EarthCube Science-Driven Workbench

A report of the Workbench Tiger Team (TT) formed by the EarthCube Leadership Council to investigate scientific workbench concepts and criteria, and report back with recommendations for future NSF directions on the topic. The TT members included:

Sara Graves (University of Alabama in Huntsville - chair)
Rowena Davis (University of Arizona)
Dave Fulker (OPeNDAP)
Ken Keiser (University of Alabama in Huntsville)
Rebecca Koskela (DataONE)
Emily Law (Jet Propulsion Laboratory)
Ouida Meier (University of Hawaii)
Ilya Zaslavsky (University of California - San Diego)

Executive Summary

This report on EarthCube Science-Driven Workbench concepts describes four main functions identified as being important to support geoscience research activities, to include Resource Discovery, Solutions/Workflows, Assessment, and Resource Platform(s). In brief, EarthCube researchers need a robust resource repository that supports the ability to record improved metadata about resources, to include interoperable data, services and applications, and subsequently support the discovery of these resources in support of science needs and use cases. The ability to then orchestrate discovered resources into solution workflows, assess the interoperability of solutions, and identify missing (gaps) components, was identified as likewise crucial to a successful workbench environment. A long-term supported resource platform(s), that includes computational and storage resources will be critical in providing usable capabilities to the community and ensuring sustainability of these EarthCube solutions for everyone's benefit. The realization of these capabilities for EarthCube will depend on the level of support that is available from NSF in upcoming solicitations, but could include approaches ranging from starting with building a more robust and complete resource repository capability that is interoperable with open workflow engines, to funding for the development, or EarthCube adaptation of, one or more complete workbench solutions. Fundamental to any solution will be the emphasis on data, service and application (resources) interoperability to ensure long term scalability of incorporating existing and future technologies, as well as the definition of EarthCube criteria for the assessment of resources to meet the program's objectives. Ideas and suggestions on how NSF/EarthCube could move ahead with defining/creating a Science-Driven EC Workbench are described in the Workbench Implementation Options at the end of this report.

Motivation

The goal of the NSF EarthCube (EC) program is to provide technology solutions that will facilitate geoscience research and develop cyberinfrastructure to improve access, sharing, visualization, and analysis of all forms of geosciences data and related resources [7]. EarthCube's success depends on integration and interoperability of data, tools, services and models, to provide geoscience solutions. EarthCube investments in architecture concepts and the development of building block technologies have been leading to the realization of the EC goal. However vertical silos of technology solutions have limited use for solving complicated problems and are difficult to integrate into the emerging NSF Geosciences data ecosystem, driven by EC. An overarching purpose for a science-driven EarthCube Workbench (EW) is to serve as a "place" where individuals are drawn to join others in (collaboratively) learning about and employing EarthCube resources for scientific problem-solving, and the motivation of an EW is to implement functionality that will ***support and improve interoperability of EarthCube resources to facilitate geoscience research***. An EW will provide scientists with web-based user interfaces to a suite of tools making it easier to connect available resources into meaningful solutions that solve their geoscience research problems as described in geoscience use cases documented by the EC Use Case Working Group [1], as well as other geoscience use cases from the community.

Two of EC's key objectives are to achieve interoperability and data integration across geoscience disciplines, and to build on and leverage existing science and cyberinfrastructure. These objectives imply that EC infrastructure supporting acquisition, management, distribution and analysis of geoscience data must consider end-to-end workflows and must accommodate multiple topologies for the physical deployment. An EarthCube workbench should focus on providing a collaborative integration environment of resources, such as data and tools that are critical to streamline the building and testing of interoperable solutions made available through the integration of EC resources and technologies, and existing capabilities provided by the broader geoscience communities. The desired outcome is that **geoscientists and students frequently will turn to the EW—and its user community—to solve problems in their own studies**.

Prior Work

A science-driven EarthCube Workbench (EW) should attempt to align with concepts depicted in a number of previously documented EarthCube solution architectures [3, 5, 8] (Figure 1). The EW concept addresses the issues of integration and evaluation methodologies, and best practices with a strong interoperability theme to advance disciplinary research through the integration of diverse and heterogeneous data, tools, models, algorithms, services, and systems. The growth of an EW will likely provide guidance for EC evolution and future

integrated solutions by enabling and encouraging the EC community to develop integrated solution prototypes, try out new technologies, reproduce results, and to share ideas, concepts, and experiments.

