Emerging Tools for Distributed Data Access and Collaborations

Glenn K. Rutledge
National Oceanic and Atmospheric Administration
National Climatic Data Center
13th Federation Assembly Meeting
Earth Systems Information Partners
Asheville, NC August 18, 2004
Briefing Overview

- Earth System Interoperability
  - Science Exploration at the Data Level
- Metadata, Catalogs, & Ontologies
- Tools and Programs
  - Too many to count?
- My World: NOMADS
- Next Steps to Collaborations
Count the black dots....

Data Interoperability...
A moving target?
Yes.

XML
SOAP
OWL
OPeNDAP

ESML
http/TCP
Globus
SWEET

Adapted from L. Olsen
What are the goals facing the GeoScience community?
- Is it just access to high volume data (satellite, radar, and model)?

How will Agencies and Institutions address interoperability?
- Should it be system, data or both?

Have the scientific requirements been adequately defined?
- Do top down approaches adequately promote science?

How can Agencies and institutions develop partnerships while allowing for attribution, with diverse goals and agendas?

Data interoperability is the key: Scientific Data Stewardship
Program Management at the Data Level

Predictability
Earth Systems
Aerosols
Solar Cycles

Atmospheric
Climate & Oceans
Air Quality
Space Wx

DATA

Rutledge NOAA NCDC 13th ESIP
Some Assumptions

- Operational Forecasting-
  - Ensemble Predictions: flow-dependant prediction of weather and climate risk - nowcasting, medium range and seasonal.

- Atmospheric and Oceanic Research-
  - Scalar and Vector processing and Workstation models
  - Model output statistics; data assimilation techniques

- Global Climate Change and Advanced Analysis-
  - Clouds, initial conditions, true coupled simulations.
  - Long term climate monitoring: in-situ analysis, trends, data homogeneity, extremes, downscaling, reducing uncertainty...
  - On-demand Data Mining and Product Generation.
Data systems based on the integration of independently developed system elements offer many more opportunities than more traditional centrally developed ones.

Peter Cornillon URI
...et al.
Collaborations: How do we get there?

• Data transport is being actively pursued: OPeNDAP, SOAP, ...

• Earth System Partners need to be able to find and use various data sets, wherever they may be, whatever format...

• THREDDS can provide dynamic access and generate catalogs

• GCMD is a major resource for metadata management for the entire GeoSciences community- this activity must evolve!

• Ontology projects such as SWEET in conjunction with THREDDS and GCMD can provide individual data sources, data variables and metadata management for the community.
• The Semantic Web for Earth and Environmental Terminology (SWEET) project provides a common semantic framework for various Earth science initiatives.

• The semantic web is a transformation of the existing web that will enable software programs, applications, and agents to find meaning and understanding on web pages.

• SWEET developed these capabilities in the context of finding and using Earth science data and information.
• Pare down large file sizes of high resolution data and products.
• (re-) Group different data sets to create needed products – such as initialization files for model development, analysis, and intercomparison.
• Subset the data sets in parameter space
• Subset the data sets in physical space
• Subset the data sets in temporal space
- Data extraction for the generation of products “on-demand”.
- Advanced data mining algorithms for pre-generation, or executed by (authorized) users also on-demand.
- Access to mined physical processes or signatures thru data mining.
- Search and location tools and metadata management.
Just several programs addressing the data access, description, and search activities:

- **CLASS** Comprehensive Large-Ararry Stewardship System
- **DAAC** Distributed Active Archive Centers
- **DIMES** DIstributed MEtadata Server
- **DLESE** Digital Library for Earth System Education
- **ECHO** The EOS ClearingHouse (middleware)
- **ESIP** Earth Science Information Partners
- **FGDC** Federal Geographic Data Committee
- **FIND** Federation Interactive Network for Discovery
- **GCMD** Global Change Master Directory
- **GOSIC** Global Observing System Information Center
- **NDG** NERC Data Grid
- **NSDI** National Spatial Data Infrastructure
- **NSDL** National STEME Digital Library
- **NMMR** NOAA Metadata Manager Repository
- **OAI** Open Archives Initiative
- **SWEET** Semantic Web for Earth and Environmental Terminology
- **THREDDS** Thematic Realtime Environmental Data Distributed Services
Leveraging Partnerships

