EarthCube Leadership Council 2019 Priorities, finalized Sep 23, 2019

Accompanies 2019 NSF solicitation guidance document: LINK

Authors: Ken Rubin, Emma Aronson, Isabel Cruz, Mike Daniels, Ryan Gooch, Denise Hills, Anna Kelbert, D. Sarah Stamps

Contents:

A. Introduction and Review

This 2019-2020 community priorities recommendations document outlines the latest thinking of the EarthCube Leadership Council (LC) for ongoing program and community priorities. It summarizes the compiled input from LC discussions and through its community-led governance subgroups. This document accompanies another one that outlines specific recommendations to the NSF for the next (early 2020) EarthCube Solicitation to make progress on these priorities [LINK]. A related document was also prepared in Spring/Summer 2019 to serve as the LC's suggestions to the Annual Program Plan for the next Science Support Office, known as ECO [LINK].

The EarthCube LC is committed to advancing community goals through the assessment and articulation of community needs, priorities, and mechanisms to accomplish them via shared infrastructure development, enabled by NSF proposals or other funding mechanisms to projects focused on mission-critical goals. Building off our 2018 priorities, mission-critical continues to be defined as the providing of advanced digital tools to a wide cross-section of the USA Geosciences community to enhance data and knowledge discovery, research workflows and outcomes, data access and data stewardship.

State of the Community and its resources: Geosciences research often involves collection and use of large, complex environmental datasets, often collected in geospatial context and streamed in real-time. All of these present special challenges for a Geosciences community infrastructure. EarthCube (EC) is the NSF program and vibrant community of researchers who aim to enhance the process and accelerate the outcomes of Geosciences research through development and use of advanced digital tools for workflow optimization, archiving, description and discovery of data, data resources, and associated human resources. The transformation in science workflows that EarthCube aims to support is part of a multi-decadal effort at NSF and elsewhere to bring digital tools and data stewardship practices to geoscientists.

Over the course of the past 5-7 years, an increasingly well-defined community of researchers (scientists and solution providers) have defined needs, priorities and possible implementation methods to enable cyberinfrastructure resources for a broad, cross-domain sector of the

Geosciences research community. A growing catalog of EarthCube and other digital tools are being employed by Geosciences researchers in their workflows, which, along with dedicated software employed by data repositories, is helping to maintain and advance the availability and use of such datasets. More recently the LC, Science Support Office, and community have zeroed in on several targeted developments (e.g., data and resource registries) to build out a core infrastructure that, along with a workbench like environment, will support coordination, access to and interoperability of existing and new cyberinfrastructure components for Geosciences research. This effort, funded by solicited subawards through the Science Support Office (SSO), envisioned and scoped by the LC and SSO Technical Officer, and advised by governance and community advisory teams, has been named “GeoCODES” (Geoscience Cyberinfrastructure for Open Discovery in the Earth Sciences). GeoCODES has been used to develop and implement registries for data repositories and data resources for the Geosciences community. While these efforts are somewhere in between pilots and production level implementations, they clearly serve important community needs and should be supported and expanded as a targeted mechanism to develop needed infrastructure that has proved difficult over the past 5 years of NSF proposal solicitations.

Alongside the aforementioned developments, broader community awareness and research into what currently limits the time to science in many cases has led to scientific community initiatives such as FAIR (which seeks to make data Findable, Accessible, Interoperable, and Reusable). Additionally, awareness of needs, tools, applications, and solutions has grown dramatically in the past 5-7 years among both working scientists and the upper echelon of NSF management, leading to next-level Foundation-wide cyberinfrastructure programs such as HDR (Harnessing the Data Revolution). While these positive developments make a lot of sense across the full range of scientific disciplines, there remains a strong need for a program like EarthCube, which is focused on the Geosciences, and their particular data and modelling needs. Finally, data management practices mandated by NSF for the research it supports, while by no means ideal, have evolved considerably over the lifespan of the EarthCube program, for the positive, such that data stewardship and persistent data access are commonly supported by GEO directorate programs, and yet the infrastructure to support this mandate is still uneven across the directorate. EarthCube can help address this.

B. Summary of Immediate Priorities

EarthCube recommends a continuation of the priorities it compiled in 2018, with some refocusing and expansion to account for successful activities over the past year, emerging trends in the field, and emphasizing that our highest priority is for projects resulting in broadly applicable developments demonstrated on domain use-cases, as opposed to domain-targeted solutions. We see high priorities in:

1. Science Advancement and Workflow Support.
2. Data and Data Resource Integration and Reuse.
3. Community Education and Outreach.