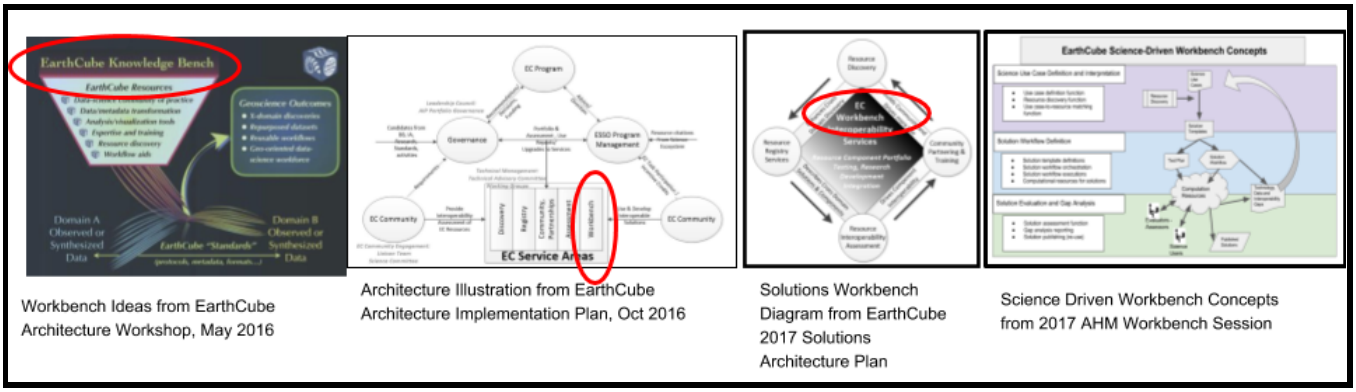

Figure 1: Past EarthCube architecture and planning documents including ideas for a science-drive workbench concept [3, 5, 8].

An EW solutions integration and assessment environment should have the potential to scale and extend to an NSF geoscience wide integration and assessment capability.

There are many challenges with building and operating a scalable, extendable and sustainable environment that supports all Earth science domains so consideration could also be given to more focused solutions that could be integrated. An EW can serve as an integration model that is responsive to needs of the geoscience community such that other communities will benefit and learn from EW systematic methodologies including integration, assessment and best practices.

Defining the Functions

At the conceptual level of abstraction (Figure 2), the basic components of an EarthCube science-driven workbench should reasonably include functions for the definition, interpretation and resolution of science use cases through *discovery and incorporation of known and properly described resources, to include data, applications and services*. *Solution workflow definition* functionality would implement the use case by stitching together the identified resources into a workable solution. A component to *evaluate/assess solutions* for correctness and adherence to EarthCube interoperability criteria would benefit the community, but could be developed later. Consideration of a *resource platform* providing computational and storage resources will be critical to support the definition, execution and management of science solutions, and address sustainability. More details on discussion and feedback from the EarthCube community on these workbench concepts collected from breakouts of the Science-Driven Workbench session

at the 2017 All-Hands meeting [6], and more recently the materials from the Science-Driven Workbench Infrastructure session held at the 2018 All-Hands meeting [8].

