Invisible Technologies for the Geosciences: the Importance of Infrastructure

Russ Rew, Unidata
Thanks to:

• GFD Dennou Club members who visited Unidata in 2004
• Research Institute for Sustainable Humanosphere
• National Science Foundation and UCAR
• Unidata Program Center staff and associated community

Masato Shiotani, Yasuhiro Morikawa, Masaki Ishiwata, Russ Rew, Takeshi Horinouchi, Ethan Davis, and Yoshi-Yuki Hayashi
Unidata

• Funded primarily by the U.S. National Science Foundation
• Mission: To provide data, tools, and community leadership for improving Earth-system education and research
• At the Unidata Program Center, we
  – Provide access to data (via push and pull systems)
  – Develop tools and infrastructure for data access, analysis, visualization, and data management
  – Support users of our technologies: faculty, students, and researchers
  – Help to build, represent, and advocate for a community
Overview

• Infrastructure, cyberinfrastructure, data access infrastructure, invisibility
• Some infrastructure in the Earth sciences
• Data push, data pull, data access, metadata
• Putting it all together: the Integrated Data Viewer
• Thoughts on the value of infrastructure
What is Infrastructure?

• The basic facilities, services, and installations needed for the functioning of a community
  – Utilities: water and power lines
  – Transportation and communications systems

• Good infrastructure is reliable, sturdy, useful, long lasting, standardized, widely used, and invisible
Infrastructure: Stones in a Wall

- Higher layers are built on lower layers
- Stones may be replaced with other stones of similar size and shape
- From the top, lower layers are invisible
What is Cyberinfrastructure?

- A big word used by NSF to describe:
  - distributed computer, information and communication technologies
  - personnel and integrating components
  - a long-term platform for modern scientific research
- Also called e-Science in Europe
- May include hardware, networks, software, and human experts
## Cyberinfrastructure: the Middle Layers

<table>
<thead>
<tr>
<th>Community-Specific Knowledge Environments for Research and Education (collaboratories, grid community, e-science community, virtual community)</th>
</tr>
</thead>
</table>

### Customization for discipline-and project-specific applications

<table>
<thead>
<tr>
<th>High performance computation services</th>
<th>Data, information, Knowledge management services</th>
<th>Observation, measurement fabrication services</th>
<th>Interfaces, visualization services</th>
<th>Collaboration services</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Networking, Operating Systems, Middleware</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Base Technology: computation, storage, communication</th>
</tr>
</thead>
</table>
Data Access Infrastructure

- Tools for analysis and visualization
- Libraries for data access
- Servers for data collections
- Formats for data storage and access
- Protocols for requesting and receiving data
- Conventions for representing meaning in data
- Standards for formats, protocols, library interfaces, conventions
Is Developing Infrastructure Rewarding?

• It’s abstract, so hard to explain at a party
• You can’t take a picture or movie about it
• If it works well, it is invisible
• End users are often not aware of it
• It doesn’t get referenced in scientific papers
• It can be expensive to evolve and support
• If not maintained, it eventually crumbles
• You can’t sell it, so you have to give it away
Earth Science Infrastructure: Bricks in a Wall of Acronyms

Developed by Unidata

Involvement by Unidata

Other technologies

XML
Unix
HTTP
SQL
C
Fortran
Java
Python, Ruby, …

Unidata decoders
NetCDF Java
Libcf
NetCDF-4
OPeNDAP
ADDE

IDV
GEMPAK
McIDAS
ArcGIS
GrADS
Ferret
NCO

IDD
CDM
TDS
VisAD
CONDUIT project
LEAD project
GALEON project

LDM
THREDDS
NetCDF
Libcf
NetCDF-4
OPeNDAP
ADDE

Unidata decoders
NcML
CF
OGC WCS
HDF5
CSML

NetCDF
Udunits
CDL
GRIB
BUFR
HDF4
GML
Visible and Invisible Infrastructure

Visible to End Users:

IDV, GEMPAK, McIDAS, ArcGIS, GrADS, Ferret, NCO

“Cloak of invisibility”

