Technology Solutions for Scientific Data Interoperability: Unidata's Perspective

October 2011

Russ Rew

Unidata Program Center
UCAR Office of Programs
P.O. Box 3000
Boulder, CO 80307-3000

Mohan Ramamurthy, Director

1 Introduction

The Unidata Program has been navigating the rapids of technological change for over twenty-five years, while building and delivering cyberinfrastructure solutions for a growing community of researchers and educators. From its perspective as a small organization providing software, services, and support to a growing geosciences community, Unidata is in a position to offer recommendations for applying technology solutions to overcome barriers that hamper interdisciplinary research.

To better understand the nature of such barriers, consider a scenario involving multiple islands, each with their own ocean model. The models are tailored for the needs of island inhabitants, so they differ in various respects, such as which physical quantities are output, how time is represented, and what spatial coordinate systems are used.

On each island, an evolving collection of software applications depend on local model outputs, satisfying specific needs of the island's inhabitants and decision makers. These applications also make use of archives of past observations and model outputs to detect trends and derive information that will be useful for estimating fish populations, seasonal wave heights, potential for beach erosion, and other such useful knowledge.

The plot thickens when a "cross-island" researcher needs to perform an analysis that requires accessing output from multiple models to create a single visualization.

Approaches to solving such interoperability problems include:

1. Develop a set of conversion tools to convert each island's model output representation into a common form, then download and convert the data.

2. Develop or choose a standard for island model outputs that is general enough for use on all the islands, and somehow get each island to commit to use that standard.

3. Wait for development of a system that uses semantic web technologies, with metadata registries for machine-readable descriptions of data and services, software that can use such descriptions to configure service providers and clients, and enough intelligence to mediate among data producers, data consumers, and registries to perform the transactions needed to satisfy user requests.

4. Develop a more modest service interface comprehensive enough to support access to each island's unique data representation. Make provision of the service interface simple for data providers, without modifying the data accessed, by deployment of a server designed for that purpose.

Approach 1 requires no effort on the part of model or application developers on each island. It might require a great deal of effort to maintain such conversion tools as island-specific data representations continue to evolve, but the cost of such efforts could be amortized over use by other cross-island researchers or projects.

Approach 2 requires an adequate standard and models modified to output results in that standard form. Unless adapted to output results in both local and standard form, this approach also requires changes to local applications to adapt to the new output standard,

changes to archives to conform to the standard, and a strong incentive to use the standard. This solution might make the cross-island researcher's job easier, but would be costly for model and application developers on every island.

Approach 3 is a description of what research may eventually provide, but most system architects would agree that more research is needed before it reaches maturity.

Approach 4 requires the definition of abstract service interfaces, the development and maintenance of an associated server, the deployment of the servers on each island, and the development of software that would make it practical to access the different model outputs on-the-fly and convert them to a common data model needed by the server.

Our researcher notices that the software for approach 4 is already available, uses it successfully, and brings this scenario to a happy ending.

This story is a simpler version of an actual recent interoperability success [Signell 2011] using Approach 4, involving:

- Various ocean and coastal models (ROMS, POM, ECOM, WW3, WRF, and FVCOM) run for regional ocean observing centers
- A variety of formats and conventions for model outputs
- Many application clients making use of web services to access, analyze and visualize the data (Matlab, ncWMS, Unidata IDV, DIVE, ERDDAP, ArcGIS, Geoportal, FERRET, GrADS, ...)

Making the model outputs interoperable required only the installation of a specific server at each coastal center and use of small text files to adapt local model outputs to a common data model. As a result, each application client could view or run analyses on each model output as if they all conformed to a common standard, *without changing the models or their outputs*. Building on this success led to a subsequent and more ambitious effort: NOAA's Unified Access Framework for gridded data, making data from numerous data centers interoperable using the same technologies: Unidata TDS data servers and NcML wrappers [Hankin 2011].

2 A Practical Vision

A practical vision for developing EarthCube begins with early delivery of technologies that provide simple solutions to a useful core subset of user requirements. The early solutions would be incomplete in several respects, for example scale of problems handled, diversity of data types, generality of the data model, and performance of some operations. Later releases would incrementally expand the range of problems solved by better scaling to handle larger problems, handling additional data types, or removing other restrictions that earlier releases required for ease of implementation and deployment. Improvements in technologies would be integrated into later implementations.

Unidata's experience in providing technology solutions for interoperability has followed this pattern, evolving simple solutions with simple implementations that satisfy a subset of user requirements, to more complete solutions with more complex implementations, continuously improving software that is backward compatible for previous uses but faster, more comprehensive, and more capable for new uses.

