Technology & Architecture Committee Feedback to the Leadership Council on the EarthCube Architecture Document

9 February 2017

Executive Summary:

The proposed architecture espouses the idea that EarthCube is a system of systems. The architecture defines the full scope of the activities and capabilities that are needed for the geoscientists and helps to centralize EarthCube activities, which have often been disorganized. It draws on resources in the broad community and the inclusion of data quality is good. It is an architecture that has the potential, with modifications, to integrate and assess across disciplines. The shortcomings of the architecture fall into these categories: 1) Lack of implementation and sustainability plan; 2) Lack of focus on geoscience research scientists; 3) The workbench s not well-defined; 4) The use of DataCite for metadata; 5) Data (rather than metadata); and 5) Need to define scope of subsystems and function. The resources that were identified as needed were a Chief Technology Officer or Developer Advocate to oversee the implementation of the architecture and software developers to do the actual implementation. It wil also be necessary to define the AIPs that were considered outside the scope of the Architecture document. Many of the responses recommended a staged implementation using rapid prototyping. The services receiving the highest priority for implementation matched those of the CDF: Resource Distribution and Access Services; Resource Discovery Services; and a Resources Registry.

Background:

The EarthCube Technology & Architecture Committee (TAC) and the Council of Data Facilities (CDF) were requested to provide input on the EarthCube Architecture as proposed by Xentity (https://www.earthcube.org/announcements/architecture-implementation-plan-documents). A survey was created to gather the input from both the TAC and CDF. TAC received 14 responses and the summary of each question is provided below. Many of the responses were quite detailed so a complete set of the responses are in the Appendix.

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 architecture espouses the idea that EarthCube is a system of systems. The architecture defines the full scope of the activities and capabilities that are needed for the geoscientists. It draws on resources in the broad community and the inclusion of data quality is good. The architecture helps to centralize EarthCube activities, which have often been disorganized. It is an architecture that has the potential, with modifications, to integrate and assess across disciplines.

Design:

- Emphasis on registry services, assessment, resource discovery and workbench is well articulated

- Focus on enablement architecture, not on specific science services or capabilities to be provided by the implemented architecture
- In line with other similar architectures
- Suitable for heterogeneous fields in general, not specific for geosciences
- Provides one viable and comprehensive description of the EC ecosystem.
- Includes aspects of discovery, aggregation, registry, assessment, interoperability testing, training and community engagement

AIP:

- Identifies the need for all geoscience resources to have a DOI, 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
- Leverages community resources in working with EC by offering to improve their metadata, exposure and usage
- Proposes development of a Workbench for testing of cyber developments
- Stresses the importance of assessing geoscience resources before including them in the EC architecture and/or Workbench

Inputs from survey respondents:

#1 Strengths: a) comprehensiveness; b) cross-domain interoperability emphasis; c) workbench to underpin cyber community-engaged R&D; d) informed by the (increasing) complexity of GEO; e) reflection of GEO's inherent need to embrace more and faster evolution and learning than most gov't organizations.

#2 "I think that some of the major strengths are at the layer below some of the highlighted items. The idea of a registry and assessment, while fine as products of the architecture, are not the key initial items to me. Rather the capacity to inspect and extract the capability of services and the affordances of data (via metadata about measurements, parameters, units, etc) are key. Some of the highlighted items like semantics and persistent identifiers (with associated metadata profiles) are the real take aways to enable this. Along with things like Linked Open Data (LOD) patterns and traditional harvesting approaches like OAI-PMH and others. A focus on enabling the architecture to advocate for these key aspects (data and service description, semantics, etc) especially drilling down to the measurement and even parameter level would enable a great deal upward on the stack. These capabilities are what enable a registry, assessment, and other high level functions. Basing this approach on web architectures approaches, especially those defined or in the process of definition by groups like W3C and OGC among others is key".

#3 "It maintains the autonomy of the scientist (absence of workflows but service-based), and yet provides multiple options to connect with EC components".

#4 "The architecture is very ambitious, including many functional components. If successful, it could be useful to many people. But the ambition is itself also a potential problem. There is no single architecture that will support every geoscience cyberinfrastructure need, which is why the community-led building block approach made sense".

Q2 What are the major shortcomings of the proposed architecture? Include missing elements you think should be identified. Please be specific and if possible, identify potential mitigation strategies for each.

The Solutions Architecture and separate Implementation Plan document lay out general guidelines to help EarthCube move forward and provide a helpful starting point for next steps. However, these documents specifically consider many of the necessary next steps as being out of scope, ostensibly because they require additional work by the EarthCube community and its governance. The respondents put much effort into their responses to question 2. The complete responses, including mitigation strategies are included in the Appendix. Classes or or even specific comments on concerns tended to be repeated by multiple respondents.

- Lack of implentation and sustainability plan
    - The biggest issue, by far, is that there is no sustainable plan to actually build this new architecture and manage it into the future.
    - The architecture also assumes an "enterprise" world view that is not representative of what will be found in NSF and in the geosciences community. While such an approach might be appropriate for mission driven agencies like NASA, NOAA, and USGS, it doesn't seem aligned with the realities of an NSF ecosystem. The NSF environment tends to be filled with far more groups that spin into and out of existence. There are a fair number of resources that "end" while still having a user community and how to address that potential within the proposed EarthCube architecture doesn't seem at all considered in the architecture. I think ignoring this aspect of sustainability is a major miss.
    - A successful EarthCube must entrain a cadre of volunteers, but volunteers will not/ cannot be expected to plow through hundreds of pages in order to become effective contributors in the realization of the EC architecture nor do they want to have standards and demands imposed on them unless there is a very obvious community and personal benefits.
    - It lacks specificity and will require a lot of community effort to select and adopt specific standards and associated APIs
- Lack of focus on research scientists
    - Like many EarthCube projects, it focuses too heavily on the development of architecture without fully integrating scientists into the development process (starting from this survey, which was originally ONLY sent to TAC and CDF and therefore explicitly ignored scientist inputs).
    - The solutions document does not provide examples; that is, no worked-through use case is presented. It would be good to walk through a few specific geoscience research scenarios, to see how components of the architecture make such scenarios possible
    - The architecture is too complicated, and relies on many assumptions about how scientists work. Any one of the five main components will be challenging to build across the geosciences.
    - Wrong priority on data discovery rather than data handling and access. The initial focus should be on workflows, transformations, mediation and visualization. Most of the interesting capabilities are in Tier 2. Priorities should be bridging across disciplines through tools, though the above data management should also be the priority. The risk is that the recommended implementation will not differentiate EC from other government supported architectures enough.

- Workbench not well-defined
    - The workbench is the key and it is not well enough defined to know what is going to be done, particularly in the early phases of implementation. It is not clear what benefits a read only registry will add to the community. From the perspective of schedule, the community may not want to wait another few years and some form of agile architecture development will be important.
    - The workbench in the document is for developers not end-users. In my opinion, this would be a major mistake and not help the credibility of EarthCube. The document has several missing components - the biggest is data. DataCite may lead you to a repository but a researcher would have no idea of what types of data are available there. A metadata catalog without access to datasets is not what is needed. Then to add insult to injury - now the metadata will be held to some standard (not defined). This is definitely a goal but in the meantime, EarthCube needs to provide a research environment for the researchers and dealing with existing data and metadata is a reality. Education and community outreach can definitely help improve the quality of metadata but the authors of the document do not seem to have any familiarity with the status of metadata in this environment. The architecture as presented in this document is just a testbed for the Building Blocks and Integrative projects - it does nothing for the end-user.
    - One major shortcoming of the proposal is the concept of the “workbench”. I’ve yet to see anything even remotely like this succeed. I think the idea of a general workbench usable across all of geoscience is not only unachievable but a dangerous distraction. Rather a focus on providing numerous “elements” of services, architecture and data that can be composited together into useful “molecules” for the user community is more attractive to me. Aspects of the architecture to support a workbench would be better exposed so that various groups could create workbenches for their communities of practices. These “workbenches” might be mobile apps, web based, Jupyter or R notebooks or whatever. The real point is that a goal of a unified and all serving workbench seems highly dangerous to me to the success of EarthCube in general.
    - I have trouble understanding the "workbench" concept. It is described as something that allows scientists to "test" interoperability solutions. Scientists don't want to test things, they want to do things. What is it that the workbench will help them do? This needs to be articulated much more clearly.

