9.3

THematic Real-time Environmental Distributed Data Services (THREDDS)

Ben Domenico[*]

Unidata Program Center
University Corporation for Atmospheric Research
Boulder, CO 80307

1. ABSTRACT

Last summer, Unidata received word that its THREDDS (THematic Real-time Environmental Distributed Data Services) initiative would become part of the NSDL (National Science, Math, Engineering, Technology Education) sponsored by the National Science Foundation's Division of Undergraduate Education.

The overarching goal of THREDDS is to provide students, educators, and researchers with coherent access to a large collection of real-time and archived datasets from a variety of environmental data sources at a number of distributed server sites. The datasets will be conveniently accessible from a collection of THREDDS-enabled data analysis and display tools. These range from "thin" web-based clients that allow the learner to browse and manipulate data using the processing power on the servers, to "thick" clients that harness the computing power and flexibility of the user's own workstation while accessing data from a collection of remote servers. THREDDS will provide real-time data delivery via reliable, event-driven "push" technology as well as transparent access to datasets using "pull" systems that make it possible to access data on remote servers as if they were on the users' own computers. The system will be built on a set of software applications and data servers, most of which either are already in operation or are under development.

At the heart of THREDDS is metadata contained in publishable inventories and catalogs (PICats). Based on the eXtensible Markup Language (XML), PICats can be created in many different ways. Crawlers will be implemented to create PICats by traversing existing retrospective data collections. Sites receiving real-time environmental data will instrument decoders to create PICats describing data products as they arrive. Since PICats do not have to reside on the data servers, researchers will be able to create PICats for research publications that point to datasets residing on several data servers. Educators will incorporate PICats of illustrative datasets into educational modules that also include tools for data analysis and visualization. Just as they now use HTML with URLs to point to relevant documents, students will eventually be able to use PICats to point to datasets related to their research projects. Since PICats are text-based, they can be "harvested" and indexed in digital libraries using specialized tools that use the internal structure and semantic content as well as by tools similar to those used by existing document search engines.

2. COLLABORATIONS

THREDDS is a highly collaborative project. The community consists of three main groups: a set of data provider sites; a group of software developers working on systems for data analysis and display; and set of experts in metadata relating to Earth systems data collections.

2.1 *DATA PROVIDERS*

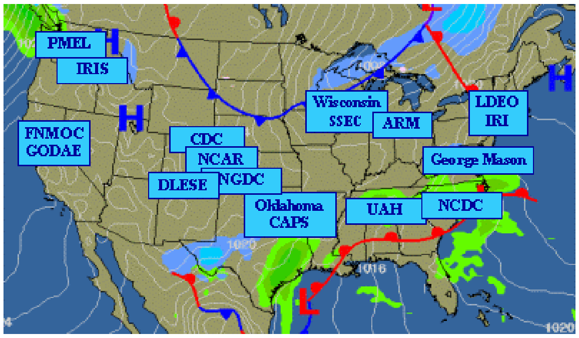


Figure 1. Data provider sites

The following institutions have agreed to be data provider partners (themes and contacts noted parenthetically):

- NCDC, the National Climatic Data Center (climate, Ben Watkins); NOMADS (NOAA Operational Model Archive and Distribution System, a unified climate and weather data archive, Glenn Rutledge).
- NGDC, National Geophysical Data Center (geophysical, Ted Habermann);
- SSEC, the Space Science and Engineering Center at the University of Wisconsin-Madison (GOES satellite data, Steve Ackerman and Tom Whittaker);
- IRI/LDEO, International Research Institute/Lamont Doherty Earth Observatory (climate, oceanographic, Benno Blumenthal);
- PMEL, the Pacific Marine Environment Laboratory (oceanographic, marine, Steve Hankin);
- NCAR, the National Center for Atmospheric Research (atmospheric, oceanographic, Don Middleton);
- CDC, the Climate Diagnostic Center (climate, Roland Schweitzer);
- FNMOC, Fleet Numerical Meteorological and Oceanographic Center (oceanographic, Dave Dimitriou);

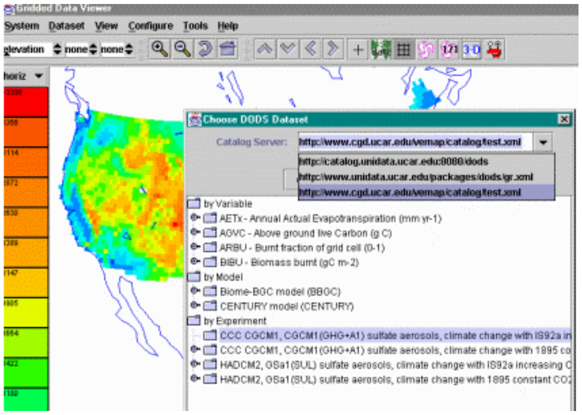
[*] *Corresponding author address:* Ben Domenico; Unidata Program Center; University Corporation for Atmospheric Research; P.O. Box 3000; Boulder, CO 80307; ben@unidata.ucar.edu.