Additionally, the Leadership Council wishes to highlight some subtopics that support Science Advancement and augment this general framework:

Goal	Sub goals	A	C	E	S
Strengthen and increase participation in the EC community	Integrate EC PIs into EC community activities				x
	Provide funding to coordinated education and outreach activities			x	x
	Support engagement activities during EC project execution	x	x	x	x
Increase adoption, establish links to outside organizations	Foster mechanisms to engage more scientists to use EC capabilities	x		x	x
	Require adoption of FAIR principles in new projects	x	x		
	Encourage data centers to adopt and implement the standards and practices that underlie current GeoCODES tools	x	x		
	Promote EC alignment with related US and international projects			x	x
Accelerate contribution to the EC cyber-infrastructure	Expose workbench specifications through the documents repository and develop a form to expand as needed.	x			
	Incentivize information population and user adoption of in-development registries		x	x	
	Encourage development of consistent documentation and illustration with common use cases	x	x		

Activity codes: A=Adoption of EC guidelines, C=Cyberinfrastructure development, E=Education and outreach, S=Strengthen EC community

The LC furthermore stresses that there is a continued need to move EarthCube activities into a more user-focused mode, provide more leveraging of existing digital tools, data centers (especially those in the EarthCube Council of Data Facilities), and parallel funding initiatives at NSF, and to enable FAIR principles for geodata and data resources (see EarthCube 2019 FAIR policy statement). There is also a need to build a better collective understanding of what has been produced by EarthCube funded projects to date and what state it is currently in, which might be addressed through workshops and/or a governance Tiger Team.

Finally, we stress the importance of developing and employing a professional, coordinated, outreach campaign for the Geosciences community about data stewardship practices, digital workflow support, and EarthCube's evolving role in these areas. Such a dedicated outreach campaign to help more working geoscientists discover and use new digital tools in their research would likely accelerate tool adoption, and EarthCube program goals.

Funding mechanisms: The primary mechanism of funding via NSF solicitation and peer review has led to the majority of EarthCube accomplishments. However, alternate funding mechanisms (SSO subawards via RFP, targeted pilots via supplemental funds) also contribute to the accomplishment of community goals by allowing progress on strategic developments or implementations that are difficult to achieve by open NSF proposal competitions. We believe there is a continued, perhaps even expanded role for such alternate funding moving forward, with the proviso that awards include strong oversight beyond the PI group via EC community and Governance-populated advisory teams.

Subawards funding: Recent work accomplished through subawards via the EarthCube Science Support Office, such as P418 and others, have shown the value of short-term, targeted funding for specific components of EarthCube infrastructure with direct office and governance oversight. Having these efforts coordinated by the EarthCube Office has proven to be a good model, especially for features that help form the infrastructure for GeoCODES. This funding mechanism has required less overhead and is faster to put in place, such that the components can be built and deployed quickly, while maintaining oversight through advisory teams formed among members of the EarthCube community. Other examples of ways this mechanism could be used include workshops, training and hackathons, targeted collaborations on tool use to further demonstrate, build out or integrate EarthCube tools. As GeoCODES develops further, it will be important to retain this subaward funding method to quickly and effectively develop infrastructure.

C. Science Advancement - End-to-End Workflow Support

Implementing strategies for scientific workflow support remain very important, including: (1) promoting to geoscientists the existence and use of existing EarthCube components used to help them identify data, software and databases; (2) promoting FAIR principles with respect to NSF-supported GEO data, tools and data resources; and (3) developing a staged-priority structure for further cyberinfrastructure development. The LC encourages high-impact and low-effort research activities for immediate attention within the range of topics addressed within the NSF-GEO directorate.

In addition to our next solicitation recommendations [LINK], we suggest these additional governance-based priorities:

- Further develop the Council of Funded Projects working group with one PI rep from each project - mandatory participation.
- Connect with existing outreach activities aimed at scientists beyond current EarthCube reach (i.e., with webinars) defined in Section D below.
- Provide financial support for utilising and/or extending one or several existing EarthCube resources for scientific exploration on IDENTIFIED and widely applicable data needs within the NSF-GEO community (via supplements or office awards).
- Further develop presentations about EarthCube at relevant science workshops and a few slides for scientists funded by EarthCube to use in their science talks.
- Put in place a scientist outreach and education plan, with a professional from the office to carry it out.
- Support scientists who wish to employ or expand EarthCube resources in their original research through:

    - Scientist workshops and demos for GeoCODES and Workbench products.
    - Technical and scientific support network development, either through the office, an RCN, or an ad hoc group managed by EC governance.