- Use of DataCite for metadata
    - It is assumed that DataCite will be the metadata source, and it will also provide a path to get details of each resource that it exposes. Is there a mechanism to ensure that content provided by DataCite will indeed be sufficient for using workbench services efficiently? Specifically, is DataCite committed to providing additional content via the API and related identifier URL requests. It isn’t the intent of their current DataCite schema. The metadata improvement process seems to involve submitting improvement requests to DataCite, and they would then update their records. What is the estimated turnaround time and mechanism? Also, DataCite is itself an aggregator – and possibly they will relay such requests to their providers. What is the expected turnaround on that?
    - There are already projects funded through EC that crawl the web for specifics sources, extract information from them, generate metadata using semantic technologies (eg the Polar Deep), or automatically enhance metadata descriptions harvested from catalogs

or contributed by researchers (CINERGI), or extract data and metadata from literature (DeepDive), or improve model descriptions to make them interoperable (ESBridge).
    - There is an over-emphases on formal metadata and human curation, with too little recognition of Google-like approaches to resource discovery and valuation (by inference).
- Data
    - Although work has been described in the solutions document to process data and to create workflows, I don't see commentary on how data may be 'queried'. There is discussion of SQL queries on the metadata but not on the data itself. (First, for metadata, many existing Websites already are architected to search on metadata fields to find data sets. Is this architecture plan going to follow those?)
    - The document has several missing components - the biggest is data. DataCite may lead you to a repository but a researcher would have no idea of what types of data are available there. A metadata catalog without access to datasets is not what is needed.
    - **It** is not clear how real-time production data will fit into the architecture. Real-time data in the workbench is described, but access to new sources of data during experiments can often occur at the very last moment and it is not clear the the architecture can support the very rapid turnaround on services that this may require.

- It is important to better define the scope of subsystems and functions.
    - For example, "Minimal metadata enables minimal functionality/capabilities, such as discovery." (p. 14). We collected a fair number of discovery use cases and they have varying metadata requirements – but far from minimal.
    - Key technologies or "subsystems" that are **not** addressed are:
        - "deep description" metadata standards for different primary resource types
        - how to acquire missing metadata (e.g. harvesting, scraping, crowd-sourcing, gamification, wizard-style online forms, etc.)
        - standardized interfaces/APIs that allow machines (and the workbench) to control, query and extract information from different types of resources, whether local or online
        - ontologies and controlled vocabularies needed to support the use of Semantic Web and linked data technologies for packaging/exposing standardized metadata
        - mediators or brokers that use standardized metadata to automatically perform necessary transformations to reconcile differences between resources coupled into workflows (e.g. spatial regridding, reprojection, unit conversion, temporal alignment, semantic mediation, etc.)
        - workflow engines or "resource coupling frameworks"
        - how "exchange items" (typically values of measured or modeled variables on spatial grids) are to be stored and passed between resources in a workflow
        - workflow output analysis tools (like Sandia's DAKOTA, https://dakota.sandia.gov)
        - visualization tools (like Lawrence Livermore's VisIt, https://wci.llnl.gov/simulation/computer-codes/visit)
        - what to include in the workbench GUI.

Q3 a What resources would be needed from the EarthCube architecture to make this connection?

The financial resources that were listed included:

- The best approach may be to fund critical parts of the architecture, such as the workbench, semantic technologies and registry as separate funded projects with a mandate to work together.
- Funding to have a basic architecture realized
- Funds to pay for more than governance - you have to have something to govern; allocate funds to bring the pieces together to build a Knowledge Workbench for researchers not developers
- Support for BB, past and current, to interface. More support for ECITE
- Perhaps a funded tech-hands workshop or a series of smaller meetings focused on specific subsystems (registry, assessment, WB, etc.) and how their capabilities can be implemented with existing EC technologies, COTS, other projects.
- Rapid prototyping
    - Less than $500K for the above rapid prototyping.
- Pay some people to work through example use cases
- Several NSF grants (e.g. ~$1-2m)

This suggestion covers both financial and human resources:

- 3 FTEs

Technical resources included:

- Metadata standards and APIs are critical and must be evaluated and selected by the EC community ASAP
- Improved mediation and cross discipline support
- Interfaces:
    - Specify an initial list of interfaces (or criteria for such interfaces) to be included in EC in the first phase
    - APIs to connect to existing resources; if a testbed is needed, use the NDS Labs
    - Registry API, standards library (for web services and data types), ontology library/repository
- Beyond the readily-available, open-source tools mentioned above (Jupyter Hub & GitHub), perhaps two things: 1) an alternative notebook environment, possibly R-Studio and b) some cloud-based notebook-creation environment for EC users unwilling to install such on their own computers.
- Some simple prototype provided
- Development platorm

This combines some of the above:

- Whatever is needed for several example use cases and for a GUI prototype so people can see what 'EarthCube' might look like

The majority of the human resources that were listed included a Chief Technology Officer (or Developer Advocate) and developers. The human resources that are needed include:

- EarthCube should have a Chief Technology Officer with experience working on community-driven cyber-infrastructure to support geoscientists. This CTO should oversee a software development team with at least 2 (junior-level) people.
- Chief Technology Officer; developers
- Personnel to develop connections to the registry, standards and APIs
- People to work through examples from beginning to end and do an initial prototype
- PIs and architects from each BB and IA

- 2-3 FTEs, probably with Python, R, GitHub and (perhaps) system-architecture experience.
- Semantics expertise
- Support from office in helping us understand the architecture
- Really useful could be dedicated full-time specialist to facilitate scientist-technologist integration

In the "Other" category of resources needed were the ability to make decisions in a timely fashion and the financial, technical, and human resources listed above to concretely determine what is needed.

Q4 Are you familiar with the May Architecture workshop? The report can be found at https://www.earthcube.org/sites/default/files/doc-repository/archws2016_finalreport_3.pdf
Since you are familiar with the workshop report how well does the Solutions Architecture reflect the outcomes of the workshop? Again, please be specific.
The responses were bimodal, which is a bit unusual and shows lack of clarity in community perceptions on both sides.

For those familiar with the Architecture Workshop,

- The Architecture Workshop Report was much better written and organized and the AIP did not really go much further, despite appearing to be highly detailed. The EC architecture will necessarily be based to a very large extent on existing Web, Web Service and Semantic Web standards, but this does not come through in the AIP report.
- The workshop recommendation was focused on a workbench capability for data handling and use. The current architecture proposal focuses on discovery and does not address, except as a second tier all the capabilities recommended by the Workshop
- The workshop didn't go into sufficient detail, compared to the document. The overall design and the key components and workflows reflect the workshop outcomes well
- Reasonably well. I consider inclusion of the workbench and notebook concepts to be positive examples of such reflection. As a workshop participant I'm a bit disappointed that points I made about simplifying abstractions and assumptions are not more evident in the AIP.
- It does not - Figure ES-2 from the workshop report somehow was morphed to Figure 1.1.1.1 in the Architecture Document. The workshop figure makes sense - the meaning is totally lost in the Architecture figure. This is one example but there are more.
- I think it reflects though the workshop was much broader. We did start with technology architecture of connecting BB's but through the workbench put the scientist at the center, which was important.
- Elaborates on the fundmental principles laid out in the workshop, improves upon them

Q5 Other comments:

Since its inception, the EarthCube initiative has prioritized governance over technical coordination to its own detriment. While a lot of money has been spent on governance-related meetings, and end-user workshops, a relatively small amount of money has been spent to help the EC funded projects coordinate their efforts and make the building blocks "fit together" into something bigger. The one meeting with this goal, the Tech Hands Meeting, was very well-received by the "technical people" who need to coordinate their efforts in order to build a coherent EarthCube "architecture". I think part of the problem may be that "non-technical" people do not fully grasp the complexity and level of effort required to develop software and to make technologies interoperable.

Another aspect or perhaps symptom of this issue is a counter-productive dynamic in which funded project PIs are sometimes seen as “technicians” or “nonscientists” who are simply using the EarthCube program to pursue their own technology agendas and interests. In this perception, the PI is characterized as being “out of touch” or unresponsive to the needs of the geoscientists, and as “just doing their own thing”. Sometimes this point of view is articulated (negatively) with a quote from the Field of Dreams movie: “build it and they will come”, to indicate that the “techies” are just building what they want and not what geoscientists need. From our experience, there is little evidence to support this viewpoint. A large fraction of EarthCube PIs are geoscientists who at some point in their careers became interested in the technologies that can make things easier for geoscientists, and over time found themselves spending more time on the technology than the science. This subset of scientists is analogous to people who become interested in how their car works and begin tinkering with or maintaining their own vehicles. It is also analogous to scientists who become involved in the development of sensors or instruments that are needed to advance science. Many of these people become involved in cyberinfrastructure projects that serve their particular geoscience domain, and in those roles there is a very strong emphasis on addressing specific community needs and a necessity of being responsive and answerable to those communities. However, it is often the case that the technologies needed to work “under the hood” to provide the capabilities the geoscientists have requested are technical and complicated. The reason these technologies are necessary to provide what was requested are often not clear to the geoscientist, and they typically take time to develop. Again, the car analogy is helpful here: the community wants a car, but then asks why the technical folks are building carburetors and camshafts. The same could be said of the community wanting their heterogeneous resources (datasets, models, services, etc.) to be interoperable but then ask why the technical folks are working on mediators, metadata, APIs and ontologies. As part of the broad communication issues outlined above, EarthCube governance should try to dispel the misperception that those building the technology to solve their problems don’t understand their needs, because in the majority of cases they do and they are already answerable to or have oversight from a geoscience community.

