Submission in Response to NSF CI 2030 Request for Information

DATE AND TIME: 2017-04-05 15:26:57
REFERENCE NO: 261

Author Names & Affiliations

- Patricia Soranno - Michigan State University
- Katelyn King - Michigan State University
- Autumn Poisson - Michigan State University
- Joseph Stachelek - Michigan State University
- Claire Boudreau - Michigan State University
- Nicholas Skaff - Michigan State University
- Nicole Smith - Michigan State University

Contact Email Address (for NSF use only)

(Hidden)

Research Domain, discipline, and sub-discipline

Ecology, Limnology

Title of Submission

Cyberinfrastructure support for collaboration and open science in ecology

Abstract (maximum ~200 words).

Two of the central goals of ecology for the coming century are to better understand species responses to climate change as well as to better resolve the role that ecological systems play in modulating global cycles of important elements such as carbon and nitrogen. Taking on these goals requires tackling emerging challenges associated with curating and analyzing environmental data at fine scales and across large spatiotemporal extents. The unprecedented volume and scale of modern research data requires support for new technologies capable of handling such large and diverse data sets as well as support for new classes of researchers trained in analyzing this data with scalable programming-based workflows. Specific technology needs include expanded services for public curation and storage of geospatial data products. These services should maximize reuse and discoverability by encouraging the use of open licensing and non-proprietary file formats. Specific researcher-support needs include expansion of technical training opportunities, support for distributed science networks, and recognition of software and data packages as primary research outputs in grant applications.

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.

The environmental sciences, and ecology in particular, are increasingly tackling research challenges that leverage massive amounts of

data being collected at unprecedented space and time-scales. Ecologists must address two classes of challenges in the coming century. First, ecologists must generate better predictions of how species will respond to a changing climate. This is a significant challenge as it requires not only detailed knowledge (at fine scales) of a particular species' climate response but also knowledge of how that species interacts with other species in increasingly novel habitats. Data needs are complex because understanding and predicting species change is not a simple matter of understanding species response in isolation. Rather, individuals, populations, and communities respond to a complex suite of ecological, geological, climatic, human, and other factors. Thus, data on all such factors are required. Data structures need to accommodate different types and rates of data collection because some factors change rapidly (e.g., weather, and some human activities), while others change slowly, if at all (e.g., geology). The challenge of generating better predictions is further complicated by the need to extrapolate knowledge gained at the local scale to regional and continental scales in order to facilitate predictions of species range shifts. Such predictions are especially important for species that have the potential to spread pathogens and disease to humans as well as among themselves. Developing the needed knowledge to make such predictions requires cyberinfrastructure to support the collection, integration, storage, and access to the necessary data for building such predictive models across space and time.
The second class of research challenge faced by ecologists is to resolve our limited understanding of the roles that ecological systems (such as stands of forest, or lakes and rivers) play in modulating global cycles of important elements like carbon or nitrogen. Previous research has demonstrated that ecological processes influence the carbon cycle at regional and sometimes global scales. However, it is still not common practice to incorporate all of the potentially important ecological processes in global climate models. Part of the reason for this is a lack of data across broad spatial extents to build the needed understanding necessary to incorporate ecology into regional, continental, and global simulation models.
Scientists need new models of conducting science to collect the necessary data to make reliable predictions at large spatial scales. In part, this is because such predictions require consideration of complex ecological interactions at a hierarchy of spatial scales from the local scale, to the regional, continental, and global scale. For example, increasingly 'science networks' are being created where scientists work together on broad-scaled research problems in an environment where it is assumed data are shared (at least within the network, and sometimes beyond). By working together at unprecedented scales, data can be collected more broadly than before, thus building the needed broad scale knowledge. The emerging field of macrosystems ecology is helping to support such research, but it requires cyberinfrastructure tools to support the integration of both fine-scaled site-based research with broad-scaled, global research. Ultimately, ecologists will need this support to actively participate in further integrated nexus-approaches addressing issues related to the intersection of social, natural, and economic systems in the context of energy production, food production, and ecosystem services.

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 research challenge(s).

Addressing the two central research challenges in ecology for the coming century, which involve quantifying species' responses to climate change and the role of ecological systems in global biogeochemical cycling, increasingly requires the use of spatially-explicit databases. These databases have a number of technical challenges related to their creation, curation, and discoverability. In particular, it is often time-consuming to connect disparate sources of information to support individual research questions. This process could be accelerated with the use of centralized (and curated) services to make individual datasets discoverable and aggregatable. Although US federal agencies have created services for their own data (e.g. The Water Quality Portal, https://www.waterqualitydata.us/), few analogous services exist for university researchers and citizen science groups. This need is especially urgent for groups producing geospatial data which require minimum (yet flexible) metadata standards regarding data collection and quality assurance protocols.
As more and more research data is being made publically available, ensuring that data remains discoverable and has stable access is becoming more challenging. Link-rot and version misidentification are very real issues. One model that NSF should consider that would address these issues is the "mirroring" strategy of redundant distributed backups employed by the open source community. This model has been used with great success in efforts such as the Comprehensive R Archive Network (http://cran.r-project.org). One recent innovation in data mirroring that could aid this effort is the Dat project (https://datproject.org/) which has greatly reduced the barrier to entry for people and organizations to share and distribute their research data.
Geospatial data sharing infrastructure for research products could be strengthened by encouraging the use of modern, open storage format standards over older, proprietary formats still commonly used for sharing these data (e.g. shapefiles). Incentives to not only publish data in these formats, but also to make the datasets easily discoverable via standardized metadata should be matched with networking of data repositories such that data consumers can metasearch across data repositories, finding more results relevant to their topic regardless or whether the data originated with a government, university, individual, etc. One example of an ongoing effort in this area that deserves

continued NSF support is the DataOne project (https://www.dataone.org/).

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.

Tackling the major research challenges in ecology requires a workforce of researchers empowered with the tools and training required to maintain and analyze large datasets. Although the dominant tool for data analysis in the field of ecology is the R statistical program (https://www.r-project.org/), it can be difficult to use for people unfamiliar with programming because R syntax must be learned simultaneously with programming skills. One approach for lowering the barrier-to-entry is to increase funding for technical training. Two initiatives that have been very successful in cost-effective technical training of ecologists are the Data carpentry (http://www.datacarpentry.org/) and Software carpentry (https://software-carpentry.org/) projects. A second approach for lowering the barrier-of-entry is to further develop (increase the user-friendliness of) open-source tools for data analysis. NSF should support software development activities to this end by funding the creation and maintenance of R packages, data entry platforms, and visualization tools such as GIS software (http://www.osgeo.org/). NSF should promote the use of open source software because it removes cost barriers for disadvantaged groups and can be openly modified by the research community to fit their own emerging needs.
One cross-cutting theme that touches on nearly all aspects of data-driven research is collaborative software development. The increasing extent to which ecology is using large databases and quantitative models to address important research questions means that there is an increasing need to support scientists whose primary output is not peer-reviewed journal articles but rather peer-reviewed software and data packages. One term that is being used to describe this new type of researcher is Research Software Engineer. In the UK, these researchers are represented by the Software Sustainability Institute (https://www.software.ac.uk/) but no equivalent organization exists for US based researchers. NSF should support this group of researchers by making allowances for their positions in grant applications as well as recognizing software development as a primary research output and accepting it as such in grant applications. One model for crediting research software engineers is to seriously consider alt-metrics such as Depsy (http://depsy.org/) scores and Impactstory (http://impactstory.org) achievements as analogous to journal articles and citation metrics.