Building and Sustaining International Scientific Partnerships Through Data Sharing

Mohan K. Ramamurthy, Tom Yoksas, and Linda Miller

Unidata, University Corporation for Atmospheric Research, (mohan@ucar.edu, P. O. Box. 3000, Boulder, CO 80307 USA, Phone: 303-497-8661, Fax: 303-497-8690)

Introduction

Increasingly, the conduct of science requires strong international scientific partnerships and sharing of knowledge, information, and other assets. This is particularly true in the geosciences where the highly coupled nature of the earth system and the need to understand global environmental processes and their regional linkages have heightened the importance of strong collaborations across national and continental boundaries. The climate system, for example, is far too complex a puzzle to be unraveled by individual nations. As science becomes increasingly global in nature, it is critical that focus is placed on full, open, and timely access to and sharing of earth system science data.

For the past two decades, the NSF-sponsored Unidata Program Center (UPC) of the University Corporation for Atmospheric Research (UCAR) has been providing data, tools, and support to enhance Earth-system education and research. In an era of increasing data complexity, accessibility, and multidisciplinary integration, Unidata provides a rich set of data services and tools. Beginning as a collection of US-based, mostly atmospheric science departments, the Unidata community has grown to include government agencies and private sector entities, and today that community also transcends international boundaries. The primary reason for the community broadening, which has in large part occurred organically through the free and open exchange of near real-time geo-scientific data and related software, is a recognition that most of today's formidable scientific problems in the geosciences are inherently multidisciplinary and global in character. As articulated in the draft NSF Strategic Plan: FY 2006-2011 (NSF, September 2006), "discovery increasingly requires expertise of individuals from different disciplines, with diverse perspectives, and often from different nations, working together to accommodate the extraordinary complexity of today's science and engineering challenges." The document further states, "…the ability to develop collaborations that create new value for the partners is often the limiting factor for progress in critical areas of science, engineering and technology."

The Internet and its myriad manifestations, including the World Wide Web, have amply demonstrated the compounding benefits of a global cyberinfrastructure and the power of networked communities as institutions and people exchange knowledge, ideas, and resources. The Unidata Program recognizes those benefits, and over the past several years it has developed a growing portfolio of international outreach activities, conducted in close collaboration with academic, research and operational institutions on several continents, to advance earth system science education and research. The portfolio includes provision of data, tools, support and training as well as outreach activities that bring various stakeholders together to address important issues, all toward the goals of building a community with a shared vision. The overarching goals of Unidata's international activities include:

- democratization of access-to and use-of data that describe the dynamic earth system

- building capacity and empowering geoscientists and educators worldwide
- strengthening international science partnerships for exchanging knowledge and expertise
- effectuating sustainable cultural changes that recognize the benefits of data sharing, and
- building regional and global communities around specific geoscientific themes.

The Unidata Program continues to place high value on the transformational changes and the increasing importance of international scientific partnerships and proposes to continue fostering such collaborations and related efforts toward the building of a globally-engaged community of educators and researchers in the geosciences. The vision for Unidata and the strategic plan currently under development are informed by these trends and emphasize the need for continual organic growth of the community both internationally and into other geoscience disciplines.

Data Access and Distribution

A critical component of successful scientific inquiry includes learning how to collect, process, analyze, and integrate data from myriad sources, and geo-science education is uniquely suited to making science relevant by drawing connections between the dynamic Earth system and societal impacts. In this section, we briefly describe some of Unidata's successful efforts in facilitating data access, use, and integration in the geosciences.

MeteoForum and International Data Distribution

The importance of sharing locally-held data was recognized in the earliest Unidata planning documents of the mid-1980s. Development of NSFnet and its successors provided the substrate on top of which a multi-way communications system could be built. The Unidata-developed Local Data Manager (LDM) evolved to be the vehicle that enabled the multi-way sharing of data through a project known as the Internet Data Distribution (IDD) system. Although Unidata has long fostered and maintained international interactions, an initiative starting in 2001 called MeteoForum was its first organized and natural extension into an international arena. Funded by the UCAR Office of Programs' STORM Funds, the MeteoForum pilot project, a joint effort between Unidata and COMET, had the following overarching goals:

The MeteoForum pilot project will include a small group of educational institutions (some universities and some WMO RMTCs) that are motivated to enhance the contributions of modern meteorology in their regions. Participants will be expected to have relatively fast Internet access, appropriate computers, and suitable personnel. Some of these personnel will be trained to run MeteoForum software on their computers so as to access real-time data, training materials, and other resources. Where practical, participants in the MeteoForum pilot also will *contribute* real-time data and educational resources to the effort. By integrating these elements, the pilot project will serve as a model on which to build a full-scale international MeteoForum. Initially, the MeteoForum pilot project will build upon capabilities now offered in the U.S. by the government-sponsored COMET and Unidata programs.

