Toward the Development of a Comprehensive Data & Information Management System for THORPEX

Mohan Ramamurthy, Unidata
Steve Williams, JOSS
Jose Meitin, JOSS
Karyn Sawyer, JOSS

UCAR Office of Programs
Boulder, CO

First THORPEX International Science Symposium
December 8, 2004
Montreal, Canada
THORPEX Science Drivers

- Predictability and Dynamical Processes;
- Observing Systems;
- Data Assimilation and Observing Strategies;
- Societal Impacts and Economic Applications;

To combine basic and applied research to benefit operational forecasting and decision support systems.

Activities related to these objectives are very data intensive!
Data and Information Management System

• Data Management is a critical cross-cutting element of THORPEX

• A comprehensive, seamless, and an end-to-end data and information management system (DIMS) is required

• It must address the needs in all four areas
  – Unlike many past projects, a THORPEX DIMS must also facilitate societal impacts and human dimension components of the program.

Goal: To ensure that data and products needed for research and education are readily available to users in a convenient form
Technology Confluence: Enabling A New Generation of Data Services

- Web services – modern mantra for data services
- Extensible Markup Language (XML)
  - Will do for data for html has done for the text
- Commodity microprocessors
- Object-oriented programming
- Open standards
- Global, high-bandwidth and wireless networks
- Digital libraries
- Collaboratories
- Grid Computing
- Data Portals and Distributed Servers
- Open Geographic Information Systems
- Data mining and knowledge discovery

To borrow a tag line from an automaker, this is not your father’s DIMS
Challenges

- Heterogeneity and complexity of distributed observing, modeling, data, and communication systems
- Nature of data coverage: diversity and multiple spatial and temporal scales
- Use of legacy and contemporary technologies and tools
- Lack of standards and interoperability
- User community not monolithic
- Political, technological and cultural and regulatory barriers
- Data integration with GIS

Myriad data distribution systems currently in place: GTS, NOAAPort, EUMetcast, MDD, LDM, FTP, OPeNDAP, ... (Push and pull systems)
Key point: Need to get started early!
Data Management Considerations

- **Data policy, protocol, and plan**
- Data types (observational, model, real-time, archived, case study)
- Data Availability, Acquisition, Transmission, and Collection
- **Data Formats, Standards, and Documentation**
- Data Classification: Adopt FGGE classification system?
- **Key Dependencies**
- Real-time Data Requirements
- Data Processing, Quality Control, and Quality Assurance
- Data Archival
- Data Access methods and protocols
- Data Distribution methods and protocols
- **Metadata creation and inventory for data discovery and integration**
- **Coordination with other Programs**
- **Data Integration and multidisciplinary synthesis**
- **Scalability, flexibility and interoperability of data management system**
- **Tools: processing, decoders, analysis and visualization**
- Support

There are many considerations, but the details need to be agreed upon and worked out.
Desired Characteristics of the Data Management System

• Facilitate rapid exchange of and easy access to large volumes of data, products and services

• Integrated system that delivers a suite of products to support myriad data user communities – researchers, educators, forecasters, policy and decision makers, etc.

• Process to assure data quality and uncover errors in data early, to allow proper interpretation and use
Some Guiding Principles

• A well-established set of guiding principles are needed for a successful data management plan
  – Free and open sharing of data and software
  – Timely availability of data
  – Compliance with WMO resolutions
  – Along with data, good (complete) metadata provision
  – Scalability, via distributed computing
  – Build on existing national and international data systems
  – Heterogeneous system of integrated systems
  – Adaptability to changes in technology, data availability, and user needs
  – Tailor it to the needs of communities rather than individual users
  – Minimal data processing required on user end

In addition to data, provision of tools and support are important
Broad Data Categories

• THORPEX must provide a seamless, end-to-end and access to
  – Real-time data
  – Archived data
  – Field and Demonstration Project and Regional Campaign data
  – Episodic (Case Study)
  – GIS databases

• The DIMS should be well integrated across all these data categories across atmospheric and related sciences

• In addition to data, a broad set of tools should be provided for the use of data in education and research
There are over 150 university sites in North and South America, Europe, and Asia that receive real-time data using the Unidata Local Data Manager; Plus there are over 300+ LDM sites in NWS, NOAA, NASA, KMA, Taiwan, and Spain that are not part of the “open” IDD.

