program management); and 3) contributing NOAA technology to important efforts such as the U.S. Integrated Ocean Observing System (IOOS), and the Global Ocean Observing System (GOOS).

Problem Statement:

Data integration is the current 'holy grail' of scientific data management, notably in fields that are of central concern to NOAA such as meteorology, oceanography, and climatology. Numerous organizations have been born whose stated purpose is the integration of data streams from wide ranging sources. High profile examples include the U.S. Integrated Ocean Observing System (IOOS) for which LAS is a recognized "pre-operational" component, and the Global Ocean Observing System (GOOS) of which the Global Ocean Data Assimilation Experiment (GODAE) is a pilot project using LAS to perform Web-based data comparison. IOOS and GOOS bring together international, Federal, state, regional, municipal, academic and commercial data providers. The data served are geographical (e.g. coastal land use), economic, biological, physical and chemical across time scales from real-time to climatological. The integrated data are relevant to research, education, ecosystem management, public health, recreation and national security.

The data management systems and data interchange formats in use within this community are as heterogeneous as the community membership. They are unlikely to become significantly more uniform in the near future. If data integration across such a heterogeneous set of data providers is to succeed four things must happen: 1) web services (or other middleware) must exist and be installed that allow data in different formats with different metadata standards to be retrieved with one (or a small number of) predictable syntax; 2) analysis and visualization software must exist that can usefully fuse (difference, co-plot, etc.) data provided by these web services; 3) a discovery mechanism must exist that allows users to find these data sources; and 4) the data sets must be configured to conform to a uniform and comprehensive semantic (geo-spatial) data content model. No such uniform semantic standard exists today, so data providers independently utilize many incompatible (so-called) data standards (e.g. WMO: GRIB, BUFR; NetCDF: CF, Argo, WOCE; HDF: EOS, v4, v5, RDBMS: many, GIS: many). As of today a manual configuration step that harmonizes these conflicting standards is required to achieve even minimally consistent 'use metadata' (coordinates, units, etc.) for purposes of data integration.

The goal of this proposal is to provide easy publication and discovery of preconfigured LAS datasets through the use of Thematic Realtime Environmental Data Distributed Services (THREDDS) catalogs (see further discussion under Proposed Solutions). THREDDS catalogs will allow fusion (e.g. differencing) of datasets hosted on independent LAS servers. This development, if successful, will immediately harness the energies of the many institutions that have adopted LAS as a web portal to important datasets. All of these data sets will be accessible through a single web service sharing uniform semantic metadata standards.

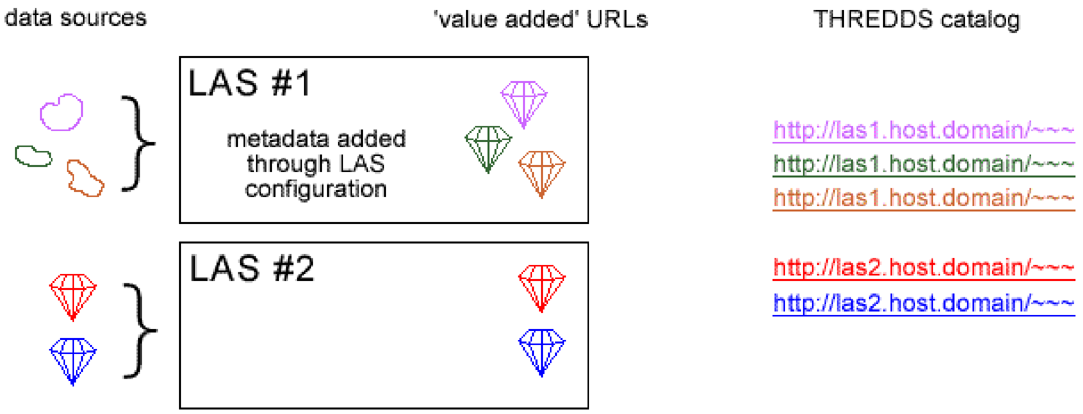
Relationship to HPCC objectives:
This proposal directly addresses the primary goal of NOAA HPCC: "to provide greater access to its vast holdings of real-time and historical information to users in a more complete, more usable form". This project is also a direct application of digital library research as THREDDS is funded as part of the National Science, Mathematics, Engineering, and Technology Education Digital Library (NSDL) program.

