Remaking the Scientific Data Enterprise

*Strategic priorities for realizing the potential of scientific data into societal benefits and US competitive advantage*

Science underpins many important areas of national concern, including our infrastructure, jobs, economy, education, health, and security. Increasingly, the basis for much of science is data. Scientific data provide value beyond the science for which they are collected and translate into opportunities for society by providing a credible basis for decision making, creating opportunities for businesses, and enabling new insights and discoveries. Scientific data are a major contributor to a thriving economy, yet we have barely begun to unleash the potential power of our nation’s data assets.

We live in a world rich with data. But due to insufficient planning, management, and resources, the potential benefits of all these data are not realized. Scientific data are lost, inaccessible, unreadable, too big to handle, undocumented, and more. Currently our scientific data enterprise is evolving and maturing in an unmanaged fashion. The problem will grow as data volumes and sources increase.

At the same time we are facing pressing national challenges, including outdated and failing infrastructure, adaptation to new technologies, shifting world economic landscapes, health care imperatives and the impact of climate change on the environment.

In order to be well positioned to manage these challenges, our nation needs an overarching, unifying strategy for managing scientific data across domains and throughout the data lifecycle. We need to envision, predict, invest, and develop capabilities to build a modern, competitive scientific data infrastructure in order to capture the potential of our data.

**The Need to Act Now**

Vast new troves of information are continually becoming available as new sensors are deployed, networks are built, and computational tools proliferate. Earth scientists, data scientists, business leaders, and the U.S. government have made important strides toward tackling problems on three fronts: the emergence of big data, changing computational paradigms, and sociological changes in the practice of science. These efforts are necessary, but they are insufficient. The US has been a leader in the promotion of open data but has offered limited strategic guidance to stakeholders that effectively and efficiently channels the potential of all that data into societal benefits and U.S. competitive advantage.

We are seeing the impacts of climate change in forms of drought, heat waves, flooding, freezing temperatures, super storms and other dislocating events. Insurance companies are including climate change predictions in their forecasting [Smithsonian]. Coastal areas are expected to experience sea level rises that will impact lives and development [NOAA] [NATL GEO]. Data are essential in order to create accurate assessments and predictions so that policy makers can make informed choices in planning and response.

Our national energy, defense, and security policies too depend on scientific data. For examples of ways that scientific data can impact these areas consider the broad list of topics that the scientific advisory group JASON has studied [JASON reports]. The Department of Energy is interested in enhancing recovery from geothermal systems [JASON 2013]. The Department of Defense asked about the impacts of DNA sequencing over the next decade [JASON 2010]. And, the Department of Homeland Security cares about space weather on the electric grid [JASON 2011] and the conflict between wind farms and radar [JASON 2008]. The basis of these questions is scientific data.

The potential of data is being recognized in other domains, such as business and society. Other countries, such as China and India, are making advances in their scientific and technical expertise and offering related services. The underlying data is a large part of that process. [MURRAY ]

Innovation in the digital world could encourage job growth and stimulate the economy. A recent Brookings Institute workshop sought ideas for developing “new solutions to our continuing economic and political woes,... [including] the identification of promising reform ideas and ways to encourage growth through innovation.” They report:

*“We need smarter policies in order to take full advantage of the digital economy and strengthen our capacity to build society, generate jobs, and improve long-term economic growth. This focus should be front and center for policy makers as they wrestle with social and economic challenges.” “Our overriding theme … is how to move from ideas, norms, structures, and regimes developed during an industrial period to institutions and policies for the digital world”.* [IBE 2012]

**The Consequences of Not Acting**

In deciding whether to act, one also needs to consider the consequences of not doing so. One consequence of inaction around the scientific data enterprise is the unrealized ‘generative value’ of the data, that is, the “capacity to produce unanticipated change through unfiltered contributions from broad and varied audiences” [Wilbanks, 2010]. In a study on “The Generative Mechanisms of Open Government Data”, the authors address the the ability of Open Government Data (OGD) to generate both economic and social value. They found that

*“... opening government data is not in itself sufficient for value generation. A number of barriers have to be overcome in order to enable the mechanisms that allow for value generation. Accordingly, we propose that the key enablers for OGD value generation are as follows: open access to data, data governance procedures and technical connectivity.”*  [ECIS 2013]

The economic value of scientific data raises the issue of return on investment. Experts found that the return on investment of the British Atmospheric Data Center is on the order of 400 - 1200% [Ashley 2014]. It is possible that in many cases the data produced by and for science will pay for itself many times over, if properly managed.

Inaction contributes to lack of access. Many data sets are hidden or off line, their existence known only to a few. Or they are in a format that renders them unusable to most, or are undocumented. Wilbanks has stated that in the medical field, “information is available only to a small set of people and they can pervert the process. We need data autonomy and portability”. [IBE 2012].

Inaction also contributes to inefficiency. Among numerous other examples of inefficiency, Wilbanks and Rossini observed that, “Lack of availability of data promotes “factionalization where we should be seeking efficiencies of scale, and centralization where we should be promoting diversity” [Wilbanks 2009]

Many studies have been conducted by agencies, foundations, and other organizations around better capturing the value of data. (Appendix X contains a partial list.) Presumably some of those recommendations have been implemented, and yet we still find that the challenges and threats to our scientific data enterprise are growing faster than our solutions.

**Call to Action**

These challenges are potential opportunities to achieve progress in science, innovation, the economy, and broader society. To actually capture the value of our data, the Federation of Earth Science Information Partners (ESIP Federation or ESIP) calls upon the National Research Council (NRC) to conduct a study to determine strategic priorities for the scientific data enterprise. NRC surveys are considered the gold standard for advice on research programming [DSSS 2007] and offer an authoritative and unbiased assessment for strategic scientific investments. This study would inform and guide decision makers in the government, academia, and industry in helping to improve their practices and priorities for managing scientific data, giving the U.S. a boost in all impacted arenas.

This study should:

* Synthesize and analyze prior work in data management/infrastructure, such as, what was successful, what was not successful, and why has this not been sufficient?
* Take a broad perspective of the value of the scientific data enterprise and the infrastructure that supports it from the perspectives of societal benefit, economic competitiveness, and other important values
* Provide a vision of what might be, then prioritize with conclusions and recommendations.

**References**

[ASHLEY 2014] Kevin Ashley’s presentation to the ESIP Winter 2014 meeting.

[DSSS 2007] Decadal Science Strategy Surveys: Report of a Workshop, <http://www.nap.edu/catalog/11894.html>

[MURRAY ] The Globalization of Science, <http://videolectures.net/mitworld_murray_globalscience/>

[IBE 2012] Building an Innovation-Based Economy [on my disk].

[JASON reports] http://www.fas.org/irp/agency/dod/jason/

[NATL GEO] <http://www.google.com/url?q=http%3A%2F%2Focean.nationalgeographic.com%2Focean%2Fcritical-issues-sea-level-rise%2F&sa=D&sntz=1&usg=AFQjCNEZKql4koDVooWniUC8ifj3rgzfTQ>

[NOAA] <http://www.google.com/url?q=http%3A%2F%2Foceanservice.noaa.gov%2Ffacts%2Fsealevel.html&sa=D&sntz=1&usg=AFQjCNFNbkbxq-6ypSXS-9_bQK8KIsN3wg>

[Wilbanks 2009] Wilbanks, Rossini, “An Interoperability Principle for Knowledge Creation and Governance: The Role of Emerging Institutions”, MINDS conference on Strategic Responses to Globalization, Nov 3 - Nov 9, 2009.