IDD, CDM, TDS, VisAD, CONDUIT project, LEAD project, GALEON project

LDM, THREDDS, netCDF Java, Libcf, NetCDF-4, OPeNDAP, ADDE

Unidata decoders, NcML, CF, OGC WCS, HDF5, CSML

NetCDF, Udunits, CDL, GRIB, BUFR, HDF4, GML

XML, Unix, HTTP, SQL, C, Fortran, Java, Python, Ruby, …
Organizing the Bricks

Data Push
- IDV
- VisAD

Data Pull
- OPeNDAP
- TDS

Data Analysis, Visualization
- THREDDS
- CF
- Libcf
- Udunits

Data Access
- netCDF
- netCDF-4
- CDM
- CDL
- NcML

Metadata
- Unidata decoders
Distributing Near Real-Time Data

Data Push
- IDD
- LDM
- Unidata decoders

Data Pull
- OPeNDAP
- TDS

Data Access
- netCDF
- netCDF Java
- netCDF-4
- CDM
- CDL
- NcML

Data Analysis, Visualization
- IDV
- VisAD

Metadata
- THREDDS
- CF
- Libcf
- Udunits

Unidata software components:
- CDL
- netCDF
- netCDF-4
- OPeNDAP
- TDS
- THREDDS
- CF
- Libcf
- Udunits

Unidata project:
LDM (Local Data Manager)

- Protocols and software for capturing, distributing, and organizing data in near-real time using reliable, event-driven data distribution
- Supports subscriptions to near real-time data feeds
- Suitable for pushing many small products, as well as large products
- Highly configurable: can inject, distribute, capture, filter, and process arbitrary data products
- The heart of the IDD
IDD (Internet Data Distribution)

Pushes data from multiple sources using cooperating LDMs

Over 170 institutions on 5 continents and growing
Real-time Data Examples
Real-Time Data Flows

**In the Beginning...**

- "a dizzying volume of information – on the order of 100 MB/day" (Davis and Rew, 1990)

<table>
<thead>
<tr>
<th>Service</th>
<th>Data Rate (Bits Per Second)</th>
</tr>
</thead>
<tbody>
<tr>
<td>NAFAX</td>
<td>Analog</td>
</tr>
<tr>
<td>DIFAX</td>
<td>2400</td>
</tr>
<tr>
<td>Watches &amp; Warnings</td>
<td>1200</td>
</tr>
<tr>
<td>U.S. Surface/Upper-Air</td>
<td>4800</td>
</tr>
<tr>
<td>NMC &amp; ECMWF Grids</td>
<td>4800</td>
</tr>
<tr>
<td>Wisconsin Channel</td>
<td>9600</td>
</tr>
<tr>
<td>International (GTS) Data</td>
<td>1800</td>
</tr>
<tr>
<td>FAA 604</td>
<td>1200</td>
</tr>
<tr>
<td>Lightning Data &amp; Others at Non-Discounted Prices</td>
<td></td>
</tr>
</tbody>
</table>

**Now**

- LDM-6 bandwidth 21 TB/week and growing
- TIGGE test from ECMWF to NCAR sustained 17 GB/hour for 5 days
- 30 data feeds provide radar, satellite, text bulletins, lightning, model forecasts, surface and upper air observations, ...
Unidata IDD
North American data delivery and sharing network

IDD-Brasil
South American peer of North American IDD

IDD-Caribe (planning)
Central American peer of North American IDD

Antarctic-IDD
Support of US Antarctic research community

Participants
United States
Canada
Puerto Rico
Costa Rica
Barbados
Venezuela
Chile
Brazil
Argentina
England
Portugal
Spain
Austria
Russia
Vietnam
China (Hong Kong)
South Korea
Antarctica (incipient)
Serving Data Remotely

Data Push
- IDD
- LDM
- Unidata decoders

Data Pull
- OPeNDAP
- TDS

Data Analysis, Visualization
- IDV
- VisAD

Data Access
- netCDF
- netCDF Java
- netCDF-4
- CDM
- CDL
- NcML

Metadata
- THREDDS
- CF
- Libcf
- Udunits
- Open-source Project for a Network Data Access Protocol, see opendap.org
- A discipline-neutral protocol to get remote scientific data and metadata (not files)
- Allows requests for subsets and aggregations
- Software reference implementations for many kinds of data: netCDF, SQL (databases), HDF, FITS, JGOFS,
- Helps make format invisible
- In use in earth sciences, astronomy, medicine, …
- IPCC model output
• Several OPeNDAP servers available: pyDAP, FDS, GDS, DAPPER, **TDS**
• OPeNDAP clients include: Ferret, GrADS, Matlab, IDL, ArcGIS, **netCDF-Java**, IDV
• Protocol uses URLs and HTTP
• Unidata provides OPeNDAP support
• OPeNDAP version 2 now a NASA standard
• Version 4 under development with a test version available: adds XML, new types, new functions, THREDDS catalogs, SOAP, outputs in HTML and ASCII
THREDDS Data Server (TDS)

