Improving the Interoperability of Earth Observations

An EarthCube White Paper

Jeffery S. Horsburgh[1], David G. Tarboton[1]

October 16, 2011

1. Introduction

In the history of science, many significant advances have resulted from new measurements. Despite the growing volume and sophistication of scientific theorizing of the past several decades, the ultimate source of information in many scientific disciplines is field observations and measurements. What is emerging today is an era of new data collection in the context of larger scale hydrologic and environmental observatories and in response to calls from leaders in the scientific community for new observing systems (e.g., data networks, field observations, and field experiments) that recognize the spatial and temporal heterogeneity of earth processes. These new data collection efforts are focused on precisely representing earth environments with data and advancing our understanding of its functional behavior (both natural and built) in efforts to, as Kirchner (2006) puts it, "get the right answers for the right reasons," referring to the fact that as conditions shift beyond our range of prior experience, improving our predictions for operational and management purposes may be completely dependent on our understanding of important processes.

In this era of observatories, digital technologies, instrumentation, and pervasive networks through which data are collected, new scientific advances are there is also a push toward cross-disciplinary, synthetic research using both new and existing data resources. Indeed, the future of science will inevitably be more data intensive. For example, advancing understanding of the functional behavior of watersheds and encoding it within the next generation of predictive hydrologic models requires synthesis of observations from multiple measurements, at multiple scales, across scientific disciplines, across environmental observatory or other experimental sites, and from multiple sources. According to NSF:

"Transformative approaches and innovative technologies are needed for heterogeneous data to be integrated, made interoperable, explored and re-purposed by researchers in disparate fields and for myriad uses across institutional, disciplinary, spatial and temporal boundaries" (NSF, 2011).

This is a problem of data fusion, where the manner in which data are organized, encoded, and described either enables or inhibits their scientific analysis (Pokorný, 2006; Tomasic and Simon, 1997). Indeed, we are now at a point where our ability to collect data far outstrips our capabilities to analyze it using existing technologies and where the tools available for describing and sharing data are becoming much more important as the scale of data collection involves more people working collaboratively. Gone are the days when all of the data collected as part of a scientific study could be included within a table or appendix in a peer-reviewed publication that described the earth environment within which the data were collected, the measurement methods used, the analyses completed, and the resulting knowledge gained. As a result, our current system for publishing scientific knowledge contains only a small fraction

[1] Utah Water Research Laboratory, Utah State University, Logan, UT

of the data we collect and, subsequently, the knowledge we have gained. It is clear that the business of research is changing, and the scale of these emerging problems and opportunities is consistent with a new paradigm in scientific research within other disciplines based on intensive data collection (Hey et al., 2009). Better infrastructure is needed for supporting the full range of scientific activities, including data capture, curation, and analysis, as well as publication.

In this white paper we first articulate a partial vision for EarthCube related to interoperability of earth observations and then describe opportunities for improving the suite of tools that is available for the current and next generation of engineers and scientists to advance our ability to understand and predict the interactions between changing climate, hydrology, ecology, and human activities.

2. A Vision for EarthCube

One vision for EarthCube is that it would be a seamless organization and repository of the best available data that holistically represent the earth's geo-systems. Supporting this vision would be a suite of software and middleware tools that enable scientific investigators to both populate the data repository and then, on the client side, discover and access data needed for their scientific purposes.

When we read research results in peer reviewed publications, we are in many ways able to stand on the shoulder of giants because we can structure our own research to advance the work that others have done. The same, however, is not currently true for data. Only a small portion of the data that are collected are ever published, and we may be missing many collaborative and synthetic opportunities for advancing scientific research and our understanding of earth processes using existing data resources because we are not sharing data as freely as we could be (in many cases because we do not know how or where to do so and lack the resources to share effectively) and because the data that we do share is made available in primitive formats that: 1) are hard for others to find; 2) are difficult for others to interpret; and 3) do not express to others the knowledge and insights of the data collector that could be applied to the next study that uses the data – as does a written publication.

We envision EarthCube as a national data infrastructure to which scientists could pose complex queries for datasets that they might include in synthetic studies. Example might look like the following:

- I am interested in studying the effects of aquatic nutrient concentrations on stream metabolism - show me locations where high frequency discharge, water temperature, and dissolved oxygen data have been collected in second order streams that are within one mile of a weather station that measures solar radiation and for which samples of nitrogen and phosphorus have been collected.
- I am interested in studying the effect of antecedent soil moisture on runoff generation in headwater streams – show me catchments within which high frequency soil moisture, streamflow, and precipitation have been measured simultaneously.