The process is slow to implement. EC should address one or more rapid prototypes to engage the geoscientist near term.

I found the documents so full of errors and unfinished content (e.g. "XXXX”) that it compounded the already onerous task of wading through 100+ pages of unconsolidated and poorly organized text. There are many important principles and directives in the Solution Architecture document that should have been distilled for the executive summary, but weren't.

That last question about ranking is extremely hard to answer at this point. It also depends on the goals of the solution architecture. What is being solved? This needs to be made more clear before any prioritization is done.

I just want to make sure that scientist voices are properly heard so that all of this technology is useful. I do really appreciate all of the hard work of TAC and CDF and TigerTeam that put together the AIP!

Worked through examples are needed in the solutions document.

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 9 possible services were combined into four groups because of the small sample size. The two figures below show the results with the highest priority having 3 rankings followed by the result with the lowest priority having 3 rankings. The separated rankings are included in the Appendix.

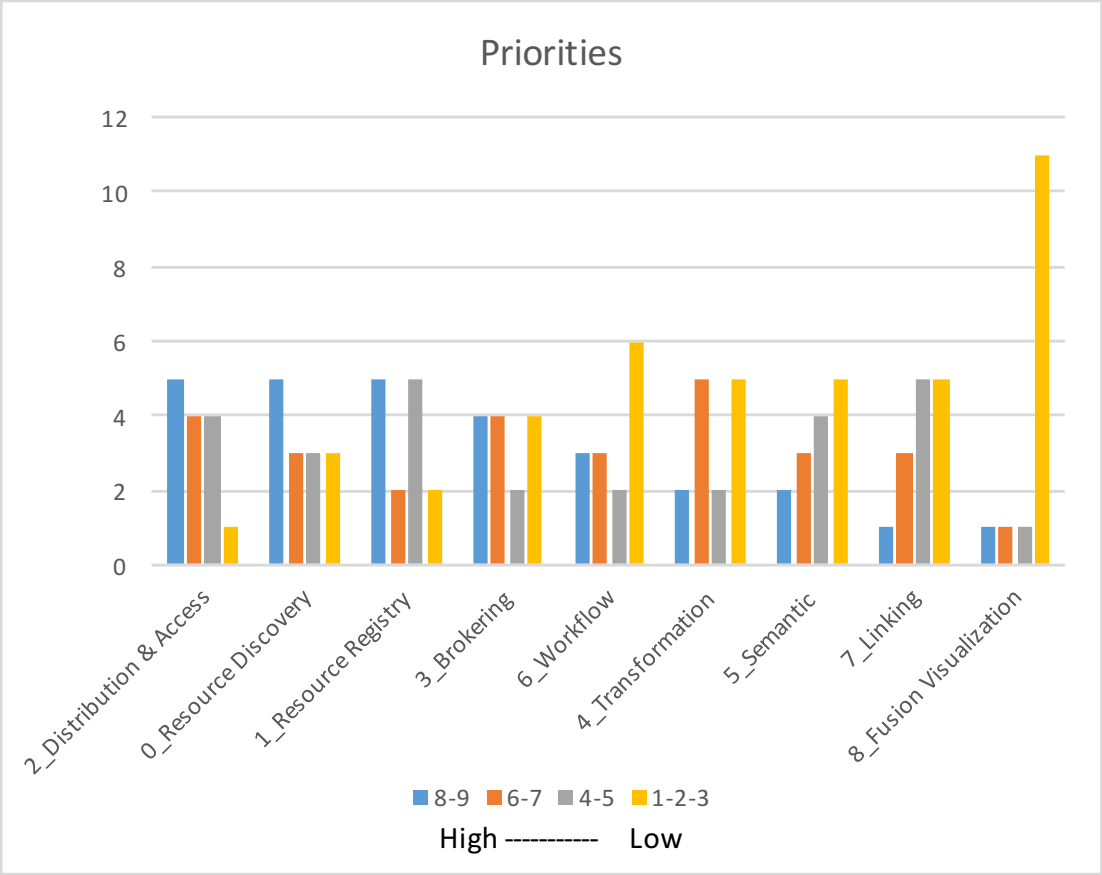

Figure 1 Priorities of services for a phased implementation of the architecture. Ratings of 8-9 indicate most important and ratings of 1-2-3 the least important.

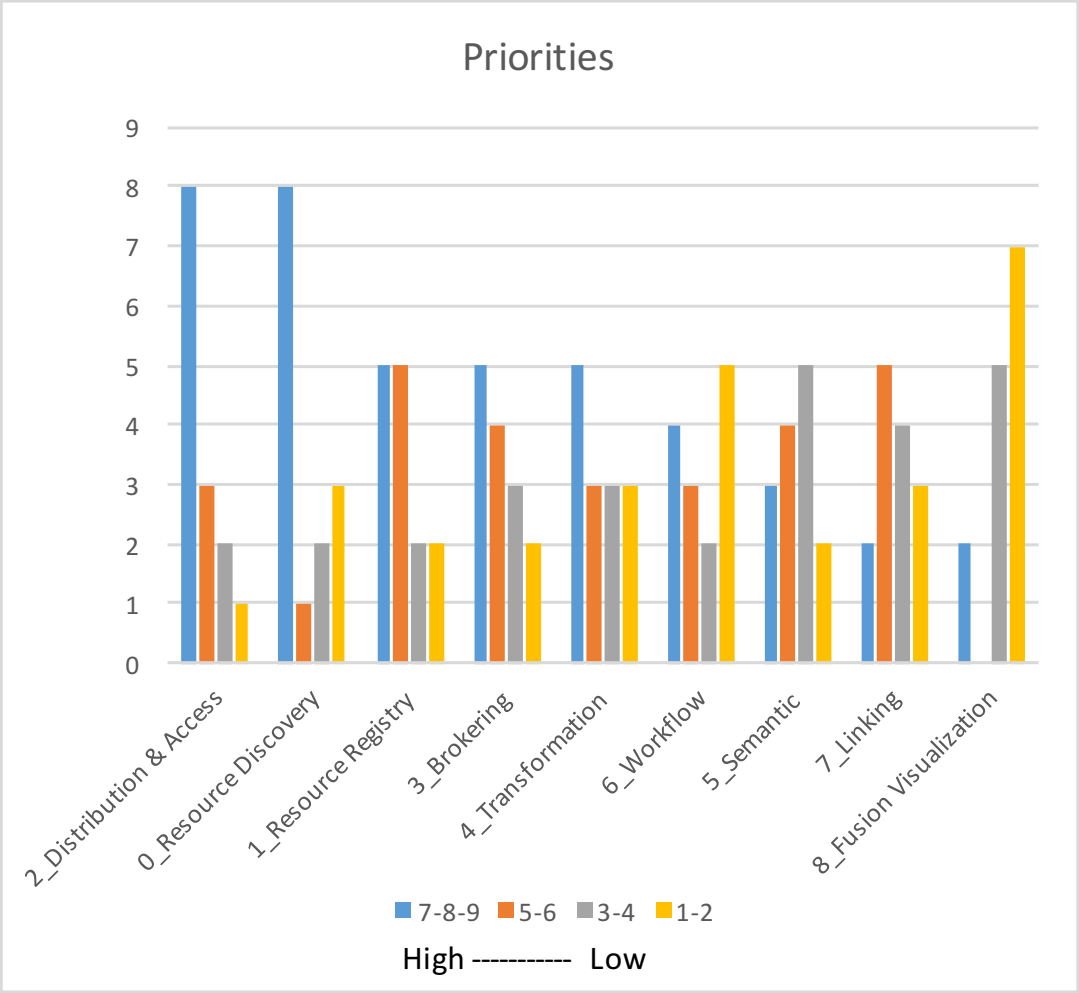

Figure 2 Priorities of services for a phased implementation of the architecture. Ratings of 7-8-9 indicate most important and ratings of 1-2 the least important.

APPENDIX 1: TAC RESPONSES

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.

-In my view, the Architecture Implementation Plan (AIP) does not go nearly far enough in terms of making specific recommendations that build on similar cyberinfrastructure work previously funded by the NSF. However, the AIP does identify the need for all geoscience resources to have a DOI, which implies there would be associated, standardized, resource-type-specific metadata to go with it. Also, the AIP proposes development of a Workbench, although again, the purpose and capabilities of the Workbench are poorly articulated. The AIP also stresses the importance of assessing geoscience

resources before including them in the EC architecture and/or Workbench, which I think most geoscientists would agree is a priority.