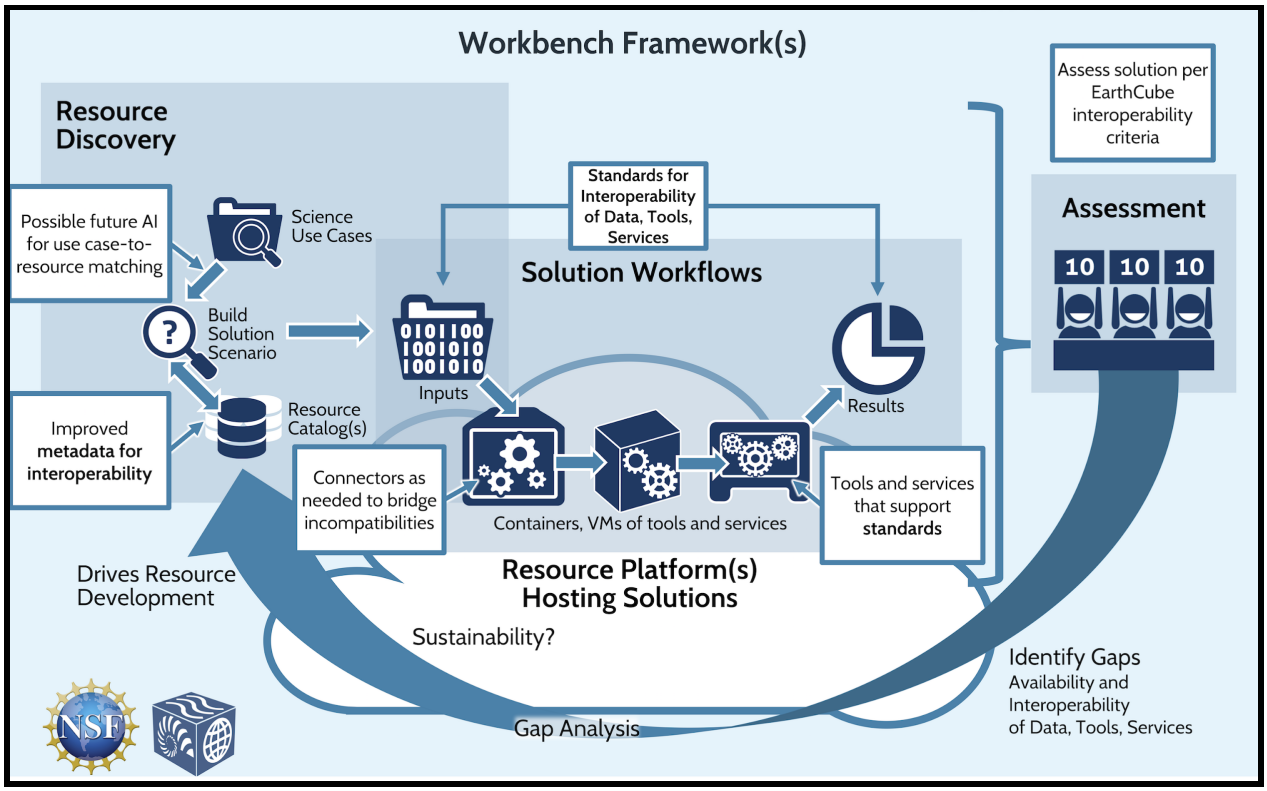

Figure 2: Illustration of EarthCube Workbench conceptual functional components [8]

Below are additional details about what might be included in the suggested main workbench functional components.

Resource Discovery

Based on geoscience use cases or problems, the EW should provide users with tools necessary to discover resources (data, applications, services, models) from an EC registry (Figure 2). The needs of a geoscience use case should be matched to data and computing resources available in an EC registry, but the EW should also facilitate the utilization of data, services and application resources that have not yet been registered with EC, making it easier to experiment with resources that have not yet been fully tested. Registering resources would allow users to discover matching resources, explore their metadata, and include these resources in EW workflows, assessing their fitness for use in the user scenarios. EW should capitalize on the registry of resources being developed through, for instance, the CINERGI and the EC Data Discovery Hub Building Blocks, among possibly others. CINERGI has been highlighted as an EC registry prototype in the EC architecture documents previously reviewed, and has demonstrated the use of expanded metadata to support interoperability of resources. The current P418 project has demonstrated the use of Schema.org expanded markup to produce

improved indexes of existing data registries to improve discoverability. An EW could expand the EC registry by registering additional datasets required for implementing the interdisciplinary use cases outlined above, and by extending the underlying registry's ontology to include semantic constructs needed for metadata augmentation. The purpose of semantic augmentation would be to assign initial semantic types to dataset characteristics which will be later used and further refined as users iteratively compose workflows in an EW. In addition, an EW should extend the registry by including model codes and data transformations and other services to be made available to the EW. An EW activity could be to explore ways to identify semantic and structural mismatches and gaps in use case workflows, which will provide a source of refinement of both the registry and the workflow composition techniques.

Workflows/Solutions

The *Solutions* function (see Figure 2) needs to support EW tools that produce a workflow solution for the science use case objectives. A *Workflow Engine* should utilize resource enhanced metadata to understand the necessary inputs and outputs, and interface specifications, allowing the logical connection of resources for a user specified scenario. The **quality and completeness of resource metadata** will be important to allow the *Workflow Engine* to streamline the process of constructing of a high-quality *Interoperable Solution* result. Gaps in resources and interoperability, such as the output of one resource is not interoperable with the inputs of another resource, with no appropriate translation resources available, should be documented in some type of *Gap Analysis Report*. A *Gap Analysis Report* generated during the Solutions processes can serve multiple purposes; (1) providing feedback to the EC community on additional technology resources that are needed to be added to the *Registry*; (2) documenting deficiencies in specific existing resources in the *Registry*, such as resource "A" does not support the necessary standard data format. The result of a successful *Workflow Engine* session would be an executable *Interoperable Solution*, including all the resources, the specified sequence of processing, the inputs and outputs, etc. The *Solution* could be executed, shared, copied or edited as needed by users to create new solutions or edit an existing solution.