- Serves data, THREDDS catalogs, and metadata
- Reads and serves several kinds of data through a uniform CDM interface: netCDF, OPeNDAP, HDF5, GRIB, NEXRAD, ...
- Adds Earth-location coordinate systems to data
- Provides OPeNDAP access and subsetting of any data readable with NetCDF-Java library
- An integrated server provides data access through the OpenGIS Consortium Web Coverage Service (OGC/WCS)
- Easy to install, 100% Java, freely available
- Supports dynamic generation of catalogs
THREDDS Data Server

HTTP Tomcat Server

Catalog.xml

THREDDS Server

- OPeNDAP
- HTTPServer
- WCS

NetCDF-Java library

Datasets

hostname.edu

Application

IDD Data

NetCDF-Java library

unidata
Data Representation and Access

Data Push
- IDD
- LDM
- Unidata decoders

Data Pull
- OPeNDAP
- TDS

Data Analysis, Visualization
- IDV
- VisAD

Data Access
- netCDF
- netCDF Java
- netCDF-4
- CDM
- CDL
- NcML

Metadata
- THREDDS
- CF
- Libcf
- Udunits
• A simple **data model** for scientific datasets
• A **format** for portable, self-describing data
• A **programming library** that uses efficient direct access and efficient subsetting of multidimensional arrays
• Several programming interfaces: C, Fortran, C++, Java, Python, Perl, Ruby, ...
• Support for appending, sharing, and archiving data
The NetCDF-3 Data Model

A file has named variables, dimensions, and attributes. Variables also have attributes. Variables may share dimensions, indicating a common grid. One dimension may be of unlimited length.
NetCDF Usage

• Used in over 60 open source packages for analysis, visualization, and data management and 15 commercial packages
• Basis for popular **CF Conventions** for climate and forecast data
• Used to archive all model output for the IPCC Fourth Assessment Report: 23 models, 30 TBytes, 70,000 files
• Used in many other archives (NOAA, NASA, USGS, DoE, NCAR, BADC, CSIRO, …)
• Other uses in geology, chromatography, mass spectrometry, neuro-imaging, biomolecule trajectory simulations
• C and Fortran netCDF Users Guides have been translated into Japanese at Kyoto University!
NetCDF’s Future

- **NetCDF-4** integrates netCDF with HDF5, another major standard format and data model.
- Parallel netCDF has proved suitable for high-performance computing.
- NetCDF-4 data model (CDM) improves interoperability with other scientific data representations.
- **NetCDF-Java** has advanced features, including access to remote data.
NetCDF-4 Features

Address limitations of netCDF-3

- User-defined compound types (portable structs)
- User-defined variable-length types
- Groups for nested scopes
- Multiple unlimited dimensions
- String type
- Additional numeric types
- Unicode names
- Efficient dynamic schema changes
- Multidimensional tiling (chunking)
- Per variable compression
- Parallel I/O
- Reader-Makes-Right conversion
Commitment to Backward Compatibility

Because preserving access to archived data for future generations is \textit{sacrosanct}

- NetCDF-4 provides both read and write access to all earlier forms of netCDF data.
- Existing C, Fortran, and Java netCDF programs will continue to work after recompiling and relinking.
- Future versions of netCDF will continue to support both data access compatibility and API compatibility.
NetCDF-Java

- 100% Java library has advances compared to C-based interfaces
- Prototype implementation of Common Data Model for access to netCDF-4, OPeNDAP, HDF5
  - Provides netCDF interfaces to other formats: Grids (GRIB1, GRIB2), Radar (NEXRAD, NIDS, DORADE), Satellite (DMSP, GINI), Point Observations (BUFR)
  - Provides uniform coordinate systems layer
- Access to THREDDS inventory catalogs
- Implements virtual access through NcML
Goals of the Common Data Model

• Look at the landscape of scientific datasets from a few thousand feet up
• What semantics are needed to make these useful?
  – georeferencing
  – specialized subsetting
Payoff

N + M instead of N * M things on your TODO List!
Metadata Catalogs and Conventions