- GMU/COLA, George Mason University/Center for Oceans Land Atmosphere (hydrologic, Menas Kafatos and Ruixin Yang);
- University of Alabama Huntsville (satellite and hydrology, Sara Graves and Rahul Ramachandran); and
- The ADDE servers in the Unidata community (real-time atmospheric data, Tom Yoksas).
- IRIS DMC, Incorporated Research Institutes for Seismology Data Management Center (seismic, Tim Ahern);
- University of Oklahoma (radar, Kelvin Droegemeier)
- ARM (Atmospheric Radiation Measurement, Chris Klaus); and
- University of Florence (European satellite data, Stefano Nativi).

The data being offered range from climate and weather data to oceanographic, marine, and satellite data. These collections are already in place but there is no methodology for a unified search across all the independent servers. As partners in the THREDDS NSDL collections initiative, these data providers have agreed to:

- Supply the hardware and system administration required to run the THREDDS servers;
- Provide access to their characteristic datasets via DODS and/or ADDE client/server protocols in addition to more traditional methods (e.g., FTP, tapes);
- Provide browser/server access to certain datasets via Web-based thin clients such as PMEL's Live Access Server (LAS);
- Using LDM/IDD technology, make real-time data available on the server where appropriate; and
- Work with Unidata to incorporate systems for expanded metadata to make it easier for users to find datasets and to use them once found. This is the key component that will tie the server systems together, enable remote clients to find and access the data, and connect the servers with the DLESE discovery system.

2.2 *CLIENT ANALYSIS AND DISPLAY TOOLS*



Choose DODS Dataset

Catalog Server: http://www.cgd.ucar.edu/vemap/catalog/test.xml

http://catalog.unidata.ucar.edu:8080/dods
http://www.unidata.ucar.edu/packages/dods/gr.xml
http://www.cgd.ucar.edu/vemap/catalog/test.xml

by Variable
AETx - Annual Actual Evapotranspiration (mm yr-1)
AGVC - Above ground live Carbon (g C)
ARBU - Burnt fraction of grid cell (0-1)
BIBU - Biomass burnt (gC m-2)
by Model
Biome-BGC model (BBGC)
CENTURY model (CENTURY)
by Experiment
CCC CGCM1, CGCM1(GHG+A1) sulfate aerosols, climate change with IS92a in
CCC CGCM1, CGCM1(GHG+A1) sulfate aerosols, climate change with 1895 co
HADCM2, GSa1(SUL) sulfate aerosols, climate change with IS92a increasing C
HADCM2, GSa1(SUL) sulfate aerosols, climate change with 1895 constant CO2

Figure 2. Accessing the data discovery system from within a THREDDS analysis and display application.

The THREDDS prototype will provide examples of a wide variety of working applications that use our metadata framework to find, analyze, and display data from server sites. This will demonstrate an end-to-end system for data access and visualization. The following developers will incorporate our client-side data-access components (class libraries and metadata access) into their own data manipulation tools:

- **Live Access Server** (LAS, PMEL, Steve Hankin). LAS illustrates the use of a Web-based (thin) client with the bulk of the analysis and display generation done on the server side.
- **INGRID** (IRI/LDEO, Benno Blumenthal). This is another example of a system enabling analysis and display of data via a Web browser. As with LAS and GDS, INGRID provides substantial data-analysis capabilities.
- **WXWise applets** (the University of Wisconsin-Madison, Tom Whittaker). These applets illustrate the use of Java to embed data-analysis and display tools directly into educational modules on a Web site.
- **The Virtual Geophysical Exploration Environment** (VGEE, formerly The Virtual Exploratorium, the University of Illinois, West Chester State, DLESE, and NCAR; Don Middleton). This application incorporates the educational functions directly into the data analysis and display tool itself.
- **Data Discovery Toolkit and Foundry** based on **EDMI** (Earth Data Multimedia Instrument, New Media Studio, Bruce Caron). These are a set of data-analysis and display tools based on IDL and Macromedia Director. They can be used to generate very elaborate educational modules. This work is being funded by an NSDL "services" grant.
- **MetApps** (Unidata Program Center, Don Murray). A set of pure Java, platform-independent, two- and three-dimensional data-analysis and display tools—based on the VISAD infrastructure.
- **VISAD** infrastructure from SSEC (Bill Hibbard of the University of Wisconsin-Madison in conjunction with the Unidata Program Center).