-The architecture defines the full scope of the activities and capabilities that are needed for the geoscientists. It draws on resources in the broad community, but there is risk (see next response). The inclusion of data quality is good, though experience has shown it is a hard topic

-This is a solid architecture design, with sufficient detail to guide implementation. I believe the emphasis on registry services, assessment, resource discovery and workbench is well articulated. The focus on enablement architecture, not on specific science services or capabilities to be provided by the implemented architecture - is a critically important (and difficult) perspective to maintain throughout the design. The document does it well. The design is in line with other similar architectures. A big plus is attention to the credits model (in the implementation plan) - which appears related to the NIH Commons credits model. Geoscience research types vary greatly by sub-disciplines, so the main goal - to support the diversity of approaches, data types, tools, etc, and enable interoperation across domains. The design is suitable for highly heterogeneous fields in general, not specific for geosciences.

-Strengths: a) comprehensiveness; b) x-domain interoperability emphasis; c) workbench to underpin community-engaged R&D. Differentiators: strength b) informed by the (increasing) complexity of GEO; strength c) as a reflection of GEO's inherent need to embrace more and faster evolution and learning than most gov't organizations.

-The major strength of the DOCUMENT is it provides a strawman - something in writing that for reviewers to respond.

-Though I am not an architecture expert, the proposed architecture seems to provide one viable and comprehensive description of the EC ecosystem. It includes aspects of discovery, aggregation, registry, assessment, interoperability testing, training and community engagement.

-I think that some of the major strengths are at the layer below some of the highlighted items. The idea of a registry and assessment while fine as products of the architecture, are not the key initial items to me. Rather the capacity to inspect and extract the capability of services and the affordances of data (via metadata about measurements, parameters, units, etc) are key. Some of the highlighted items like semantics and persistent identifiers (with associated metadata profiles) are the real take aways to enable this. Along with things like Linked Open Data (LOD) patterns and traditional harvesting approaches like OAI-PMH and others. A focus on enabling the architecture to advocate for these key aspects (data and service description, semantics, etc) especially drilling down to the measurement and even parameter level would enable a great deal upward on the stack. These capabilities are what enable a registry, assessment, and other high level functions. Basing this approach on web architectures approaches especially those defined or in the process of definition by groups like W3C and OGC among others is key.

-The proposed architecture espouses the idea that EarthCube is a system of systems. It maintains the autonomy of the scientist (absence of workflows but service-based), and yet provides multiple options to connect with EC components. I must admit I have not seen many government-supported architectures but what I see is that the architecture is begin designed based on latest technology as is being used/adopted in building blocks. Government-funded architectures due to legacy must often adopt simpler solutions or make simplifying assumptions.

-The AIP wisely avoids building entirely new infrastructure and instead leverages the assets in existing community catalogs and data repositories. It also engages these community resources in working with EC by offering to improve their metadata, exposure and usage.

-The architecture is very ambitious, including many functional components. If successful, it could be useful to many people. But the ambition is itself also a potential problem. There is no single architecture that will support every geoscience cyberinfrastructure need, which is why the community-led building block approach made sense.

-Good that it is interoperable with other resources. Good that it pulls together all existing EarthCube products so they can be more easily located and used. Designed (in theory) to be more easily accessed (and assessed) by scientists. Helps to centralized EarthCube activities, which have often been disorganized.

-This document is hard to read, with spelling and grammatical errors and no examples to follow. Sometimes the meaning of sentences is obscure. Nevertheless, it is apparent that a lot of thought has been given to many details of the architecture.

-It is an architecture specifically developed to integrate and assess across disciplines.

Q2 What are the major shortcomings of the proposed architecture? Include missing elements you think should be identified. Please be specific and if possible, identify potential mitigation strategies for each.

-

Scope of the Solutions Architecture and Implementation Plan and Next Steps

The Solutions Architecture and separate Implementation Plan document prepared by Xentity lay out general guidelines to help EarthCube move forward and provide a helpful starting point for next steps. However, these documents specifically consider many of the necessary next steps as being out of scope, ostensibly because they require additional work by the EarthCube community and its governance. (See separate subsection below.) Most of these critical next steps require that the community:

(1) evaluates and selects one or more technology alternatives when they exist or

(2) develop required technologies when existing options are deemed inadequate.

Key among these technologies or "subsystems" that are **not** addressed in these documents are:

1. "deep description" metadata standards for different primary resource types
2. how to acquire missing metadata (e.g. harvesting, scraping, crowd-sourcing, gamification, wizard-style online forms, etc.)
3. standardized interfaces/APIs that allow machines (and the workbench) to control, query and extract information from different types of resources, whether local or online
4. ontologies and controlled vocabularies needed to support the use of Semantic Web and linked data technologies for packaging/exposing standardized metadata
5. mediators or brokers that use standardized metadata to automatically perform necessary transformations to reconcile differences between resources coupled into workflows (e.g. spatial regridding, reprojection, unit conversion, temporal alignment, semantic mediation, etc.)
6. workflow engines or “resource coupling frameworks”
7. how “exchange items” (typically values of measured or modeled variables on spatial grids) are to be stored and passed between resources in a workflow
8. workflow output analysis tools (like Sandia's DAKOTA, https://dakota.sandia.gov)
9. visualization tools (like Lawrence Livermore's VisIt, https://wci.llnl.gov/simulation/computer-codes/visit)
10. what to include in the workbench GUI.

Referring to item (10), the workbench GUI must allow users to do the following:

- browse available primary resources by resource type,
- examine documentation and metadata for resources,
- select the resources they may want to use in some workflow,
- determine the level of available metadata for each selected resource and whether it is sufficient to support the mediation steps that must occur at connections in the workflow,
- have the option of entering/discovering additional/missing metadata to raise resources to higher levels,
- prepare input data files required by computational models (GUI assisted)
- execute workflows, which typically requires connecting two or more resources in a "chain" and requires a “workflow engine” or “resource coupling framework”
- visualize data sets and output from workflows
- save workflows for re-execution and modification.

In considering software/technology alternatives for the subsystems listed above, I believe that the following overarching design criteria are helpful for ranking alternatives.

- open-source software
- support for “big data” (i.e. scalability and ability to utilize multiple processors)

- maturity
- long-term funding support and dedicated development team
- good documentation
- active user and developer communities
- leverage and extend existing software whenever possible
- Python APIs when possible (widely used)

Scoping of the Solutions Architecture and Implementation Plan

According to the Solutions Architecture document (see p. 3), its intent or scope is to:

1. "document a business context in order to prioritize services consistent with needs."
2. "describe these business and technical concepts in a manner that directly aligns their development and implementation to EC program goals."
3. "provide requirements and context for EC processes, data, applications/services, technology, management controls, and security requirements, and other elements central to the Architecture Implementation Plan."
4. "provide measurement guidance for the sponsor to help establish implementer selection criteria." (p. 4)

However, excerpting from the document, the Solutions Architecture does not:

1. "architect the Science Resources and Workbench Capabilities themselves, ... but the collaborative and community framework in which to do so." (p. 3).
2. prescribe which standards are to be supported (or are appropriate) for metadata or interfaces/APIs.
3. describe how to implement the "actual Science Resources, the actual assessments, nor the Workbench Capabilities". (p. 4)
4. include "builder physical and operational views."
5. discuss how Semantic Web technologies are to be utilized.

The Solutions Architecture document also states that: "After Scientist's integrated resource solutions are tested, they could be interoperable resourced and transferred to external existing scalable operational environments for full scale and repeated execution." However, existing modeling framework "workbenches" (e.g. CSDMS Web Modeling Tool, HydroShare) actually support both operational use and development on the same high-performance cluster, with development occurring on a "dev branch" that is not exposed to users until ready. Users connect to the cluster via browser-based client applications that run on their own desktop or laptop computer.

In addition, the Solutions Architecture document states that: "The implementer should be sure to demonstrate with their implementation how their design meets all the service components, process

flow, and data flow requirements, and allow governance to approve exceptions or deferring scope. The implementer would use their own approach and templates for demonstrating physical process, data, service, and technology architecture work products." (p. 27)

In the following sections, four major and specific shortcomings of the AIP are discussed in detail as "Next Steps" for EarthCube. This discussion is informed by having served as the Chief Software Architect for a large, successful cyber-infrastructure project called CSDMS (Community Surface Dynamics Modeling System), which involved assessing a wide variety of technologies to support interoperability and designing several critical CSDMS subsystems, including the Basic Model Interface, the CSDMS Standard Names and much of the CSDMS modeling framework. The following sections are also informed by having collaborated with other software architects over a 10-year period from around the world, and from in-depth communications and collaboration with numerous EarthCube PIs as a lead PI or co-PI on 7 different EarthCube funded projects. I also led the Tech Hands Meeting and attended 2 of the 3 EarthCube Architecture workshops.