THREDDS Data Providers

- University of Alabama Huntsville (Sara Graves, Rahul Ramachandran, Steve Tanner, Ken Keiser)
- ARM (Atmospheric Radiation Measurement, Chris Klaus)
- CDC, the Climate Diagnostic Center (Roland Schweitzer)
- COLA, Center for Oceans Land Atmosphere (Joe Wielgosz)
- University of Florence (Stefano Nativi)
- GMU, George Mason University (Menas Kafatos and Ruixin Yang)
- IRI/LDEO, International Research Institute/Lamont Doherty Earth Observatory (Benno Blumenthal)
- ESG, the Earth System GRID (Luca Cinquini, NCAR/SCD)
- IRIS DMC, Incorporated Research Institutes for Seismology Data Management Center (Rob Casey)
- NCAR, the National Center for Atmospheric Research (Don Middleton)
- NCDC, the National Climatic Data Center (Ben Watkins)
- NGDC, National Geophysical Data Center (Ted Habermann)
- NOMADS, NOAA Operational Model Archive and Distribution System, (Glenn Rutledge, NCDC)
- University of Oklahoma (Kelvin Droegemeier)
- PMEL, the Pacific Marine Environment Laboratory (Steve Hankin)
- FNMOC, Fleet Numerical Meteorological and Oceanographic Center (Phil Sharfstein)
- SSEC, the Space Science and Engineering Center., U. of Wisconsin-Madison (Steve Ackerman, Tom Whittaker)
- Unidata Community ADDE servers (Tom Yokeltas, Unidata Program Center)
- CIESIN (Consortium for International Earth Science Information Network, Bob Downs)
- CUAHSI (Consortium of Universities for Advancement of Hydrologic Science, David Maidment)
- ESIG / NCAR (NCAR Environmental Societal Impacts Group, Bob Harriss)
- Earthscope (UCAR UNAVCO, Chuck Meertens)
- GEON (GEOphysical Network, Chaitan Baru, UCSD San Diego Supercomputer Center)
- ESRI GIS Community (ESRI, Inc., Jack Dangermond, President)
Leveraging Partnerships

THREDDS Collaborators

• **ADDE**, Abstract Data Distribution Environment (University of Wisconsin – Madison, Tom Yoksas)
• **DIMES**, DIstributed MEtda System (George Mason University, Ruixin Yang)
• **DODS/OPeNDAP/Aggregation Server**, Distributed Oceanographic Data System/Open source Project for a Network Data Access Protocol (University of Rhode Island, Unidata, Ethan Davis)
• **DLESE**, Digital Library for Earth System Education (Rajul Pandya)
• **ESML**, Earth System Markup Language (University of Alabama-Huntsville, Rahul Ramachandran)
• **ESRI**, Environmental Science Research Institute (various)
• **GCMD**, Global Change Master Directory (Gene Major)
• **OGC and ISO Standards** (University of Florence, Stefano Nativi)
• **ADL** (Gazetteer Services The University of California, Santa Barbara, Linda Hill and Michael Goodchild)
• **DLESE Evaluation Services** (The University of Colorado CIRES, Susan Buhr)
• **DLESE Data Services** (Tamara Ledley)
• **DLESE Program Center** Digital Library for Earth System Education (Mary Marlino)
• **ESRI** (Jack Dangermond, President)
• **OPeNDAP** (The University of Rhode Island Open source Project for a Network Data Access Protocol -- formerly DODS, Peter Cornillon)
• **LAITS** (Laboratory for Advanced Information Technology and Standards, Liping Di, George Mason University)
• **NSDL Evaluation Services** (University of Colorado, Tamara Sumner)
• **OGC** (Open GIS Consortium, David Schell, President)
• **SWEET** (Semantic Web for Earth and Environmental Terminology, Rob Raskin)
Evolve GCMD?

GCMD DODS/OPeNDAP Portal

Find DODS/OPeNDAP-Related Data Sets by Topic:

- Agriculture (forestry, soils...)
- Atmosphere (precipitation, temperature...)
- Biosphere (vegetation, zoology...)
- Climate Indicators (air temperature, drought...)
- Human Dimensions (environmental impacts, land use...)
- Hydrosphere (snow & ice, water quality...)
- Land Surface (land cover, erosion...)
- Oceans (circulation, salinity...)
- Paleoclimate (tree rings, land records...)
- Snow and Ice (snow, snow cover...)
- Spectral / Engineering (radar, visible imagery...)

- GCMD (Search the entire GCMD database)

http://gcmd.gsfc.nasa.gov/Data/portals/dods/

http://gcmd.gsfc.nasa.gov/Data/portals/dods/freetext/ft_search.html

Rutledge NOAA NCDC 13th ESIP
The ODC
Collaborations require long-term maintenance of both the data and descriptions of the data; i.e., metadata.

