

In Data Driven Computational Science the Systems Issues Are the Pressure Point to Information Integration

Beth Plale, Associate Professor
Department of Computer Science, Indiana University

Information integration in data driven computational science takes a complex form because multiple levels of interactions must be supported if the advancement of science is to continue. In its most general form, information integration is the ability of entities to communicate with one another about data. Integration often requires discussion and agreement on methods, formats, schemas, and languages. Our position is that information integration has to occur at multiple levels of abstraction, and that the model of abstraction we present here is reasonable. Integration strategies that ignore one or more levels are incomplete. Strategies that favor higher level interactions over lower level ones are insufficient, and run the risk of inhibiting scientific progress. We argue that the model we propose holds for computational science, and show one approach, namely, our work in the NSF funded large ITR, Linked Environments for Atmospheric Discovery (LEAD) project.

Computational science is an emerging science that uses advanced computing capabilities to understand and solve complex problems. It fuses modeling and simulation software with developments and optimizations in advanced system hardware, software, networking, and data management components needed to solve computationally demanding problems. Data-driven computational science adds data generated in real time from instruments, sensors, and other devices.

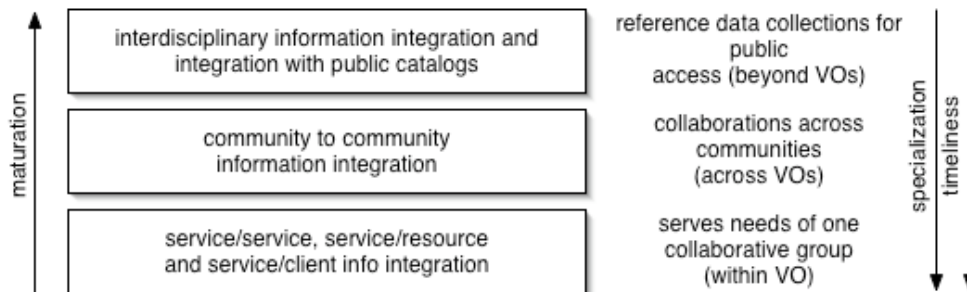


Figure 1. In computational science information integration must occur at and between all three levels of interaction.

1.0 Integration Model

Because of the relative immaturity of computational science, communities or Virtual Organizations (coined by Ian Foster) are still organizing their own efforts to take advantage of advances in technology and resource availability. This is shown as the lowest level of the model, as shown in Figure 1. Information integration at this level is the exchange of information between computational services, and the enabling of computational services to interact with reference or community data collections¹. Tasks communicate with one another and with collections to accomplish goals that are the goals of the one project. While the collaborative project might be very large, such as LEAD, BIRN, or OSG, the needs of the members of the group largely define the driving use cases. Integration at the lowest level, while still extremely challenging, is easier in the sense that all parties share a vested interest in success of the project.

The middle level is the alignment of multiple communities. This might consist of an infrastructure developed in one community extended to offer services to another community. Or both communities may have evolved independent integration standards, and the information integration effort is all that is needed for the two groups to accomplish

¹ Collections fall into one of *reference data collection*, *community data collection*, or *research data collection* depending on the maturity of the collection and other factors [National Sci Board 2005 report].

their science. At the top level is where broader, longer term issues take on importance. It is here that the project reaches a maturation point where long-term preservation and broad accessibility (to schools and libraries) can now be important.

Information integration must occur across the levels of the model itself. From lowest layer upwards, one moves from specialization to generalization. The information about data at the lowest layer must be abundant, precise, timely, and domain specific. As one moves up the levels, the information can be more general. Detailed information is less essential, hence agreement with global catalogs can be easier to achieve.

2.0 Information Integration Across the Levels in LEAD

LEAD is a large-scale collaborative project between meteorologists and computer scientists to improve dynamic forecast capabilities and to democratize weather forecasting research. It has adopted the service-oriented architecture and has built a framework of services and resources to reach its goal. The data subsystem provides a personal workspace, rich search capabilities, data virtualization, and provenance collection among other things.

In comparison to many other collaborative projects that share data, the LEAD project encompasses a variety of complexities that challenge information integration. There are multiple catalogs holding information at different scales of metadata richness. Second, large portions of data being used in LEAD are generated at external sources outside LEAD control, thus creating metadata generation and interoperability issues. Finally, the resources to be described by the metadata schema range from geospatial data to data mining generated statistics files to workflows used to run numerical model simulations. The approach agreed upon is to define the LEAD metadata schema (LMS). LMS leverages the Federal Geographic Data Committee (FGDC) defined standard for geospatial data. We created an FGDC profile for LEAD by trimming down the FGDC schema, restructuring it to handle LEAD's notion of resource collections, and adding additional elements while maintaining namespaces for the separate LEAD catalogs. Version 1.0 of the schema was released mid-June 2005.

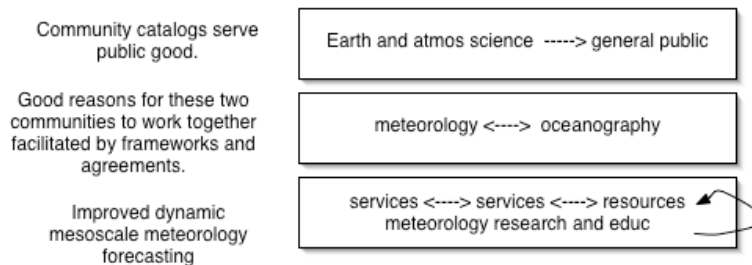


Figure 2. Information integration model applied to mesoscale meteorology.

The LEAD data subsystem is designed to be expandable so as to be able to include new catalogues as they come into development. The conceptual framework for integrating new catalogues into the data subsystem is **LEAD Sandbox** and **crosswalks**. LEAD data providers within in sandbox 1.) communicate by means of grid and web service protocol and invocation methods, and 2.) convey metadata information about data products using LEAD data product schema. Data repositories not adhering to communication protocol and markup language format of “sandbox” are ported in by means of a metadata crosswalk, or mapping between the metadata descriptions of the collection’s native format and the LEAD metadata schema. LEAD is exploring a minimal API such as the API defined in the Open Geospatial Consortium’s Web Coverage Service (WCS 1.1) to further interoperability.

3.0 Summary

The design decisions in LEAD address all levels in the model. The LEAD metadata schema aids integration at the bottom layer (Figure 2). Crosswalks and the sandbox enable extensibility at the middle (across VO) level. Conformance with FGDC and minimal APIs is targets interoperability at the top community catalog level. These system-level design decisions have had a broader impact on the CS challenges as well. Provenance collection can assume one representation while within the sandbox so is eased. The major complexity of query rewrite is avoided because it is pushed out to the edge in the form of a crosswalk. Specialized XML shredding solutions into a relational database can be considered because of the known bounds on the metadata space. Three years into the LEAD project, we feel these solutions directly address the most pressing aspects of integration.