COUNCIL OF DATA FACILITIES FEEDBACK TO EARTHCUBE LEADERSHIP COUNCIL RELATED TO THE EARTHCUBE ARCHITECTURE
30 January 2017

Background:
In response to the EarthCube Leadership Council's (LC) request to the CDF and the TAC, CDF and TAC worked together to canvas our constituents regarding the community's thoughts as they relate to the proposed EarthCube Architecture. CDF members were provided links to the various EarthCube Architecture documents for background. As the LC understands these were difficult and dense documents to review. A total of 22 responses (13 completed, 9 attempted) were received from 16 of the 35 members of the CDF. This summary is intended to give a distillation of the recurring responses that could be gleaned from CDF responses.

There were five questions posed by the LC and the following text addresses each one individually

Q1 What are the major strengths of the proposed architecture? Please be specific and if possible, identify strengths that support geoscience research and differentiate this architecture from other government-supported architectures.

The proposed EC architecture describes a unique platform to provide key cyberinfrastructure (CI) resources to the geoscience community including data and model search, interoperability, data transformation, service assessments and tools for scientific analysis and visualization, ways to contribute resources, and finally to publish data, workflows and results and utilize EC through a comprehensive workbench and to provide training. This is a very ambitious project that, when realized, would set the standard for CI community resources.

All proposed components identified in the architecture, and in particular the workbench, would be useful and supported by the community. The workbench represents a fresh approach that may facilitate new recognition of what works and what does not in assembling resources and CI to solve a problem. The workbench also lends itself to a phased implementation of capabilities. This allows new infrastructure to be released in a phased manner allowing the community to start seeing benefits in the shorter term with no need to develop a complete end to end system. As part of the CDF survey we have identified some input as to what components might be candidates for early development.

The proposed architecture clearly identifies needed capabilities that are needed in the EarthCube system. Members of the CDF repeatedly stressed the importance of some of the key aspects of the proposed architecture. Those items that were repeated multiple times included:

- The importance of defining standards and protocols consistent with efforts in groups such as W3C
- The identification and use of best practices in CI such as persistent identifiers, DOIs, OrcIDs, ARK, and CRM)
- The integration of existing CDF capabilities and infrastructure already in operation at CDF members as well as building links to other federal agency repositories
- An agile approach allowing prioritization of key services
- User engagement and assessment are included and can help identify support for the emerging new architecture.
- Key components can be deployed in an EarthCube cloud environment enabling new capabilities such as contributed workflows, transformation tools, visualization capabilities to name a few.
- The ability to programmatically extract the capability of the various resources in an EarthCube registry through APIs
- Vetting the value and correctness of metadata enabling trust in the data/model repositories in the CDF membership

  There is another government-supported architecture built by NASA that has some but not all of these components. The NASA system is intended for NASA projects but has some systems that could be emulated by EC. Having active participation by EC community in ESIP and Earth Science Data System Working Groups (ESDSWG) should be encouraged to stay on top of NASA, NOAA and USGS-supported cyberinfrastructure developments. To be more broadly useful the planned connections to data resources at these other government facilities is essential for EC to be successful.

Q2 What are the major shortcomings of the proposed architecture? Please be specific and if possible, identify potential mitigation strategies for each.

The architecture and related implementation plan are poorly written and can not be used to start implementing any significant capabilities. Additionally it is not clear how current and previous funded projects can modify anything to be usable by the proposed architecture. The implementation plan may be too ambitious for the time allowed and does not allow enough time for feedback from the partner/scientific contributors, external developers and users. It appears to be more of an enterprise approach rather than one that will rely on an agile approach of short development cycles and learning as new components are developed and not pretending that the end-to-end system is clearly understood. In this way the strength of the plan (the goal of creating an integrated and extremely capable EC set of resources) is also its weakness by setting what are probably too high expectations. If we don’t have something to show in the short term, for all the effort and dollars going into EC the program will surely die.

