17.2

NetCDF to GIS: the Good, the Bad, and the Ugly

John Caron* and Stefano Nativi[†*]
*Unidata Program Center [†]University of Florence

1. Abstract

Unidata, along with many other groups, is exploring the use of Geographic Information Systems (GIS) with scientific datasets. We have created a prototype Web Coverage Server (WCS) that can read datasets through the netCDF API and translate them into GeoTIFF files, a common image format for GIS. WCS is used to implement an interoperability service between a scientific distributed framework (i.e. the THREDDS system) and general purpose GIS applications. This gave us the opportunity to test WCS in a real and complex environment. In this paper we give the context for this work, and summarize a few of our findings; an interesting example is the data granularity issue. A more detailed description of the results can be found at http://www.unidata.ucar.edu/staff/caron/papers/AMS-04/

2. Scientific Information Systems

Scientific data are stored in a wide variety of file formats, some highly specialized for a specific discipline or data type. Important general purpose and portable formats include *netCDF* and *HDF*, developed by Unidata and NCSA respectively. Another class of formats are the WMO standards for meterological data transfer such as *GRIB* and *BUFR*. *OpenDAP* (University of Rhode Island) and *ADDE* (University of Wisconsin at Madison) are client-server protocols which fill the same function, provinding access to scientific data in a portable and standard way. All of these are examples of what can be called *Scientific Information Systems* (SIS).

All SIS may be understood at a high level through their *abstract data model* and more concretely through the *application programmer's interface* (API) provided by the software libraries that implement the model and provide data access for application programs.

SIS data models generally start with the simple *multidimensional arrays* that are also the main data structures of the FORTRAN programming language. The netCDF data model mostly consists of just multidimensional arrays, which can be seen as both a strength and a limitation. The HDF5 and OpenDAP data models both add support for *structures* and variable length arrays called *sequences*. Structures were introduced to programming languages through Algol and Pascal, and are an important part of all modern languages, including C, Java, and Fortran-90. Sequences correspond most closely with relational tables returned from database queries.

Multidimensional arrays, structures, and sequences are at a rather low level, sometimes called the *syntactic* level. A higher level charactorization of scientific data types would include *gridded data* from numerical models, *image data* from satellites and radar instruments, *point data* from metars, buoys, stations, etc., and *trajectory data* from soundings, profilers, aircraft tracks, etc.

The information we add to SIS to clarify its meaning is often called *semantic* or *use metadata*. The most important semantic metadata needed to implement a WCS server is geolocation information, typically in the form of georeferencing coordinate systems. This information however is not always contained in the dataset themselves, rather it may be encoded in analysis and display software routines developed specially for the numeric model, satellite instrument, or scientific project. When this information is present, it is typically in the form of general purpose name/value attributes that annotate the data, rather than being represented in the data model itself. This leads to many different attribute conventions for representing coordinate information.

3. Geographic Information Systems

Geographic Information Systems (GIS) can be informally characterized as relational databases with spatial capabilities. The primary spatial data types are the *feature* (a point or area-enclosing polyline), the *map* (a 2D image), and recently the *coverage*, a generalization of a map.

The OpenGIS Consortium (OGC) is a consortium from industry, government and academia to create open standards for the GIS community. Recently they completed the Web Coverage Server (WCS) specification. This is a client-server protocol with enough generality to handle gridded and image type scientific datasets

GeoTIFF is an extension of the Tagged Image File Format (TIFF), that adds georeferencing metadata. It is a widely used data format for GIS.

4. Prototype WCS Server

To understand the strengths and weaknesses of the WCS specification, and to compare it to existing client-server protocols for accessing scientific data,

* *Corresponding author address:* John Caron, UCAR Unidata, P.O.Box 3000, Boulder Co. 80307-3000; email: caron@ucar.edu. The Unidata Program Center is sponsored by the National Science Foundation and managed by the University Corporation for Atmospheric Research.

we implemented a prototype WCS server. The server is written completely in Java using servlets, and we run it within the Tomcat web server. The server accepts requests using either CGI style key-value pair encoding, or SOAP messages.

