

# Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 18:55:01

PAGE 1

REFERENCE NO: 307

This contribution was submitted to the National Science Foundation as part of the NSF CI 2030 planning activity through an NSF Request for Information, [https://www.nsf.gov/publications/pub\\_summ.jsp?ods\\_key=nsf17031](https://www.nsf.gov/publications/pub_summ.jsp?ods_key=nsf17031). Consideration of this contribution in NSF's planning process and any NSF-provided public accessibility of this document does not constitute approval of the content by NSF or the US Government. The opinions and views expressed herein are those of the author(s) and do not necessarily reflect those of the NSF or the US Government. The content of this submission is protected by the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>).

## Author Names & Affiliations

- Emily Law - JPL/ESIP
- Christine White - Esri/ESIP
- Bruce Caron - ESIP
- Stace Beaulieu - Woods Hole Oceanographic Institution
- Wade Bishop - University of Tennessee
- Jim Bowering - College of Charleston
- Sky Bristol - USGS
- Dave Jones - StormCenter Communication
- Leslie Hsu - USGS
- Soren Scott - Ronin Institute
- Ruth Duerr - Ronin Institute
- Erin Robinson - ESIP

## Contact Email Address (for NSF use only)

(Hidden)

## Research Domain, discipline, and sub-discipline

Geoscience, Earth Science Informatics

## Title of Submission

Earth Science Information Partners, Vision for the Future of Cyberinfrastructure

## Abstract (maximum ~200 words).

ESIP is a community of data and information technology practitioners who collaborate across Earth science organizations. Our response comes from the perspective of maximizing technical and organizational cooperation among diverse entities. First, the challenge of overcoming technical, institutional, and cultural barriers for pursuing transformative, interdisciplinary Earth and Space science. Another challenge is building semantic understanding across different research communities seamlessly. Fully taking advantage of the analytical resources available through big earth science data is another challenge, and the last is better support for research code and information provenance to improve and expedite research reproducibility. Regarding specific cyberinfrastructure needs, the first is support for data driven knowledge development via collaborative analytics and visualization. Another recommendation is that NSF leverages existing infrastructures and opportunities for partnering instead of duplicating or creating competing resources. The last need is for project support and sustainability, to transition projects into a next funding phase as appropriate. Our response concludes with an additional consideration regarding education. All stakeholders benefit from education emphasizing aspects of data curation, management, and data and software stewardship. ESIP recommends a multidisciplinary approach to cyberinfrastructure training to support current, as well as future researchers.

# Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 18:55:01

PAGE 2

REFERENCE NO: 307

**Question 1** Research Challenge(s) (maximum ~1200 words): Describe current or emerging science or engineering research challenge(s), providing context in terms of recent research activities and standing questions in the field.

**Transformative Interdisciplinary Science:** Breakthroughs in understanding Earth and Space phenomena are made when we pursue transformative interdisciplinary science. Grand science challenges beyond a single discipline require integrated study to better understand the Earth and Space Systems to predict outcomes, consequences and impacts. Building Earth and Space data infrastructures in innovative ways that enable data, model and system interoperability is key to integrate diverse scientific data and observations stemming from various fields of research and transform them into knowledge that advance scientific discovery. Today, most science and engineering research is done independently relying on siloed discipline-oriented data and software infrastructures. Even though organizations such as ESIP, Research Data Alliance (RDA), The Open Geospatial Consortium (OGC) and others work to address interoperability, many technical, institutional and cultural barriers continue to hinder researchers in conducting transformative science. Our vision for Cyberinfrastructure is a collaborative environment/infrastructure with analysis, modeling, and visualization capability for integrated study, with a culture where sharing is rewarded and rewarding.

**Semantic Understanding Across Communities:** One of the major inhibitors of interoperability is the semantic heterogeneity of the language and concepts used within different disciplines. Language differences developed over decades as each discipline progressed on its separate path. Often the terms used gloss over the significant underlying assumptions and tacit knowledge within that discipline, concepts that are simply taken for granted within that field. As such, simply learning to communicate with researchers in other disciplines can take considerable time. These language differences make it difficult for researchers to find, assess and properly use data originated by another discipline. To facilitate transformative interdisciplinary science, interdisciplinary semantic infrastructure is needed. Ideally one day infrastructure would assess the knowledge base of each user and translate from the semantic knowledge base of the data producer to that of the user to appropriately tailor the services provided (e.g., data search, browse, assessment and access capabilities). While progress is being made in semantics through several paths including both traditional ontology development and Natural Language Process and Machine Learning techniques, the first step is to ensure that the semantic resources created for various fields are curated and maintained for the indefinite future.

**Big Earth Science Data:** In addition to challenges in integration and communication, there are challenges in analysis and modeling with the higher volume and frequency of data used by today's Earth scientists. The Internet-of-Things is enabling more frequent and ubiquitous observations of the environment. Analyses and modeling of atmospheric, ocean, and solid Earth phenomena are needed in greater frequency than ever before to forecast not only the weather but to predict earthquakes, aquatic algal blooms, air quality and other atmospheric chemistry changes, and so on. These analyses and models require highly efficient and reproducible workflows with integrated data and software infrastructure to enable disparate organizations to work together in real time. Every research challenge in the coming decade must plan to accommodate and harness data streams that push the boundaries of volume, velocity, variety, and veracity.

