FORUM

Realizing the Value of a National Asset: Scientific Data

PAGES 477–478

"We have a shared responsibility to create and implement strategies to realize the full potential of digital information for present and future generations," according to the Electronic Geophysical Year (*e*GY) Declaration [*CoBabe-Ammann et al.*, 2007]

Seven years have passed since the *e*GY declaration, and, despite national mandates and agency-wide policies on data sharing, we remain mired in the challenges of dealing with scientific data. A coordinated effort at the highest level is needed to allow the Earth sciences to fully capitalize on the unfolding data revolution and to best serve society. This Forum outlines the challenges and opportunities related to Earth science data and proposes a path forward: creating a National Research Council study that will provide high-level strategic guidance to effectively and efficiently address the grand data challenges.

Science Data Infrastructure: Challenges and Opportunities

Data, the core of any scientific endeavor, are valuable national assets. Over the past 2 decades, Earth scientists, data scientists, business leaders, and the U.S. government have made important strides in tackling scientific data problems by harnessing big data, changing computational paradigms, and instituting sociologic changes in the practice of science data management.

Harnessing Big Data for Societal Benefits

"Pasteur's quadrant" describes the class of scientific research motivated by both societal and scientific needs. Earth science research falls into this category because it has implications for applications that can yield profound societal benefit. Sensors, satellites, drones, and other devices are cheaper and more ubiquitous than ever before. As the world is increasingly monitored, enormous data collections are accumulating in government, academic, and industry databases. NASA's Global Change Master Directory alone describes more than 31,000 Earth science data sets and services from NASA and other agency and international sources. Some of these data sets could be valuable for research far beyond their original point of collection and for purposes other than their original intent. However, the difficulty of finding and using data properly is a barrier to realizing these benefits.

Maximizing the usefulness of these collections presents challenges throughout the entire data life cycle, including planning, collection, storage, documentation, maintenance, and preservation. Focusing on data from the beginning could yield significant payoffs in important areas such as security of energy, water, and food. The new decadal survey missions from NASA are a positive step. As part of mission planning, NASA held several applications workshops informing the applications community of the upcoming capabilities and preparing users for data from the missions.

Capitalizing on Computational Advances

We are making progress in technologies for finding, understanding, mining, integrating, analyzing, and sharing data. Our computational and storage capacity has exploded as processors become increasingly powerful and innovative new software is deployed. However, the impact is limited because of interface mismatches in data formats, time representation, terminologies, and vocabularies, absence of metadata standards and practices, etc. Establishing and infusing the right combination of tools, standards, and best practices in a coherent way into diverse disciplines, including the various science domains, as well as computer, library, information, and social sciences, could mitigate much of this impedance. For example, the establishment of domain ontologies could greatly improve cross-domain data discovery and understanding.

One compelling vision is that of an "executable publication," where readers can follow links in a scientific publication to acquire primary data and execute code to verify research results [*Giordani*, 2013]. Sharing data and metadata as "nanopublications," components of the scientific process, would enable this vision while providing a mechanism for attributing credit to data creators or editors.

Changing the Practice of Science

Scientists work in increasingly fluid funding environments on projects that span traditional discipline and organizational boundaries. Solitary science is giving way to large-scale collaborations seeking breakthroughs not attainable by individual or simply additive efforts. Free and open data, open source software, open access publishing, citizen science, and crowdsourcing provide new entry paths to science. Transparency, reproducibility, and accountability have become highly important as scientific claims are made and challenged. Recent directives aimed at "opening up" many government data sources reinforce this trend.

One fundamental barrier is the lack of an economic model that sustains data-related activities, resulting in tension between performing science and creating science data infrastructure. At the 2013 AGU Fall Meeting, NASA Chief Scientist Ellen Stofan illustrated this point, saying that NASA was expected to do more with less. Data center managers across agencies have anecdotally expressed the same perception.

Important Efforts to Date

Many organizations, initiatives, and advisory groups have made important strides toward solving our data challenges, including the following:

- The Blue Ribbon Task Force on Sustainable Digital Preservation and Access
- Cooperation EU-US (COOPEUS)
- Data.gov
- EarthCube
- Federation of Earth Science Information Partners (ESIP)
- National Research Council (NRC)
- NASA's Earth Science Data System Working Groups
- NOAA's Environmental Data Management Committee
- Research Data Alliance
- Sustainable Digital Data Preservation and Access Network Partner (DataNet) and its funded projects, such as DataONE and Terra Populus
- National Consortium for Data Science (NCDS)

However, these necessary efforts are insufficient. Data management and stewardship problems continue to be addressed piecemeal. Each organization responds to its own needs with its own data standards and policies, leading to an unorganized duplication of effort while not comprehensively addressing economic, cultural, transdisciplinary, and cross-sector issues.

