Charting a Course through a Rapidly Changing Scientific Paradigm

Anne Wilson, Robert R. Downs, W. Christopher Lenhardt, Carol Meyer, William Michener, Hampapuram Ramapriyan, Erin Robinson

Data—the core of any scientific endeavor—is at once a great asset and an impediment to scientific progress. Vast new troves of information are becoming available all the time as new sensors are deployed, networks are built, and computational tools proliferate. But we need high-level strategic guidance to effectively and efficiently channel the potential of all that data into real benefits for science and society.

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 sociologic changes in the practice of science. These efforts are necessary, but they are not sufficient. Members of the Federation of Earth Science Information Partners (ESIP) and representatives from the National Research Council (NRC) have met regularly since January 2013 to discuss the possibility of a high-level study to provide the unifying vision we need to coherently and effectively address our data grand challenges. A plenary discussion at the Summer 2013 ESIP meeting brought these issues into focus as panelists considered the need and feasibility of establishing an NRC study on data developments, management, and stewardship in the Earth sciences realm. This paper builds on the ongoing ESIP/NRC conversation to outline the data opportunities we face, ongoing efforts in this realm, and a potential path forward.

**The Vision: Riding the Data Tide**

As the Earth science community is well aware, the data sets available today are vastly larger, more numerous, and more complex than ever before. Recent directives aimed at “opening up” many government data sources will likely reinforce this trend (Obama 2013; Holdren 2013). It is our shared responsibility to maximize the return on collective investments such as data sets and models produced by government agencies and the scientific community. Optimizing the use of data is no simple feat, however. The exponential growth of data is akin to a giant tide; as we rise to the tide’s crest, we can either channel its energy into major scientific advances or drown in the flood.

*Harnessing Big Data*

Enormous data collections are accumulating in government, academic, and industry databases. For example, NASA's Global Change Master Directory alone describes over 29,000 Earth science data sets and services from NASA as well as other agency and international sources. As sensors, satellites, drones, and other devices become cheaper and more ubiquitous, data pours in from all corners of the world. Some of these data sets are highly likely to be used far beyond their original point of collection and for purposes other than their original intent. Maximizing the usefulness of these collections presents challenges throughout the entire data life cycle, including planning, collection, storage, documentation, maintenance, and preservation.

*Capitalizing on Computational Advances*

The past decade has seen remarkable progress in the tools available for mining, integrating, analyzing, and sharing data. Our computational capacity has exploded as processors become increasingly powerful and innovative new software is deployed. But as new tools emerge, interface mismatches proliferate and the lack of interoperability among systems and software becomes more troublesome. In addition to the challenges of dealing with multiple data formats, we face fundamental challenges in data representation. Such issues must be addressed from a cross-sector and interdisciplinary perspective.

*Changing the Practice of Science*

The evolving technological landscape is changing the way science is done. Scientists work in increasingly fluid funding and employment structures and must constantly navigate across disciplines and organizations. At the same time, science has become more “open.” Trends toward freely-available data, open-source software, open-access publishing, citizen science, and crowdsourcing broaden the scientific community and allow more people access to information and analytical tools. As a result, reproducibility has become more important; as more scientific claims are made and challenged, there is a growing need for transparency and accountability.

One compelling vision for the future is that of an “executable publication,” which would allow readers to follow links in a scientific publication to acquire primary data and execute code to verify research results (Giordani 2013). Building on this idea, sharing data as a “nanopublication” in the middle of the scientific process would provide a mechanism for giving appropriate credit to data creators or collectors while allowing multiple researchers to work in parallel to generate insights and discoveries from shared data; the Global Biodiversity Information Facility offers a model of this approach.

**Important Efforts to Date**

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

* The Blue Ribbon Task Force on Sustainable Digital Preservation and Access
* Data.gov
* EarthCube
* National Research Council
* 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

All of these efforts are promising and necessary, but they lack an overarching strategy. Data management and stewardship problems continue to be solved piecemeal, with each organization responding to its own needs with its own data standards and policies. This leads to duplication of effort and makes it difficult to comprehensively tackle trans-disciplinary and cross-sector issues. Strategic guidance is needed to optimize the use of our time, funding, and creative energy.

