ELSEVIER

Computers & Geosciences 31 (2005) 1104–1118

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

Design and implementation of netCDF markup language (NcML) and its GML-based extension (NcML-G_{ML})

Stefano Nativi[a,b,*], John Caron[c], Ethan Davis[c], Ben Domenico[c]

[a]*University of Florence at Prato, Piazza Ciardi, 25, 59100 Prato, Italy*
[b]*IMAA-Italian National Research Council (CNR), Italy*
[c]*Unidata Program Center, University Corporation for Atmospheric Research, Boulder, CO 80307, USA*

Received 14 May 2004; received in revised form 11 October 2004; accepted 31 December 2004

Abstract

The Network Common Data Form (netCDF) is one of the primary methods of self-documenting data storage and access in the international geosciences research and education community and beyond. NetCDF was designed for use in a networked environment. The recent evolution toward web services approaches to data exchange has focused attention on communication via messages in the defacto standard XML language. XML is a text-based language while netCDF is based on a binary file storage mechanism; thus NcML is a natural augmentation of netCDF with extensions encapsulating descriptions of the structure and content of netCDF objects in an XML form. Since netCDF was designed to be self-documenting, the XML representation of internal netCDF documentation is a natural augmentation of the original netCDF concept. In fact, the netCDF Markup Language (NcML) and NcML-G (NcML-Geography) extensions described in this article have applications beyond merely representing the internal netCDF documentation in the XML language. The NcML coordinate system makes it possible to describe the coordinate system used to represent the netCDF dataset. Furthermore the NcML dataset is a tool for describing "virtual netCDF" files that may be aggregations of data from several existing netCDF files, or it can represent a target dataset to be created by transforming existing netCDF files into a new form described in the NcML language. The NcML-G extension provides a means for fusing the data models of the traditional netCDF atmospheric science community with those of the GIS community which is of the utmost importance. Bringing the data models and data systems of those communities together will foster an era of interdisciplinary research and education within the geosciences subdisciplines. It will also encourage closer interactions between the geosciences and the societal impacts community. The design and software implementation of the core NcML specification and its extensions are presented and discussed.

Keywords: Scientific data markup language; NetCDF; GML

1. Introduction

The Network Common Data Form (netCDF) software is widely used throughout the world as a mechanism for storing and accessing scientific data—especially datasets related to the geosciences. The

*Corresponding author. University of Florence at Prato, Piazza Ciardi, 25, 59100 Prato, Italy.

E-mail addresses: stefano.nativi@pin.unifi.it (S. Nativi), caron@ucar.edu (J. Caron), edavis@ucar.edu (E. Davis), ben@ucar.edu (B. Domenico).

0098-3004/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2004.12.006

Unidata netCDF web pages[1] list several means for estimating netCDF usage: for example, the netcdfgroup@unidata.ucar.edu mailing list currently contains over 500 addresses in 32 countries. Over 2000 distinct hosts in 55 countries have downloaded the netCDF software distribution since May 1997.

A few of the earth science studies institutions that employ netCDF are: NOAA's Climate Diagnostics Center (CDC), NASA's Halogen Occultation Experiment (HALOE), The global ocean modeling effort at Los Alamos National Laboratory (LANL), Lamont-Doherty Earth Observatory (LDEO) of Columbia University, the National Center for Atmospheric Research (NCAR), the Commonwealth Scientific and Industrial Research Organization (CSIRO) Division of Atmospheric Research in Australia, The Earth Scan Lab, a High-Resolution Picture Transmission (HRPT) ground station at the Coastal Studies Institute, Scripps Institution of Oceanography (SIO), and Sandia National Laboratory.

Beyond the research and education communities, several commercial analysis and data visualization packages have been adapted to access netCDF data. Moreover, netCDF is the vehicle adopted by the Analytical Instrument Association (AIA) to implement the Analytical Data Interchange Protocols for chromatography and mass spectrometry. In addition, the Positron Imaging Laboratories and the Neuro-Imaging Laboratory of the Montreal Neurological Institute have selected netCDF as the data format for their medical image files.

The advent of the web services approach to making data available has increased the emphasis on eXtensible Markup Language (XML) as a means for conveying data and information about data available on the web. NcML was developed to encode information about (though not the data contained by) netCDF files and provide a standard XML dialect by which this information can be shared. NcML does not encode the data from a netCDF file only the metadata about a netCDF file. Thus, NcML provides a powerful complement to the self-documenting nature of netCDF which employs binary file formats and transport mechanisms. NcML has evolved beyond its original goal of simply describing netCDF files, introducing extensions to: explicitly encode part of dataset semantics, generate virtual datasets, and introduce GIS community semantics. NcML-G is an extension of NcML that introduces conventions for specifying information in netCDF files that is characteristic of GIS data (i.e., georeferencing and coverage information); NcML-G_{ML} implements such conventions using the OpenGIS Geography Markup Language (GML) grammar.

2. NetCDF background

This article is a general description of the netCDF Markup Language (NcML) and a special extension of NcML leveraging GML grammar (NcML-G_{ML}). To understand NcML, it is important to understand the basics of netCDF which are described in the next section which borrows liberally from the frequently asked questions about netCDF web page.[2]

NetCDF is an Application Programming Interface (API) that provides methods for accessing array-oriented data and a freely distributed collection of software libraries for C, FORTRAN, C++, Java, and perl that implement this interface. The software was developed by Glenn Davis, Russ Rew, Steve Emmerson, John Caron, and Harvey Davies at the Unidata Program Center in Boulder, CO, USA and augmented by contributions from other netCDF users (see footnote 1). The netCDF libraries define a portable format for representing scientific data. The interface, libraries, and format support the creation, access, and sharing of scientific data.

NetCDF data is:

- *Self-Describing*: A netCDF file includes information about the data it contains.
- *Architecture-independent*: A netCDF file is represented in a form that can be accessed by computers with different ways of storing integers, characters, and floating-point numbers.
- *Directly accessible*: A small subset of a large dataset may be accessed efficiently without first reading through all the preceding data.
- *Appendable*: Data can be appended to a netCDF dataset along one dimension without copying the dataset or redefining its structure. The structure of a netCDF dataset can be changed, though this may require copying the dataset.
- *Sharable*: One writer and multiple readers may simultaneously access the same netCDF file.

The XML language and related technologies have similar characteristics: XML is self-describing and architecture-independent; whereas XPath, XQuery, XUpdate, and XPointer, among others support direct access and modifications. Therefore, XML technologies proved to be ideal candidates for netCDF encoding, and

[1]Unidata netCDF Group, "netCDF Usage", *Unidata software: netCDF*, available at http://my.unidata.ucar.edu/content/software/netcdf/usage.html.

[2]Unidata netCDF Group, "NetCDF Frequently asked questions", *Unidata software: netCDF,* available at http://my.unidata.ucar.edu/content/software/netcdf/faq.html#whatisit.