AIP Tiger Team Response to ESSP Architecture Implementation Plan

November 9, 2016

Contents

Introduction

This is a report from the EarthCube Architecture Implementation Plan Tiger Team (AIP-TT) appointed by the EarthCube Leadership Council. The Charge to the Tiger Team was to provide guidance and review of the EarthCube Science Support Program's (ESSP) Architecture Implementation Plan (AIP). Due to the short lead time allowed for this review, this AIP-TT Group Response is limited to discussing the Solution Architecture and the Implementation Plan, though not in great detail. These are the last two major portions of an extensive set of materials in the Final Work Products google drive folder. However, the TT's guidance and responses to the earlier materials are largely included in various contexts within these last two volumes. While two members of the Leadership Council are on the Tiger Team, this represents the views of the Tiger Team members, not necessarily the Leadership Council as a whole.

The AIP-TT members represent the EarthCube community, as follows:

TAC – Scott Peckham, University of Colorado
SC – D. Sarah Stamps, Virginia Tech (also on Leadership Council)
CDF – Bob Arko, Columbia University
LC – Janet Fredericks, Woods Hole Oceanographic Institution (also on Leadership Council)
At-Large – David Arctur, University of Texas (chair of AIP-TT and Architecture Workshop, May 2016)

The TT members brought into consideration "the established goals, vision statements and roadmaps of the EarthCube Leadership Council, Council of Data Facilities, Science Committee, and Technical Architecture Committee." As part of the review process, the AIP-TT met with leaders of the AIP development team several times during the project, and provided updates to members of their respective governance bodies. We also assembled numerous materials for consideration in an architecture library folder with a reference index, which we have opened up to the EarthCube community to view and comment.

This AIP-TT Group Response is considered a companion document to the AIP final report from Xentity. Due to the short timeframe allowed for both the architecture development and its review, this could not be a thorough review of the proposed architecture, but rather a commentary on several aspects of the AIP report (about 100 pages in the last two volumes) from the perspective of EarthCube stakeholder communities. This is consolidated and condensed from initial notes from each of the TT individuals listed above, and from conversations between the TT and Xentity. It is organized around a set of themes, with links in context to the initial notes, the Solution Architecture and the Implementation Plan documents. We hope this will help aid others' reading and understanding of the AIP, in some cases by simplifying and clarifying concepts and terms, and in other cases by challenging or extending what is presented in the AIP. Some of these topics discuss cross-cutting issues for multiple sections of the AIP.

Regarding the intended audience for the architecture and this AIP-TT group response: these are really written and directed toward those who will be building the architecture, i.e., cyberinfrastructure implementers. This drives the use of certain terms that may not be obvious to many geoscientists in the EarthCube community. For example, the most commonly used term in these reports is "resource". To a geoscientist, this would normally refer to something like natural resources, or possibly even financial resources. However, in these documents, resources are anything that could be used by and within the EarthCube architecture: datasets, models, publications, web services, registries, and any number of enabling technologies for these. Please be mindful that if something you read in these documents does not make sense, it could be due to differences in meanings of common terms and concepts.

Topic 1. Process of EarthCube Architecture Development

As we started this AIP process in September 2016, we were already five years into EarthCube without an agreed architecture. Numerous Building Blocks and Integrative Activities have been awarded and completed already, many of them providing aspects of an architecture, but without a sustainability plan. It has been tempting to jump to defining concrete lists of standards, ontologies, API's (application programming interfaces), etc., but instead this architecture defines a framework and foundation for agile and self-regulating development, in which evolving models, data, and workflows can be created and managed through EarthCube governance, ie, working groups. We feel this is the right emphasis.

One result of this design is a distinction between components that implement the registries, assessments, communications, and workbench, vs. components that provide content for the registries, assessments, communications and workbench. Example candidates for implementing EarthCube core components include CINERGI, GeoSoft, GeoDataspace, and others. Example candidates for contributing content include "partner aggregators" such as CHORDS, BCube, GeoLink, OntoSoft, and potentially many non-EarthCube catalogs of resources hosted by NASA, NOAA, EPA, USGS, etc.