The degree of system interoperability is determined by the associated metadata and the quality of that metadata.
FIND combines the capabilities of two search systems (the Global Change Master Directory and Mercury). It provides users with a rich set of options to locate ESIP Federation data, services, and information. FIND is accessible from the Federation Home Page.

- Data search
- Topical keyword search
- Data tools and services search

- Federation-wide search
- Web/data free text search
- Advanced data search w/data order links

http://www.esipfed.org/find
• NESDIS Metadata Working Group
  - A good first start.
  - Community wide audience needed.

How?
GeoScience Technology Forum (GTF)
• NSF Cyber Infrastructure
• GEO ??
• NOAA’s Scientific Data Stewardship (SDS) well conceived.

• CLASS requires more community involvement and they are actively seeking feedback. The time is now to design interoperability into CLASS. Re-engineering difficult.

• Many efforts now exist from which to leverage: GCOS, IOOS, US Oceans, DMAC, NVODS, WCRP, IPCC, WMO, GCMD, THREDDS, (more on this)…

• NOAA’s Office of Project Planning and Implementation now formed.
Earth Observation Summit

13th Federation Meeting

- Affirmed need for timely, quality, long-term, global information as a basis for sound decision making.

- Recognized need to support:
  1) Comprehensive, coordinated, and sustained Earth observation system or systems;
  2) Coordinated effort to address capacity-building needs related to Earth observations;
  3) Exchange of observations in a full and open manner with minimum time delay and minimum cost; and
  4) Preparation of a 10-year Implementation Plan, building on existing systems and initiatives by European ministerial in late 2004

- Established ad hoc Group on Earth Observations (GEO) to develop Plan

- Invited other governments to join.
Active Agency Participation with IWEGO

- A system of systems can be designed with active involvement with existing data managers system managers and scientists: ESIP Role?

- Leveraging intra-Agency activities with GEO 10-year plan as the driver.

How is our community addressing the needs of GEO?
To overcome a deficiency in model data access, some of the Nations top scientists are actively engaged in a grass-roots framework to share data and research findings over the Internet.

NCDC, NCEP and GFDL initiated the NOAA Operational Model Archive and Distribution System.

NOMADS is a distributed data services pilot for format independent access to climate and weather models and data.
• provide distributed access to models and associated data,

• promote model evaluation and product development,

• foster research within the geo-science communities (ocean, weather, and climate) to study multiple earth systems using collections of distributed data,

• develop institutional partnerships via distributed open technologies.
The users experience is often frustrating—

- What data of interest exist?
- Are they going to be useful to me?
- How can I obtain them in a usable form?

Time and effort are wasted on data access and format issues.

As a result atmosphere/ocean/climate data are under-utilized. Model inter-comparison nearly impossible.
NOMADS simplifies scientific data networking, allowing simple access to high volume remote data, unifying access to Climate and Weather models:

**Data access (client)**
- Access to remote data in the user’s normal application
  - IDL / IDV / Matlab / Ferret
  - GrADS (GRIB/BUFR w/ GDS)
  - Netscape / Excel / http (wget)
  - CDAT (PCMDI)
  - Any netCDF application (i.e., AWIPS)
- Don’t need to know the format in which the data are stored.

**Data publishing (server)**
- Can serve data in various formats
  - netCDF / GRIB / BUFR / GRIB2
  - HDF (3-5) / EOS
  - SQL / FreeForm
  - JGOFS / NcML
  - DSP
  - ascii, others...

- Spatial and temporal sub-setting and host side computations on the fly.
<table>
<thead>
<tr>
<th>Collaborating Programs</th>
<th>DOC DOE EPA State Dept</th>
<th>NCAR</th>
<th>NOAA Representative</th>
<th>NOAA Representative</th>
<th>(in progress)</th>
<th>Founding Member</th>
<th>Science Advisory Board</th>
<th>Advisory Committee</th>
<th>Member</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAP Climate Action Partnership</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDP Community Data Portal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CEOS Committee on EO Satellites</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CEOP Coordinated Earth Obs Period</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EPA Air Quality Models</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>GO-ESSP Earth Science Portal</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NASA GCMD</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NERC DataGrid</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSF Cyberinfrastructure</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSF LEAD &amp; Geo-Science Tech Forum (GTF)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NVODS / US GODAE / GOOS</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Unidata THREDDS, NSDL, DLESSE</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WCRP World Climate Research Program</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Advancing Collaborations

Rutledge NOAA NCDC 13th ESIP
Multiple paths to format independent data access:

- Direct Client Access
  - GrADS, Ferret, MatLab, IDL, IDV, Web browsers or any OPeNDAP enabled client

- Web/Grid Catalog
  - THREDDS, GCMD SWEET, DIMES...