2.3 *METADATA EXPERTISE*

As noted earlier, the technological core of this proposal, the crucial component under development, is a system for adding the semantic description of scientific datasets necessary for data manipulation and discovery. It must interoperate with data providers, data servers, data clients, catalog servers, discovery systems, and other middleware components. Investigators will select key scientific datasets and semantic descriptions developed for an end-to-end demonstration of the utility of this approach. THREDDS staff will work closely with DLESE to ensure that the resulting metadata system will interoperate effectively with NSDL.

Expert partners collaborating on matters of metadata and interoperability are:

- The University of Alabama-Huntsville on the Earth System Markup Language (ESML,);
- George Mason University on the DIstributed MEtadata System (DIMES);

- The University of Rhode Island on the DODS aggregation catalog server;
- DLESE;
- The University of Florence will act as a liaison with the international metadata standards community.

3. TECHNICAL APPROACH

The THREDDS approach builds on its community's strengths. It also provides a mechanism for extending the DLESE discovery system (Sumner et al., 2001) to embrace the metadata in the previously described PICats. To this end, we will build two essential new components: a formal definition for PICats and software to facilitate their use. PICats will be built using XML transported via HTTP (i.e., on the Web), and will refer to datasets that are usable via DODS, ADDE, or other direct-access methods. Needed software includes tools to create standards-compliant PICats, plug-ins, or server-side visualizers to enable the use of PICats in browsers and components that help developers incorporate PICats into applications.

A more detailed description of the THREDDS technological approach is contained in *THREDDS Technical Underpinnings* on the Unidata web site and in *THREDDS: A Geophysical Data/Metadata Framework*, a paper by Ethan Davis and John Caron in the Distributed Data and Metadata Access session at this conference.

4. DEFINING SUCCESS

To define success, we have to state up front what a completed THREDDS will provide – not only for the end users, but also for the many participants and, perhaps more importantly, those who will consider joining the effort once the prototype is complete.

Table 1 captures what THREDDS will mean to these communities in terms of the three primary functional categories:

- Creation and publication
- Discovery
- Access and interaction

Participant	Creation and Publication Functions	Discovery Functions	Access and Interaction Functions
End Users **Research Scientists** **Educators** **Learners**	Publish online research papers, educational materials, class assignment reports in online journals, educational web sites, courseware -- with references to datasets and tools for viewing and interacting with those datasets	Using central discovery sites, browsers, and data analysis applications, find datasets of interest (on THREDDS servers) associated with a particular time, place, subject topic, phenomenon, etc.	Having found datasets of interest, readily interact with them using THREDDS-enabled applications, applications associated with the datasets, server-based analysis tools, or browser-based thin clients.
Data Providers	Install and run automated THREDDS tools for generating PICats, entering PICats into central collections, or making the PICats available for harvesting by crawlers.	Generate hierarchy of catalogs for contents of local site and others that are related, so users of local site can readily find datasets that reside elsewhere.	LAS, INGRID, GDS sites provide interaction via browsers. All sites provide client/server access via THREDDS-enabled clients applications
Analysis and Display Applications Builders	Include hooks for initializing apps from within online publications.	Have search facilities built into client applications to find data of interest on THREDDS servers from within the applications.	Primary mode of interaction via both thin (web-based) clients and thick, full-blown local application clients.
Central Discovery Providers, e.g. DLESE	Facilitate integration of pointers to datasets and applications into publications	Facilities for harvesting and indexing information in PICats. Provide a programmatic interface for contributing PICat at time of creation.	Found documents point to datasets and interaction tools

Table: THREDDS Success from Participant Viewpoints

5. SUMMARY AND CONCLUSION

THREDDS is a very ambitious project. As a proposal reviewer pointed out, the risks are great but the potential payout is correspondingly dramatic. The collaborating team is strong and committed, so we look forward to achieving this vision within the two-year time frame of the grant.

6. REFERENCES

References to this paper are available online at: http://www.unidata.ucar.edu/projects/THREDDS/PublicationsAndPresentations/BensAMS/References.htm