Workshop Report: Planning for a Community Study of Scientific Data Infrastructure

Executive Summary

Our world is rich with scientific data. These data hold great potential for progress in science, innovation, the economy, and broader society.

And yet, much of that potential lies unrealized due in large part to insufficient planning, management and resources. Scientific data are lost, inaccessible, unreadable, too big to handle, undocumented, and more. Our national scientific data enterprise is evolving and maturing in an uncoordinated fashion. As data volumes increase exponentially—along with complexities and sources—the problem continues to grow.

Our treasure trove of scientific data needs and deserves elevated status and priority to allow for its greater potential to bloom. Our scientific data is a national resource that must be fully exploitable in order to steer policy and research decisions that support the utmost levels of social welfare, progress and innovation. An overarching, unifying strategy is urged, which collaboratively addresses the management of scientific data across domains and throughout the data lifecycle. To capture the fullest potential of our data, a modern, competitive scientific data infrastructure is necessary. Key stakeholders must envision, predict, invest and develop capabilities to achieve this aim.

The ESIP Data Study Working Group supports the undertaking of a National Research Council study on science data infrastructure. This report summarizes recommendations from a 2014 workshop exploring this avenue of inquiry.

Workshop participants envisioned a sustained Science Data Infrastructure (SDI) and associated technical and cultural shifts to better enable science in the face of major challenges now and into the future. To achieve this and to capture the fullest value of data investment, a study is needed to investigate the costs and benefits of providing a sustainable infrastructure for the long-term management and stewardship of scientific data. Workshop participants identified important aspects of scientific data in which study and guidance is needed, including: the economics of scientific data; provision of sustainable infrastructure for the long-term management and stewardship of scientific data; cultural changes needed to realize value; research in relevant domains such as computer, information and library, and data sciences; improved education in scientific data management and stewardship; and creating policy that achieves these goals in a sustained manner.

Further, the National Research Council (NRC) is proposed as the logical entity to oversee such an effort in order to ensure an authoritative and unbiased assessment of requisite, sustained investments in science data infrastructure. This assessment would inform and guide decision makers in the government, academia, and industry in helping to improve their practices and priorities for providing sustainable infrastructure for scientific data, giving the U.S. a boost in all impacted arenas.

Workshop Overview

Experts from various U.S. academic and research institutions convened on January 7, 2014 in Washington, D.C. to participate in a workshop on *Planning for a Community Study of Scientific Data Infrastructure*. The workshop was sponsored by the Foundation for Earth Science, on behalf of the Federation for Earth Science Information Partners' Data Study Work Group, and facilitated by Dr. William Michener, University of New Mexico. See Appendix 1 for a list of workshop attendees.

The workshop format framed a community inquiry of challenges and opportunities associated with the U.S. scientific data infrastructure. The group's scope encompassed several elements:
- Define the primary emphases of an Academy study (domains, practice, priorities for research and funding, infrastructure)
- Identify some of the grand challenges in scientific data infrastructure
- Articulate why a study of these issues is needed now
- Define the stakeholders of the study

Workshop Proceedings

The one-day workshop was comprised of five sessions. Key questions were utilized to focus on identifying broad or common themes. See Appendix 2 for workshop agenda and Appendix 3 for notes of session discussions.

Within this report, the acronym *Science Data Infrastructure (SDI)* is deemed to mean a ubiquitous, reliable, and easy to use system for publishing, finding, understanding, using and accrediting scientific data. Also, the phrase *science data enterprise* used within this report refers to all relevant management and stewardship aspects of scientific data throughout the data lifecycle.

Session 1: Scientific Data Infrastructure Challenges

What are three of the "grand challenges in scientific data infrastructure" from either your personal perspective or that of a "stakeholder" of your choosing?

The group collectively identified challenges and then individually ranked them. Raw notes from this section are found in Appendix 3.

An analysis of Sessions 1 and 2 produced these broad categories of challenges and goals in somewhat overlapping areas: Economics, Cultural Values and Awareness, Data Science Research and Goals (including technological and societal/cultural challenges), Challenges in Education, and Legal, Ethical, and Policy Challenges.