To achieve the above goals, Unidata carried out the following MeteoForum activities:

- Facilitate data access to a broad spectrum of observations and forecasts
- Coordinate a data-relay network that collects and distributes data in real-time at no cost to educators and researchers
- Build a community where data, tools, and best practices in education and research are shared
- Support faculty at research and educational institutions in the use of Unidata systems

MeteoForum is a success story of organizations that leveraged their expertise in a collaboration that resulted in the creation of a data distribution system for South America, the ***IDD-Brasil***. This success was built collaboratively among the UPC and several Brazilian institutions including the Universidade Federal do Rio de Janeiro, the Universidade de São Paulo, the Universidade Federal do Pará, and the Centro de Previsão de Tempo e Estudos Climáticos (CPTEC, a division of INPE). The data relay infrastructure established in Brazil coupled with the North American ***IDD*** represents the beginnings of a hemisphere-wide network that acts as conduit for multi-way sharing of international, national, and locally-held environmental datasets. For the first time, previously unavailable observational data and high resolution model output for Brazil are now being made available to both Latin American ***IDD-Brazil*** and North American ***IDD*** participants in near real-time. Real-time atmospheric science data delivered to Latin America by the ***IDD***/***IDD-Brasil*** has helped initiate teaching innovations in multiple geo-science disciplines in Argentina, Brazil, and Chile.

Currently, over 160 institutions of higher education in North, Central and South America, the Caribbean, Asia, and Europe are participating in the IDD. Profound and transformative impacts have already been noticed since the distribution system expanded beyond North American borders some three years ago. For example, data delivery to Central and South America has initiated teaching innovations at many universities in those regions, including the University of Costa Rica, the University of Buenos Aires, and Universities of Rio de Janeiro and São Paolo. Integration of real-world data has provided opportunities for active, student-centered and inquiry-based learning, infusing the excitement of discovery into geo-science courses at these institutions.

The ***IDD*** can be an invaluable tool for learning more about differences in atmospheric phenomena and processes in different geographical regions. For example, every tropical meteorology textbook states that hurricanes do not occur in the South Atlantic Ocean. Imagine the befuddlement of the meteorological community when forecasters followed the development of the first ever recorded hurricane off the coast of Brazil in March, 2004. Hurricane Catarina was significant for two reasons: a) it is forcing the reevaluation of conventional wisdom; and b) it could potentially be a climate change signal. Researchers believe that the South Atlantic region is one of the areas to watch for increased tropical cyclone activity in a warmer global climate. Looking to the future, the ***IDD*** will provide an important resource to scientists across the two hemispheres who can investigate, among other things, geographic differences in atmospheric and oceanic processes and circulations.

The democratizing and transformative effects of free and open access to data on atmospheric science research and education cannot be overstated. For example, the ***IDD*** system is providing important benefits to the Antarctic meteorological community. Because of

communication and logistical difficulties, the provision of data to Antarctic researchers, educators and forecasters has been a significant challenge, and these challenges are being overcome by the ***Antarctic-IDD***, which carries surface and upper air observations, satellite imagery, and forecast model output to an increasing number of participating nodes, including one at the US McMurdo Station. The availability of observations from polar areas is especially crucial for documenting the nature and extent of climate change, for those are the very regions that are projected to experience the most significant warming in climate simulations and as such most vulnerable from an Earth system science perspective.

TIGGE, (THORPEX Interactive Grand Global Ensemble) is a key component of THORPEX--a framework for central access for the complete set of forecast ensembles that will be combined into a single ensemble system. NCAR, in collaboration with ECMWF and the Chinese Meteorological Agency (CMA), is seeking to establish identical international data repositories for the TIGGE data which will be output from THORPEX. Data from the project needs to be received and archived and accessible to ensemble forecasts from several international numerical weather prediction centers. The data have to be quickly available through a portal and secure long-term archives and services need to be implemented. NCAR's CISL/SCD will be one of the long-term archives. TIGGE will use the LDM to enable the global data receipts to build the TIGGE archives and probably will be used to fulfill subscription data requests.

When fully implemented, the TIGGE archive will provide researchers and educators access to global model output from ten different operational numerical weather prediction centers, including ECMWF (Europe), UK Met Office (UK), NCEP (USA), MSC (Canada), JMA (Japan), CPTEC (Brazil), BMRC (Australia), CMA (China), KMA (Korea), and FNMOC (US). Continued collaborations resulting from the availability of these data will result in greater understanding of a range of geo-scientific problems that include advances in climate change science, weather and ENSO predictions, and hydrologic processes. This will provide richer global analyses of the state of the planet. Notably, what once was the province of a few in elite US universities is now available to many, thanks to innovations made possible by Unidata's sustained cyberinfrastructure, development and deployment, and support for international outreach.