- NCDC Web Interface & ftp

- CEOS-Grid

- Earth System Grid

Advancing Collaborations

The NOMADS Philosophy

- NOMADS Archive Interface

- NCDC Archive

- Dual Ingest & QC

- Web Browse & Compute

- NCDC Archive and Distribution System

Rutledge NOAA NCDC 13th ESIP 13th Federation Meeting
• NOMADS uses the Open Source http based OPeNDAP.

• OPeNDAP is a binary-level protocol designed for the transport of scientific data subsets over the Internet. Provides server side data manipulation on-the-fly (e.g., GrADS-DODS).

• Data formats: GRIB, GRIB2, BUFR, HDF, NetCDF, ascii...
Conventions: COARDS, CF, FGDC, DIF....libraries built as necessary.

• APIs: JAVA-OPeNDAP, C++-OPeNDAP, NetCDF, GRIB, BUFR, THREDDS, Python.
A grass roots effort has formed by data managers called the Global Organization for Earth Systems Science Portals (GO-ESSP)

GO-ESSP  [http://esportal.gfdl.noaa.gov](http://esportal.gfdl.noaa.gov)

- Unidata
- ESG (NCAR, LLNL)
- OPeNDAP
- COLA
- NOMADS (GFDL, PMEL, NCDC, NCEP, others)
- NASA/GCMD
- BADC, BODC
- WMO
The Global Organization for Earth System Science Portal (GO-ESSP) is a collaboration designed to build the infrastructure needed to create web portals to provide access to observed and simulated data within the climate and weather communities.

The infrastructure created within GO-ESSP will provide a flexible framework that will allow interoperability between front-end and back-end software components. GO-ESSP is an international collaboration involving software developers from both Europe and the United States.
<table>
<thead>
<tr>
<th>Institution</th>
<th>Data Availability</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC</td>
<td>Reanalysis, climate weather models, in-situ</td>
</tr>
<tr>
<td>GFDL</td>
<td>Coupled Models, Control and Perturbation Integrations and historical 20th century simulations using solar, volcano, GHG and aerosol forcings.</td>
</tr>
<tr>
<td>FSL</td>
<td>MADIS mesoNets, Hi-Res RUC-II</td>
</tr>
<tr>
<td>NCAR</td>
<td>Community Climate System Model / Land Surface CO2 predictive models (VEMAP), Reanalysis / Eta</td>
</tr>
<tr>
<td>NCDC</td>
<td>Archive for NCEP model input/output / Select NCDC Observation datasets, Ocean/Ice WAVE, NARR, SST’s...</td>
</tr>
<tr>
<td>NCEP</td>
<td>Real-time Input/Output, Reanalysis (I&amp;II), Ensembles, Sea Ice Ocean, CDAS, Hourly Eta, Climate Forecast Models...</td>
</tr>
<tr>
<td>LLNL</td>
<td>AMIP / Probabilistic information</td>
</tr>
<tr>
<td>PMEL</td>
<td>Ocean and Climate datasets</td>
</tr>
</tbody>
</table>

Rutledge NOAA NCDC 13th ESIP
NCDC and NCEP Data

- NCDC NOMADS Archive
  - POR: 2002 to Real-Time
  - Eta (12km); GFS (1 degree); GDAS; NARR 12km 30yrs
  - RUC-II 20/40km; Ocean and Ice WAVE Models
  - NCDC Reference Data Sets (Reynolds SST’s, GHCN...)
  - NCDC Mirror site to NCEP NOMADS for Eta & GFS

- NCEP Real-Time NOMADS
  - Global Forecast System GFS 1 degree
  - Hourly Eta at 12km
  - Regional Spectral Model (RSM) and Ensembles
  - Climate Data Assimilation System (CDAS)
  - AMIP Climate Monitoring, Climate Forecast Model
  - NCEP/NCAR Global Reanalysis 1&2
**NOMADS Archive and Users**

**Data Philosophy and Retention**
- Data are free.
- NWP forecast data are retained for five years.
- Analysis, Reanalysis, observations, and GDAS model input are retained for long term stewardship.

