FORUM

Science and Cyberinfrastructure: The Chicken and Egg Problem

PAGES 458–459

In September, I participated in a general scientific discussion regarding the U.S. National Science Foundation Directorate for Geosciences (NSF GEO) Priorities and Frontiers 2015–2020 document. One of the key issues raised in conjunction with this document was the issue of science versus infrastructure. Although there was overwhelming agreement on the need for infrastructure to do our science, there was much concern about the corresponding balance of investment.

The argument one way is that if we invest in large infrastructure, the selective funding process then favors the kinds of science that make use of the new capabilities, with the eventuality that the infrastructure could end up driving the science and not the other way around. Thus, the argument continues, investments in the infrastructure should be the minimum required to accomplish a specific scientific goal. The argument the other way is that large infrastructure creates new scientific possibilities that could not have been imagined before such infrastructure existed.

Both of these viewpoints are valid, of course, and the right balance is hard to strike. It is nevertheless always useful to define the minimum infrastructure that would serve existing scientific needs. Here I try to explore this issue in the context of EarthCube.

NSF's EarthCube cyberinfrastructure initiative describes itself as "community driven" and "community governed." This model assumes that the Earth science community makes informed choices. However, precisely because of its fluidity and nonprescriptive foundations, EarthCube has not generated as much enthusiastic engagement as it should have. Many end users are lost in the talk of agile architecture, middleware, and ontologies.

The biggest concern is that EarthCube is shaping into a toy of information technology that will be so general that it is no longer useful for any real science. As one end user put it, "They are making it too complicated! They should just *simply* make a geo-referenced database and make it easy for us to upload data." Of course, what this end user does not fully appreciate is that simply making a database will not create a sustainable multidisciplinary solution to our problems.

What are these problems, and how much of an infrastructure do we really need to address them? Those seem to be the first questions to ask ourselves as we are working on formulating the goals and scope of EarthCube.

What Should EarthCube Be?

If someone asked me to describe the minimum extent of the technology that we are lacking, I would say that we need a visual, georeferenced, semantically enabled repository for scientific software and data. By this I mean that every scientific component would be tagged with its time and spatial location, if relevant, and with sufficient metadata to point to its scientific discipline and interdisciplinary purpose in words that nonspecialists could relate to.

We would want to have integrated cloud and remote high-performance computing (HPC) capabilities and workflow management tools. We would also want it to function similar to a successful social network, with capabilities for streamlined workflow sharing and creative reuse, usage tracking, and open discussion. Now, why do I want all that? Wouldn't a simple database do just as well? Am I guided by real scientific needs?

Answering these questions requires a close look at the factors that hinder our scientific discovery.

Proposed EarthCube Goals and Scope

Modern geosciences have several rate-limiting factors, the realization of which necessitated the existence of the EarthCube initiative in the first place and got some enthusiastic followers (like me!) to jump onboard.

I suggest that if EarthCube positions itself as an initiative focused on addressing specific key, but very broadly posed, technical bottlenecks of scientific exploration, it will be well poised both for community outreach and for setting the science versus technology balance just right. Indeed, progress in alleviating these bottlenecks could also serve as the metric of success. Every EarthCube building block could be evaluated on the basis of (1) how it helps address one or more of the bottlenecks and (2) how it fits in with the other components. This hands-on approach would prevent the infrastructure from becoming "overweight" while keeping it science governed.

Below are four bottlenecks with four plans to address them that EarthCube could adopt.

Problem 1: Scientific Reproducibility and Scientific Benchmarking Capabilities

Poor workflow recording makes it exceedingly tricky to share the intricacies of our work with others for independent verification or to create an improved version of an old result. Instead, all workflows that resulted in a published conclusion need to be readily available for critical review, replay, and creative modification. Visualization tools need to be treated as optional end-member components in the complete data interpretation workflow chain.

Thus, the peer-reviewed manuscript, although of utmost importance, needs to be placed in its due context. Ensuring workflow modularity (such that upon the cloning of a workflow chain to a user's personal workspace, any data or software component could be replaced by a functional alternative) would automatically enable performance metrics benchmarking, and cross-validation of numerical codes.

Problem 2: Interdisciplinary Communication

Currently, sharing digital information across disciplines is complex to the point of being prohibitive (because of diverse data formats, the lack of appropriate metadata, and domain-specific terminology). However, only an integrated multidisciplinary approach can help us address and communicate Earth system issues as complex as the energy crisis, sustainable water resources, earthquake prediction and damage mitigation, and climate change.