What happens after this review? A phased implementation approach is being developed, see section II.2 EC Planning Guidance of the Implementation Plan. ESSO, LC and NSF will review the AIP report and TT's companion document. They will consider whether the proposed AIP approach is a "good start", a "bad start" or a "non-starter". That may take 2 or more weeks from time of submission. They will then further consider best ways of moving forward.

Topic 2. Motivation for EarthCube Workbench

These are topics that should be identified and articulated as being met with the proposed design [Scott in AIP-TT notes p.3]:

- Atoms of science: supporting the key role of variables as the core elements around which geoscience happens.
- Importance of relationships: nothing happens in isolation; resources need to be understood in terms of their roles and effects on other resources in context of workflows and other applications.

[Scott-Ouida in AIP-TT notes p.11] "From my experience as a geoscientist, if I were to assign percentages of project time that are currently spent on various tasks, I'd say: resource discovery (5%); resource access (5% or less, depending on how this is defined); resource assessment (5%); understanding, transforming/preparing, using, composing, etc. (85%). This is when my role is "resource user" vs. "resource creator", where the tasks are different."

"EarthCube Representative Scenario" in Solution Arch, p.8: this is a common usage pattern we need to support: searching for data and models, connecting them, transforming inputs/outputs as needed, performing the science analysis and visualization, and publishing results.

Topic 3. Geoscientist's Response to Proposed Architecture

[Sarah, AIP-TT notes pp.1-3]
Under Objective 1 on p.9 in Solution Architecture:

- What is available from other domains?
- How useful is it?
- How has it been used?
- How does one filter metadata based on the roles of a scientist or cyber infrastructure specialist?
- How have others assessed it?
- What assessments have been done?
- Are assessment ongoing and at multiple levels?
- Is the metadata machine interpretable?
- How can it be improved?
- Are resource links & descriptions still current?
- Running?
- Providing correct results?

My response: To me, it seems these objectives are at the cutting-edge of cyberinfrastructure. Standards are in development, assessment criteria have yet to be developed, and interoperability is needed to make all of the other objectives possible. This goals of the EC WB are laudable.

Under Functional System Decomposition, p.10:

"Newly created resources metadata will need to be fully described and submitted through the author's metadata supplier to ensure the EC's read-only registry and the metadata are curated appropriately."
My response: Note that "encouraging, recognizing, and rewarding" the use of EC WB will be highly dependent on the acceptance by tenured colleagues and upper administration at Universities. I'm glad that there is an emphasis on EC community training. Perhaps in the transition phase more emphasis could be placed on educating upper administration at Universities.

Additional note regarding "EC's read-only registry" highlighted above: it's important that scientist (data provider) can be sure the metadata is their own, not tampered or augmented in any way.

Scott notes: OntoSoft project proposes crowdsourcing metadata, so that qualified users could augment other providers' metadata, with sufficient governance. This is not so different from open source software development. Github for example, allows the primary author to accept collaborators who wish to help with development. Wikipedia also leverages crowdsourcing for edits, with moderators.