Next Step 1:

Assessment of Data/Metadata Standards Currently Used by Data Facilities

For each data facility in the CDF (and ideally for major federal data providers as well), EarthCube governance needs to request a description/synopsis of the types of data they provide and which standards (i.e. metadata schemas, formats, protocols, APIs, cataloging, etc.) they currently use. This should include recognized strengths and weaknesses, the degree of adoption of linked data and Semantic Web technologies (e.g. RDF, SKOS, OWL, PROV, etc.), adequacy of ontologies and controlled vocabularies, APIs for machine access to holdings, etc. In some cases it may be feasible for a data facility to adopt new standards or to provide an alternate, standardized API to their data holdings. However, we anticipate that it will usually be necessary for EarthCube to develop adapters that convert between currently-used standards and EC-supported standards. This will require close coordination between technical staff at the data facilities and technical staff responsible for implementing the EC architecture and/or Workbench. The Adapter Pattern is a classic concept in software engineering that is particularly well-suited to the integration of heterogeneous resources and systems into a “system of systems”. It typically requires minimal changes to existing systems and resources but often requires appropriate metadata. See Wikipedia’s article: Adapter Pattern.

Next Step 2:

Identify "Deep-Description" Metadata Schemas for Different Resource Types to Support Advanced Capabilities and Greater Automation in the Workbench

The Solutions Architecture document specifically mentions the DataCite project as an existing technology that can be leveraged by EarthCube. DataCite's tagline is "Find, access and reuse data". DataCite provides persistent identifiers (DOIs) for "datasets, workflows and standards". They "support the creation and allocation of DOIs and accompanying metadata. The DataCite Metadata Schema is described here: https://schema.datacite.org. According to their site: “The DataCite Metadata Schema is a list of **core metadata** properties chosen for an accurate and consistent identification of a

resource for citation and retrieval purposes, along with recommended use instructions." The metadata schema has a mandatory field called **ResourceType**, with a subfield **resourceTypeGeneral** that takes values from the following controlled list:

1. Audiovisual, (2) Collection, (3) Dataset, (4) Event, (5) Image, (6) InteractiveResource, (7) Model, (8) PhysicalObject, (9) Service, (10) Software, (11) Sound, (12) Text, (13) Workflow, (14) Other

See p. 15 of this PDF:
DataCite Metadata Working Group. (2016). DataCite Metadata Schema Documentation for the Publication and Citation of Research Data. Version 4.0. DataCite e.V. http://doi.org/10.5438/0012.
This may be viewed as a listing of **primary resource types**, defined as resources that geoscientists want to find and work with. For clarity, it is recommended that the term **secondary resources** be used to describe the types of resources that are meant to work "behind the scenes" to support user-desired capabilities. Collectively, these secondary resources are what we refer to as “cyberinfrastructure” and they are precisely the types of resources that are being developed within EarthCube Building Block projects. (While every metaphor has limitations, a helpful metaphor for communication is to view the secondary resources as the parts of a car under the hood, while the car itself could be the proposed Workbench or some other set of user capabilities that are enabled by these additional resources.) Secondary resources include:

1. standardized interfaces or APIs to query, control and extract information from primary resources
2. ontologies and controlled vocabularies for describing each type of primary resources
3. metadata that describes specific resources via RDF files, using appropriate ontologies
4. metadata acquisition tools (e.g. harvesters, scrapers)
5. catalogs and registries (which may include both primary and secondary resources)
6. mediators or brokers (for conversion, transformation, regridding, etc.)
7. tools to acquire metadata for primary resources (e.g. by harvesting or direct entry)
8. graphical user interfaces
9. visualization software
10. analysis software (e.g. for uncertainty analysis)
11. low-level web services and other utilities.

Note that DataCite's Metadata Schema primary purpose is to support discovery/access and citation of primary resources (as is also true about Dublin Core) and can be viewed as a very basic or **least common denominator metadata schema** applicable to a broad variety of resource types. *However, a significant amount of additional resource-type-specific metadata is needed to support more advanced workbench capabilities such as automated mediation when heterogeneous resources are connected into workflows.* For several specific geoscience resource types -- including most types of geospatial data, computational models and other geoscience software -- "deep-description metadata schemas" already exist or are under active development (e.g. ISO 19115, GeoTIFF, CF Conventions, OntoSoft, CSDMS Standard Names). These often employ ontologies and the same deep metadata may be packaged in a variety of different formats, such as XML, RDF, JSON, etc. Led by EarthCube governance, the EarthCube

community will need to examine, evaluate and make selections from the "deep metadata schemas" that are currently available, and will likely also need to devote resources toward extending them as needed to support specific capabilities that are desirable for the Workbench. The importance of metadata for this purpose is well-known and was emphasized in the Solutions Architecture document. However, acquisition of high-quality metadata poses significant challenges and cannot be fully automated. Domain experts will need to be involved to ensure that information is captured correctly.

In view of these observations, the EarthCube community needs to (1) determine appropriate, standard metadata schemas for different resource types to support desired workbench capabilities and (2) find effective methods for acquiring the high-quality metadata indicated in the schema for individual resources of interest to the geoscience community. Finally, it is important to note that metadata entries should use standard terms from ontologies and controlled vocabularies as opposed to "free form" text strings -- otherwise the metadata will be human readable but not machine actionable.

Next Step 3:

Identify Which Mediators are Required for the Workbench, then Evaluate and Select Ones that meet Design Criteria

Mediators (also called brokers or service components) play a critical role in "resource composition frameworks" such as the proposed Workbench. They are employed by all modern model coupling frameworks (e.g. CSDMS, ESMF, OMS, OpenMI) and are invoked automatically by these frameworks to reconcile differences between heterogeneous resources, especially with regard to the exchange items (e.g. values of variables, often on spatial grids) that must be passed between coupled resources in a workflow. However, the frameworks can only invoke them automatically when needed if the required resource metadata is available (e.g. grids, units, georeferencing, variable names) and can be accessed by the framework (e.g. through a standard API). The types of mediators needed follow directly from answering the question: "In what specific ways do the resources in my workflow differ from one another?" Examples of powerful, well-established, open-source mediators are Unidata's UDUNITS for unit conversion and the Earth System Modeling Framework (ESMF) regridder for spatial interpolation. Both of these mediators can be accessed through Python APIs -- UDUNITS via the cfunits Python package and the ESMF regridder through ESMPy. Python APIs are highly desirable and greatly simplify integration into a "system of systems". Virtually all modern GIS (Geographic Information System) applications now provide a Python API (e.g. ArcGIS, GRASS, QGIS), as does the powerful VisIt visualization package. Python is therefore also a logical choice for the middleware (or "connective tissue") that will be used under the hood in the Workbench to pull together all of the component technologies.

The EarthCube Technology and Architecture Committee (TAC) itself, or a separate working group should be appointed to determine which mediators will be needed to support desired workbench functionality and to compile a list of candidate mediators to be evaluated for inclusion. While many excellent

mediators already exist, new ones can be developed as needed but some advanced planning will help to ensure that the metadata required for their automatic invocation is collected and stored.

Next Step 4:

Identify Appropriate Interfaces/APIs for Interacting with Different Resource Types

Standardized interfaces --- in the sense of APIs not GUIs --- will be critical to developing the desired functionality of the proposed Workbench. The middleware that underpins the Workbench will make heavy use of these APIs in order to (1) request data or metadata from resources, (2) post data or metadata to resources, and (3) control resources (as for time-stepping models). The "request" functionality is implemented with **getter functions** in the API (alternately known as retrieving, extracting or reading), while the "post" functionality is implemented with **setter functions** in the API (alternately known as placing or writing). Due to the vast number of resources that are expected to be made available to users in the Workbench, ***standardized*** APIs are essential --- development and maintenance efforts quickly spiral out of control if each individual resource (or even data facility) is allowed its own API and must be supported. A wide variety of standardized APIs already exist for interacting with different resource types. For data sets, there are many, such as APIs for NetCDF files. For computational models, several good ones have emerged in recent years (e.g. the Basic Model Interface and OpenMI) and some projects (e.g. the EarthCube-funded Earth System Bridge project) have developed adapters that can convert between different APIs for models (e.g. BMI to OpenMI, BMI to OMS, BMI to ESMF/NUOPC). For web services, there are also many, but the service APIs listed on this OGC implementation standards page are probably the most mature and widely used. In the context of multi-processor computation, MPI (Message Passing Interface), OpenMP and CUDA (for GPUs) are some of the key API standards.

It is important to note that once EarthCube decides on and announces which APIs (and other standards) will be supported or used by the Workbench and its subsystems, this will create the opportunity for anyone --- including EarthCube funded project PIs and others in the EarthCube community --- to prepare their resources (primary or secondary) for easy inclusion in the Workbench or architecture. It is a key step toward having an open development environment in which everyone is welcome to contribute and can do so in a straightforward manner.

-My work as a geoscientist is more in data handling than focused on discovery. The architecture proposed in the early (and critical phase) is emphasizing discovery. The initial focus should be on workflows, transformations, mediation and visualization. Most of the interesting capabilities are in Tier 2. The priority is wrong. Bridging across disciplines through the above data management should be the priority. The risk is that the recommended implementation will not be differentiate EC from other government supported architectures enough. The workbench is the key and it is not well enough defined to know what is going to be done, particularly in the early phases of implementation. It is not clear what a read only registry added to the community. The community may not want to wait another few years.