Data Push
- IDD
- LDM
- Unidata decoders

Data Pull
- OPeNDAP
- TDS

Metadata
- THREDDS
- CF
- Libcf
- Udunits

Data Access
- netCDF
- netCDF Java
- netCDF-4
- CDM
- CDL
- NcML

Data Analysis, Visualization
- IDV
- VisAD
Thematic Real-time Environmental Distributed Data Services (THREDDS)

- Provides catalogs to help find data
- Catalogs are XML documents (metadata) describing and pointing to datasets accessible via client/server protocols (OPeNDAP, ADDE)
- Datasets may be found by discovery centers (master directories, digital libraries, data portals) via catalogs
- Catalog hierarchy provides places to hang common metadata
- Unidata coordinates THREDDS activities, community implements servers
- Many partners as data providers, tool builders, interoperability experts from academia, government, industry
THREDDS Examples

http://motherlode.ucar.edu:8080/thredds/catalog.html

http://nomads.ncdc.noaa.gov:8085/thredds/
## Dataset

- List of THREDDS catalogs
- Fleet Numerical Meteorology
  - Fleet Numerical Meteorology and Oceanography Center server 1/
  - Fleet Numerical Meteorology and Oceanography Center server 2/
  - Fleet Numerical Meteorology and Oceanography Center ARGO server 1/
  - Fleet Numerical Meteorology and Oceanography Center ARGO server 2/
- IRI/LDEO Climate Data Library/
- NCAR Data Portal/
- NOAA/NCDC NOMADS
  - NCDC NOMADS Data Server/
  - NCDC NOMADS Data Server 2/
- Ocean Watch SUFSC/Environmental Research Division/
- Satellite-Derived Oceanographic Data Sets/
- University of Alabama Huntsville POND server/
<table>
<thead>
<tr>
<th>Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCDC NOMADS GrADS Server 2</td>
</tr>
<tr>
<td>GDAS-BUFR/</td>
</tr>
<tr>
<td>GFDL-CM2.1-DATASETS/</td>
</tr>
<tr>
<td>IGRA_MONTHLY_RAOBS/</td>
</tr>
<tr>
<td>IGRA_STATION_RAOBS/</td>
</tr>
<tr>
<td>NCEP_GFS/</td>
</tr>
<tr>
<td>NCEP_GFS_ANALYSIS/</td>
</tr>
<tr>
<td>NCEP_NAM/</td>
</tr>
<tr>
<td>NCEP_NAM_ANALYSIS/</td>
</tr>
<tr>
<td>NCEP_NARR-A_MONTHLY/</td>
</tr>
<tr>
<td>NCEP_NARR-A_MONTHLY_3hr/</td>
</tr>
</tbody>
</table>
NCEP NAM Individual Run

Dataset: File_Access/NAM_CONUS_40km_conduit_20070225_1800.grib1

- **Data format:** GRIB-1
- **Data size:** 455.0 Mbytes
- **Data type:** Grid
- **Naming Authority:** edu.ucar.unidata
- **ID:** fmrc/NCEP/NAM/CONUS_40km/conduit/files/NAM_CONUS_40km_conduit_20070225_1800.grib1

**Documentation:**

- **summary:** Individual data file, which comprise the Forecast Model Run Collection.
- **summary:** Model runs are made at 00Z, 06Z, 12Z, and 18Z and have analysis and forecasts every 3 hours out to 84 hours.
- **summary:** NCEP North American Model : AWIPS 212 (R) Regional - CONUS - Double Resolution. Horizontal = 185 by 129 points, resolution 40.63 km; Lambert-Conformal projection. Vertical = surface, 1000 to 50 hPa pressure levels, layers, and depth.
- **summary:** NCEP Nonhydrostatic Mesoscale Model (NMM) and Gridpoint Statistical Interpolation (GSI) analysis, running in the Weather Research and Forecasting (WRF) infrastructure.
- **COMET MetEd (Meteorology Education and Training) documentation**
- **NCEP Model documentation**
- **rights:** Freely available
- **processing_level:** Transmitted through Unidata Internet Data Distribution.
- **processing_level:** Read by CDM Forecast Model Run Collection.

