2017 EarthCube Architecture Refinement Workshop Report

Completed, August 11, 2017

Prepared by the EarthCube Architecture Refinement Workshop Organizing Committee:
Mike Daniels (Chair)
Bob Arko
Leslie Hsu
Rebecca Koskela
Jay Pearlman

Executive Summary

During 2016 and 2017, a tremendous amount of effort went into the process of defining an architecture for the EarthCube (EC) initiative. These efforts produced key documents and reports from workshops, the Architecture Implementation Plan and Solutions Architecture documents and a Tiger Team response and a report describing EC community feedback related to the proposed architecture. The community feedback report contained results of surveys that identified the highest priority architecture elements for Phase One of the EC Architecture development are: 1) Resource Discovery, 2) Resource Registry and 3) Resource Distribution and Access. In order to plan out next steps for developing these elements of the EC architecture, an Architecture Refinement Workshop (ARW) was approved by the EC Leadership Council on April 19, 2017.

As a first step in preparing for the ARW, members of the EarthCube science and technical communities defined common needs in these three architecture areas across multi-disciplinary science use cases at the 2017 EarthCube All-Hands Meeting. These highest priority needs served as a foundation for the ARW which was held July 10-12. The ARW participants translated these needs into requirements and described existing services and interfaces. As a representative use case, the group converged on defining the services and interfaces necessary to perform a space-time query of data holdings across geoscience data repositories that hold data from the Hurricane Sandy event. A Project Plan was created that laid out the technical steps and milestones to effectively build the services to accomplish this search and more fully develop the three architecture components and their interfaces. Building the services to accomplish a space-time query for data across multiple repositories will then serve the fundamental needs of the EC community more generally.

Participants in the 2017 EarthCube Architecture Refinement Workshop. Front (Left to Right): James Davies (ESSO), Mike Daniels (NCAR), Mohan Ramamurthy (ESSO), Eric Lingerfelt (ESSO), Emily Law (NASA/JPL), Edwin Skidmore (CyVerse), Matt Mayernik (NCAR), Ilya Zaslavsky (SDSC), Back (Left to Right): Tim Ahern (IRIS), Dave Vieglais (DataONE), Michael Bell (CSU), Chuck Meertens (UNAVCO), Jay Pearlman (J & F Enterprises), Siri Jodha (NSIDC), Lynne Schreiber (ESSO). Organizers and Participants not pictured: Bob Arko (UNOLS/R2R), Doug Fils (Ocean Leadership Consortium), Leslie Hsu (USGS), Shantenu Jha (Rutgers) and Rebecca Koskela (DataONE).

Background

During 2016 and 2017, a tremendous amount of effort went into the process of defining an architecture for the EarthCube initiative. These efforts produced key documents and reports from workshops, an Architecture Implementation Plan and Solutions Architecture documents, a Tiger Team response and a report describing community feedback on the proposed architecture. After careful consideration by the EarthCube community and leadership, it was decided that a good next step would be to conduct an Architecture Refinement Workshop (ARW) to focus specifically on the three highest priority architecture components (Resource Discovery, Resource Registry and Resource Distribution and Access) and layout a project plan for completing work over a 10-12 month period. An organizing team comprised of EarthCube leaders was assembled and the proposal for the workshop was approved on April 19, 2017.

ARW Preparation at the 2017 EarthCube All-Hands Meeting

In preparation for the workshop, the organizing team scheduled an Architecture session at the EarthCube All-Hands Meeting. During this session, needs in the three architecture component areas were identified in the context of three multi-disciplinary use cases. Participation at the All-

Hands meeting was very good. The team received over 35 pages of feedback which served as a backdrop for the ARW itself.

Workshop Proceedings

The ARW was held on July 10-12 at NCAR in Boulder, CO. Eighteen participants with a wide range of expertise, both within and outside of EarthCube, were selected to attend. A workshop agenda and other architecture materials were disseminated the week before the workshop. Each participant began by giving brief presentations of the projects they are involved with, the technologies they use to address discovery, registry and access, and their future needs. The presentations served to uncover overlap and commonly used technologies across the projects. After this session, Dr. Michael Bell, a CSU researcher, described his vision for science needs in the areas of discovery, registry and access. A recap of the science needs identified at the 2017 AHM was also presented to the group.

The afternoon of the first day concluded with a discussion of common needs and services that address them. The group then discussed how to appropriately match the scope of what could be accomplished in a 2 ½ day workshop with the need to develop an actionable project plan for the three architecture components. In particular, the list of resources was very broad and so the group decided it would be best to focus on *software* and *data* only. Notes from the Plenaries and a summary of key points from the first day were captured and recorded.