Assessment

Assessment should be the final step (Figure 2) of ensuring that a user's Solution actually fulfills the intended Scenario Definition (in other words solves the original science problem), and that the Solution is meeting interoperability criteria, expected to be defined by the EC community. In some cases "self-assessment" by a project team using EC-defined criteria may be sufficient, but in other cases 3rd party assessment may be warranted - say when evaluating funded projects for interoperability. Assessment of solutions is only feasible if EC defines community interoperability criteria and mandates that EC resources adhere to those practices.

Resource Platform

Underlying the EW functions should be an EarthCube *Resource Platform* (Figure 2). Whether built on existing capabilities or customized for EarthCube, the platform should provide the

computational and storage resources necessary for the hosting of EW tools, managing the user profiles, and storing the *Solutions* and *Assessment* results. The *Resource Platform* should provide tools to administer the computational and storage resources, monitor the progress of processes, and create and manage libraries of geoscience virtual machines that can be constructed for the various *Interoperable Solution*s (see Figure 2). The *Resource Platform* should support: (1) Resource scheduling and selection of computing resources from the *Platform* to maintain a spatiotemporal status of chainable EW resources, providing an algorithm to match and allocate resources, and to calculate the cost of cloud resource usage; (2) Support for containerization providing better portability and extensibility in the *Resource Platform*, giving scientists the freedom to effortlessly move their application from local to cloud resources, and supporting higher levels of user interactions and specific geoscience use cases for fast and scalable science use case deployment. A modularized architecture would help extend existing components to meet various demands by high level geoscience use cases; and (3) support interoperability to help mediate service inputs and outputs when orchestrating workflows across the EW, the EC registry, other EC resources, and geoscience resources. One of the key unresolved problems in creating an online workbench environment is the efficient provisioning of executable containers that include all components and dependencies required to design, refine and execute research workflows. An EW effort could experiment with several approaches made possible through recent open source projects which make reusable science environments possible through generating and provisioning custom Jupyter containers that combine code in users' GitHub repositories with external services, such as databases, analysis tools, or interactive front-ends. Container images could be version-controlled and registered in the EC registry, making it easier for geoscience end-users to create and execute container images referencing data they discover in EC registry and retrieve into *Resources Platform* storage. The EW Resource Platform is likely a good candidate to be an EC project office function that can be sustained and support enforcement of interoperability criteria.

Conclusions and Recommendations

Members of the Workbench Tiger Team conducted activities at the recent 2018 All Hands Meeting (AHM) [8] to solicit input on perceptions and needs from the EarthCube community of what an EarthCube science-driven workbench might look like. Figure 2 above was used as an illustration in both a workbench-themed poster and follow-on breakout session at the AHM, in an effort to communicate the suggested high-level concepts of workbench functions that should be considered, specifically:

- Resource Discovery
- Solutions/Workflows
- Assessment
- Resource Platform

As might be expected from a community with vast geoscience research and analysis experience, the perspectives and suggestions on workbench technologies was unbounded, with no clear consensus on one specific solution for all users. For some participants, resource

discovery was the only needed functionality, while others focused more on the successful definition and execution of solutions workflows, etc. There was agreement, however, that the richness of metadata would be key to successful interoperability between resources and workflows, but most existing resource registries do not have sufficient metadata available so the ability to define enhance metadata is crucial, and the ability to include users' data and tools is needed. It was recognized that the current P418 project demonstrates improved indexing and discoverability of data resources using schema.org approaches, but does not address the needed enhanced metadata to support the orchestration of solutions. A more comprehensive EarthCube registry solution was considered to be a key requirement, but also the definition of interoperable interfaces between a Discovery/Registry system and the Solutions/Workflow system was recognized as a necessity by the AHM session participants.