**Data Users**
- Resolution of IP addresses indicate a broad range, and consistent use of NOMADS available data:
  - U.S. Agencies, Academic Institutions: K-12 to Research
  - International governments, (Italy, Japan, countries within South America and Africa. Many others).
  - Private Sector and Non-Government Organizations NGO’s
  - World Bank, United Nations (FAO), others.
NCDC Ingest Volume
Tb/Yr

NOMADS Archive and Users (cont.)

NOMADS
Download / month

Existing and Projected Volume

* 5-YR retention of fcsts. Long term for anal.

Rutledge NOAA NCDC 13th ESIP
Three primary methods for data access:

- Web Interface
- OPeNDAP
- ftp w/ on the fly Grib subsetting

On-line or Off-line (archive)

- Server-side data computations...

Model Resources

- About NCDC’s Model Resources
- Model Data Inventories
- Get / View Model Data
- Publications
- Other Model Data Resources
At left is the complete script for generating mean and sdev at 500mb analyzing 18 years of “Climate of the 20th Century” over the Internet:

Traditional vs. NOMADS methods:

Data volume transported: 100Gb vs. 2Kb
Time to access data: 2 days vs. 60 sec
Code development: days vs. minutes
Fortran based LOC: 1000 vs. 50 LOC
NOMADS leverages efforts across the community.
NCDC NOMADS ingests 150K grids day. POR 2002 to present.

Any one of these accessible in seconds
Via: OpENDAP
GDS
ftp
Web Plotter
LAS (soon)
Promoting Model to Obs. Intercomparisons

NCDC Reference Datasets

NCDC reference
And others
datasets also available:

- CARDS (IGRA)
- GHCN
- NARR
- Ocean WAVE
Value added retailers who make value added products can use NOMADS GDS to get the meteorological data they need without downloading entire files.

Users (forecasters) of NDFD can create their own products using GDS server accessing only data they need.

GDS reduces the bandwidth needed to create products in weather service operations.

For internet-2 bandwidth, servers at Regional Centers can distribute data to WFO’s for their operations.
Enabling private sector access: An example

NOMADS Ensemble Probabilities on the fly

No need for image generation of ensembles...

OPeNDAP constraint expression

URL is: http://nomad3.ncep.noaa.gov:9090/dods/enshores/archive/ens20040809/ensc0_00z_1x1.ascii?pratesfc[3:3][125:125][277:277]
• Under NOAA’s Scientific Data Stewardship (SDS) programs the NOAA Comprehensive Large Array-data Stewardship System (CLASS) will act as the main portal for NOAA/NESDIS environmental data, providing physical archive, access, and distribution capabilities for large array data sets.

• The NOMADS team and its collaborators are working with CLASS as the system progresses through its phased implementation plans for access to weather and climate models via OPeNDAP and OPeNDAP Servers (GDS/LAS). Metadata management must be addressed at the Agency level.
#1

- **Leverage the resources as goals outlined by the Group on Earth Observations – GEO (& Earth Observation Summit) through appropriate Agency working groups and representatives:**

  - Interagency Working Group on Earth Observations (IWGEO) Data and Information Systems (OWGDIS)
#2

- Ensure that NGO’s, University, or Institutional partners are involved in this process e.g.,
  - COLA
  - EOGEO
  - many more...
#3

- Agencies partially fund (5%?) data management for each program. This should not be considered a separate activity.
#4

- Engage and leverage from existing efforts and organizations especially NASA, NOAA, NSF, etc.

- NSF CyberInfrastructure (Ad Hoc Committee June 2004)
- GO-ESSP
- LEAD GeoScience Technology Forum (GTF)
- THREDDS / GCMD / SWEET / FIND /...
- NERC Data Grid (Europe)
- WMO CBS
Next Steps to Collaborations

#5

- Advance Agency Program Management at the Data level.
#6

- Advance the building of Ontologies at Data Centers and Providers, (with SWEET), to interact with an enhanced “THREDDS & GCMD” effort for data search and access at the “variable” level*.

* Using OPeNDAP enabled clients and Servers
For NOMADS Program Information see:
http://www.ncdc.noaa.gov/oa/climate/nomads/nomads.html

For NOMADS Model Data Access:
NOAA NCDC Main Page → Climate → Model Resources
http://nomads.ncdc.noaa.gov

Or contact:
Glenn.Rutledge @ noaa.gov

Selected Publications on distributed data access and NOMADS:
http://www.ncdc.noaa.gov/oa/model/publications/publications.html

QUESTIONS ?