Multidisciplinary research could be greatly facilitated by establishing the technology for sharing data and workflows within and across disciplines, as well as for discovery of and access to information across disciplinary boundaries. The language barriers necessitate a controlled vocabulary approach and a semantic web. International, interagency, and academic to industry communications are additional significant bottlenecks that could be only partially addressed with technology; policy changes are needed.

Problem 3: Integration of Data and Models

Everyone would benefit from a tighter integration of data and models. Numerical models allow controlled hypothesis testing. Modeling results ("exact" and "continuous") could be used for data exploration and to inform further data collection. The data (inexact and sparse) are, in turn, invaluable for model validation and calibration.

An intuitive four-dimensional virtual globe application programming interface (API) could facilitate this convergence. However, the front-end capacity of this tool needs to be strongly emphasized. To achieve sustainability, interdisciplinary data discovery and access components need to be developed and maintained as stand-alone capabilities. Big data storage and transfer solutions would constitute other critical stand-alone components.

On the basis of the type of selected data resources and the user's preferences, corresponding scientific tools capable of visualization, analysis (including uncertainty quantification), and modeling of the data would be evoked. These tools could utilize either local or remote HPC resources to enable the "modeling for all" paradigm.

Problem 4: Data and Software Management

The data management plan requirement, introduced by NSF in 2011 (see http://www.nsf.gov/eng/general/dmp.jsp), is an important step toward scientific reproducibility but is lacking in several respects. First, very few disciplines have standard and easy to follow procedures and/or depositories for such data submissions. Help with metadata management and a standard archival strategy are needed. Second, the requirement is currently a hurdle for scientists, taking their time and giving nothing back, at least not directly.

To address this, EarthCube should aim to provide an incentive by being useful early on at the data interpretation stage. Further, a shift in credit attribution practices is needed, which would be facilitated through citable workflows in a social network–based environment.

The Role of Use Cases

How do we know that our problems have been adequately addressed by the cyberinfrastructure? This is where the concept of "use cases" comes in. According to the Unified Modeling Language definition, "a use case shows the interaction between the system and 'actors,' which may be human users or other systems."

In EarthCube, use cases will be based on real scientific scenarios and will be used to capture the requirements and test drive the cyberinfrastructure. Which use cases to focus on and the role of use cases in the development of EarthCube have been much debated. I believe that the ultimate choice and testing of use cases is better left to the domain experts, perhaps as a funded activity. This inclusive strategy would allow scientists from a variety of research communities to get closely acquainted with EarthCube's functionalities and influence its development from the onset. Such a strategy would also avoid alienating those whom we intend to serve.

As part of the governance plans of EarthCube, the Science Standing Committee and the Architecture and Technology Standing Committee have now been formed. Perhaps the Science Committee could crystallize some important use cases out of the final reports developed by scientists who attended EarthCube's many End User Workshops. The committee could group these use cases into sets around common technological requirements. It could proceed to form and lead interdisciplinary working groups—official ones recognized by EarthCube's charter—for each use case set. These groups would act as bridges between the technology development and the wider scientific community. To ensure exposure, representatives from every relevant scientific domain could be recruited. Such an approach would allow the use cases to come directly from end user scientists, to be tested in close collaboration with them, and to form part of our outreach early on.

The technological requirements communicated by the interdisciplinary working groups assigned to each use case set could then, together, be used to refine the goals and scope of EarthCube beyond what is proposed here, in close alignment with EarthCube's image of a community-driven and community-governed, dynamic and strategic effort.

Timely and Relevant Outcomes

In evaluating the system, our end users will not necessarily be concerned with its inner workings; they will be primarily concerned with whether EarthCube is intuitive, functional, and bug free. Some components of this functionality, such as a user-friendly interface, have not even been funded yet.

Once we are ready to test the system in the wider scientific community, EarthCube educators could be funded to assist end users with custom use case evaluations. Until then, we might do best to focus not on the use cases but on aligning our goals and scope with science-guided technical hurdles, as proposed above, and providing the technology to overcome them.

Acknowledgment

A.K. is funded by NSF EarthCube grant ICER-1343811.

—Anna Kelbert, College of Earth, Ocean and Atmospheric Sciences, Oregon State University, Corvallis; email: anna@ceoas.oregonstate.edu;