As described above, the **conceptual functions** for an EarthCube Workbench should strive to provide solutions to Earth science research problems include (1) an easily accessible/usable **Resource Discovery** capability that supports finding the data, tools and service resources necessary to fulfill a science use case, and provides sufficient metadata to support the interoperability of resources when building solutions, (2) a **Workflow/Solutions** capability that employs commonly used workflow creation and execution technologies to execute the defined scenarios consisting of multiple data, tools and service resources. (3) (possibly) an **Assessment** capability that provides tools for determining if the generated solution is correct, solves the original science problem, and is meeting EC-defined interoperability criteria, and finally (4) an EC **Resource Platform** that provides computational and storage resources to support the definition, execution and management of Earth science solutions. The focus for an EarthCube science-driven workbench, rather than targeting specific technologies, should be the definition of interoperable interfaces between the functions that allow for adapting resources and technologies from multiple sources to successfully realize science solutions.

Implementation

Ideas and suggestions on how NSF/EarthCube could move ahead with defining/creating a Science-Driven EC Workbench are presented below. A Workbench Implementation Options section follows. Recommendations could include: (1) solicitation(s) on defining, implementing and demonstrating all the main functions, but perhaps with an initial focus on interoperable interfaces between a Resource Registry and the Solution/Workflow functions. The functions could be awarded as separate functions, but should be required to demonstrate interoperability which would need to be defined and stated in the RFP; (2) awards for multiple workbench projects (could be existing technologies for complete system or functional components) that will compete to demonstrate the best fit for EC, for instance possible Resource Platform solutions - as stated earlier, could be a good candidate to be treated as a project office function in order to insure sustainability of resources for the community; (3) possibly solicit for the assessment of the vast existing workbench tools and technologies for adoption by EC, but this is potentially an unbounded activity that would never be complete.

Workbench implementation options are presented below:

1. NSF could have a solicitation for the development of a complete workbench solution, based on stated EC objectives and principles of interoperability. To include...
    a. Evaluation of all possible existing technologies for the 4 stated main functions (**Registry/Discovery, Workflow/Solutions, Assessment, and Resource Platform** or however many parts of the workbench NSF wants to pursue), to possibly include evaluation of existing workbench frameworks used for other disciplines.
    b. Development of a workbench framework that will allow for the plug-n-play of alternate (interoperable) technologies for each of the functions
    c. Implementation of one or more technologies for each of the targeted functions
    d. Testing and assessment with science team participants to rank the usability of the workbench and the individual technologies - which would require the definition of EC criteria for the testing and assessment.

2. NSF could have a solicitation for only the evaluation of technologies for the stated functions - possibly with decisions on the best solution(s) between multiple awards for each function
    a. Follow with a solicitation for building a framework(s) that would provide the glue between the previously identified and vetted functions

3. NSF could have one or more solicitations for projects for the individual (interoperable) functions - would need to have a target framework for them to be compatible with.
    a. **Discovery** - interoperable interfaces that support the passing of discovered resources from the registry of data, tools and services, to the Solutions/Workflow system, and the return of usage metadata back to the registry
    b. Interoperable and standards-based **Solutions** system that prepares, tests and executes solutions workflows
    c. **Assessment** function that evaluates the correctness and interoperability of solutions
    d. **Resource Platform**, to provide computation, storage resources for workbench solutions, and management/publishing/reuse of solutions
    e. **Gap Analysis -** may not be a defined function, but could include the ability to track data/functionality gaps across solutions to provide feedback to EC on what functionality or levels of interoperability are missing, and to provide a collaborative environment for information sharing between workbench participants

Key considerations for NSF are:

- Science-driven (task specific, discipline specific, interdisciplinary, etc)

- Coordination by PI(s) and/or project office
- Determine adaptability of existing components for EC needs
- Plans for sustainability (e.g. suggestion of the Resource Platform possibly being a project office function)
- Definition of requirements for Interoperability of resources, e.g. how tools, applications, services interoperate between themselves and with data resources.

References:

[1] *Repository of EarthCube Use Cases.* 2017
[2] *EarthCube Integration and Testing Environment (ECITE).* 2017
[3] EarthCube, *EarthCube Architecture Workshop Report*. 2016-2017.
[4] EarthCube Leadership Council, *EarthCube Funded Projects Requirements: Tiger Team Report,* 2016.
[5] EarthCube, *EarthCube Architecture Implementation Plan (AIP)*. 2016.
[6] Synthesis of Results from Science-Driven Workbench Breakout Sessions at the 2017 EarthCube All-Hands Meeting, 2017.
[7] EarthCube Website “About” Information.
[8] Community input from the EarthCube Science-Drive Workbench Infrastructure session at the 2018 EarthCube All-Hands Meeting, 2018