The Economics of Scientific Data

The top economic challenges identified by the group are (1) developing an economic model for sustained infrastructure without competing with research dollars, and (2) commoditizing the SDI. Economic issues impact every aspect of the science data enterprise. Organizations are continually expected to manage more data using better practices and tools and with fewer resources. Funding for data management is minimal, irregular, and of limited time and scope. Groups are generally asked to pay for data management out of funds budgeted for research.

Interest in economic models, cost models, and return on investment of research data is growing. Although the economic value and impact of scientific data has not been adequately studied, efforts are underway to begin to measure impact and consider generative potential. Altmetrics—which attempts to measure the impact of academic literature from web downloads, views, storage, links, bookmarks, electronic conversation, and other novel 'traces'—is heralded as an example of new ways to measure value [Altmetrics]. [1]

Reliable, long term funding is needed, as well as financial incentives and rewards for safeguarding and stewarding data. Data management expenses must not be afterthoughts taken from scientific research funds. New business models for data use and reuse would benefit not just science but also serve national

---

[1] Subsequent to the workshop we learned that Jisc, a champion of the use of digital technologies in UK education and research, recently completed a study of the value and impact of three well established research UK data centers [Jisc 2014]. A goal of the study was to "contribute to the further development of impact evaluation methods that can provide estimates of the value and benefits of research data sharing and curation infrastructure investments." (The analysis found a significant return on investment and value has been realized.) We also learned that Jetzek, et al [Jetzek 2013], provide a strategic framework with archetypical generative mechanisms, and offer a conceptual model that provides a systematic way of articulating and examining the ability of Open Government Data (OGD) to generate both economic and social value.

security, sustainability, and growth. Innovative economic models might treat the SDI as a commodity similar to existing commodities, like highways or a utility.

Cultural Values and Awareness

Cultural values must change, and challenges are associated with moving toward change:
- New academic values must be set for data management and stewardship.
- Cultural incentive and reward structures do not reflect the importance of quality data management.
- Attractive, rewarding career paths must be established for system builders in science.
- Data stewardship must be integrated into ways that science is taught.

The current reward system does not recognize the value of creating a high quality or high impact data set, Similarly, good data management and stewardship practices are often not understood, and are generally barely supported or rewarded.

Varying cultures across disparate communities, organizations and domains raise challenges for coordinating data-related efforts. Today's societal needs require research at the intersection of what were once disparate domains. New research questions can be studied by integrating scientific data shared and used across disciplines. For example, disaster response and planning for climate change requires Earth science data as well as biological, GIS, population and sociological data.

Policy makers, academic institutions, scientists, and the public do not fully understand the importance of data infrastructure (or the lack thereof) to their work. Raising awareness of the significance and potential value of managing scientific data is advised, with equal consideration given to communicating the clear risks of not doing so. Stakeholders in the SDI must understand the value of scientific data so that resources will be allocated for its management and stewardship.

A cultural divide exists between domain scientists and data scientists who develop and work with systems that manage data. Domain scientists need to appreciate best practices for data management and stewardship, and data scientists need to better understand scientific needs and goals in designing and building tools and infrastructure.

Data Science Research, Goals

The deliberate management and stewardship of scientific data is not only a matter of unrealized potential but also a matter of defensive need, as we face a tsunami of data with no end in sight. How will we manage this sea of data so that scientific resources can be used effectively in the future and new scientific efforts can build on the legacy of past studies?

Additional data science challenges include:
- What information, best practices, and tools are needed along the stages of the data lifecycle?
  - How to measure the value and impact of a data set?
  - How to identify which data sets to throw away? Which to keep? For how long? At what cost? At what resolution? In what format?
  - What information about the data should be kept and for how long?
  - What attribution, provenance tracking, citation and curation tools and practices are needed and at what points in the data lifecycle?
- What new practices and tools are need to be developed? How can they be infused?
- What cultural and social changes are needed?
- How to facilitate future access to current data?

- Is there a set of essential SDI components or functionality that can be identified?
- Is there a fundamental common data model that could aid in data interoperability and fusion? Is there a mathematics of data?

Metadata is a significant challenge in the world of information. What metadata fields are needed? At what resolution? How can potential users discover data that meets their needs? "Magic metadata" is the term coined to describe not too little, not too much, but just enough of the right information. Achieving magic metadata depends on unintended uses of the data that arise and new potential uses occurring over time.