The plan does not prioritize the capabilities to be developed nor how to reduce, rescope, and reprioritize when things inevitably fall behind. Is discovery and interoperability the priority? Workflows? Semantics? As for the registry the plan of the CDF to investigate the use of and possibly enhancing re3Data to meet EC needs seems like a better idea than creating yet another registry. The Workbench is great to tackle specific problems, but will leave many problems waiting to be addressed. This inevitably leaves much effort happening outside the EarthCube context and may leave EarthCube behind if a relevant breakthrough paradigm shift happens elsewhere. We expect the development of the EC architecture will be expedited by concentrated bursts of centralized development during phase I, but wonder if the fast ramp up and down of dozens of developers and project managers, as called for in the various documents, is realistic

Some of the specific weaknesses identified include:

- The EarthCube architecture may be too large to implement in a reasonable period of time and not be able to show earth scientists that EarthCube can actually do something for them so they see the benefit.
- The text of the architecture documents is not clear, not well written, and obfuscates too many of the concepts.
- There are a multiplicity of registries being discussed. EarthCube is even funding registry development but does not seem to be well coordinated with other international registry efforts. There is support in the CDF to pursue the Re3data registry due to its international presence and relevance.
- There is no clear understanding of how metrics can be used for ways to assess the effectiveness of the EarthCube infrastructure.
- Research style developments in many of the funded projects are most likely not ready for a production environment
- The promise of the workbench may be overstated. While stated already in other projects no actual successful implementations have been identified attesting to the difficulty of the problem.
- The architecture does not acknowledge or consider many existing capabilities that exist within the NSF ecosystem (DataOne is specifically identified). There is concern that the architecture is not cognizant of the previously NSF funded ITR effort GEON.
- The widespread use of undefined terms makes the documents fairly opaque and confusing. There is a very strong need for a glossary.
- The document confuses what is needed with how to implement components.

Q3 What additional information would be needed to connect existing and planned capabilities inside and outside of EarthCube, including building blocks, with the architecture as described?

- Clear specification and prioritization of standards and protocols are the development of systems to mitigate differing standards and protocols.

- Building blocks deemed worthy need to be operationalized and provided with long-term support.
- Building blocks that can be leveraged to be helpful to the EarthCube architecture need to be identified. Not all funded building blocks are likely to be proven to be helpful.
- How will controlled vocabularies or ontologies be developed and where will that effort take place if not done in a central location.
- Domain specific knowledge must be brought into the development of the EarthCube system. How can domain repositories connect to the EarthCube ardchitecture?

Q 3a What resources would be needed from the EarthCube architecture to make this connection?

- Almost all of the CDF members responding indicated that there was a need for a technical development team at a central site, most likely ESSO.
- Some indicated the need for centralized hardware but did not specifically mention cloud, but central location
- Webinars helping CDF members to understand how to link to the central site resources

Q3b What resources might be needed from specific external resources (Building Blocks, CDF members, etc) to connect the external resource to the architecture?

CDF members consistently thought there was a need to support the CDF members with technical, financial, and computational resources to connect with a EarthCube central system

- Targeted funding for the resource providers to implement the EarthCube protocols, standards, etc where needed.
- Cloud computing to support shared environments
- Support for technical people familiar with the EarthCube architecture and its components to assist connecting CDF/resources to EarthCube.
- Consulting resources to connect with the resource providers

Q3c For your work, will the timing of the availability of the architecture have an impact on your work. Can you give a specific example please.

- An early definition of EarthCube standards, protocols and the like will make connecting future building blocks to the EarthCube centralized architecture more efficient and effective.
- Building blocks and other efforts that are complete most likely are too far along to have the EarthCube architecture affect their development.
- Without knowing more specific details of the architecture some found it impossible to know how to answer this question.

The vast majority (77%) of CDF responders indicated that the timing will not impact their work.