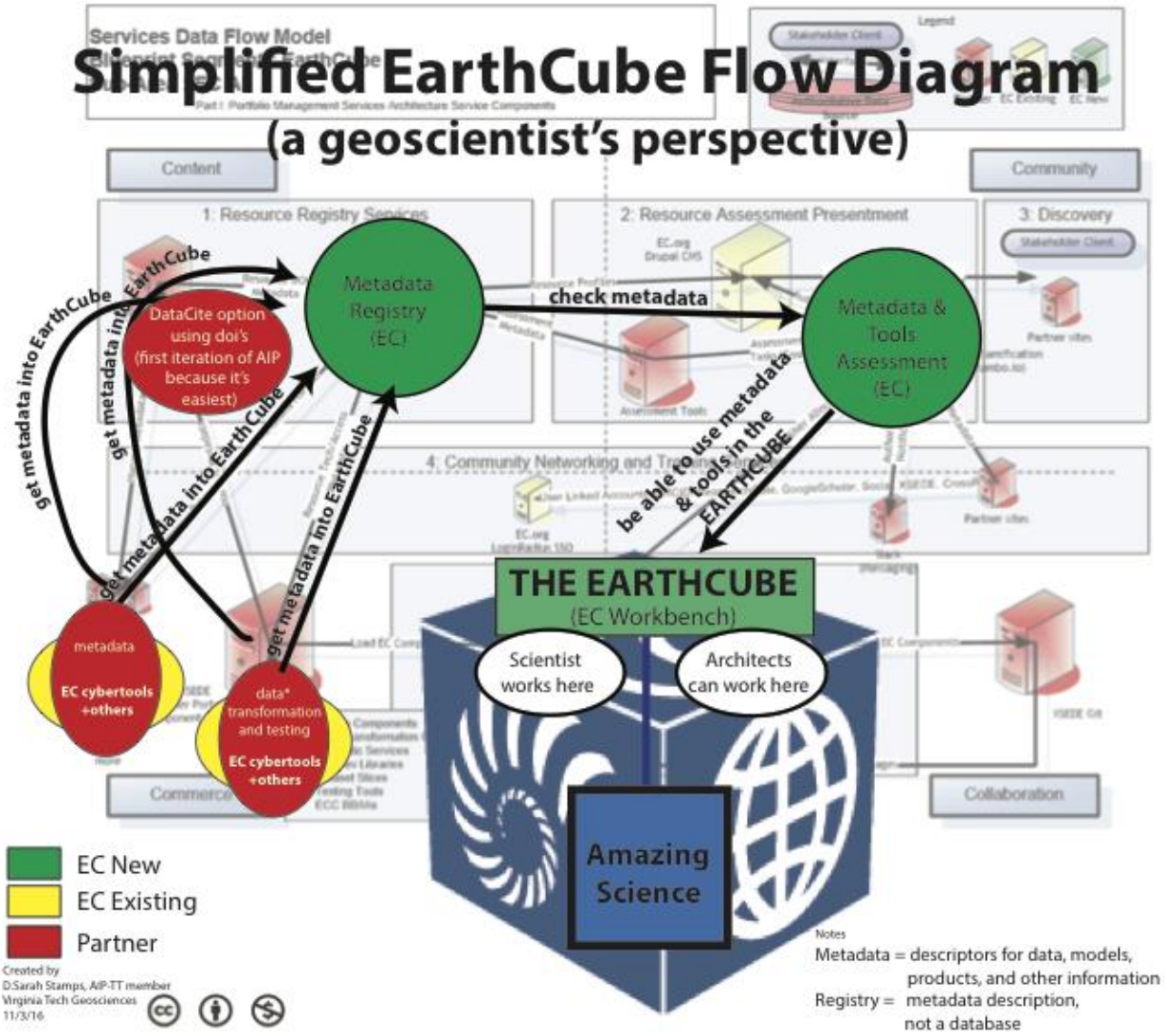
"The program will need to train the EC community on the workbench capabilities, metadata management approach, discovery and assessment methods and any implications to grant applications requirements."
My response: "It's good to see training is part of the plan because it's critical to get the EC geoscientists, data scientists, and architects using the cyberinfrastructure."

GENERAL COMMENT in response to Reverse Site Visit section 2.2
My response: We are getting to the point of succeeding with EarthCube. As a geoscientist I see this EC WB as becoming what I need for the transdisciplinary, cutting-edge science I aim to perform.

GENERAL COMMENT in response to Reverse Site Visit section 2.2 1st recommendation
My response: We now have ESSO and are succeeding with this recommendation.

GENERAL COMMENT in response to Reverse Site Visit section 2.2 2nd recommendation
My response: We now have the suggestion to consider UNAVCO, NSF's Geodesy Facility. UNAVCO started at UCAR and evolved into its own independent organization. The governance structure of UNAVCO could be considered as a model governance structure.

Below is a simplified interpretation of the Services Data Flow Model in Section 1.1.3.1 High-Level System Architecture (p.22) that is intended to be for a geoscientist. From the geoscientist's perspective they will login to the EarthCube workbench and work within this specific part of the entire EarthCube cyberinfrastructure. The registry, assessment tools, and other cyber tools/infrastructure that are funded through the EC program are understood to be the behind-the-scenes parts of the EarthCube as defined by a geoscientist.