Data that is interoperable across domains, groups, and organizations and tools is a continuing technical challenge. What frameworks or mathematical solutions might help? How can high performance computing be leveraged? Having both too much and not enough data at the same time presents the paradoxical challenge of finding a needle in a haystack. Often cited is the 80/20 rule when working with data—scientists spend 80% of their resources getting data into a usable format and 20% doing actual science. Finally, networks appear to be used far less than optimally, and it is likely that major gains could be achieved by better network management practices.

Coordinating data practices across domains is essential to more fully realize the generative value of data. The identification of and infusion of those practices is another challenge. The nature of scientific review is changing, with traditional publishing experiencing increasing competition from 'crowd' and social networking forums. For example, a mechanism that allows unintended users of a data set to give feedback to data providers and tool developers (who would presumably have resources enabling them to respond) could increase the value of that data set.

Education Challenges
Developing current and next-generation workers is a critical need and must be held as a national priority. Data availability and integrity is a cornerstone for educators and a lynchpin in developing a highly functional workforce. eScience is changing the very nature of doing science. Accessibility of science data in our classrooms is paramount, both for K-12 and graduate Science Standards. Getting data to teachers in ways they can easily use in lessons is an increasing challenge for educators.

Legal, Ethical & Policy Challenges
Licensing, transparency and verifiability present challenges for those involved with managing data. The ethical use of data is also of utmost consideration in the scientific realm. Science progresses through opportunities for repeatability and verification. Costs, benefits and tradeoffs of good data management must be understood to make informed decisions about science infrastructure.

Science is evolving into more cross-disciplinary activities. However, current funding vehicles contribute to disparate, potentially duplicative, and possibly non-interoperable efforts around data. Vision and funding for a SDI on a national scale rather than an organizational scale is warranted. The contemplation of an entirely new entity, e.g., a National Science Infrastructure Foundation, is one possible avenue to accomplish sustained funding. At the very least, organizations that are part of the scientific enterprise must prioritize funding and related actions to better support data management and stewardship for sustainable science.

Session 2: Potential Outcomes of a Community Study
What are the two or three key recommendations or actions that you would like to see emerge from a study of the "grand challenges in scientific data infrastructure"?

A community study is deemed timely and warranted, with anticipated positive societal and scientific benefits. Potential key recommendations and actions that could emerge from a study are summarized as follows:

1. A Ubiquitous Science Data Infrastructure (SDI) is needed
2. Economic Goals
    a. Don't take funding away from science research for infrastructure
    b. Sustained funding
    c. Economic models, cost models, ROI
    d. Measures of progress/sharability/impact
    e. A way to evaluate the value of a data set , e.g. to decide to what extent to fund maintenance, such as altmetrics.org
    f. Financial incentives and rewards
3. Cultural Issues/Changes
    a. Improved cultural incentives, rewards
    b. Elevated awareness of importance of data
4. Research in Data Science, Data Science Goals
    a. Interoperability, possibly via a common data model
    b. Transparency
5. Improved Education
    a. Education: All ages (K - grave), all domains, training current researchers
    b. Elevated awareness of importance of data
6. Policy
    a. Establishment, movement of an agency, council, or office
7. General Concerns
    a. Need for sufficient specifics to be useful
    b. Dealing with commonalities vs domain specific needs
    c. Alignment of solutions for U.S. or global network

The content of the worksheets resulting from this session and their categorization is available in a spreadsheet format linked in Appendix 3.

Session 3: Identification of Community Stakeholders
Who are the stakeholders that should be invited to participate in a study of "grand challenges in scientific data infrastructure"?

Potential stakeholders span multiple dimensions, including data lifecycle, economic sector and organizational type. Categories identified within each dimension form a functional outline of affected key stakeholders. Raw notes from this session are linked in Appendix 3.

Data Lifecycle
- Consumers or end-users
- Long-term curators
- Data Providers
- Data Producers/Analyzers - Data Scientist/Informaticist
- Social Scientist
- Data Infrastructure

Economic Sector
- IT Industry (for and not-for profit)
- University Operations
- Education (K-20)
- Government Funders and Advisors
- Foundation Funders
- Science Policy

Adjacent Organizations
- Data-related Collaborative Initiatives
- Professional Societies and Publishers

Session 4: Engaging the Community in a National Study
How would your team design an effective study of "grand challenges in scientific data infrastructure" that could be carried out within a two-year period?