Through the MeteoForum Project, real-time data is now being distributed to universities and WMO RMTCs in S. America and the Caribbean
SuomiNet in Action

- A network of GPS receivers to provide real-time atmospheric precipitable water vapor measurements and other geodetic and meteorological information

- SuomiNet collects data from 100+ GPS receivers distributed throughout the world.

- The observations are sent to Boulder, CO for processing and analysis and then redistributed to the community using the LDM.
OPeNDAP/DODS

- Open Source Project for a Network Data Access Protocol
- OPeNDAP is the discipline independent core infrastructure for remote data access

  - One can design a system of distributed data servers using the OPeNDAP protocol
  - OPeNDAP servers deal with data population and provision.
  - This has been done for real-time and retrospective operational model data (NOMADS, THREDDS, Community Data Portal)
  - Set up an International Virtual THORPEX Data System?
  - Thin and thick client applications can access and use data on distributed OPeNDAP servers
  - The data can be cataloged using THREDDS and published via digital libraries
OPeNDAP/DODS Servers
Thematic Real-time Environmental Distributed Data Services (THREDDS)

To make it possible to publish, locate, analyze, visualize, and integrate a variety of environmental data.

Combines “push” with several forms of “pull” and digital library discovery.

Connecting People with Documents and Data
The First Global Integrated Data Sets of the Water Cycle

Model Outputs by Numerical Weather Prediction Centers

Surface Observational (in-situ) Data from the 33 CEOP Reference Sites

Satellite Remote Sensing Data

In-Situ Data Archiving Center at UCAR (Center at University Corporation for Atmospheric Research) of USA
http://www.ucar.edu/

MODEL Output Data Archiving Center at Max-Planck Institute of Germany
http://www.mpg.de/

Data Integrating/Archiving Center at University of Tokyo and NASDA of Japan
http://monsoon.t.u-tokyo.ac.jp/ceop/

Global Land Data Assimilation System at NASA Goddard Space Flight Center of USA
http://ldas.gsfc.nasa.gov/

Observations

4DDA

Model

New Products

Input of Observed Data Into Model

Model Integration
Integrated Data Viewer

• Unidata’s newest scientific analysis and visualization tool – freely available
• Provides 2, 3 and 4-D displays of geoscientific data
• Stand-alone or networked application, providing client-server data access via multiple protocols
Next Steps?

• Organize a Data Management Working Group ASAP
• Plan and organize a Data Management Workshop, assembling all stakeholders
• Form subcommittees to deal with focused issues (e.g., data, tools, support, coordination)
• Develop policies, protocols and timelines
• Ask the DMWG and workshop participants to draft a data management plan (both strategic and tactical) and begin designing the THORPEX Data and Information Management System

Bottom Line: Get organized ASAP and begin addressing the issues!
Possible topics for a Data Management Workshop

- Overall data management strategy
  - Data acquisition, processing, distribution, archiving, sharing, quality assurance
- Develop use case scenarios and based on that define requirements and functionalities
- Compile data inventory, list of available tools
- Begin designing the overall architecture
- Develop plans for governance, management, and coordination of activities
Generic Data Management Plan Outline

1.0 Introduction/Background
1.1 Scientific Objectives
1.2 Data Management Philosophy

2.0 Data Management Policy
2.1 Data Protocol
2.2 Data Processing/Quality Control
2.3 Data Availability
2.4 Data Attribution
2.5 Community Access to Data

3.0 Data Management Functional Strategy/Description
3.1 Data Archive and Analysis Centers
3.2 Investigator Requirements
3.2.1 Data Format Conventions
3.2.2 Data Submission Requirements
3.3 Data Collection Schedule
3.3.1 On-line Field Catalog
3.4 Data Processing following the Field Phase
3.5 Data Integration
3.6 Data Archival and Long-term Access

4.0 Data Sets
4.1 Data Collection/Processing
4.2 Status Update Procedures
4.3 In-field Data Display and Analysis Requirements
4.4 Coordination with other Programs

APPENDICES
A. Research Data Sets
B. Operational Data
C. List of Acronyms