Proposed Solution:

Over the past several years HPCC has funded proposals from our group and to address the needs described above in requirements 1 (web service interface), 2 (data fusion) and 4 (data set configuration). The success of these efforts can be seen in the widespread adoption of LAS. For requirement 3 (data discovery), however, LAS currently lacks a systematic solution. And for requirement 4 (data set configuration) we are seeking approaches that are more highly automated and scalable. An innovative and high visibility solution to both problems in the NOAA science community is the Thematic Realtime Environmental Data Distributed Services (THREDDS) effort that is lead by UCAR/Unidata and funded by NSF (http://my.unidata.ucar.edu/content/software/thredds/). THREDDS envisions a linked web of "catalogs" (pointers to data sets, web services and other catalogs) that are encoded in XML. Client applications navigating this distributed catalog can discover the appropriate protocol to use to access data and metadata.

The heart of THREDDS consists of high level metadata contained in XML based Publishable Inventories and Catalogs (PICats). This metadata does not include the 'use metadata' (syntactic and semantic description of the data) needed for data fusion but instead provides links to web services that can provide this information. It provides what programmers call a 'layer of abstraction', so that client software interacting with the THREDDS catalog can communicate via published interfaces to access both the use metadata and the data, itself.

Through the THREDDS 'web of catalogs' users and software components will have access to both LAS configuration metadata and 'value added' services (e.g. graphics). Using this catalog new LAS user interfaces can (in principle) be created on-demand with guaranteed levels of access to independently managed datasets. This guaranteed level of access is required before we can successfully fuse (intercompare) data. The following graphic depicts a three tiered view of our solution (left to right): 1) source datasets in multiple, incompatible standards; 2) the same datasets as configured in LAS servers; and 3) THREDDS catalogs listing all available datasets and services.


data sources
'value added' URLs
THREDDS catalog
LAS #1
metadata added through LAS configuration
LAS #2
http://las1.host.domain/~~~
http://las1.host.domain/~~~
http://las1.host.domain/~~~
http://las2.host.domain/~~~
http://las2.host.domain/~~~

It is not LAS sites alone that will benefit from the proposed LAS-THREDDS combination. Indeed, the vision of the Web Service approach is to enable widely varying applications to harness remote data and computational power in unanticipated ways. We predict that the uniform access to data that is provided through these catalogs will also be of use to search and visualization software being developed by other groups who are collaborating with THREDDS. This vision is shared by the THREDDS project leadership as demonstrated in the **letter of support** (attached) for this proposal from Unidata, submitted separately through our HPCC representative.

Implementation Plan

LAS will be enhanced to add a web service interface that provides THREDDS catalogs as output upon request. We will maintain a THREDDS catalog at PMEL (a catalog of catalogs in the spirit of THREDDS) containing linkages to the individual LAS THREDDS catalogs. This site will act as a clearinghouse for information on distributed LAS data holdings and associated services. Both end users and LAS installers will make use of the THREDDS catalogs to discover sources of preconfigured data. For LAS installers, we will provide tools that automatically configure a new dataset in LAS, given the THREDDS catalog entry of that dataset. Finally, we will provide a demonstration Web site where datasets from the THREDDS catalogs can be configured on-the-fly into an LAS interface.

The capabilities provided by this work are a natural consequence of past and current LAS developments. Our FY 2001 proposal, "Distributed Collaborative User Interface Components …", funded the ability of LAS user interface servers to provide access to multiple LAS product servers. Our FY 2002 proposal, "From Web Servers to Web Services …", formalized the interface to LAS product servers as an XML based web service. We are already moving forward on the FY03 proposal "Bi-Directional Coupling of OPeNDAP and LAS" to support the OPeNDAP web service interface to pre-configured data at the "binary" level through LAS.