THREDDS

It has long been realized that the **push** data delivery method employed in the ***IDD*** is not applicable when the volume of data to be moved exceeds the capacity of a site's network capacity, or when each site desires different subsets of data from a large distributed data collection. Since the volume of geo-scientific data slated to become available in the near future will be several orders of magnitude greater than what is available today, Unidata embarked on a project aimed at providing programmatic remote access to collections of data. The THREDDS (Thematic Realtime Environmental Distributed Data Services) project is developing middleware to bridge the gap between data providers and data users. The goal is to simplify the discovery and use of scientific data and to allow scientific publications and educational materials to reference scientific data.

THREDDS is truly international in scope. Contributors to THREDDS development and/or deployment include the UK's British Atmospheric Data Center (BADC), and organizations

in France, Greece, Japan, Italy, and Australia. Users of the prototype THREDDS server hosted by the UPC at UCAR include Brazil and China as well as those contributing to development and deployment.

Tools

In addition to data provision, Unidata develops, maintains, and supports a variety of software packages for data access, management, analysis, and visualization. They include analysis and display software such as GEMPAK, McIDAS, and IDV, as well as other middleware like netCDF, UDUnits, and THREDDS Data Server. Most of these packages are developed at the UPC, while a few others originated externally, but are modified, maintained, and supported at the UPC. Software provided by Unidata is available at no charge to users worldwide. All these tools are downloaded and possibly used broadly by the international community[1], although their use varies markedly from application to application and country to country. While netCDF, Unidata's most widely used software, is used in 78 countries, even a relatively new application like the Integrated Data Viewer has been downloaded by users in over 60 countries.

Cyberinfrastructure Leadership

The UPC has also been active in organizing meetings to bring communities together to address important cyberinfrastructure challenges and opportunities. For instance, the UPC staff proposed, organized, and held special sessions at recent European Geophysical Union (EGU) and American Geophysical Union (AGU) meetings:

It should be added that there is a serious push to make the Earth and Space Sciences Informatics (ESSI) session international by promoting active Latin American and European participation.

> "***Earth and space science Cyberinfrastructures: Data, Tools, Distribution and Forecast Systems for International Collaboration"***

The underlying idea of this session, organized with NSF sponsorship and collaboration with members of the community, was to help build a stronger and broader space and Earth science cyberinfrastructure community through international collaboration. The session encouraged papers on bio-geophysical, meteorological, ecological, remote sensing, and related intelligent and adaptive systems for data collection, processing, and knowledge discovery and classroom dissemination. Of special interest were presentations on data, tools and forecast systems that foster collaborations among the research and education communities in the Western Hemisphere.

The UPC strongly encouraged participation of Latin American scientific and higher education organizations in this session.

[1] Downloads may not equate to usage.

Some Thoughts on Future Activities

An increasing number of US and international institutions rely on the UPC's expertise in providing access-to and use-of a wide variety of data; provision of tools to analyze and visualize those data; provision of support in the use-of data; and in community leadership and advocacy. The UPC has furthered its reputation for excellence through successful efforts in a variety of domestic and international projects. Most of the international activities that have been undertaken to date have been *ad-hoc*; they have been undertaken on a low priority basis in addition-to fulltime support and advocacy for the US education and research community. Given the availability of limited resources, we expect this modus operandi to continue, although UPC will continue to facilitate the ongoing deployment of the ***IDD-Brazil*** throughout South America and extend ***IDD-Caribe*** to serve Central American and Caribbean research and educational institutions. Other potential activities include:

- informing institutions of the existence of the data sharing through expanded outreach and advocacy activities at international conferences
- assisting in the installation and configuration of application servers akin to ones installed at the Universidad de Costa Rica and the Caribbean Institute for Meteorology and Hydrology as part of the MeteoForum project
- assisting sites in the use of real-time data by providing the same display and analysis applications used throughout the existing North American Unidata community at no cost
- assisting in connecting to the networks during site visits and conducting short training sessions that will allow sites to effectively operate their data ingest/relay nodes
- providing remote troubleshooting if and when problems arise
- assisting in the development of regional training workshops for use of Unidata systems in research and education
- assisting institutions in expanding their data distribution roles within their country and to other countries, especially to the U. S.
- entraining users into the greater Unidata community through their active participation in email discussion groups and workshops
- promoting the sharing of locally held datasets of interest with the entire Unidata community
- working with sites to develop THREDDS data cataloging and remote data access services
- in collaboration with the community, convening and organizing regional workshops and meetings to bring communities together to address issues of importance to Unidata
- promoting strategic partnerships centered around cyberinfrastructure and data sharing efforts

Clearly, the extent to which Unidata will engage in the above activities is dependent on the availability of resources and other programmatic and strategic considerations.