

Improving Access to a Climate Data Repository Using Unidata Tools (Final Report for 2011 Unidata Equipment Award)

Catherine Smith and Don Murray
University of Colorado/CIRES

The University of Colorado at Boulder's Cooperative Institute for Research in Environmental Sciences (CIRES) Climate Diagnostics Center (CDC) maintains a repository of climate datasets that is used daily by researchers and educators at CIRES and around the world. These datasets are being used to answer questions about the Earth's climate system, such as the cause and nature of extreme climate events like the 2010 Russian Heat Wave. We received funds from the Unidata Equipment Grant to purchase a new server to enhance and expand the existing THREDDS Data Server (TDS) capabilities and establish a RAMADDA server at the CDC in order to provide end-to-end data services that facilitate research and education in the climate sciences.

With these funds, we purchased a Supermicro A+ Server 2022-URF 16-core server with 96GB of memory to run both TDS and RAMADDA. We chose to run each server in a separate virtual machine instead of purchasing separate hardware for each server. Running the system in a virtual machine provides extra security and allows us to better deal with changing CPU and memory requirements for TDS and RAMADDA. If either server needs more resources, they can easily be allocated to the virtual machine without buying more hardware.

CDC had several objectives in acquiring the new equipment from Unidata. The chief objective was to move the TDS to its own system because the existing server's performance was being taxed by TDS and the other processes running on it. Additionally, the old system was running a version of Tomcat that could not be upgraded and has a known problem which caused the access URLs automatically generated by TDS to be incorrect. With the new server, we were able to move the TDS to its own virtual machine which solved the performance issue, and upgrade to Tomcat version 6, which solved the URL problem. The new system was made operational on August 8, 2011 and is available at <http://www.esrl.noaa.gov/psd/thredds/>.

Since the upgrade, CDC is making available ~33 Gbytes/month on average via the TDS. Our most popular datasets are the NCEP/NCAR Reanalysis I, the Twentieth Century Reanalysis (V2), the NCEP/DOE Reanalysis 2 and the NASA GPCP Precipitation dataset. IPs of the institutions using OPeNDAP are from over 30 countries with the biggest users being Woods Hole Oceanographic Institute, NOAA's PMEL, and the University of Amsterdam. A few institutions including PMEL and the Integrated Climate Data Centre at Hamburg, Germany base some of their use of their Live Access Server "LAS" application on our OPeNDAP URLs.

Another objective was to create timeseries aggregations for datasets that are separated into multiple files by time (to avoid having extremely large files). We

have not yet been successful doing this with TDS, though we are working with Unidata to try to solve the problem. In the mean time, we have successfully implemented timeseries aggregation of several of our datasets using RAMADDA.

A third objective was to run a publicly accessible RAMADDA server locally at CDC. We have been running an internal RAMADDA server on the new server since Fall, 2011. For the publicly accessible server, we plan on running a read-only version of RAMADDA that shares the same database as the internal server starting in May 2012 (<http://www.esrl.noaa.gov/psd/repository/>). We use RAMADDA's access control to keep some portions of the database for internal collaboration and datasets that cannot be distributed and other portions for external viewing and access. We are in the process of acquiring an additional 22 TB of storage that will be used for sharing climate datasets and accessible through RAMADDA.

Overall, we believe given the number of datasets that CDC makes available, improved OPeNDAP access has made the files more useful and our statistics support that. In addition to the timeseries aggregation fix, future plans include an improved catalog with more human readable dataset names, and adding additional types of data available via OPeNDAP (satellite, timeseries...). Additional services will be added to RAMADDA to provide interactive climate data analysis capabilities. Don Murray will do a presentation on access to climate data using RAMADDA and IDV at the upcoming 2012 Unidata Users Workshop in July.



Don Murray standing next to the video display wall in the NOAA/CDC Climate/Weather Laboratory. IDV displays of SST and NCEP reanalysis data accessed through TDS are used to monitor the climate using bundles stored on RAMADDA.