-

1) The architecture does a good job describing a system supporting varying needs, data types, etc by domains - it is not clear what makes it geospatial, besides referring to geo data providers, OGC/INSPIRE services, mentioning spatial data types. It would be good to walk through a few specific geoscience research scenarios, to see how components of the architecture make such scenarios possible
2) I would give a bit more attention to semantic services. The document mentions reliance on controlled lists and ontologies, but in reality there is little consensus about them, especially across domains. So the system may need to maintain multiple versions of such controlled lists, and mappings.
3) I have a few concerns about the registry portion - though these are relatively minor (and I suspect, primarily result from my limited time digesting the document).

Nevertheless:
a) It is assumed that DataCite will be the metadata source, and it will also provide a path to get details of each resource that it exposes. Is there a mechanism to ensure that content provided by DataCite will indeed be sufficient for using workbench services efficiently? Specifically, is Datacite committed to providing additional content via the API and related identifier URL requests. It isn't the intent of their current DataCite schema, AFAIK.

b) I actually tried to search DataCite to imagine how it may be used in EC context:

i. Search for IRIS. 77357 records returned, but our IRIS is not among the first 100 records (maybe more). If we search for "iris seismology" the second record is the re3data entry for IRIS. Clicking through to re3data, we may eventually get to an IRIS page listing available APIs - but this is a multi-step process going via a few non-machine actionable documents.

ii. Search for CUAHSI. The first result is CUAHSI HIS, from re3data (good!). The DOI link sends to re3data (with some incorrect metadata, eg a claim that it is a repository with 5.1 bil points) and a link to HISCentral (but not a link that would get us machine-actionable metadata). The second result is four records pointing to the same 2013 poster on Figshare. Then we have a big block of almost identical records – over 90 (!!) records pointing to the same Baltimore service (different DOIs though). This is the bulk of 106 records returned on this search.

iii. Search for discharge: 14452 results. But they don't include either CUAHSI HIS or USGS NWIS endpoints (though both are indexed in re3data). Makes me think that Datacite is not intended to support this type of data discovery. It appears that getting via DataCite to actionable metadata isn't possible in some cases. Of course, DataCite will improve, hopefully with guidance from EC: it would be good to add a discussion of mechanisms and directions for such improvement to make it useful for EC.

c) As I understand, the metadata improvement process will involve submitting improvement requests to DataCite, and they updating their records. What is the estimated turnaround time and mechanism? Also, DataCite is itself an aggregator – and possibly they will relay such requests to their providers. What is the expected turnaround on that?

d) There are already projects funded through EC that crawl the web for specifics sources, extract information from them, generate metadata using semantic technologies (eg the Polar Deep Web presented at the last TAC call), or automatically enhance metadata descriptions harvested from catalogs or contributed by researchers (CINERGI), or extract data and metadata from literature (DeepDive), or improve model descriptions to make them interoperable (ESBridge). This additional information automatically generated through harvesting/crawling/digitizing and subsequent semantic enhancement – provides additional support for discovery. The architecture document mentions (in 1.1.1.1.5.3) that eventually such information may be included in the EC Registry (as part of Resource Interoperability Assessment? Does assessment imply enhancement?) – but it is left to the future, and the mechanism is unclear. Yet, these projects exist already.

e) It is important to better define the scope of subsystems and functions. For example, there is a statement "Minimal metadata enables minimal functionality/capabilities, such as discovery." (p. 14). We collected a fair number of discovery use cases and they have varying metadata requirements – but far from minimal.

f) Not sure I follow 1.1.1.2.10.1. The partner aggregator will provide DOI-ed records. What updates are referred to here? Any specific mechanism for dynamic data citation here?

-Major shortcomings: a) daunting complexity, reflecting attempt to be comprehensive in the face of GEO's (present) complexity; b) shortage of simplifying abstractions and/or assumptions; b) over-emphases on formal metadata & human curation, with too little recognition of Google-like approaches to resource discovery & valuation (by inference).

Possible mitigation: reorder AIP, altering its emphases to i) prioritize workbench implementation; ii) recognize notebook creation as the _primary_ workbench function; iii) build in greater potential for automated methods--exploiting machine-actionable notebooks, e.g.--to replace human-intensive approaches to resource discovery & curation.

Notes: mitigation ii) reflects a simplifying assumption based on trends (evident at the Fall AGU mtg, e.g.); mitigation iii) reflects a Google-informed abstraction that tends to erase the data/metadata distinction; mitigation i) assumes the correctness of mitigations ii) and iii).

Final note: exacerbating my concern re AIP complexity are two factors -- 1) a successful EarthCube must entrain a cadre of volunteers, but volunteers will not/ cannot be expected to plow through hundreds of pages in order to become effective contributors in the realization of the EC architecture; 2) the absence of simplifying assumptions in the AIP reminds me of the terribly ill-fated first version of NASA's EOSDIS.

-Unfortunately, I don't think Xentity understood the problem. EarthCube has been in existence for many years but what exists that end-users can use to do research that they've never done before? The workbench in the document is for developers not end-users. In my opinion, this would be a major mistake and not help the credibility of EarthCube. The document has several missing components - the

biggest is data. DataCite may lead you to a repository but a researcher would have no idea of what types of data are available there. A metadata catalog without access to datasets is not what is needed. Then to add insult to injury - now the metadata will be held to some standard (not defined). This is definitely a goal but in the meantime, EarthCube needs to provide a research environment for the researchers and dealing with existing data and metadata is a reality. Education and community outreach can definitely help improve the quality of metadata but the authors of the document do not seem to have any familiarity with the status of metadata in this environment. The architecture as presented in this document is just a testbed for the Building Blocks and Integrative projects - it does nothing for the end-user.

-It is not clear how real-time production data will fit into the architecture. Real-time data in the workbench is described, but access to new sources of data during experiments can often occur at the very last moment and it is not clear the the architecture can support the very rapid turnaround on services that this may require,

-I fear that one major shortcoming of the proposal is the concept of the “workbench”. I’ve yet to see anything even remotely like this that succeeds. I think the idea of a general workbench usable across all of geoscience is not only unachievable but a dangerous distraction. Rather a focus on providing numerous “elements” of services, architecture and data that can be composited together into useful “molecules” for the user community is more attractive to me. Aspects of the architecture to support a workbench would be better exposed so that various groups could create workbenches for their communities of practices. These “workbenches” might be mobile apps, web based, Jupyter or R notebooks or whatever. The real point is that a goal of a unified and all serving workbench seems highly dangerous to me to the success of EarthCube in general.

The architecture also assumes a “enterprise” world view that I think if not representative of what will be found in NSF. While such an approach might be appropriate for mission driven agencies like NASA, NOAA, and USGS, it doesn’t seem aligned with the realities of an NSF ecosystem. The NSF environment tends to be filled with far more groups that spin into and out of existence. Obviously this is not the case across the board in NSF with many long living facilities key aspects of the NSF landscape. However, there are a fair number of resources that “end” while still having a user community and how to address that potential within the proposed EarthCube architecture doesn’t seem at all addressed in the architecture. I think addressing this aspect of sustainability is a major miss.

-The architecture does not place emphasis on visualization which are often the interfaces that users interact with The architecture does not place emphasis on feedback, and user studies

-effective discovery based on curated metadata has to rely on more than keyword matches on the common set of fields harvested from all geoscience domains. While the architecture refers to "extended metadata" that will be kept at the source, not in the EC registry, it was not made clear how this metadata would be utilized in doing intelligent searches. this is hugely important, because users won't

come back to EC if they don't find what's useful to them, and can subsequently be worked with in EC's R&D environment. These seems to be a large "magic happens here" component in the architecture.

-In general, i think the architecture is too complicated, and relies on many assumptions about how scientists work. Any one of the five main components will be challenging to build across the geosciences. With regard to more specific issues, first, any architecture that relies on DOIs to be present for every asset is immediately setting itself up to be at best a partial solution because not everything has (or will have, or need to have, or can have) DOIs. The potential mitigation strategy is to encourage and support but not require DOIs. The second shortcoming is that the "Resource Assessment Presentment" will be challenging to implement because gathering assessment data that is consistent, reliable, and trustworthy will be extremely difficult. The geoscience data world is not Wikipedia, which can rely on thousands and thousands of contributors to achieve some trustworthiness and self-correction. I don't know of a mitigation strategy for this. Finally, I have trouble understanding the "workbench" concept. It is described as something that allows scientists to "test" interoperability solutions. Scientists don't want to test things, they want to do things. What is it that the workbench will help them do? This needs to be articulated much more clearly.

-The biggest issue, by far, is that there is no sustainable plan to actually build this new architecture and manage it into the future. Even if ESSO hires a additional staff, this is a job that will take thousands of person-hours and therefore require several full-time people will specific expertise on the technology and scientist requirements. Also, like many EarthCube projects, it focuses too heavily on the development of architecture without fully integrating scientists into the development process (starting from this survey, which was originally ONLY sent to TAC and CDF and therefore explicitly ignored scientist inputs).