D. Data and Data Resource Integration and Reuse

EarthCube has put a priority on incorporating and extending FAIR concepts for data to all of the elements of a data intensive scientific workflow (e.g., analysis tools, repositories, cloud services). The FAIR principles posit that data (and data resources) should be Findable, Accessible, Interoperable and Reusable. In a recent position paper, the EarthCube LC concluded that such resources should also be open-access and sustainable to provide the highest benefit to the scientific community. EarthCube continues to promote a strategy wherein EarthCube infrastructure and resources are targeted towards alignment with FAIR principles. The LC continues to endorse activities that will lead to great exposure and use of EarthCube resources in a FAIR-optimized way.

In addition to our next solicitation recommendations [LINK], we suggest these additional governance-based priorities:

- Develop a strategy and timeline for building out GeoCODES with the Technical Officer, adding capabilities, reaching new communities for adoption, etc.
- Use scenarios to guide workbench pilot and full implementation, deliverables schedule, and success metrics.

D1) Registries

The EarthCube Registry (or group of registries) will help geoscientists answer questions like:

1. "What EarthCube and other tools work with the data set I want to use?”
2. “What EarthCube software components can I use to get my application working?”
3. "How can EarthCube help me improve my workflows?"

The challenge of linking datasets and software to more quickly visualize, analyze, or otherwise utilize data is a problem faced by the entire Geosciences community. The **ultimate goals** for EC Registries are to: (1) enable software clients that utilize dataset descriptions from the registry to seamlessly connect a user with that data in a working environment where they use the data; and (2) allow developers who are building and geoscientists who are using such software to find and reuse existing components to accelerate time to science.

This concept, initially put in place with the Earthcube Data Resource Registry (i.e., P418/P419) has been extended to the in-development EarthCube Resource Registry that will list tools software components, semantic resources, models and repositories, with descriptive information to facilitate evaluation for reuse and, where feasible, to directly connect data with services, tools and other capabilities, and to facilitate working with them. If accomplished it will be a major outcome of the program and a major resource to geoscientists.

The deliverables for the current EC Resource Registry project at the end of October 2019 include an implemented document database and triple store with example resource descriptions from previous completed EarthCube projects, demonstrations with sample queries, documentation of the information model, and an interchange serialization using JSON-LD. The Resource Registry will be able to utilize the same GeoCODES technology platform developed by the the data registry projects (P418/P419), and that the GeoCODES graphical user interface, developed to allow users to upload and search for dataset descriptions, can also be configured to work with the resource registry.

The **next steps** needed to make EarthCube Registries fully functional and useful to the Geosciences community are listed below. Since these are not research projects per se, funding through EC Office subawards might be the best mechanism to get the work done. Implementation of a production-quality user interface for entering new resource descriptions. All interfaces work should be based on sound usability design, and include iterative user testing and feedback plans.

1. Implementation of production-quality, well-tested, functional and accessible user interface for searching registry content and viewing results.
2. Implementation of production-quality, well-tested APIs to enable the community to build new software applications using registry resource descriptions.
3. Work with commercial search providers (Google, Bing, Yahoo) to verify that the Schema.org JSON-LD resource documentation is successfully indexed and accessible through commercial search sites.
4. Expansion of the data registry to allow discovery of data holdings in addition to metadata about holdings.
5. Support for training and outreach to work with the community to populate the registry and establish a community of practice to sustain the registry (this is the biggest job).
6. Support for outreach to work within specific scientific communities to assess domain specific requirements (e.g., searchable concepts), and iteratively extend the system to meet those needs.
7. Establishment of sustainable governance processes to manage vocabularies necessary for interoperability, and the evolution of the registry information models, interchange serialization, and API.
8. More details are available here.

One additional goal for the registry is to help geoscientists with their data management needs, for instance with respect to the NSF Data Management Plan (DMP). A one-stop access for NSF DMP compliance would go a long way towards establishing FAIR practices for NSF-funded Geosciences research. The EarthCube Resource Registry will serve as a portal to discover data and data resources produced by EarthCube, and more generally within the Geosciences. The EarthCube Resource Registry could be further developed to act as a dispatcher for Geosciences resources for enablement of FAIR practices in the Geosciences and to help release the DMP burden on NSF award recipients. While collecting the high-level metadata from the scientist, it could also help them to choose an appropriate data provider and/or data resource and redirect the products to that location. Additionally, the Registry could be connected to a generic resource storage facility (a concept that NSF should consider how to support, either through the EC program, or more broadly) that would serve as a placeholder data provider for those data and data resources that are not currently associated with an existing data storage solution.

We note that none of the solutions described above address the larger challenge of long-term storage of the information in the registries. We encourage NSF through EarthCube and other aligned efforts at the Foundation to explore mechanisms to lower the cost barrier to long term storage solutions through the implementation of shared resources, tools, standards and strategies, thereby developing an economy of scale that could serve the entire foundation, not just the Geosciences Directorate.

