EARTHCUBE
TRANSFORMING GEOSCIENCES RESEARCH

EarthCube Community Feedback on the Proposed Architecture and Next Steps

March 20, 2017

Executive Summary

The Leadership Council (LC) sought feedback from the Technology & Architecture Committee (TAC) and the Council of Data Facilities (CDF) regarding EarthCube's (EC) architecture plan. The input received from TAC and CDF informed this response. While none of the recommendations of the CDF were absolute, the overall consensus was that EarthCube should develop capabilities at a central site such as the EarthCube Science Support Office (ESSO). Many of the TAC responses recommended a staged implementation using rapid prototyping and an Agile development approach. The services receiving the highest priority for implementation from TAC matched those of the CDF.

Consensus indicated that three of the most important services that should receive attention are:

- Resource Discovery Services
- Resources Registry
- Resource Distribution and Access Services

Early indications point to the following in a second phase of development:

- Transformation Services
- Semantic Services
- Linking Services
- Brokering Services

A final round could then include:

- Fusion and Visualization Services
- Workflow Services

The proposed architecture espouses the idea that EarthCube is a system of systems and attempts to define the full scope of the activities and capabilities that address the requirements and desires expressed by geoscientists through a comprehensive, community-driven effort. It helps to provide the central hub for integration and coordination of EarthCube activities that so far has been lacking. The unifying architecture leverages resources in the broad community and has the potential, with modifications, to integrate and assess data and other resources across disciplines.

Community concerns are evident in the feedback received and must be addressed, specifically:

- Lack of a true implementation plan and a sustainability plan, e.g. standards and technologies are not defined;
- Lack of focus on geoscience research scientists;
- "Workbench" is not well-defined;

- Use of DataCite, a general rather than domain-specific registry, for metadata;
- Should be able to access data, rather than just metadata.

The community identified the need to invest resources in a Technology Officer or Developer Advocate to oversee the implementation of the architecture and Software Developers to do the actual implementation. It will also be necessary to define the Application Programming Interfaces (API) that were considered outside the scope of the Conceptual Architecture document.

Introduction

The Council of Data Facilities and the Technology and Architecture Committee worked together in response to the Leadership Council's request to canvas their members through a survey to aggregate opinions with regards to the proposed EarthCube Architecture.

The Leadership Council understands that the Architecture documents and Implementation Plan were difficult and dense documents to review. A total of 22 responses (13 completed, 9 partial) were received from 16 of the 35 members of the CDF. The TAC received 27 responses (14 completed, 13 partial). Leveraging a summary of the recurring responses from both the TAC and CDF, this LC review is meant to provide recommendations to the NSF for EarthCube directions and next steps for implementing the EC Architecture Plan.

Major Strengths of the Proposed Architecture

The proposed EC Architecture espouses the idea that EarthCube is a system of systems. The architecture defines the full scope of the activities and capabilities needed by geoscientists and draws on resources in the research community, helping to connect EarthCube activities, which have often been disparate and unconnected. It is an architecture that has the potential, with modifications, to integrate current repositories and tools across disciplines.

Architecture Design strengths:

- Places a clear emphasis on registry services, assessment, resource discovery, and "workbench" for testing component interoperability;
- Emphasizes the importance of utilizing standards and protocols consistent with efforts in groups such as W3C;
- Suitable for heterogeneous fields in general, not specific for geosciences;
- Provides one viable description of the EC ecosystem;
- Includes the identification and use of best practices such as persistent identifiers.

Architecture Implementation Plan strengths:

- Identifies the need for all geoscience resources to have a Digital Object Identifier (DOI) or other applicable unique and persistent identifiers, which implies there would be associated, standardized, resource-type-specific metadata to go with it;
- Works to leverage the assets in existing community catalogs and data repositories;

- Proposes development of a Workbench for testing of cyber developments and identifying possible services for interdisciplinary geoscience;
- Stresses the importance of assessing geoscience resources before including them in the EC architecture and/or Workbench.

