Data integration lessons learned at the IRIS Data Management Center

Timothy Ahern, Chad Trabant, Mick Van Fossen, Mike Stults, Robert Casey, Bruce Weertman, and Yazan Suleiman

IRIS Data Services, DMC, Seattle, United States (chad@iris.washington.edu)

Building on a foundation of web services for data access the IRIS Data Management Center (IRIS DMC) has been involved in a number of multidisciplinary data integration activities over the last couple of years. In this presentation we will summarize the key lessons we have learned that are useful for a variety of Earth science data integration activities. In our experience, the degree of difficulty in integration changes significantly with data similarity and format, access mechanism, federation frameworks and end-user usability. In our own discipline of seismology we are lucky to have an internationally recognized data format and web service access standards. Having shared standards reduces the data integration problem, allowing the creation of data discovery and end-user tools capable of leveraging federated access. We will describe the IRIS Federator system and steps we have taken to help users access data among globally distributed seismological data centers. To address the challenges of interdisciplinary data integration, we worked with our geoscience data center partners in the GeoWS project, supported by the US National Science Foundation's EarthCube program, to develop a generic template for web service APIs and a simple text data format called GeoCSV. These developments provide common ground to ease development of both end-user software and federation/integration systems for a wide array of scientific datasets. To address the broad challenge of data discovery, we are working with partners in the EarthCube Council of Data Facilities to explore strategies such as publishing JSON-LD (Linked Data) online content to describe resources at data repositories. Such linked data descriptions can potentially be harvested by scientific data discovery systems or even data-agnostic popular search engines. As a common thread through all of these activities, we strive to build components, including what would commonly be considered middleware, that is accessible and usable by scientific data end-users. By involving and facilitating end-users whenever possible we find that we solve more use cases than originally envisioned and remain focussed on the ultimate goal of getting scientific data to researchers.