**Charting a Path Forward**

“We have a shared responsibility to create and implement strategies to realize the full potential of digital information for present and future generations.” –*e*GY Declaration (CoBabe-Ammann, et al. 2007)

Though six years have passed since the *e*GY declaration, we remain mired in the challenges of dealing with scientific data. The NRC is the logical coordinator of a much-needed effort to develop a unified vision for transforming our data challenges into scientific opportunities. As the operating arm of the National Academy of Sciences, the NRC is ideally positioned to ensure the concerns and needs of all stakeholders—from the private sector to academic researchers to 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, the NRC is uniquely capable of drawing upon the top echelon of scientific leaders to guide visionary science and chart a path to achieving it through targeted research investments, cultural changes, and strategic coalitions. The NRC also has a strong track record of informing priorities in the federal agencies, executive branch leadership, and Congress.

An NRC-led study would identify major gaps in data management knowledge and practices, set research priorities for scientific data management and stewardship, and offer a consensus view of opportunities to retain the role of the United States as a global scientific leader. A workshop planned for the Winter 2014 ESIP meeting will further refine the vision and goals for such a study.

A coordinated effort at the highest level is needed to allow the Earth sciences to fully capitalize on the unfolding data revolution. We hope you will join this conversation by contributing your needs, challenges, and ideas to our ongoing dialogue. A site for community input is under development and will be available at esipfed.org in the near future. We look forward to working together to solve our common problems and channel our creative energy into a new wave of discovery and innovation.

**References**

CoBabe-Ammann, E., Peterson, W.K., Baker, D. Fox, P., & Barton, C. (2007). “The Electronic Geophysical Year (2007–2008): eScience for the 21st Century.” *Geophysics*, 26:1294-5.

Giordani, A. “Scientific publishing 2.0: moving the compute to the data rather than moving the data to the computers”, soapboxscience, Nature.com Blogs, Blog post February 6, 2013.

Holdren, J.P. “Memorandum for the Heads of Executive Departments and Agencies—Increasing Access to the Results of Federally Funded Scientific Research.” February 22, 2013.

Obama, B. “Executive Order—Making Open and Machine Readable the New Default for Government Information.” May 9, 2013.

**Acknowledgements**

We greatly appreciate the panelists of the ESIP July 2013 Workshop for their time, input and support: Dan Baker, LASP, Stan Ahalt, RENCI, Todd Vision, UNC and NESCent, Michael Tiemann, RedHat. We thank the ESIP Data Study Working Group for input, review, and ongoing support. We thank Anne Johnson, freelance writer, for her excellent editorial contributions. Finally, we thank the EOS editors and reviewers for helpful suggestions and comments.

Support for Robert Downs was provided under NASA Contract NNG13HQ04C for the Continued Operation of the Socioeconomic Data and Applications Center (SEDAC). Anne Wilson gratefully acknowledges the support provided by the Laboratory for Atmospheric and Space Physics (LASP). Ramapriyan worked on this paper as a part of his official duties as a U.S. government employee. William Michener is supported by the U.S. National Science Foundation (Grant [#ACI-0830944](http://nsf.gov/awardsearch/showAward.do?AwardNumber=0830944) and #IIA-1301346). Any opinions, findings, and conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of their employers or funders.

**Authors**

Anne  Wilson, Laboratory for Atmospheric and Space Physics (LASP), University of Colorado, Boulder, Colorado.

Robert R. Downs, Center for International Earth Science Information Network (CIESIN), Columbia University, Palisades, New York.

W. Christopher Lenhardt, Renaissance Computing Institute (RENCI), UNC Chapel Hill, NC.

Carol Meyer, Foundation for Earth Science, Raleigh, North Carolina.

William Michener, The University of New Mexico, Albuquerque, NM.

Hampapuram Ramapriyan, NASA Goddard Space Flight Center, Greenbelt MD.

Erin Robinson, Foundation for Earth Science, Raleigh, NC.