Taken together, these features of LAS will allow for the complete integration of remote datasets. LAS sites will become 1) "aware" of data at independent remote LAS product servers; 2) able to send requests for data products to the remote server; 3) able to send requests for regridded subsets of data to the OPeNDAP web service on the remote server in order to perform data comparison operations (fusion). All the data coming to the fusion LAS site will be guaranteed to be semantically consistent and available for access and data intercomparison. The current proposal greatly multiplies the success of this work by enabling discovery of LAS data and services and simplifying the inclusion of 'discovered' data into existing and novel LAS installations.

Analysis:

Scope

The need for accurate data access with consistent and complete semantics ("use metadata") is universal. LAS is among the most popular packages within NOAA for providing this to users of data. As the number of LAS installations has grown it is becoming increasingly important to provide a discovery path through which many LAS sites may be searched as a whole. The deliverables from this proposal will have a scope that includes all LAS sites already in existence -- in NOAA line offices, other Federal agencies, the DOE (Globus) "Grid", academia and internationally – as well as anticipated new sites to be installed by widely dispersed data providers within the US Integrated Ocean Observing System (IOOS). This proposal will, in fact, become part of the NOAA contribution to the US IOOS. The web services that make this integration possible will be published and available to other web enabled applications.

Leverage

This proposal leverages the use of 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 THREDDS has received funding from NSF at UCAR/Unidata. Each of these projects is receiving increasing attention from providers of data as a solution to the needs of their data users. This proposal will leverage funding from ONR, NOPP, NASA, and multiple NOAA sources that is anticipated to be of Order($500K) for FY04. The work funded by these sources will broaden the scope of data types handled well by LAS to include curvilinear models, satellite swaths, realtime model outputs and large *in-situ* data collections. HPCC funding of this proposal will benefit both from the work of the LAS core development group at PMEL and from the work of individuals within the LAS community at many established LAS sites.

Alternatives

One of the key components in the success of Internet based data sharing is the adoption of interoperable software packages developed by other institutions. Whenever possible, one should look to invest energies in existing or emerging systems that have built some momentum behind them. The THREDDS project satisfies this requirement well. THREDDS has a higher level of generality and abstraction than do alternative metadata catalogs like NOAAServer, NASA's Global Change Master Directory (GCMD) and the NASA ESIP Federation's Mercury project. Rather than requiring that all participants utilize a single, uniform 'middleware' solution the THREDDS approach knits multiple solutions together through a very flexible approach to the emerging "web services".

Cost/Benefit

The cost/benefit ratio achieved by adding functionality to established, popular software systems such as LAS is extremely favorable. This proposal will apply cutting edge technology that is being implemented in the scientific and educational communities to improve the ease and robustness of access to remote data. The problem of inadequate semantic and syntactic 'use metadata' has been a significant barrier to interoperability for many years. We believe the technical solutions presented here will efficiently harness the human energy already being directed at this problem by LAS sysadmins at many separate institutions. Importantly, our solution will not require any additional work on the part of data providers; it will merely enable better use of work that is already being done by them.

Performance Measures:

Milestones

Month 03 – milestone 1: LAS web service to provide THREDDS catalog.
Month 05 – milestone 2: LAS with THREDDS service installed operationally at PMEL
Month 09 – milestone 3: LAS web service to provide packaged configuration metadata.
Month 11 – milestone 4: Installer tools that utilize linkages gleaned from the THREDDS catalog to automatically configure fusion-ready LAS datasets.

Deliverables

1 – Publicly available version of LAS with web services that produce THREDDS catalogs and respond to requests for LAS configuration information.
2 – Publicly accessible THREDDS master catalog of LAS datasets and web services.
3 – Web site through which data sets hosted by independent LAS servers can by configured on-the-fly into a new LAS server for purposes of data fusion.