Major Shortcomings of the Proposed Architecture

The Solutions Architecture and the Implementation Plan documents lay out general guidelines to help EarthCube move forward and provide a starting point for next steps. However, these documents specifically consider many of the necessary next steps as being out of scope, because they require additional work by the EarthCube community and its governance. Reviewers found the documents dense and the widespread use of undefined terms, as well as inconsistent use of terms, made the documents fairly opaque and confusing. There is a very strong need for a glossary. There is confusion between what is needed to meet requirements and how to implement components. Additionally, it is not clear how current and previous funded projects can be modified to connect with the proposed architecture. The architecture does not acknowledge or consider capabilities that exist within the NSF ecosystem such as CyVerse, DataONE, or XSEDE. The Implementation Plan may be too ambitious and does not allow sufficient time for feedback. In addition, it takes an enterprise approach, rather than an Agile approach that would entail short development cycles and continuous feedback as new components are developed. EarthCube must embrace the emerging nature of this system-of-systems development and inclusion of its geoscience end users. EarthCube should also consider paths within the architecture for rapid prototyping to directly engage geoscience users earlier in the development process.

The workbench described in the Architecture Design does not match previous workbench concepts discussed within EarthCube. The workbench in this document is envisioned as a test bed for the Building Blocks (BB) and Integrative Activities (IA) for use by developers, not as a working environment for researchers.

The document assumes that DataCite will be the metadata source providing a path to get details of each resource that it exposes. However, it is not clear that content provided by DataCite will be sufficient for using workbench services efficiently. There are already projects funded through EC that 1) crawl the web for specifics sources and extract information from them; 2) generate metadata using semantic technologies; 3) automatically enhance metadata descriptions harvested from catalogs or contributed by researchers; 4) extract data and metadata from literature; 5) improve model descriptions to make them interoperable. It is not clear that the Architecture Design would use or benefit from any of these capabilities. These capabilities could be an important differentiator of the EC architecture from the geoscience perspective.

The document has several missing components - *the biggest is access to data*. DataCite may lead users to a repository but does not indicate what types of data are available there. A metadata catalog without access to datasets is not what is needed. The document includes a discussion of SQL queries on the metadata but not on the data itself. It is not clear how real-time production data will fit into the architecture.

While this exercise in developing a Solution Architecture and Implementation Plan hasn't provided us with a clear path forward it has been a useful exercise in identifying priorities, requirements and gaps in understanding. Possibly the most important realization has been that this effort requires full-time committed expertise to facilitate and promote the architecture development.

Additional Requirements or Needed Information

Additional requirements include:

- Clear specification and prioritization of standards and protocols or the development of systems to bridge differing standards and protocols;
- Assessment and identification of Building Blocks that can be leveraged by the EarthCube architecture. Those identified need to be operationalized and provided with long-term support. Research-style development in many of the funded projects may not be ready for a production environment;
- Identification and/or development of controlled vocabularies or ontologies;
- Specification of an initial list of interfaces or criteria for such interfaces to be included in the first phase of EC; APIs to connect to existing resources; identification of a testbed for interoperability testing and operational maturity.

All additional requirements have technical, financial, and human resource components. Funding will be needed to develop some capabilities at a central site. The LC has identified the need to support technical capabilities at the ESSO by hiring an Executive Director, Technology Officer and software developers.

Does Solutions Architecture Reflect Outcome of May 2016 Earthcube Architecture Workshop

Feedback suggests that the report from the Architecture Workshop (link) that was held in May 2016 was much better written and organized. The Architecture Implementation Plan, although appearing to be highly detailed, did not go much further. The Workshop Report started with an architecture to connect the BBs, but the workbench described in that report had the capability for data handling, analyses and other uses and put the scientists at the center, which is important.

Recommendation for a Phased Implementation

The CDF/TAC feedback to the LC focused on results from a survey of the EarthCube community and CDF members. Part of the survey asked respondents to identify and rank the importance of a variety of services identified for the workbench. These capabilities grouped nicely into three groupings. The numbers in parentheses below provide the relative importance of the services as assigned by the respondents, where 8 is high and 1 is low. The Leadership Council recommends a phased approach to implement these capabilities as follows.