**Access:**

1. **OPENDAP:**
   
   http://motherlode.ucar.edu:8080/thredds/dodsC/fmrc/NCEP/NAM/CONUS_40km/conduit/files/NAM_CONUS_40km_conduit_20070225_1800.grib1
2. **HTTPServer:**
   
   http://motherlode.ucar.edu:8080/thredds/fileServer/fmrc/NCEP/NAM/CONUS_40km/conduit/files/NAM_CONUS_40km_conduit_20070225_1800.grib1
3. **WCS:**
   
   http://motherlode.ucar.edu:8080/thredds/wcs/fmrc/NCEP/NAM/CONUS_40km/conduit/files/NAM_CONUS_40km_conduit_20070225_1800.grib1
4. **NetcdfServer:**
   
   http://motherlode.ucar.edu:8080/thredds/ncServer/fmrc/NCEP/NAM/CONUS_40km/conduit/files/NAM_CONUS_40km_conduit_20070225_1800.grib1
Catalog of catalogs in IDV
(Catalog from within a Client)
CDL (Common Data Language)

- A schema language for netCDF data and metadata
- A text representation for netCDF data
- Presents a high-level view of data: dimensions, variables, and attributes
- Notation for examples in **CF Conventions**
- Tools `ncgen` and `ncdump` convert between CDL and netCDF
NcML (NetCDF Markup Language)

• An XML representation of netCDF metadata, similar to CDL
• A schema language for Earth science data
• To get NcML from netCDF data, use ncdump or Java ToolsUI program
• To create netCDF from NcML, use ToolsUI or (eventually) ncgen
Climate and Forecast (CF) Conventions

- A widely used metadata standard for atmospheric, ocean, and climate data, based on netCDF
- Specifies coordinate systems used in models, data cell properties and methods, packing, standard names for quantities, and grid mappings
- CF-aware software can automatically determine space-time location of data variables
- Originally intended for climate model output conventions, but use has broadened to weather and ocean models and observational data
- Community governance structure now in place for maintaining and advancing the CF conventions, WMO Working Group on Coupled Modeling (WGCM)
Libcf

• Purpose: ease creation and use of datasets conforming to the CF Conventions
• In early stages of development and testing
• C and Fortran interfaces available from Unidata in alpha release
Udunits (Unidata Units)

- Library for manipulating units of physical qualities.
  - Conversion of unit specifications between formatted and binary forms
  - Arithmetic manipulation of unit specifications
  - Conversion of values between compatible scales of measurement
- C, Fortran, and Java interfaces
- Required by **CF conventions**
Putting It All Together: The Integrated Data Viewer

- IDD
- LDM
- Unidata decoders

Data Push

- OPeNDAP
- TDS

Data Pull

Data Access
- netCDF
- netCDF Java
- netCDF-4
- CDM
- CDL
- NcML

Data Analysis, Visualization
- IDV
- VisAD

Metadata
- THREDDS
- CF
- Libcf
- Udunits
Integrated Data Viewer (IDV)

- Unidata’s newest scientific analysis and visualization tool
- Freely available 100% Java framework and reference application
- Provides 2- and 3-D displays of geoscience data
- Stand-alone or networked application
- Integrates data from disparate sources
- End-to-end test for Unidata technologies
Some IDV Features

- Client-server data access from remote systems
- Suite of data probes for interactive exploration (slice and dice)
- Animations (temporal and spatial)
- HTML interface for pedagogic materials
- XML configuration and bundling allows collaboration with other educators
- Java-based framework supports Extensions built via plug-ins: e.g. for geosciences network (GEON) solid earth community
A Few Last Thoughts on Infrastructure …
What Is Good Infrastructure?

• Provides a useful service
• Makes abstractions at the right level
• Cloaks invisible details with a simple interface
• Binds loosely to other infrastructure
• Behaves reliably
• Adapts easily to changes
An Example of Great Infrastructure: Popular Programming Languages

- Base of huge collection of higher layers of infrastructure
- People continue to build on top of this infrastructure
- The opportunity to create a long-lasting and popular programming language is rare
- Jim Backus (Fortran), John McCarthy (Lisp), Dennis Ritchie (C), Bjarne Stroustrup (C++), James Gosling (Java), Yukihiro “Matz” Matsumoto (Ruby)
- Other great infrastructures: Unix, TCP/IP, HTTP, …
Rewards of Developing Infrastructure?

• It “raises the level” for other developers
• Beautiful and useful new layers and applications are built on top of it
• You can feel a part of everything it supports
• If it’s long lasting and widely used, you have made a difference for future generations
• So, it’s one way to get closer to immortality
• Infrastructure is abstract, but rewards can also be real
• … like this trip to Japan!
For More Information

support@unidata.ucar.edu
russ@ucar.edu
http://www.unidata.ucar.edu/