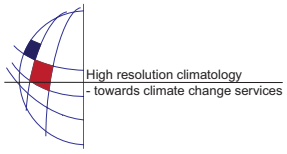
EMS Annual Meeting Abstracts
Vol. 7, EMS2010-80, 2010
10th EMS / 8th ECAC



The Research Data Archive at NCAR: A System Designed to Handle Diverse Climate Datasets

B. Dattore and S. Worley

National Center for Atmospheric Research, Boulder, Colorado, USA (dattore@ucar.edu, worley@ucar.edu)

Improving discovery of and access to high quality climate data is critical to advancing our understanding of the Earth's climate system and how it changes over time. The Research Data Archive (RDA), managed within the Computational and Information Systems Laboratory at NCAR, is a data resource designed to meet the needs of climate and weather research. First efforts to build the RDA began in the late-1960s when datasets were typically small in size and did not have global coverage, e.g. surface observations for a single country or balloon data from a single experiment. Further, many distinct specialized data formats were common. Today datasets are generally larger, have global coverage, and are often stored in common data formats. The RDA is now over 600 distinct datasets comprised of more than 4 million files totaling 400 TB with 10% readily available online.

Major changes in metadata collection have transformed the RDA into a climate data flow management system that is efficient, adaptable, and provides many user data access benefits. We will describe the uniform methods and tools used in the workflow of the RDA, including: the transformation of legacy metadata into the new system using a XML structure that allows for diversity and richness of a long-term archive, a GUI for entering and editing constrained discovery metadata, tools for collecting file level metadata for common formats (GRIB1, GRIB2, netCDF, BUFR, IMMA, and several legacy formats) and handling file archiving I/O, and the usage of a relational metadatabase to hold all the results. Most importantly, we will then show how this integrated metadata foundation enables several different approaches for data discovery that are always up-to-date, accurate definition of multiple data access and extraction methods that can vary from dataset to dataset, interfaces where user driven constraints will isolate the data files needed from within terabyte size collections, automatic construction of codes to aid in downloading, a standard way to request that deep archived data be automatically restaged to online disk, and the facilities to share metadata through the OAI-PMH protocol. The content in the RDA data management system is rapidly growing and annually serves more than 6000 unique users worldwide.