Q4- If you are familiar with the architecture recommended by the EC Architecture Workshop in May 2016, how well does the Solutions Architecture reflect the outcomes of the workshop? Again, please be specific.

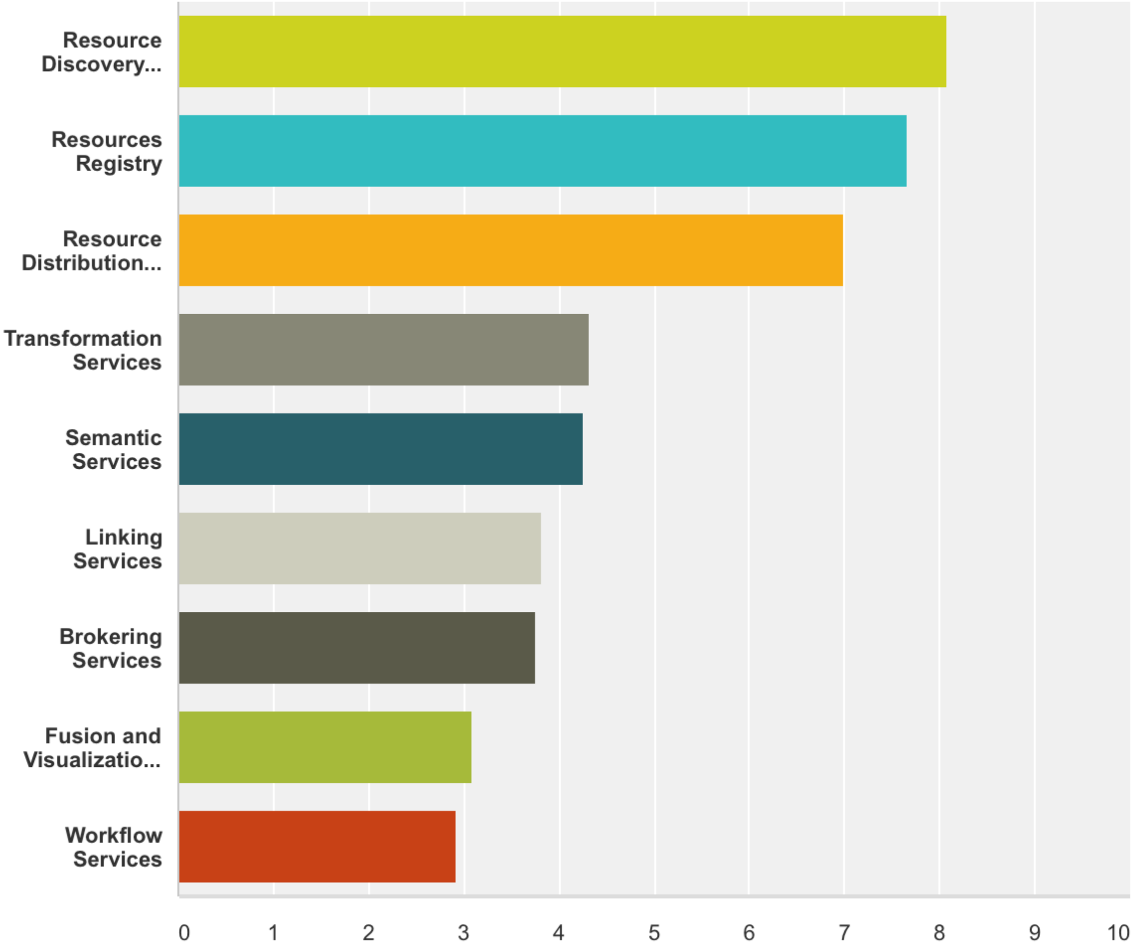
About 62% of the 13 answering this question were familiar with EC Architecture Workshop in May 2016.