Although this white paper can't possibly cover all of the aspects of a system that would support this type of investigation, here we focus on three major needs that we assert must be addressed for this vision of EarthCube to be realized: 1) a common information model for describing and encoding earth observational data so they can be interpreted; 2) linking observational data to a digital representation of the earth features to capture their geospatial context and support; and 3) capturing the knowledge content of data – e.g., information that specifies why data were collected, for what purposes they have been used, and insights about the knowledge that they contain.

3. A Common Observations Information Model

"It is time to integrate these data and technologies in an open, adaptable and sustainable framework (an "Earth-Cube") to enable transformative research and education in Earth System Science; ***foster common data models*** *and data-focused methodologies; develop next generation search and data tools; and advance application software to integrate data from various sources and advance knowledge"* (NSF, 2011).

In the EarthCube Dear Colleague letter, attention is drawn specifically to fostering the use of common data models. Abstracting from any particular physical implementation, there is certainly a need for developing common agreement about the information needed to describe observational data so they can be discovered and interpreted by investigators other than those who collected the data. The physical encodings of this information model will follow, including database schemas for storage, data transfer schemas and XML encodings for data exchange, etc.

Semantic and syntactic heterogeneity are major hurdles to be overcome – especially across data types and scientific domains. There has been good work along these lines within some disciplines. For example, within the Consortium of Universities for the Advancement of Hydrologic Science, Inc.(CUAHSI) Hydrologic Information System (HIS), the Observations Data Model (Horsburgh et al., 2008) and Water Markup Language (WaterML) (Zaslavsky et al., 2007) both implement a common information model and community vocabulary for hydrologic observations collected at fixed monitoring sites (e.g., stream gages, water quality sampling sites, weather stations, etc.). Within the solid earth geochemistry domain, there is EarthChem (http://www.earthchem.org) and EarthChemML for observations made from geochemical samples of the solid earth environment. The Open Geospatial Consortium Observations & Measurements (Cox, 2010) is an information model has also gained a lot of traction in different domains for earth observations and now has several XML profiles serving domains in geosciences. These models are all focused on relatively narrow data types, and we acknowledge that there is a need to handle data of many other types – e.g., geospatial datasets, time varying fields, etc.

There are now some very good options and examples in this space, but they are still mostly domain specific and work is still needed to achieve consensus and work toward standards-based approaches for modeling earth observational data across domains. For example, in related EarthCube white papers, Rick Hooper discusses using OGC's O&M information model to improve data interoperability, Ilya Zaslavsky et al. describe how one of the highest levels of data interoperability can be realized with shared domain information models, and David Tarboton et al. describe the importance of standardized data models in the potential architecture for EarthCube. We propose that there is a need to bring together scientists and technologists from disciplines within the Geosciences to develop the high-level information model for earth observations and then to work toward standard technology implementations for data storage, publication, cataloging for discovery, and data exchange.

4. Linking Data to the Geo-Environment

The location at which observations from Geoscience disciplines were made is often described only by a set of latitude and longitude coordinates. The point coordinates may define the absolute location of the data collection device, but they rarely capture the geospatial context, support, or "feature of interest" for the measurements. For example, the coordinates may be the point location of a streamflow gage. It is generally left up to a data consumer to determine that the gage lies on a particular river reach, measures the outflow of a particular catchment, is downstream of another stream gage, is located at

the same location as a water quality monitoring site, and is located near a weather station – yet all of this information is needed by an investigator who is building a model of the watershed and takes precious time to develop. In another example, a point location may identify where a weather station is physically located, but it does not convey the geographic area over which observations from that weather station are representative.

Scientists generally know how to interpret the geospatial context of data created by data collection activities common within their domain – whereas scientists from outside a particular domain may struggle. Creating a geospatial data framework that can serve as the context for observational data is a reasonable next step. Observational data could then be linked to a geospatial "feature of interest" to which it applies. For example, a stream gage would be represented as a point location, the point would be associated with a line representing a stream reach, and the point and line would be associated with a polygon representing a catchment boundary. Each of those geographic features could be linked not only through their physical proximity on a map, but also through specific relationships encoded within feature attributes. Here we could build on existing geospatial datasets and data models like the National Hydrography Dataset, Arc Hydro (Maidment, 2002), Arc Hydro Groundwater (Strassberg et al., 2011), and others. Locating observational data within such a geospatial "fabric" provides a way to evaluate complex geospatial queries such as those posed above and enables an explicit linkage between observations and the earth features that they represent. In a related EarthCube White Paper, Alva Couch and Alex Bedig discuss issues of data trust and plausibility in building digital representations of physical entities within the context of a "digital watershed."