**Reproducible Science and Information Provenance:** The ability to reproduce and build upon prior research is critical to advancing Earth Science. Two opportunities for technology to support the research process are in growing researchers' code and software capabilities and improving information provenance. Data, software, and code are required artifacts for other scientists to reproduce and build upon Earth science research. ESIP supports scientists in sharing their code and has provided ESIP Software Guidelines (<https://esipfed.github.io/Software-Assessment-Guidelines>) as a community resource to ensure high quality and reuse. In addition, ESIP recommends that the scientific review system - through publishers, journals, and scientific institutions - continue to demand the norm of data and code being fully coupled in the release of scientific findings. The paper of the future should be fully executable. With these advancements in code-driven science comes a challenge in managing and maintaining the scientific provenance of discovery and new knowledge. With so much information available, how do researchers determine the most appropriate pathway and what resources to use and trust? Continued research and development are needed in scientific provenance methods and technologies, and continuing education for scientists and scientific repository managers is required to implement sound and sustainable techniques.

**Question 2** Cyberinfrastructure Needed to Address the Research Challenge(s) (maximum ~1200 words): Describe any limitations or absence of existing cyberinfrastructure, and/or specific technical advancements in cyberinfrastructure (e.g. advanced computing, data infrastructure, software infrastructure, applications, networking, cybersecurity), that must be addressed to accomplish the identified

# Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 18:55:01

PAGE 3

REFERENCE NO: 307

research challenge(s).

**Collaborative Resources for Analytics:** Cyberinfrastructure that supports data driven knowledge development via collaborative analytics and visualization to support decision making is needed to address research challenges and to drive future mission and acceleration of scientific discovery. Advancement requires architecture, development, research and operation of scalable data and software infrastructures that address challenges presented in the entire science data and software lifecycle (from collection, through long-term management, discoverability, distribution, processing, analysis, visualization, and curation). Specific technical advancements needed in such cyberinfrastructure include: information modeling (e.g., using semantic technology), advanced discovery and exploratory analytics (e.g., using natural language processing, pattern recognition), cognitive computing (e.g., a combination of automation, machine learning), advanced analysis tools and visualization (e.g., immersive, interactive visualization), and the ability to sustain and enhance these resources over time.

**Leverage Existing Infrastructures and Opportunities for Partnering:** There are many existing data centers, repositories, and web-enabled resources to support the scientific community. Several ESIP member organizations specialize in such data, information products, infrastructure, tools, and collaboration mechanisms. Therefore, ESIP encourages NSF to partner with existing or upcoming programs to advance innovation and discovery through cyberinfrastructure applications, and avoid building duplicative or competing resources. Building new cyberinfrastructure should be a careful and well-researched endeavor where the resulting resources, capabilities and technologies truly fill gaps that existing resources could or would not fill. Coordination of existing repositories and resources, and training for domain scientists on how to share their data in existing repositories, would be a valuable contribution.

**Sustainability:** The NSF is a significant driver of research software creation, whether through dedicated cyberinfrastructure programs or indirect products of the scientific programs. The importance of these technologies (Hettrick, 2014 - 1) for improving research outcomes and opening new research areas is undeniable. Effective sustainability of these projects, however, is less evident. We often turn to open source practices, but recent research has raised concerns that the Open Source Software model is not sustainable (Eghbal - 2). This concern is compounded by the nature of some of our grant-funded projects, those that may support science but find no financial support or mainstream adoption in industry.

We suggest the NSF consider an approach that supports research into our research software ecosystem and its impacts across the breadth of funded activities and that bridges the public/private communities to encourage new sustainability practices. This may require a funding model not currently found within any NSF directorate to provide some measure of financial sustainability for projects providing value to the scientific community and with an understanding of project lifespans. NSF programs like SBIR-STTR or the supported I-Corps provide frameworks for migrating research activities into industry. We can envision similar efforts to address the operationalization and longer term support for domain adoption, with a focus on the workforce development needed to address operationalization needs and to address sustainability through community management and feedback. This may also include new funding vehicles and models between the NSF and universities, between funded research groups or between industry partners, the NSF and research groups to allow for ongoing support to those projects that are operational within the research community and that are not well-served by traditional grant funding opportunities.

1 Hettrick, Simon, Neil Chue Hong, Les Carr, David De Roure, Giacomo Peru, Stephen Crouch, Aleksandra Nenadic, et al. "UK Research Software Survey 2014." Zenodo, December 4, 2014. doi:10.5281/zenodo.14809.

2 Eghbal, Nadia. "Roads and Bridges: The Unseen Labor Behind Our Digital Infrastructure." Ford Foundation. Accessed July 15, 2016. <http://www.fordfoundation.org/library/reports-and-studies/roads-and-bridges-the-unseen-labor-behind-our-digital-infrastructure/>.

