NCAR's Research Data Archive: OPeNDAP Access for Complex Datasets IN11C-3628

Robert Dattore* and Steven Worley*

*CISL/DSS, National Center for Atmospheric Research, Boulder, CO <dattore@ucar.edu>, <worley@ucar.edu>

http://rda.ucar.edu/

New RDA Data Service

Now you can use your OPeNDAP-aware analysis and visualization tools to access data from complex datasets hosted by the Research Data Archive at NCAR. The RDA has developed a service that lets users create "customized" aggregations and then access them via OPeNDAP.

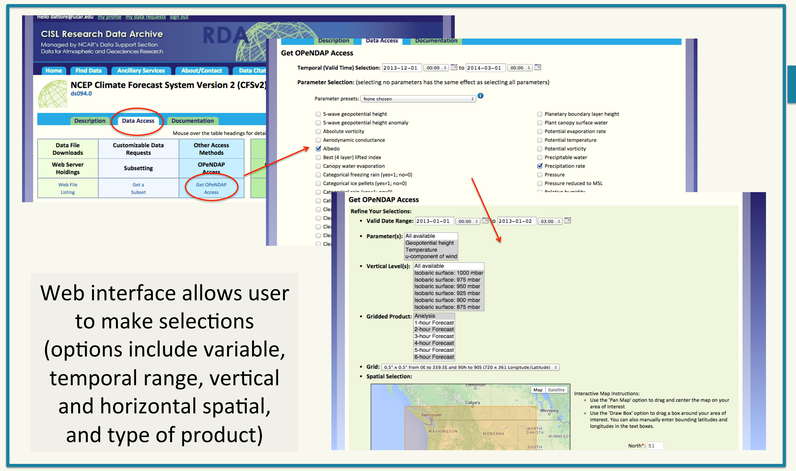
Many modern datasets contain numerous variables at multiple vertical, temporal, and geospatial resolutions, and these products are spread across large numbers of physical files whose structures are dictated by the processes that produce the data. Most users are likely to want only a subset of such a dataset, with the subset probably spanning some number of dataset files, and the data needing to be accessible to analysis and visualization tools. The RDA hosts several complex datasets and provides a subsetting service where users can make selections and receive only the data they want from the much larger datasets, but the resultant files must still be downloaded to local computers before tools can access the data.

OPeNDAP is a powerful protocol for delivering data directly to tools, but for providers of complex datasets, deciding how to aggregate can be a challenge. It is difficult to anticipate all possible uses of the data, and supporting only specific aggregations could limit other potential uses of the data. One solution is for the data provider to not aggregate at all and instead allow the users to define for themselves the aggregations that serve their needs. By leveraging the already-existing subsetting infrastructure, RDA users can make selections that define the subsets they want, and instead of downloading files locally, they can now use OPeNDAP-capable tools to connect to the RDA and access the data directly.

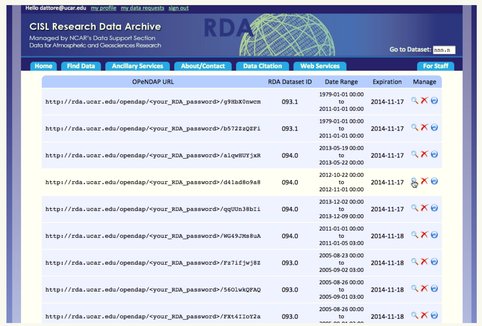
Acknowledgments

The RDA content and infrastructure are supported by a team of data specialists from the Data Support Section of the Computational and Information Systems Laboratory at NCAR. In addition to the authors of this poster, the team includes: Cecilia Banner, Joey Comeaux, Tom Cram, Hua Ji, Grace Peng, Doug Schuster, Chi-Fan Shih, and Dave Stepaniak.

User Workflow

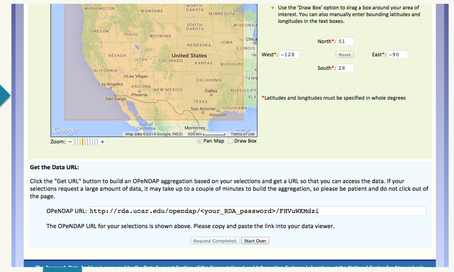


Web interface allows user to make selections (options include variable, temporal range, vertical and horizontal spatial, and type of product)

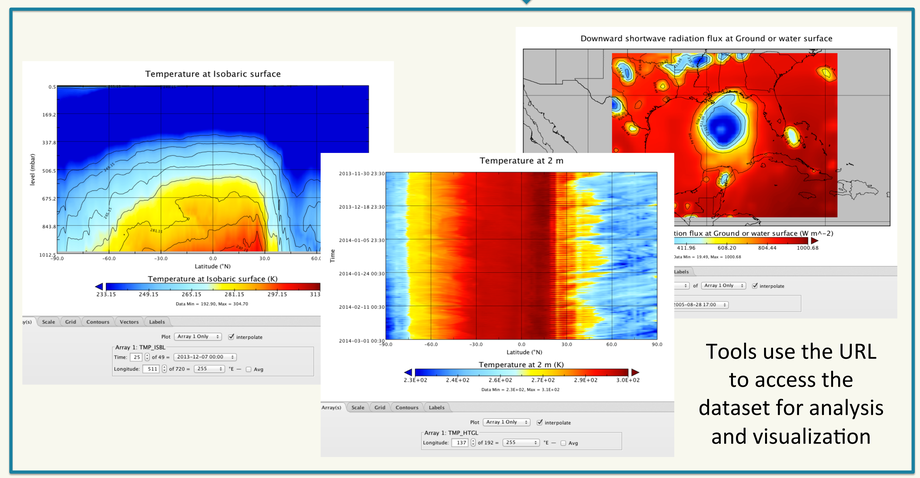


"mydap" application allows users to manage their OPeNDAP aggregations: view full details, delete, and extend expiration date

Backend process builds the aggregation in real-time and assigns it a unique ID connected to the registered user

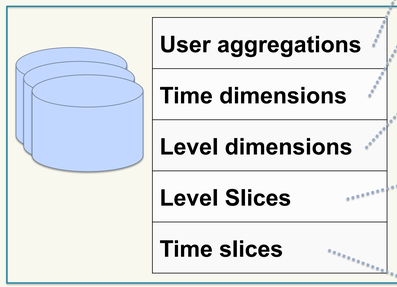


Once the aggregation is built, the user receives a URL that can be used to open the dataset remotely



Tools use the URL to access the dataset for analysis and visualization

Technical Description for Database Managers



New MySQL database to support OPeNDAP access

Row for each aggregation: unique ID, user selections, user identity, date of last refresh

Row for each time dimension: aggregation ID (FK), name of dimension (e.g. time), size of dimension

Row for each level dimension: aggregation ID (FK), type of level (e.g. isobaric), name of dimension (e.g. level), size of dimension, unique ID

Row for each level slice: aggregation ID (FK), level type (FK), internal level code, slice index number, unique ID (FK)

Row for each time slice: aggregation ID (FK), date/time of data, variable name, internal product code (e.g. analysis), internal level code (FK), internal ID of disk file containing the data, time slice index number

Define the das and dds for each aggregation

Identify disk file

Row for each disk file in the RDA dataset: file ID (FK), RDA dataset number, start and end dates, internal ID code

Identify data record

Row for each record in a disk file: internal ID code (FK), byte offset, byte length, date/time of data, internal product code, internal grid resolution code, internal level code

dods output

RDA data files

Data file inventory (by variable)

Existing databases already supporting dataset subsetting