Phase 1: **Discovering and Accessing Individual Geoscience Resources**

- **Resource Discovery Services** (Combined priority score - 6.7):

Discovery of resources available to geoscience researchers includes development of services that discover tools and data from existing repositories. These services may or may not reside on the EC infrastructure.

- **Resources Registry** (6.5):
  In some cases, resources may not be available through partner organizations. The EC infrastructure must have the ability to provide a mechanism for discovery of its own (BB, AI, etc) resources, as well as any deemed significant to the broader community, but not being discoverable through the Resource Discovery Services.
- **Resource Distribution and Access Services** (6.4):
  As data and software become available through web services, EC must facilitate access to resources and distribute resources through standards-based services.

Phase 2: **Making Resources More Usable and Interoperable -- Mediation**

- **Transformation Services** (4.7):
  By providing transformation across a select group of adopted standards, EC Infrastructure will enable data providers and consumers a mechanism to select one standards based framework for distribution and access, while the EarthCube TS brokers the content across standards (e.g., SWE-> netCDF or visa-versa).This transformation service is typically included as a service in a Broker framework. Other transformation services can be translation of units; geo-spatial subsetting or gridding tools, etc.
- **Brokering Services** (4.6):
  Brokering Services including mediation are implemented in a Brokering Framework that maps discipline specific data attributes (metadata, standards and data formats) from one discipline to another. This mapping eliminates the need for scientists to learn the details of data instantiation of other disciplines. It includes standards and metadata transformations.
- **Semantic Services** (4.6):
  The W3C Semantic Web (RDF, OWL, LOD) implementation facilitates better discovery and understanding. As the EC community develops domain-specific and cross-domain registries of terms and associated ontologies, the EC Infrastructure must provide the services to create, manage and utilize the resources.
- **Linking or Interoperability Services** (4.4):
  Integration of Services and Resources.
  This includes the use of standard APIs and standardized metadata. With sufficient metadata, automatic mediation (i.e. brokering) for interoperability is possible.

Phase 3: **Combining Resources in Scientific Workflows to do Research**

- **Workflow Services** (4.0):
  Workflow services provide provenance for geoscience research, as well as reproducibility.
  As WorkFlows are registered, standards-based workflows enable reuse of well-documented workflows with input from the broader community via Resource Assessment Services.
- **Fusion and Visualization Services** (3.2):
  These services will be a key but final component of The Workbench. As the discovery and access services are developed, the FVS will provide a user-friendly interface for the utilization of the resources.

The following were not prioritized in the TAC/CDF review but the LC thinks these services should also be included.

- **Resource Assessment Services**
  Services that enable interactive, constructive assessment of tools and data that are discovered and utilized will provide the most significant contribution to the broader geo-science community.
- **Community Networking and Training Services (CNTS)**
  Some examples of community networking and training services include instantiating single-sign-on across partner organizations and formalizing training for creation/utilization of community resources for EC registered members

The Path Forward

The LC believes significant progress can be made by beginning to centralize some EC social and technical capability at the ESSO. The addition of a Technical Officer is being taken as the first step and will help to begin the implementation of prioritized services. The LC also recommends the hiring of an Executive Director to oversee the integrative activities and relationships required of the program. The TAC and CDF recommend a phased implementation of specific services that are needed to develop the first components of the EarthCube system that will allow integration of capabilities in the funded projects and the members of the Council of Data Facilities and initial use by geoscientists.

The LC proposes developing the EC system in an Agile manner, implementing key components first, identifying additional needs, and iterating toward an EC architecture that meets the needs of the EC scientific community. The TAC/CDF prioritization leads EC forward in an Agile manner.

The LC is actively exploring mechanisms to solicit community input to further define services and requirements needed to move us towards an EarthCube infrastructure. Funded projects, both previous and current, and data facilities will provide better definition on how to connect and integrate existing resources, with the specification of adopted standards and required services. Input from the EarthCube scientific community will guide in defining requirements and capabilities needed to assess services and to build an exemplar WorkBench that facilitates easy access to interdisciplinary tools and other resources that facilitate providing reproducible research outcomes.