D2) Integrated Tool Platform (“Workbench”)

The concept of an application environment that serves as a data and digital tool integration platform, wherein geoscientists can seamlessly develop interactive workflows using existing cyberinfrastructure tools (EarthCube and otherwise) to improve the time to science and/or initiate new discoveries, has been an interest to the EarthCube community for several years, albeit with a range of capabilities and uses. The LC does not wish to limit any possible future uses or extensions to such a community-resourced workbench, but suggests that to start the development off, the effort be framed around specific targeted use-scenarios, promotes tool interoperability through semantics and APIs, reuses existing components if appropriate, and demonstrates robust outcomes within 12 to 18 months of funding. Use scenarios could include:

- Find, access, integrate and analyze together three geospatial data sets from at least 2 NSF Geosciences divisions Earth, Ocean, Atmosphere, Polar for a science driver.
- Find, access, integrate and analyze together three temporally bounded or real-time data sets from at least 2 NSF Geosciences divisions Earth, Ocean, Atmosphere, Polar) for a science driver.
- Find, access, and use 2 geospatial or temporally bounded data sets for use with at least one external model (geodynamic, thermodynamic, compositional, etc.) preparing the data for proper scaling, density, and coverage to achieve a statistically significant model outcome.

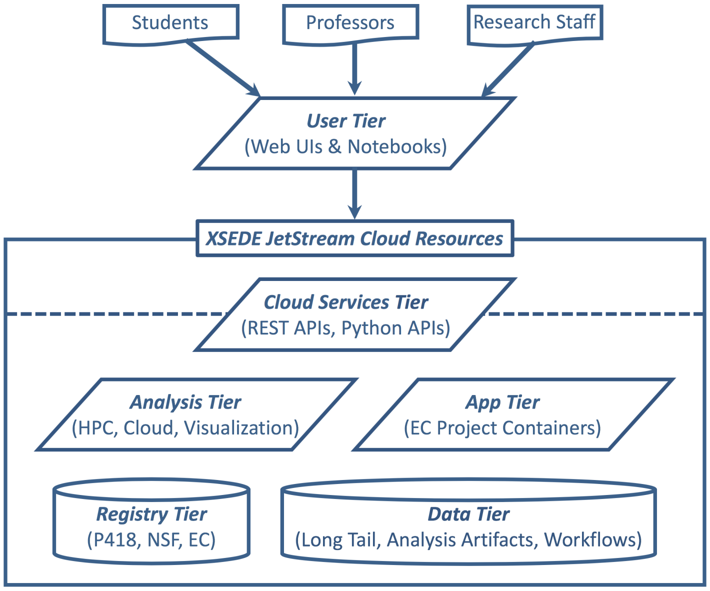
These activities could be initiated as one or more pilots via the solicitation or subaward support through the office.

Possible EarthCube “workbench” implementations have been discussed for several years and has generated several high-level planning documents:

1) the Leadership Council Guidance on the EarthCube workbench,
2) the Workbench Tiger Team report and
3) the SSO Technical Officer workbench assessment (see the figure).

The concept of an N-tier workbench, as specified in Document 3 above, is to build and execute a modular workflow support environment that leverages existing or new technologies and services and reuses, adopts, and adapts previous and current innovative approaches and solutions to Geosciences computational and data needs, and provides a collaborative environment to foster close working relationship between earth scientists, computer scientists, data scientists and resource providers.

Since the time these documents were created, components of EarthCube's GeoCODES architecture have been or are currently under development, providing some kernels of what could be an eventual n-tier workbench functionality, including:



- The schema.org (P418/P418GUI/P419) dataset and data repository discovery service.
- Access NSF-GEO data repositories.
- Access tools in the Resource Registry.
- Discover/use data sets and compatible models.

In-development components provide the ability to find datasets, repositories and other resources shared among the EarthCube community, but they do not as of yet support specific use scenarios or identified workflows. The EarthCube community needs an expand tool integration/workbench functionality within an N-tier architecture. This could be fast-tracked by funding pilot projects that demonstrate the use of components via a reproducible science workflow, demonstrating interoperability among several tools (e.g., to address the "I" in FAIR) and allowing these components to be assessed and refined. Implementations of tool platform components and workflows within notebook environments (e.g., RMarkdown, Jupyter, MATLAB live scripts) are desirable. The awards should leverage existing workbench functionality from other major cyberinfrastructure efforts as described in the SSO Technical Officer's assessment document. In addition, projects and workflows demonstrating workbench functionality should become artifacts (e.g., new resources) for future researchers to examine and build upon. These awards could also be carried out by way of supplements to existing science grants.