3 Useful Current Technologies

Unidata has experience with developing and supporting several current technologies that might be useful as examples in planning an ambitious undertaking such as EarthCube:

- A simple scientific data model for uniform access to self-describing data
- Virtual dataset wrappers providing subsets, aggregations, and additional metadata for data collections
- Subscription services to real-time, self-managing data streams
- A framework for 4-D interactive integration, visualization, and analysis of selected subsets of remote data
- A discipline-specific conventions layer above discipline-independent infrastructure

A simple scientific data model: A suitable abstract data model facilitates access to data in various forms, through a single unified application programming interface (API) and associate services to achieve interoperability. Higher layers of the data model can support access to coordinate systems and to scientific feature types, such as gridded data, time series, and observational data on discrete sampling geometries. The Unidata Common Data Model is an example, unifying the netCDF, OPeNDAP, and HDF5 data models to support access to many file formats for which plug-ins have been developed.

Virtual dataset wrappers: Logical views of scientific datasets can be implemented using small virtual dataset wrappers that reference actual data, add metadata to achieve standards compliance, or aggregate existing datasets for ease of use, all without modifying the referenced data. A non-standard collection of existing data can be made to appear compliant to a standard by use of such wrappers, when accessed through appropriate software. NcML, an XML dialect for the Common Data Model, is an example of such a technology.

Subscription services for data streams: Timeliness of data delivery is important for assimilating observations into forecast models. The ability to subscribe to specified subsets of available data streams and to notification services is crucial for providing decision makers with timely and relevant information. Subscriptions to information about changes in data archives supports efficient incremental processing of derived analyses. Unidata has developed and supports event-driven software for the Internet Data Distribution system, which provides near real-time observational data and timely forecast model outputs to universities, US government agencies, and other organizations and projects worldwide.

Integrated visualization and analysis of remote data subsets: An advanced visualization and analysis framework is needed for interdisciplinary geoscience data, assisting in exploration of a wide range of local or remote data, such as satellite imagery, gridded model results, and observations, within a unified interface. It should support accessing subsets of large remote datasets, providing multiple 2-and 3-D data displays within a common display, as well as a rich set of analysis capabilities and interactive or script-based generation of products such as animations in various formats. Such an advanced application could help determine gaps in geoscience cyberinfrastructure and provide a context for testing real use cases. Unidata has developed an example of such an application, the Integrated Data Viewer (IDV).

A discipline-specific conventions layer: Conventions are community agreements restricting metadata representations to limit the number of equivalent possibilities with which software must deal, to foster interoperability. Although a human may be able to ignore gratuitous differences and recognize a variety of metadata representations as equivalent, it is difficult to

write software that handles such differences. Conventions that select a single way to represent metadata make it practical to write software that "understands" the metadata. The resulting uniformity of access supports building applications with powerful extraction, regridding, analysis, visualization, and processing capabilities. Unidata has collaborated in the development of several community conventions.

4 Involvement with partners

Important collaborations and partnerships have proved of vital importance for the development of existing cyberinfrastructure for the geosciences. Unidata's partnerships with education, research, government, and commercial organizations have included:

- The University of Wisconsin Space Science and Engineering Center, NASA, and NOAA in development and maintenance of visualization and analysis frameworks
- A global community of data providers, researchers, and developers self-organized to help evolve and implement software for the Climate and Forecast Metadata Conventions
- The Open Geospatial Consortium (OGC) for developing interoperability demonstrations, web services, and data encoding standards
- The HDF Group in development of a new data model, format, and reference library
- The OPeNDAP.org organization and user community to support and implement remote access in client and server software
- NOAA NCDC in the provision of high-resolution forecast model outputs and radar data for use in research
- Developers of widely used commercial software such as ArcGIS, Matlab, and IDL for support of scientific data access
- Numerous individual developers of open-source software for data management, analysis, and visualization

5 Lessons learned

Principles and lessons gleaned from developing infrastructure for the geosciences community since 1989 include:

1. One size will not fit all
2. Nurture short user feedback loops
3. Involve developers in support
4. Leverage community efforts
5. Emphasize discipline-independence
6. Favor loose coupling among components
7. Drive development with tests

One size will not fit all: Problems in the real world are rarely amenable to a universal solution, but a number of distinct solutions can often be hidden behind a simple interface to provide the illusion of uniformity instead of complexity. This is the most basic pattern of interoperability solutions, moving the complexity of dealing with a diversity of representations into the infrastructure, so users and data providers see what appears and behaves like uniformity but is actually implemented as an extensible framework of plug-ins or agents, dealing with multiple cases that must be handled separately, behind the curtain.