5. Capturing the Knowledge Content of Data

Another constraint to the use of existing datasets for synthetic studies is that it is rare for datasets to be annotated with information such as interpretations of what the data mean, how they have been used, analyses that have been done, conclusions that have been drawn, etc. It is becoming more common see peer reviewed publications that are linked to electronic appendices containing data or linked directly to datasets published on the Internet. This is a trend that we should certainly continue, but it is insufficient as it leaves out data not referenced directly by scientific papers (likely the majority of data collected).

We propose that a system is needed for sharing data within which the data can be annotated with contextual information about the meaning of the data, how it has been used, its appropriateness for given uses, and potentially even information about its quality and limitations. A good example of this would be annotating a streamflow dataset with comments that specify when major events such as storms, floods, etc. occur. A trained hydrologist could likely infer when major events occur from the numeric values in the dataset and by overlaying ancillary datasets such as precipitation, but it would be far easier for a non-hydrologist to interpret a set of abnormally high discharge values if there was an annotation stating that a hurricane occurred.

Some of this could be done through collaborative and social networking technologies. Data consumers could comment on or rate published datasets, leaving information that can be evaluated by subsequent data consumers. Additionally, data consumers could "tag" datasets as fit for a particular purpose, which may assist in the data discovery process.

6. Conclusions

Many cyberinfrastructure systems in the Geosciences are now at a state of functionality that a savvy investigator could, with a bit of effort, find and access datasets from multiple disciplines for a synthetic analysis. However, this process is still not seamless – it may require the use of multiple software systems and techniques and there are still inconsistencies in the way the different systems describe, encode, and share data that make integration difficult. We see EarthCube as an opportunity to develop interoperability of earth observations data in a completely seamless way that is transparent to the average investigator. A common information model that results in standards for data interfaces and encodings, a geospatial fabric within which observational data can be located and organized, and the ability to capture the knowledge content of data in a searchable way are three steps toward realizing the level of interoperability envisioned for EarthCube.

We believe that these three things will drive new innovations in synthesis and modeling in the Geosciences (including water, energy, and material balances, and particularly weathering/soil processes as truly novel aspects) based on extensive data already being collected nationally by organizations like the United States Geological Survey, and more locally in research watersheds, Critical Zone Observatories, and other research sites.

References

Cox, S. (2010), Geographic Information: Observations and Measurments OGC Abstract Specification Topic 20, v2.0.0, OGC 10-004r3, Open Geospatial Consortium, Inc., 49 pp., http://portal.opengeospatial.org/files/?artifact_id=41579.

Hey, T., S. Tansley, and K. Tolle (2009), *The Fourth Paradigm Data-Intensive Scientific Discovery*, Microsoft Research, Redmond, Washington, 283 p. (Available at http://research.microsoft.com/en-us/collaboration/fourthparadigm/)

Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., Zaslavsky, I. (2008), A relational model for environmental and water resources data, Water Resources Research, 44 (W05406), 12. doi:10.1029/2007WR006392.

Kirchner, J.W. (2006), Getting the right answers for the right reasons: Linking measurements, analysis, and models to advance the science of hydrology, Water Resources Research, 42, W03S04, doi:10.1029/2005WR004362.

Maidment, D. R. (2002). Arc Hydro GIS for Water Resources, ESRI Press, Redlands, CA.

National Science Foundation (2011), Dear Colleague Letter: The "Earth Cube" – Towards a National Data Infrastructure for Earth System Science, NSF 11-065, http://www.nsf.gov/pubs/2011/nsf11065/nsf11065.jsp?org=NSF

Pokorný, J. (2006), Database architectures: current trends and their relationships to environmental data management, Environmental Modelling & Software, 21 (2006) 1579-1586.

Strassberg, G., Jones, N. L., and Maidment, D. R. (2011). Arc Hydro Groundwater: GIS for Hydrogeology, ESRI Press, Redlands, CA.

Tomasic, A., and E. Simon (1997), Improving access to environmental data using context information, ACM SIGMOD Record, Vol. 26, No. 1, pp. 11-15.

Zaslavsky, I., Valentine, D., Whiteaker, T. (Eds.) (2007), CUAHSI WaterML v0.3.0. OGC07-041r1. Open Geospatial Consortium, Inc. 76 pp. http://portal.opengeospatial.org/files/?artifact_id=21743S.