-

1. The solutions document does not provide examples; that is, no worked-through use case is presented. This creates questions and possible doubts about whether user needs will be met. There is no description of the use of the registry or workbench from a user's perspective, and this needs to be here. Actual use cases need to be presented with potential screen shots and completely worked out solutions/answers. Use cases were compiled by EarthCube, but no examples are given here as to what these use cases would actually look like when being solved using this EarthCube architecture.
Without some concrete examples using real data for real tasks, including potential screen shots and a walk-through from a user perspective, it's not possible to determine the workability of this architecture overall. That is, if it ends up being just a collection of tools, it may not be used. Also, the registry part (versus the workbench) is confusing because the solutions document seems to expect users to do general Internet searches rather than use EarthCube's compilation of registries. What's the big picture here from a user perspective?

2. I'm not sure that the analogy to movie data is relevant.

3. Providing assessments to metadata seems vague. How will this be done? What sorts of information would be additionally provided to existing metadata? In my opinion, this is a hard problem and a research issue. That is, for FGDC metadata, for example, how and what to add to actually make the data discoverable and useful for a particular problem is a research issue. Existing metadata searches on geospatial Websites, for example, can yield hundreds of data sets. Figuring out which data sets are right for a particular task currently is more in the brains of a few 'people in the know' and not apparent to most users. No examples are given as to what additional value-added assessments might be or how those would be useful.

4. Although work has been described in the solutions document to process data and to create workflows, I don't see commentary on how data may be 'queried'. There is discussion of SQL queries on the metadata but not on the data itself. (First, for metadata, many existing Websites already are architected to search on metadata fields to find data sets. Is this architecture plan going to follow those?)
But, back to querying the data itself, I didn't see that mentioned. Partially, it would involve accessing data over the Web that is in Database Management Systems, along with giving some help to the user to know the schema of the data, including data types and meanings. And, the user would either have to know SQL, or else simple interfaces would be presented to the user (but those might be limited in what is allowed). How would data in separate DBMSs be able to be queried?
Also, I know some EarthCube projects are putting data into RDF format and providing endpoints for SPARQL queries. But, few people would be able to write SPARQL queries, even if the linked data are already semantically resolved across data sets (also note this implies that different data sets have already been combined into one linked cloud). What data sets to link together and how to present easy (but not restricted and simplified) interfaces so that users can query the data is still a research issue.
As part of this, the types of data available in EarthCube need to be delineated as to whether it makes sense to query the fields or not. For the data that can be queried and related to other data, ETL type procedures and semantic resolution are needed. Again, this is still a research area, so I understand that it's not in this solution architecture, but it is absolutely needed. This is the future, and only with being able to relate and integrate data will more of the vision of EarthCube be possible.
How will querying of the data (through DBMSs or RDF networks) be part of the Workbench?

5. Will EarthCube resemble HUBzero in some way?

6. Without some concrete examples using real data for real tasks, including potential screen shots and a walk-through from a user perspective, it's not possible to determine the workability of this architecture overall. That is, if it ends up being just a collection of tools, it may not be used or be useful. Also, the registry part (versus the workbench) is confusing because the solutions document seems to expect users to do general Internet searches rather than use EarthCube's compilation of registries. What's the big picture here as to a user? Mitigation: Hire people to actually solve some of the use cases and then show how it could be done in the envisioned EarthCube.

-It lacks specificity and will require a lot of community effort to select and adopt specific standards and associated APIs. But the AIP is specific enough to provide a framework which can define a scope for planning and first stages of implementation.

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?

-In order to achieve the degree of interoperability that geoscientists want (which goes beyond simple resource discovery and download to support usage and workflow coupling) -- as articulated in numerous End User Workshops -- EarthCube will need to make a concerted effort (or provide tools and mechanisms for this purpose) to generate "deep-descriptive" and standardized metadata (and APIs) for a variety of geoscience resource types, especially computational models and data sets of all kinds. There is no other practical solution to achieving the desired interoperability with such a heterogeneous set of resources. There should first be a community effort to determine what "metadata schemas" currently exist that can fulfill this need. After that, data facilities and other resource providers (e.g. model developers, data collectors) will need to make an investment in this activity, which should be distributed and shared to the extent possible. While this is not necessarily trivial, it also doesn't need to be especially onerous. Many data providers and federal agencies and EC funded projects are already working on this.

-enough detail in the architecture capability and a change in priorities.

-It would be useful to go through an exercise of mapping BB's and IA's architecture diagrams to this architecture, identify differences, and decide what needs to be adjusted. BB's and IA's work against specific use cases, so syncing in the context of these use cases would make sense

-I suggest the rapid prototyping of an EC workbench built almost entirely around one or more of the increasingly-popular, open-source, notebook-creation environments (such as Jupyter Hub), augmented by a notebook repository (built on GitHub?) that reflects some key EC-approved guidelines and templates.

-Start over - have someone actually review the outputs of the many workshops held with the end-user community. Pay attention to the workshops that have been held by/for TAC. There are many projects in this environment - pay attention to how they accomplish what they do. Pay attention to the European projects - in particular EPOS and do not try to do this in a vacuum.

-Definition of standards and APIs for data access and interoperability, including the registry and any ontologies that may be needed. ORCIDs and the DataCite metadata are straightforward.

-While the document goes into some detail on the issue of API and other services there was not much discussion on patterns and architecture that could be implemented by funded projects to facilitate the

connection into a larger architecture environment. Issues of authentication, metadata and other standards that could give guidance from the start to funded projects seems thin. I think some work needs to take place to develop the practices and policies that could be implemented by groups to enable these connections. These might be along the lines of oauth style connections, patterns for describing data and services, patterns for application state within web based resources and other approaches.

-I do not think that this information will be evident in a straightforward manner. Since the architecture is high-level, as a first step each BB/capability must show how they map to the architecture. This will highlight common elements and interfaces in the architecture that are needed. If there is some consolidation we have to link with use cases (other activities in EC) and also see what is part of external CI that can be brought as part of the architecture.

-clearly, the vocabularies of the various domains have to come into play when discovering, accessing and using data across disciplines. this needs to be part of the core infrastructure. the community can contribute and build off of a semantic component, but the capacity to hold and utilize vocabularies has to be part of the infrastructure from the start.

-More information about the who and the what. In reading the EC Solution Architecture document, I made many notes to myself asking the same question, "Who is going to do this?" With regards to the "what", I believe I understand what a registry and discovery service might look like, and I can imagine what the community training aspect might be, but I do not have a clear view of the envisioned end state of the resource assessment presentment or workbench.

-What do scientists actually want and need, and what are they actually capable of using? By pulling all EarthCube resources together in one place, the AIP (if implemented, which is a big if), will in principle make things a bit easier. But only if developed and tested with scientists seriously involved at every step of the process. And such scientist involvement is not going to happen for free, as existing volunteers are already strained. Real long-term funding support for scientists (perhaps through the new "Integration" NSF funding stream) needs to be provided to ensure such scientist-based information at every step of the process.

-I recommend that people be hired to solve (i.e., work through) a number of the already gathered use cases such that there are concrete solutions. These solutions can then be prototyped in an initial EarthCube architecture. I think a prototype with concrete examples and screen shots will reveal if this potential architecture will work, and people can then react to something concrete.

Q3 a What resources would be needed from the EarthCube architecture to make this connection?

Financial-The best approach may be to fund critical parts of the architecture, such as the workbench, semantic technologies and registry as separate funded projects with a mandate to work together.

Technical-Metadata standards and APIs are critical and must be evaluated and selected by the EC community ASAP.
Human-EarthCube should have a Chief Technology Officer with experience working on community-driven cyber-infrastructure to support geoscientists. This CTO should oversee a software development team with at least 2 (junior-level) people.

Financial-support for BB, past and current, to interface. More support for ECITE
Technical-improved mediation and cross discipline support

Financial-Perhaps a funded tech-hands workshop or a series of smaller meetings focused on specific subsystems (registry, assessment, WB, etc.) and how their capabilities can be implemented with existing EC technologies, COTS, other projects.
Technical-Specify an initial list of interfaces (or criteria for such interfaces) to be included in EC in the first phase
Human-PIs and architects from each BB and IA

Financial-Less than $500K for the above rapid prototyping.
Technical-Beyond the readily-available, open-source tools mentioned above (Jupyter Hub & GitHub), perhaps two things: 1) an alternative notebook environment, possibly R-Studio and b) some cloud-based notebook-creation environment for EC users unwilling to install such on their own computers.
Human-2-3 FTEs, probably with Python, R, GitHub and (perhaps) system-architecture experience.