Charting a Path Forward

Members of ESIP and representatives from NRC have met regularly since January 2013 (http://commons.esipfed.org/node/695) to deliberate on a high-level study to provide the unifying vision needed to coherently address our grand data challenges. A plenary discussion at the summer 2013 ESIP meeting (http://commons.esipfed.org/node/1536) and a workshop at the January 2014 ESIP meeting [*Wilson et al.*, 2014] brought these issues into focus as panelists and participants considered a possible NRC study on data developments, practices, and economics in the Earth sciences. The workshop concluded that an NRC-led study could set research priorities for scientific data management, including sustainable economic models for scientific data infrastructure, helping the United States maintain its position as a global scientific leader.

NRC is the logical coordinator to develop a unified vision for transforming our data challenges into scientific opportunities. As the operating arm of the National Academy of Sciences, NRC can ensure that the concerns and needs of all stakeholders—including the private sector, academic researchers, government agencies, and policy makers—are heard and integrated into the overarching vision. With its longstanding role as the central forum and voice of the scientific community, NRC is uniquely capable of drawing upon the top echelon of scientific leaders to guide visionary science and chart a path forward through targeted research investments, cultural changes, and strategic coalitions. NRC has a strong track record of informing priorities in the federal agencies, executive branch leadership, and Congress.

To fully realize the value of data, it will take all of us. In your own research, we'd encourage you to consider the importance of science data infrastructure and make data management and preservation a part of your workflow. The ESIP Data Study Working Group is open to any interested Earth scientist (http://wiki.esipfed.org/index.php/Data_Study_Working_Group). At the upcoming ESIP winter meeting, we will continue to work with NRC on the next steps for this report (http://commons.esipfed.org/taxonomy/term/1482). We look forward to working together to solve these problems and more effectively channel our resources into a new wave of discovery and innovation.

Acknowledgments

We acknowledge ESIP July 2013 Workshop panelists Dan Baker (Laboratory for Atmospheric and Space Physics), Stan Ahalt (Renaissance Computing Institute), Todd Vision (University of North Carolina, NESCent), and Michael Tiemann (RedHat); the ESIP Data Study Working Group; Anne Johnson; and the *Eos* editors and reviewers. The January 2014 workshop was supported by the Gordon and Betty Moore Foundation, NCDS, and the ESIP Federation. Additional support was provided by NASA contract NNG13HQ04C (R. R. D.), the Laboratory for Atmospheric and Space Physics (A. W.), and National Science Foundation grants ACI-0830944 and IIA-1301346 (W. M.). H. R. worked on this paper as a U.S. government employee. Any opinions, conclusions, or recommendations expressed in this material are those of the authors and are not necessarily the views of their employers or funders.

References

CoBabe-Ammann, E., W. K. Peterson, D. Baker, P. Fox, and C. Barton (2007), The Electronic Geophysical Year (2007–2008): eScience for the 21st century, *Geophysics*, *26*, 1294–1295.

Giordani, A. (2013), Scientific publishing 2.0: Moving the computer to the data rather than moving the data to the computers, *soapboxscience* (blog), 6 Feb., http://blogs.nature.com.

Wilson, A., E. Robinson, W. Lenhardt, R. Downs, and H. Ramapriyan (2014), Workshop report: Planning for a community study of scientific data infrastructure, Fed. of Earth Sci. Info. Partners, Raleigh, N. C. [Available at http://dx.doi.org/10.7269/P3R49NQZ.]

—Anne Wilson, Laboratory for Atmospheric and Space Physics, University of Colorado, Boulder; email: anne.wilson@lasp.colorado.edu; Robert R. Downs, Center for International Earth Science Information Network, Columbia University, Palisades, N.Y.; W. Christopher Lenhardt, Renaissance Computing Institute, University of North Carolina at Chapel Hill; Carol Meyer, Foundation for Earth Science, Raleigh, N.C.; William Michener, University of New Mexico, Albuquerque; Hampapuram Ramapriyan, NASA Goddard Space Flight Center, Greenbelt, Md.; and Erin Robinson, Foundation for Earth Science, Boulder, Colo.