5.7 THREDDS: Publishing, Cataloging, Describing, and Discovering Scientific Datasets

Ethan R. Davis*
John Caron
Ben Domenico
Robb Kambic
UCAR/Unidata, Boulder, CO

Stefano Nativi
University of Florence, Italy

1. Introduction

The THREDDS project has developed a simple, extensible framework to provide services for published datasets. THREDDS catalogs provide a structure for cataloging and describing already published datasets and communicating that information between data providers and data users. Existing metadata standards are used for describing datasets. The catalog structure has allowed us to develop services for search and discovery of the published datasets. The ability to add information to existing datasets allows human and machine to better understand and use datasets.

2. Cataloging and Adding Information

THREDDS catalogs provide a simple, hierarchical structure in which scientific data can be inventoried. At a minimum, the catalogs contain references to datasets and a human understandable name for that dataset. Additional information about a dataset or collection of datasets can be added to the catalog in various ways. Using this simple structure, we are developing several simple search systems and providing existing, more extensive, search and discovery systems with information.

Most of our prototype systems catalog data that is available online through data access protocols such as OPeNDAP (aka DODS) [8] and ADDE [9]. We are also cataloging data available through more familiar file access protocols like FTP and HTTP.

These catalogs can be used to detail the datasets available on one particular data server, a subset from a server, the set of datasets that satisfy a query on a large body of datasets, or the datasets found about a particular topic (e.g., datasets relevant to a published paper).

Many scientific datasets are not fully described in a self-contained way. This means that additional information is often required before the general user can more fully understand and make use of the data. THREDDS catalogs may contain or reference additional information, either plain text or more structured information. The type of metadata can be identified in the THREDDS catalog so that applications that understand certain standards can access the appropriate metadata record. Several THREDDS tools recognize a number of metadata standards, e.g., Dublin Core, FGDC, ISO, OpenGIS, and NcML [6].

*Corresponding author address: Ethan Davis, UCAR/Unidata, P.O. Box 3000, Boulder, CO 80307, email: edavis@ucar.edu.

3. Search and Discovery

We are working on various search and discovery techniques. The simplest uses a THREDDS catalog's hierarchical structure to navigate the catalog and find a dataset of interest. For instance, figure 1 shows a user interface for selecting a dataset from a THREDDS catalog hierarchy.

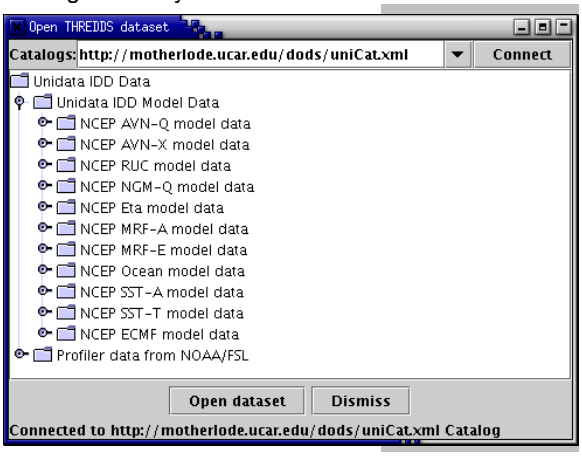

Figure 1: Example GUI Dataset Selector

However, as a user interface, a straight hierarchical browse does not scale well. Once you get beyond a hundred or so entries at each level, the hierarchical browse can become quite cumbersome. For large and/or volatile collections of datasets, it is convenient to allow clients to select subsets of the entire collection. The THREDDS DatasetQuery XML allows data servers to describe the queries clients can use to restrict the set of datasets returned. Often, very large datasets can be efficiently described by defining "product sets" of choices (i.e., choose one from column A and one from column B and so on). Dynamic data can be described using logical names ("most recent", "latest hour"), rather than explicitly listing each dataset. Based on user selections, a query is formed which the data server resolves to a list of specific datasets. With these simple

techniques, THREDDS has prototyped a Level 3 Radar server, which dynamically receives data through the LDM [5], describes it simply as a choice of station, product and logical time, and serves it via ADDE.

For more sophisticated search and discovery techniques, we are working with a variety of dataset cataloging and digital library efforts. The Global Change Master Directory (GCMD) [3] has a long history of cataloging scientific datasets. We are working with the GCMD to support ingesting dataset information into the GCMD system through the THREDDS framework. We are also working closely with both the NSDL [7] and DLESE [1] for cataloging datasets, services that access those datasets, and resources that reference those datasets and services.

4. THREDDS API and Tools in Java

The THREDDS project has developed a Java library that reads and writes THREDDS XML documents and provides an API to access and modify the information communicated through THREDDS XML documents. The library also contains a variety of tools for working with the THREDDS framework. The library includes:

1) a GUI dataset selector widget (see Figure 1);

2) a validation tool (also available at http://motherlode.ucar.edu:8080/validation/CatalogValidation.html);

3) a web interface to the THREDDS catalogs which uses XSLT to translate the XML catalogs into HTML;

4) a catalog generator for crawling large collections of datasets and producing THREDDS catalogs. Allows adding additional descriptive information to the catalog;

5) a dynamic catalog servlet for serving THREDDS catalogs on the fly.

4. Data Applications

Several data analysis and visualization applications are already using THREDDS. Unidata's IDV [4] is one of our main THREDDS client testbed.

The DODS Aggregation Server [2] was the first DODS (OPeNDAP) server to supply THREDDS catalogs by default. Other DODS/OPeNDAP servers are now starting to provide inventories of their datasets with THREDDS catalogs.

5. More Information

The technical status of the THREDDS project is at http://www.unidata.ucar.edu/projects/THREDDS/tech/. More information on THREDDS is available from the THREDDS web page, http://www.unidata.ucar.edu/projects/THREDDS/.

6. References

[1] DLESE (Digital Library for Earth Science Education), http://www.dlese.org/.

[2] DODS Aggregation Server, http://www.unidata.ucar.edu/projects/THREDDS/tech/AggServerStatus.html

[3] GCMD (Global Change Master Directory), http://gcmd.gsfc.nasa.gov/.

[4] IDV (Integrated Data Viewer), http://my.unidata.ucar.edu/software/metapps/

[5] LDM/IDD, http://my.unidata.ucar.edu/software/ldm/ and http://my.unidata.ucar.edu/software/idd/

[6] NcML (NetCDF Markup Language), http://www.scd.ucar.edu/vets/luca/netcdf/.

[7] NSDL (National Science Digital Library), http://www.nsdl.org/ and http://comm.nsdlib.org/.

[8] OPeNDAP (aka DODS), http://www.unidata.ucar.edu/packages/dods/

[9] Taylor, W., J. Benson, T. Whittaker, J. Rueden, 1995: Seamless Access to Local and Distributed Data. Preprints, 11th Int. Conf. on IIPS for Meteorology, Oceanography and Hydrology, 349-352.