Financial-Funds to pay for more than governance - you have to have something to govern; allocate funds to bring the pieces together to build a Knowledge Workbench for researchers not developers
Technical-APIs to connect to existing resources; if a testbed is needed, use the NDS Labs
Human-Chief Technology Officer; developers
Other-Ability to make decisions in a timely fashion

Technical-registry API, standards library (for web services and data types), ontology library/repository.
Human-Personnel to develop connections to the registry, standards and APIs

Financial-funding to have a basic architecture realized
Technical-some simple prototype provided
Human-Support from office in helping us understand the architecture

Financial-3 FTEs
Technical-dev platorm
Human-semantics expertise

Other-This is hard to assess at present.

Financial-Several NSF grants (e.g. ~$1-2m)

Human-Really useful could be dedicated full-time specialist to facilitate scientist-technologist integration

Financial-Pay some people to work through example use cases
Technical-whatever is needed for several example use cases and for a GUI prototype so people can see what 'EarthCube' might look like
Human-People to work through examples from beginning to end and do an initial prototype
Other-The above are necessary to concretely determine what is needed

Q3 b What resources might be needed from specific external resources (Building Blocks, CDF members, etc) to connect the external resource to the architecture?

Financial-(see my detailed previous comments)

Financial-It is unclear how much money is needed for the architecture implementation.

Financial-salary support for PIs/architects/developers to map their project's infrastructure to the architecture

Financial-Support for organizing and attending workshops and hackathons geared toward notebook-centric manifestations of as many BB & CDF services & capabilities as possible.
Human-Modest (not sure?)

Other-Ability to make decisions in a timely fashion

Technical-registry API, standards library (for web services and data types), ontology library/repository.

Financial-monies for travel to attend the workshops
Human-willingness to connect to the architecture

Other-This is hard to assess at present.

Human-As a condition for funding, all EarthCube funded projects should require a certain level of engagement by PIs in EC community activities, such as architecture development

Financial-Do it
Other-The above are necessary to concretely determine what is needed

Q3 c 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

-yes-My work will involve adding capabilities to the architecture that support interoperability, especially with computational models. Not knowing what metadata and interface standards the architecture will support is delaying this integration.

-yes-the design is a factor in prioritizing DDH features

-yes-I suppose so but am unsure what this means for us...

-no-

-yes-Yes, but we are likely to stay connected to the development of the EC architecture and can adjust aspects of our development as needed.

-no-

-yes-It will help us to decide for instance what authentication mechanism to use.

-no-

-yes-

-yes-Timing is not my concern. If it takes longer to create something better, then take the time to do this.

-no-not sure

Q4 Are you familiar with the May Architecture workshop? The report can be found at https://www.earthcube.org/sites/default/files/doc-repository/archws2016_finalreport_3.pdf
Since you are familiar with the workshop report how well does the Solutions Architecture reflect the outcomes of the workshop? Again, please be specific.

-Yes I am familiar.-I felt that the Architecture Workshop Report was much better written and organized and that the AIP did not really go much further, despite appearing to be highly detailed. The EC architecture will necessarily be based to a very large extent on existing Web, Web Service and Semantic Web standards, but this does not come through in the AIP report.

-Yes I am familiar.-The workshop recommendation was focused on a workbench capability for data handling and use. The current architecture proposal focuses on discovery and does not address, except as a second tier all the capabilities recommended by the Workshop

-Yes I am familiar.-the workshop didn't go into sufficient detail, compared to the document. The overall design and the key components and workflows reflect the workshop outcomes well

-Yes I am familiar.-Reasonably well. I consider inclusion of the workbe
nch & notebook concepts to be positive examples of such reflection. As a workshop participant I'm a bit disappointed that points I made about simplifying abstractions & assumptions are not more evident in the AIP, but to be honest, my points probably should not be considered among the workshop's primary outcomes.

-Yes I am familiar.-It does not - Figure ES-2 from the workshop report somehow was morphed to Figure 1.1.1.1 in the Architecture Document. The workshop figure makes sense - the meaning is totally lost in the Architecture figure. This is one example but there are more.

-Yes I am familiar.-I do not feel qualified to answer.

-Yes I am familiar.-Honestly they seem fairly well aligned but I have not gone back to review the two side by side.

-Yes I am familiar.-I think it reflects though the workshop was much broader. We did start with technology architecture of connecting BB's but through the workbench put the scientist at the center, which was important.

-Yes I am familiar.-elaborates on the fundmental principles laid out in the workshop, improves upon them

-Yes I am familiar.-I did not attend the workshop, but the architecture documents seem very much in line with what was presented at the EarthCube all-hands meeting this past summer.

-No I am not familiar-

-No I am not familiar-

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.

TOP THREE SERVICES IN BOLD.

Weighted average

6.00 Resource Distribution and Access Services
5.60 Resource Discovery Services
5.53 Resources Registry
5.27 Brokering Services
5.07 Transformation Services
4.87 Workflow Services
4.80 Semantic Services
4.79 Linking Services
3.27 Fusion and Visualization Services

Q5 Other comments:

-Since its inception, the EarthCube initiative has prioritized governance over technical coordination to its own detriment. While a lot of money has been spent on governance-related meetings, and end-user workshops, a relatively small amount of money has been spent to help the EC funded projects coordinate their efforts and make the building blocks "fit together" into something bigger. The one meeting with this goal, the Tech Hands Meeting, was very well-received by the "technical people" who need to coordinate their efforts in order to build a coherent EarthCube "architecture". I think part of the problem may be that "non-technical" people do not fully grasp the complexity and level of effort required to develop software and to make technologies interoperable.

Another aspect or perhaps symptom of this issue is a counter-productive dynamic in which funded project PIs are sometimes seen as “technicians” or “nonscientists” who are simply using the EarthCube program to pursue their own technology agendas and interests. In this perception, the PI is characterized as being “out of touch” or unresponsive to the needs of the geoscientists, and as “just doing their own thing”. Sometimes this point of view is articulated (negatively) with a quote from the Field of Dreams movie: “build it and they will come”, to indicate that the “techies” are just building what they want and not what geoscientists need. From our experience, there is little evidence to support this viewpoint. A large fraction of EarthCube PIs are geoscientists who at some point in their careers became interested in the technologies that can make things easier for geoscientists, and over time found themselves spending more time on the technology than the science. This subset of scientists is analogous to people who become interested in how their car works and begin tinkering with or maintaining their own vehicles. It is also analogous to scientists who become involved in the development of sensors or instruments that are needed to advance science. Many of these people become involved in cyberinfrastructure projects that serve their particular geoscience domain, and in those roles there is a very strong emphasis on addressing specific community needs and a necessity of being responsive and answerable to those communities. However, it is often the case that the technologies needed to work “under the hood” to provide the capabilities the geoscientists have requested are technical and complicated. The reason these technologies are necessary to provide what was requested are often not clear to the geoscientist, and they typically take time to develop. Again, the car analogy is helpful here: the community wants a car, but then asks why the technical folks are

building carburetors and camshafts. The same could be said of the community wanting their heterogeneous resources (datasets, models, services, etc.) to be interoperable but then ask why the technical folks are working on mediators, metadata, APIs and ontologies. As part of the broad communication issues outlined above, EarthCube governance should try to dispel the misperception that those building the technology to solve their problems don't understand their needs, because in the majority of cases they do and they are already answerable to or have oversight from a geoscience community.

-The process is slow to implement. EC should address one or more rapid prototypes to engage the geoscientist near term.

-I do not mean this to be overly critical, and I'm sorry that finding time to participate in TAC matters has been so difficult that I've contributed very little.

-I sent the same response in as both the TAC and CDF since I am part of both. So I just wanted to be open about the fact I have voted twice... vote early ... vote often!

-I do not think that I know the priority order very well. So I have just marked it in some order. All things seem important.

-I found the documents so full of errors and unfinished content (e.g. "XXXX") that it compounded the already onerous task of wading through 100+ pages of unconsolidated and poorly organized text. And the grammar wouldn't pass muster as a high school English class essay, e.g. on p. 9 of Step4-ECSolutionArchitecture.pdf, under Objective 1: "Metadata are considered foundational to EC. It's quality and utility is highly dependent" (Improper use of concatenation and subject-tense mismatch.) Worse yet (p.13): "In the end, the goal is for the capabilities developed on the EarthCube Workbench to improve, the interoperability of data and technology to create an improved System of System environment for the GeoScience community. To to this, the EC Interoperability Workbench will host a variety of supporting resources listed above, ... This will allow the community to assess resources for resource solutions for re-use and look to integrate resources in ways that have yet to be done create new value-add capabilities." Or, p. 15 "If the user sees its available at a Workbench they like, they would likely lean that way, click that and going to the Workbench experience at that site." There are many important principles and directives in the Solution Architecture document which should have been distilled for the executive summary, but weren't.

-That last question about ranking is extremely hard to answer at this point. It also depends on the goals of the solution architecture. What is being solved? This needs to be made more clear before any prioritization is done.

-Sorry for being so critical, just want to make sure that scientist voices are properly heard so that all of this technology is useful. I do really appreciate all of the hard work of TAC and CDF and TigerTeam that put together the AIP!

-Worked through examples are needed in the solutions document.