Nurture short user feedback loops: The value of users who will provide quick feedback is hard to overstate. Interested and enthusiastic users are the source of great ideas for improvements as well as bug reports and quick evaluations of bad ideas. Working within a community large enough to include such users accelerates development of high-quality software.

Involve developers in support: Some organizations attempt to shield developers from user support, because support is considered a low-level activity best handled by less highly skilled staff. In Unidata's experience, having developers involved in support leads to improved understanding of user needs and priorities, appreciation for the value of concise and accurate documentation, and novel ideas for solving problems in ways that decrease the need for future support.

Leverage community efforts: Open-source development encourages bug reports that are accompanied by bug fixes, contributions of plug-ins to framework architectures, ports to new platforms, and adaptations to new scientific problems. Making the status of plans and progress transparent encourages valuable suggestions from users and gives the community a sense of ownership and participation in large endeavors.

Emphasize discipline-independence: The temptation to optimize a solution for a specific problem area or discipline is hard to resist, but keeping infrastructure discipline-independent pays large dividends in generality and usefulness to a broader community. Choosing an appropriately high level of abstraction in designing scientific data infrastructure helps to amortize the development and support costs across many communities, long time scales, and a large variety of useful applications. A layer of standard conventions created and maintained above the data model provides customization for specific communities of practice.

Favor loose coupling among components: Loose coupling is a principal of component and service-oriented architectures that has many benefits. It requires agreement on simple, abstract interfaces before independent development of components can proceed. It helps keep components independent of each other, so that they can be flexibly composed to solve problems. It reduces complexity by eliminating dependencies among parts, which facilitates diagnosing and fixing bugs. It encourages modularity and proper layering of systems and frameworks that are necessary for durable infrastructure.

Drive development with tests: Test-driven development emphasizes creating tests before development and maintaining them along with the code. It results in a large number of automated tests to verify that software captures the intent of the developer and fixes reported bugs. Evolving a large collection of such tests improves the stability, robustness, and portability of software, and supports confidence needed to periodically refactor code to improve its maintainability. Testing and deploying software on multiple platforms improves the quality and maintainability of the code in surprising ways, because it reveals hidden assumptions that might lead to bugs on future platforms.

6 Future developments

Anticipating which future technologies will be useful in an infrastructure designed for deployment a decade or more in the future is somewhat foolhardy. A less risky plan is to enhance or replace existing working solutions in an incremental fashion, using a process of accelerated evolution and guided competition, resulting in not just survival but flourishing of the fittest technologies.

From the current perspective, it appears likely that further development of cloud computing, semantic technologies, web applications, massive parallelism, wavelet representations, and persistent memory offer promise for contributing to progress toward achieving the ambitious EarthCube vision. With data from sensors and models growing exponentially, and metadata connecting linked data growing even faster than the data it references, geoscience cyberinfrastructure will be a challenging test bed for grounded research in these and other technologies.

7 Conclusion

The concept of islands of non-interoperability, separated by distance, disciplinary cultures, and generational time spans, may be a useful metaphor for where the geosciences may be heading in the absence of intervention.

Two futures diverge in developing cyberinfrastructure for the geosciences. The first leads to growing isolation into islands of non-interoperability separated by stormy oceans across which data access and transport lose meaning or intent. A second and more promising future requires connecting isolated islands with social, cultural, and electronic networks, bridges, causeways, and scalable infrastructure that leads to sharing of data, information, and knowledge. We need to begin building the infrastructure that preserves meaning and richness in data from different disciplines and communities of practice. We're only at the beginning of knowing how to do this, and it will require community cooperation and wise leadership, such as an undertaking like EarthCube may engender.

8 References

[Signell 2011] Rich Signell (USGS, Woods Hole), The US-IOOS Modeling Testbed Cyberinfastructure: Unstructured Grid Standards and Standards-Based Tools for Analysis of Ocean, Atmosphere & Climate Model Data, presented at GO-ESSP Workshop, Asheville, NC, May 10-11, 2011, http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/Signell-2011-05-11_GO-ESSP_Asheville_v2.pdf

[Hankin 2011] Steve Hankin (NOAA PMEL), The Unified Access Framework (UAF), presented at GO-ESSP Workshop, Asheville, NC, May 10-11, 2011 http://nomads.ncdc.noaa.gov/go-essp/presentations/goessp/weds/Hankin-UAFatGOESSP2011-v2.pdf