The server is configured using a *THREDDS catalog* to list the datasets to be served. The datasets are read through the Java-netCDF library, which currently can read either netCDF files or OpenDAP datasets. We hope to add the capability to access other SIS eventually through a generalized API, including HDF5, GRIB, and ADDE.

Not all netCDF or OpenDAP datasets have georeferencing metadata using attribute conventions that our software can recognize. Those that do are converted to a *geogrid* object, essentially a netCDF variable with a georeferencing coordinate system. A geogrid object is the semantic equivalent of a WCS *coverage*. The server reads the datasets listed in the catalog, extract the geogrids, and builds a WCS *capabilities* XML document, which a client can obtain through the *GetCapabilities* request. A client can get more detailed information about a specific coverage through the *DescribeCoverage* request, and get the data itself through the *GetCoverage* request.

When a GetCoverage request is received, the data is extracted from the geogrid, and transmitted to the client as a GeoTIFF file. Currently the server can handle subsetting, but not resampling or reprojecting the data.

We are also exploring the possibility that the server could act as a proxy WCS server for other THREDDS data servers. The data server would register their catalogs with the proxy server. The proxy server would receive the WCS request, obtain the data from the THREDDS data server, and generate the response. The question of creating a global registry for such services remains open.

5. Results

Testing of the WCS server was ongoing as this paper was being written, but the following observations can be made.

The WCS 1.0 specification is brand new and implementation experience is very limited. Like any client-server protocol, much depends on implementation details on the clients. As of this writing, there are few to no WCS clients.

Currently coverages are limited to regular grids, so irregular grids will have to be resampled. Typical GIS clients will probably ask for the data to be resampled and subsetted to fit their view area. Scientists who are trying to see their data “as is” may find this annoying. On the other hand, sattelite data in its raw form often has a complex sensor-dependent coordinate system that cannot be described with any standard algorithm. The geopositioning must be either completely enumerated or generated by a special purpose routine. In this case, server-side resampling is a reasonable choice for most general purpose visualization.

GeoTIFF has done a good job in defining a robust set of tags for georeferencing. However there are some grey areas of the specification, and its not always obvious how a GeoTIFF client will interpret various tags. We sometimes have to guess how a GeoTIFF file should be written, test it in one or more GIS clients, and iterate. An important question is how to encode vertical levels and time series. While we can write multiple images into a GeoTIFF file, there is currently no way to specify the meaning of those multiple images in GeoTIFF. A WCS client will have to be quite smart in order to cleanly represent multiple vertical levels and time series data to the user. Given the 2D orientation of traditional GIS it may be a while or never before standard GIS clients add such functionality.

Scientific data is characterized by heterogeneous types of aggregations; each of them useful in specific application domains, and characterized by specialized semantics. Efforts to implement application interoperability are often frustrated by data granularity mismatches. WCS provides useful abstractions to face this issue, but it doesn't seem to be sufficiently scalable to solve the problem fully.

There are few (or perhaps better to say too many) standards for adding georeferencing information to scientific datasets. This makes it difficult or impossible to write general services to bridge SIS to GIS. Clearly the next step in the evolution of SIS data access models and libraries is to explicitly include georeferencing coordinate systems. Unidata is exploring this possibility in collaboration with both the HDF5 development group at NCSA and the OpenDAP developer community at URI and elsewhere. Another approach is to standardize on attribute conventions for representing coordinate information, and we have been focusing on the CF conventions for model output, involving researchers from NCAR, UK Met Office, LLNL, and others.

As georeferencing standards emerge in SIS, and GIS starts to think past 2D, we expect WCS and similar services to be useful to decision managers for overlaying scientific data onto the rich set of data features GIS clients provide. However it seems unlikely that scientists will replace special purpose tools with general purpose GIS any time soon.

The authors would like to thank Yuan Ho for GeoTIFF development, Giacomo Villoresi for the WCS package development, Roland Schweitzer and John Cartwright for testing, and Steve Kopp of ESRI for technical advice.