**Question 3** Other considerations (maximum ~1200 words, optional): Any other relevant aspects, such as organization, process, learning and workforce development, access, and sustainability, that need to be addressed; or any other issues that NSF should consider.

**Education:** To enable the frontiers of science and engineering to advance over the next decade and beyond, the workforce development aspects of cyberinfrastructure require further study. Scientists, engineers, and other stakeholders of cyberinfrastructure all benefit from formal and informal training. Education should emphasize the organization, access, and use aspects of data curation, data management, and data and software stewardship while including ethical discussion of the implications of each choice made. A multidisciplinary approach of curriculum development would lead to a coordinated 21st Century workforce that scaffolds from K-12 to graduate programs related to cyberinfrastructure.

# Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 18:55:01

PAGE 4

REFERENCE NO: 307

Dramatic changes in data creation, dissemination, and re-use impact many aspects of education and training for a workforce that not only creates data, but facilitates science and engineering via the ethical organization, access, and use of data. In addition, as big data trend is driving the shift of science data system from data stewardship to data-driven knowledge discovery, it is important to focus skill sets, training and retention strategy in data science and cognitive computing that advance data usability and data-to-knowledge transfer (such as, data fusion, machine learning, advanced analytics and visualization) in an ethical and coordinated manner.

With anticipated growth in the workforce, new curricula being developed to prepare individuals for these new jobs, and policy requiring investigators to ethically make their data available, these workforce development efforts would benefit from coordination.

There are many efforts already underway. For example, the academic competencies to guide future education in the geospatial sciences have received considerable attention from the University Consortium for the Geographic Information Science, which has produced several iterations of a Geographic Information Science and Technology Body of Knowledge funded by NSF (DiBiase et al., 2006 - 1; DiBiase et al., 2010 - 2). For Information Science, a number of Library and Information Science schools and Information Schools offer curriculum related to digital curation. A matrix of skills and functions for digital curation was one outcome of the Preserving Access to Our Digital Future: Building an International Digital Curation Curriculum (DigCCurr I) and Extending an International Digital Curation Curriculum to Doctoral Student and Practitioners (DigCCurr II) projects funded by IMLS (Lee, 2009 - 3; Poole, Lee, Barnes, & Murillo, 2013 - 4).

However, findings from a recent survey conducted by the Usability & Assessment Working group of DataONE determined that very little data management instruction occurs at the undergraduate level and the most taught topics are Quality Control (21.6%), File Management (20.1%), and Citing Data (19.4%). Yet, these topics are only covered by approximately one-fifth of the 134 instructors surveyed (Tenopir et al., under review - 5). Although the cyberinfrastructure enables the frontiers of science and engineering to advance, further education and training is needed to develop a workforce that focuses on the professionals tasked to maintain data curation, data management, and data stewardship activities.

The current training for individuals working in the areas of organization, access, and use of the data within the cyberinfrastructure requires the coordination of several competencies. These competencies need to be addressed in the curriculum from K-12 to graduate programs, as well as informal learning opportunities for professionals. If current professionals are involved in the creation of these competencies in a systematic way for both formal and informal opportunities, the entire future workforce will benefit.

One idea would be to develop, and fund, a new field of science - Cyberinfrastructure Science - that includes all the related sciences that now suffer in academic silos. The US academic system is not currently configured to support interdisciplinary research for tenure and promotion. This failure is widely judged to severely dampen young researchers interested in exploring the target space of NSF's 21st Century Cyberinfrastructure initiatives - the space occupied by the traditional sciences such as earth science, computer science, and human-centered sciences such as psychology and ethics.

1 DiBiase, D., DeMers, M., Johnson, A., Kemp, K., Luck, A. T., Plewe, B., & Wentz, E. (2006). Geographic information science and technology body of knowledge. Washington, D.C.: Association of American Geographers.

2 DiBiase, D., Corbin, T., Fox, T., Francica, J., Green, K., Jackson, J., ... & Van Sickle, J. (2010). The new geospatial technology competency model: Bringing workforce needs into focus. *URISA Journal*, 22(2), 55.

3 Lee, C. (2009). Functions and Skills (Dimension 2 of Matrix of Digital Curation Knowledge and Competencies). Retrieved from <http://www.ils.unc.edu/digccurr/digccurr-functions.html>

4 -Poole, A. H., Lee, C. A., Barnes, H. L., & Murillo, A. P. (2013). Digital curation preparation: A survey of contributors to international professional, educational, and research venues. Association for Library and Information Science Education. Seattle, WA.

5 Tenopir, C., Allard, S., Sinha, P., Pollock, D., Birch, B., Dalton, B., Frame, M., & Baird, L. (2015). Data management education from the perspective of science educators.

## Consent Statement

- "I hereby agree to give the National Science Foundation (NSF) the right to use this information for the purposes stated above and to display it on a publically available website, consistent with the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License (<https://creativecommons.org/licenses/by-nc-nd/4.0/legalcode>)."

# Submission in Response to NSF CI 2030 Request for Information

**DATE AND TIME:** 2017-04-05 18:55:01

**REFERENCE NO:** 307

**PAGE 5**

---

---