- The concepts and features that continue to be included in the architecture are
  - The approach of a System of Systems
  - Concept of a workbench remains
- Concepts that changed include
  - Portfolio of Services changed to a portfolio of resources
- CDF members were inconsistent about whether the Solutions architecture reflected the outcomes of the workshop. Perhaps this reflects that lack of clarity in the various documents.
- At best it is not clear if the two elements are aligned.

Q4a If the architecture implementation is done in phases, what are the most important services and should therefore be implemented early in the architecture workbench. Please rank each of the following 9 services with 9 being the most important and 1 the least important. You must have one service ranked 9, one ranked 8, etc.

The CDF members were asked to rank the importance each of the following services and capabilities of the architecture. The following list is ranked in the order from highest to lowest priority that resulted from the survey. The CDF believes that this clearly indicates that a focus on various Discovery, Registry, and Access services should be the candidates for early development.

	Service	**Weighted Average**
1.	Resource Discovery Services	[8.08]
2.	Resources Registry	[7.67]
3.	Resource Distribution and Access Services	[7.00]
4.	Transformation Services	[4.33]
5.	Semantic Services	[4.25]
6.	Linking Services	[3.82]
7.	Brokering Services	[3.75]
8.	Fusion and Visualization Services	[3.08]
9.	Workflow Services	[2.92]



The above graphic shows a clear way to prioritize development of EarthCube capabilities in a phased manner. The top three services should take precedence in a phase implementation plan. Perhaps the next four services could be considered for a second phased implementation and finally the fusion and workflow tasks done in a later phase.

Q5. Other comments

Few additional comments were supplied through the survey tool. Of the responses returned, one stated that the proposed architecture is not sufficient to realize EC's overall vision. Instead, EC should consider developing some of the specified services and capabilities in a prioritized manner.

One comment related to the content and organization of the document itself, describing dense, obfuscating terminology. An additional comment related the survey question on prioritization of architecture services (Q4) to the document organization, stating that in order to prioritize these services, connections between services listed in the survey and the architecture document should have been clearer and better cross-referenced. For example, services listed in the survey prioritization question should have had corresponding sections in the architecture document with descriptions of the proposed services, tools and capabilities and data to be delivered in that section (i.e. API, GUI, parameters/information).

Additional comments related to the timing or order of implementation, recommending EC begin development as soon as possible with identified exemplary Workbench capabilities and by engaging data facilities (CDF members) to provide initial registration of facilities and resources.

This open-ended question solicited several comments. They are included here in their entirety.

- Identify a handful of exemplary Workbench capabilities to spin up as fast as possible.

- I do not think that the existing architecture specified in these documents can be used to successfully implement a viable solution to the EC problem. Instead EC should consider developing some of the specified services and capabilities in a prioritized manner.

-

- Overall we are firm believers that EC is our one chance this decade to get a comprehensive CI for geosciences in place. We think that the best chance for “low hanging fruit” is to get the CDF facilities on board to provide initial registration of facilities and resources and to present GeoWS-like resources that are intrinsically aligned with emergent standards (e.g GeoWS csv format, simple web service, and rdf or something like it for cataloging). The data facilities would thereby lead the way for EC and provide some real tangible resources for the user community to use while the grand plan is realized and implemented.

- The complex format and obfuscating language prohibited any in-depth understanding of the proposed architecture, making it difficult to determine any strengths or weaknesses (even for developers/engineers).

- I would suggest that in order to set the priorities in the preceding section, I would be able to open the architectural document and find a section whose heading match each line, and in each section I would find the proposed API, GUI, and parameters/information needed to access and a description of the data delivered.

Conclusion:

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 ESSO. It was also clear that three of most important new services that should receive attention were:

1. Resource Discovery Services
2. Resources Registry
3. Resource Distribution and Access Services

A second phase of development might be

4. Transformation Services
5. Semantic Services
6. Linking Services
7. Brokering Services

And a final round could include

8. Fusion and Visualization Services
9. Workflow Services