Topic 4. Identity Management

The architecture discusses DOI management with the DataCite registry and API as the core resource. Xentity explained to TT that DataCite (or something similar) would be a "super-aggregator" of DOIs for resources collected from the various partner aggregators, other registries, etc. The EC Registry would be a filtered view of DataCite's registry, harvested via API. The DataCite API is public. DataCite can also accept DOIs generated from other resources. The DOI Consortium manages policies of assigning and using DOIs among assigning authorities. Note that DataCite is not "locked in" for this role, though it seems like a good fit; further discussions would be needed at an institutional level before it can be confirmed here.

Topic 5. Issues of Metadata Content, Interfaces, and Other Standards

The DOI metadata records are not meant to capture the deeper metadata and workflows needed for science and research, but rather what's needed for a platform that enables and manages such. Systems

such as CSDMS and OntoSoft, with extended metadata for representing models and their interconnections in workflows, could register their extended metadata directly with the EarthCube Registry. However, model framework metadata does not yet have widespread, mature content standards, so this is an area that will evolve in EarthCube as it evolves in the broader community. For more details, search for the string "extended metadata" within the Solution Architecture document.

Clearly distinguish between "primary" and "secondary" resources in the registry [Scott, AIP-TT notes p.8]
Geoscientists care about and want to work with the "primary" resources like data sets, models, publications. However, the advanced capabilities they are all asking for require a separate tier of "secondary resources", including mediators, catalogs, registries, ontologies, utilities, etc. The geoscientist doesn't really care about these, but they are precisely the products that are needed to deliver the capabilities they want and are also what every funded EarthCube project is developing. Putting both types of resource into a single registry for geoscientists to browse with no distinction between the two is like a department store that mixes the "Radio Shack" section with the clothing section. Keep in mind that EarthCube PIs may be geoscientists, computer scientists (and other tech developers) or both at the same time. The EC Registry should allow researchers to focus on their primary resources of interest.

Standardized and machine-actionable metadata [Scott, AIP-TT notes p.3-6]
Metadata needs to be not just human- or machine-readable but *machine-actionable*. This depends on extended metadata standards and conventions agreed on by data providers and EC WB software: not just overall metadata content formats like ISO 19115 or Darwin Core, but profiles of these, as well as taxonomies, vocabularies, and other semantics.

Specific metadata content standards, web service standards, APIs, and other conventions are not called out in the design proposal; these need to be decided through EC governance as soon as possible.

Topic 6. Methods of Assessment

This architecture places considerable emphasis on tasks and components related to assessment. A key aspect of the proposed assessment measures is stimulating community input to encourage providers to keep their metadata current and useful. This is seen as a gap in other systems of systems such as GEOSS, data.gov and others. The following is a clarification of the long documentation of the methods (and motivations) for assessment.

There are at least 3 ways assessments will happen in EarthCube:

- Middle-out approach: At registration, EC would have some EC registry-hosted tools automatically running metadata assessments. These would not be augmenting metadata but recommending fixes by the source provider.
- Bottom-up approach: (through crowd incentives) We expect there will be external assessment tools (eg, from ESIP) that could also be run to recommend fixes by the source provider.
- Top-down approach: (through portfolio governance) EarthCube WGs (Standards WG, etc) or candidate projects (BBs, etc) could develop and/or formalize resources to become workbench capabilities, that can be assessment tools.

Another key aspect of the proposed architecture is to move the assessment process from metadata creation & curation to process orchestration.

Here is an example scenario for how assessment could be carried out and leveraged:

1. A researcher works with a tool in the workbench & gets unreasonable results. The workbench enables identification of the cause of the errors, which results in requests to the resource provider to fix the problem. Community feedback and incentives (we hope) will drive providers to respond quickly & completely & well, which raises their profile and reputation. This is the "gamification process" at work.
2. Responding to community input, the provider can change their tool on the workbench, update the resource metadata, and the scenario can be tested again.
3. Gamification doesn't only incentivize the provider, but also aggregator systems who would assess their own members' contributions. The aggregators would then be rewarded to the extent that their contributors' resources are well managed.
4. Metadata harvest has little cost and little incentive to watch closely; but we can incentivize aggregators to keep their members' contributions current and useful, and to "up their game" thus enabling improvement of the sciences.
5. This could also be tapped for the grant process: (a) solicitation language, and (b) a scoresheet on PI responses to assessment requests (governance).

Interoperability and orchestration of workflows will improve as researchers and cyber developers respond to incentives to help each other and improve the quality of their products themselves.