On the morning of the second day of the workshop, participants split into three groups to define requirements for each of the three architecture component areas. The initial basis for these requirements were the needs that were identified at the AHM Architecture Breakouts. After the breakout, group facilitators summarized the requirements for Discovery, Registry and Distribution and Access.

During the afternoon of the second day of the workshop, the participants focused on technologies, interfaces and interoperability elements that could address Discovery, Registry and Access. Summaries of the breakouts for interfaces and interoperability in the areas of Discovery, Registry and Distribution and Access were produced. During the discussions, lists of existing technologies and prior work on EarthCube standards were referenced and, by the end of the second day, a description of a holistic approach to link the three architecture components emerged.

On the third day of the workshop, the group began development of a project plan to implement an EarthCube Discovery, Registry and Access service. The plan was focused on describing a system to perform space-time queries of datasets available for the purposes of studying Hurricane Sandy across solid earth (e.g. UNAVCO, IRIS), ocean (e.g. WHOI) and atmospheric (e.g. Unidata) data repositories as well those connected through a distributed cyberinfrastructure framework such as DataONE. Existing technologies discussed included CINERGI, JSON-LD and CKAN.

Workshop Outcomes and Conclusions

Several important conclusions were reached as a result of this workshop. A plan forward was created and extensive materials were captured at the workshop that will be helpful for future efforts. In particular, the breakout materials from both the All-Hands Meeting and the ARW Workshop contained an excellent descriptions of science needs and requirements as well as technology inventories and specifications for interfaces and interoperability. Other specific actions and findings from the workshop are described below:

- During the workshop, a 10-12 month Project Plan for implementing time/space queries of datasets across multi-disciplinary repositories was created and is being developed for implementation of a multi-disciplinary science use case (e.g. Hurricane Sandy). The EarthCube Technical Officer and other ESSO staff have expanded this plan and created a detailed Statement of Work (SOW) for NSF which will more fully describe the work necessary to accomplish the tasks identified for this project.

- Participants at this workshop, influenced by successful architecture models from organizations such as DataONE, described a discovery, registry and access architecture model similar to Xentity's (e.g. see the EC Solution Architecture diagrams on pages 31 & 46).

- Development of a common information model is critical to support a cohesive discovery, registry and access architecture model.

- Compared to other large cyberinfrastructure initiatives, EarthCube seems to be unique in that it is addressing multi-disciplinary needs.

- When searching for data, two common forms of query were discussed and debated: an "event" mode (e.g. "Hurricane Sandy") and a space-time query mode (e.g. data from this time period taken in this geographical region). At this time, though technology exists to perform these queries to some extent, it was felt that neither of these modes will produce a comprehensive listing of datasets. Data Facilities that are collecting ongoing data may not specifically tag an event using terms that are outside of their science domain. For example, a data facility hosting seismic network data such as atmospheric state parameters and infrasound measurements collected during Hurricane Sandy may not label these data sources as such.

- The modeling community was sparsely represented at the workshop. For some larger-scale modeling communities, data and tool discovery seemed less important to them because their resources are large and high visibility which are often well known among their communities. However, there are indications that smaller modeling communities may have different architecture needs where data discovery is much more important.

- If data are successfully discovered across repositories, key questions remain about the quality of the data found, the dataset's uncertainty characteristics and measurement

limitations and other aspects that a scientist needs in order to trust the data sufficiently to publish findings using it. Many scientists have built up trust in their commonly used datasets over a number of years through an in-depth understanding of quality and uncertainty by working with similar data previously.

- Transformation, mediation, brokering, subsetting and the like will need to be addressed as important next steps once data are found.

Other Lessons Learned

- Given a small (~20 people), short-duration workshop (~2.5 days), we had to narrow the scope of the workshop to cover only data and software, but in the end we focused most of our discussions on data. The agenda evolved and was adjusted somewhat over the course of the workshop. On the final day, in order to have an actionable plan, we further narrowed the focus to describing steps necessary to implement space-time queries across multiple repositories for the Hurricane Sandy use case.

- Some amount of inventorying technologies that are out there will always seem important, especially for those who may have been hearing about some of these technologies for the first time.

- If we are able to produce a long listing of datasets based on a space-time query, we need to work more on relevancy of the results and filtering the searches. Science teams are willing to help with this, but they cautioned about the time commitment they would have to make if there were multiple prototyping efforts.

- In terms of software discovery, there was a suggestion to store all EC software in GitHub and add “EarthCube” to code repository names (as it turns out, there are several EarthCube projects that already do this).

- The group recognized the importance of, but did not have time to sufficiently address, so-called “dark data”, or data that are not accessible on the Internet. EarthCube could provide easy-to-find resources on best practices for getting dark data online as a first step in making these data more visible.