E. Education and outreach

The digital world is changing rapidly, with major advances in data resources, data storage, cloud computing, web-based data exchange protocols, and associated semantic resources. Some Geosciences domains have kept up to date with these changes, and yet, anecdotally, many working geoscientists are still in a 1990s or 2000s siloed, disconnected data paradigm for data management and use. EarthCube can change this if it engages with more scientists.

Community engagement has been a subject of active debate in the EarthCube community for years, focusing on increasing engagement among scientists, engineers, technologists, and students with the EarthCube program, but without resources or a dedicated person to put them in place. Thus, a major goal for the next year or two is to incorporate mechanisms for funded projects to support community engagement more effectively, thereby rejuvenating and building new interest in the EarthCube Program. Currently, the majority of scientists using EarthCube resources are associated with a funded project through use cases, or via visioning exercises in one or another Governance team. Engaging more scientists to utilize EarthCube resources and to drive user-relevant resource development is critical for the advancement of science through the EarthCube initiative. Plus, engaging more working scientists will allow EarthCube to promote modern thinking about data stewardship and data resources to support FAIR data practices to much broader communities. There are currently efforts to connect EarthCube resources and make tools, data, and data resources to make them more readily available through GeoCODES. This should be leveraged by more scientists via education and outreach, as a priority through one or more NSF proposal mechanisms.

In addition to our next solicitation recommendations [LINK], we suggest these additional governance-based priorities:

- Develop an Education and Outreach functionality within the next SSO, **staffed by an outreach professional,** who can develop a coordinated outreach strategy and format for outreach products, revise or redo the EarthCube website to be external user-facing and to easily enable resource discovery, and to help funded projects coordinate their outreach efforts.
- The Outreach Coordinator should be responsive to the community via the Leadership Council.
- Redefine the role of the Engagement Team to support this person, engage in long range vision exercises, and interact with the community to help understand their needs and develop effective communications strategies.
- Promote the GeoCODES effort to identified science domain communities that can benefit from adoption.
- Analyze how well EarthCube is satisfying its own goals with respect to previously identified success metrics, and identify methods for improvement

In order to engage more scientists, active and passive modes of engagement must be improved. There are various examples of sustainable cyberinfrastructure tools in the Geosciences and elsewhere that all share a few key features, including a core of invested developers and a community built around those tools. EarthCube must seek to build communities not only in governance, but also around the tools themselves, and doing so requires cultivating avenues through which new scientists can engage with EarthCube tools.

This engagement will be difficult and requires more than a volunteer effort, yet it is critical if EarthCube is to be sustained and useful into the future.

The LC strongly recommends that these activities be driven by a single point of responsibility in the Science Support Office, whose focus is on building communities around mature and developing tools in the EarthCube ecosystem, using professional outreach methods. Much of this person's duties would be in ensuring a coherent and easily accessible flow of information about the tools and activities themselves via the EarthCube website, which would require education of PIs and team members developing consistent documentation and illustrating common use cases. The position would also oversee potential activities include training workshops, online tutorials, instructional webinars, detailed manuals, and workflow support.

F. Alignment and interaction with other funded efforts

The EarthCube community began its formation in 2011. Community building, especially for a revolutionary concept such as EarthCube, is a slow, painstaking process with many hurdles. The tools, services and other outcomes from the EarthCube community have prepared Geosciences researchers to meet new data and cyberinfrastructure challenges ahead as described by NSF's Harnessing the Data Revolution (HDR) Big Idea. In many cases, the architecture, widely used components and building blocks of EarthCube can successfully scale up to broader initiatives across the NSF. It is imperative that tools and resources are designed as interoperable and reusable components to maximize use and minimize duplication of efforts across the Foundation wherever possible.

Toward this end, new key initiatives should be leveraged with EarthCube developments including Cyberinfrastructure for Sustained Scientific Innovation (CSSI), and the NSF Convergence Accelerator. The web page at [LINK] is a good start, but if possible, NSF should work further with EarthCube governance and the Office to inform the EarthCube community about related solicitations in a timely fashion so they can take advantage of them. In addition, it would be beneficial for EarthCube technologies to be integrated with other CI efforts, and deployed in standard science grants across the GEO Directorate, which could be facilitated both in new solicitations and existing grant supplements.

EarthCube's domestic projects can also be leveraged with work abroad such as the European Plate Observing System (EPOS), the Environmental Research Infrastructures (ENVRI), European Science Cloud (EOSC) and the Australian National Data Service (ANDS).