- Scientists work here: they are making value-added, chainable capabilities on the workbench.
- Education efforts within EC community will increase the capacity of scientists such that it's possible for more scientists to leverage and extend capabilities on the workbench.

Topic 7. Issues of Registries, Linked Data, Semantic Web

There has been considerable discussion around the importance and roles of registries in EarthCube for discovering datasets, tools, and other resources. Populating and maintaining the content and quality of these kinds of registries is notoriously difficult. It is labor-intensive and error-prone to compile and maintain any registry of computational resources, because the context for computational resources can change quickly with little or no warning. Files and services can be renamed, reformatted, moved, and deleted. Whole catalogs of datasets may be shut down due to lost funding.

Nevertheless, a central registry is seen as necessary in the EarthCube architecture. The idea is to develop assessments around the results of registry requests (Topic 6 above), so that the broader community is motivated and engaged in helping maintain the currency, accuracy and completeness of content in the registries.

EC need not be limited to a single approach for discovery. One alternative to a central/registry-based approach is a "Linked Data" approach, that leverages a family of "Semantic Web" protocols (HTTP, URI, RDF/OWL, SPARQL) to create a network of content providers. This approach has been taken up by several existing EC projects including EarthCollab, GeoLink, and LinkedEarth as well as a growing network of

geoscience repositories in the Council of Data Facilities. Content published as Linked Data is easily consumed by machine harvesters and workflows, while remaining "close to the provider" where it can be most effectively curated. However, development of common ontologies across disparate providers can be difficult and labor-intensive. Nevertheless, it is likely that Semantic Web and Linked Data design patterns will be increasingly utilized. A response in this regard from Xentity was that the Linked Data approach directly supports the Discovery portion of the EarthCube system, while the EC Registry is only intended for intake of content, not really discovery. More on this is in the Solution Architecture section 1.1.1.1 Business Service Areas Overview, see Register and Discover bullet headings.

Some other issues related to EC Registry have to do with Partner Aggregators, who will handle their own content currency / accuracy / operational status. EC outsources resource aggregation. One purpose is to expose and follow the chain of responsibility for errors in data retrieval, which partner aggregators would need to fix; EC can't do that.

Topic 8. Leveraging Existing Model Frameworks and Workbenches

[Scott, AIP-TT notes p.3-6] Modeling frameworks represent a class of software systems or environments that are based on the concept of reusable, interoperable "plug-and-play" components. (These could also be called **workbenches**, especially the ones that have nice GUIs and a palette of resources to choose from.) Examples include CSDMS, ESMF, OMS, OpenMI and FRAMES. (The key papers for each of these 5 frameworks have been collected as PDF files in our Reference Library.) The components in these systems are typically modules that model a particular physical process --- that is their level of "granularity". Some have conventions for self-describing interfaces, integration with analysis and visualization tools, and all support various parts or all of the workflow represented in "EarthCube Representative Scenario" on p.7 of the proposed design.
Scott's notes continue with several example tools, languages, conventions and standards that have emerged from development of these modeling frameworks. This will be useful for next steps in the design, which call for EC governance to decide on these aspects of implementation.

Topic 9. Reviewing and Using the Solution Architecture

TAC governance would create a WG to be the functional requirements and solution architecture verifier and validator of the implemented services. Each of the five main business services (bullets below) would have an owner/representative, plus reasonable backup. The WG members would follow a service architecture pattern organizationally, for their own communications and actions.

- Registry - works with ESSO to accept the registry, handle feedback for the registry, maintain, etc.
- Assessment
- Discovery
- Communication
- Workbench

There are three levels of governance needed: IT governance, content governance, and usage governance.

AIP-TT Group Response to EarthCube Solution Architecture

IT Governance: The TAC-designated WG as a whole would share the cross-interface requirements. Eg, registry might want to change its interfaces, this needs to be a governance matter.
ESSO, LC and NSF would direct and fund the overall process to make the architecture operational; TAC WG would carry out the work.

Content Governance: A different governance WG would provide guidance for resources, assessments, and workbench capabilities. This WG would consider the resource value based on assessments and workbench capabilities in place.

Use Governance: User engagement team/WG provide guidance for gamification, outreach, and assessment tasking, ie, participation in the processes.

This is further developed in the implementation guidance.

###