Community engagement would be key to producing useful, acceptable results.   Various data related community groups could be leveraged for community outreach, including ESIP itself, the AGU's Earth and Space Science Informatics focus group (ESSI), NASA's Earth Science Data Systems Working Group (ESDSWG) and the Boulder Earth and Space Science Informatics Group (BESSIG).

Various structures and strategies are possible. A PI at 1.5 months FTE plus a steering committee of ten to fifteen people could steer the effort, as well as identify experts in the field and engage the community. The study could be launched with a large physical meeting, followed by virtual meetings with occasional other face to face meetings. Mixed approaches are suggested, including surveys, case studies, focus groups, and brainstorming meetings. Working groups could be formed around topical areas, such as social impacts of data. Broad ideation through brainstorming is needed, followed by the filtering of ideas to identify worthwhile, actionable, efforts. There needs to be balance between policy, technical, and use issues.

Session 5: Wrap-up and the Road Ahead
Produce a high level design for a study and discuss plans for moving forward.

Federal funding for SDI is strongly advised. SDI deserves a national priority similar to the cabinet level status currently accorded to science in general. Business-as-usual meetings and workshops are not adequate due to the changing nature of the field. An innovative, new approach is needed and strongly advised. Also, the cost of doing nothing needs consideration.

Multiple studies around scientific and digital data have already been performed (Appendix 4 lists some), yet we still find ourselves in this predicament.   Where did these studies provide improvements? Where did they fail and why? How to minimize the chance a new NRC study would produce useless results?   Is the return on investment worth it?   Answers to these important questions depend on the scope and depth of the study to be undertaken, as well as having some idea of the potential value to be realized.

A two-year timeline for study completion is envisioned, followed by ongoing assessment and evaluation. The first task—with a 6-month estimated timeline for completion—will involve defining the scope of the study, identifying trends and performing a gap analysis of what is needed. Also necessary is a review and meta-analysis of prior studies to identify past successes and failures. Once the study is sanctioned, has support and is feasible, themes and chapters for the proposed study will be developed with a second workshop effort.

Call to Action

The Federation of Earth Science Information Partners (ESIP) calls upon the National Research Council (NRC) to conduct an authoritative, unbiased assessment of strategic scientific data investments. The assessment's purpose will serve to inform and guide government, academic and industry decision makers in improving practices and priorities for managing scientific data. The assessment's intended result will establish the United States as a leader in data strategies to inform key decision makers in all impacted arenas.

The assessment scope recommended by workshop participants includes:
- Synthesize and analyze prior work in science data management and infrastructure, such as, what was successful, what was not successful, and why past efforts have not been sufficient.
- Incorporate a broad perspective of the value of the scientific data enterprise and the infrastructure that supports it, inviting perspectives of social benefit, economic competitiveness, and other important values.
- Provide a vision of what might be, and prioritize with conclusions and recommendations.

Next Steps

Results from the *Workshop Report: Planning for a Community Study of Scientific Data Study* will be shared with the NRC Board on Research Data and Information and broadly disseminated amongst Earth and environmental science community members. Constituents of supporting disciplines are also enthusiastically invited to review results. Other complementary initiatives such as the US GEO Earth Observatory Assessment, NSF CIF21 initiatives, and NASA ESDSWG proposed work group for Vision 2020 are encouraged to participate in a shared agenda leading to discussions and refinements of workshop outcomes.

Acknowledgements

Appendix 1: Organizers and Participants

Stan Ahalt, RENCI, University of North Carolina, Chapel Hill
Lee Allison, Arizona Geological Survey
Karl Benedict, EDAC/Libraries, University of New Mexico
Robert Cook, Oak Ridge National Laboratory
Steve Diggs, Scripps Oceanographic Institution
Robert Downs, CIESIN, Columbia University
James Frew, Bren School, University of California, Santa Barbara
Juliana Friere, New York University
Peter Fox, TWC, Rensselaer Polytechnic Institute
Sara Graves, ITSC, University of Alabama, Huntsville
Steve Gustafson, GE Global Research
Bryan Heidorn, SLIS, University of Arizona
Roberta Johnson, NESTA/ State University of New York at Albany
Chris Lenhardt, RENCI, University of North Carolina, Chapel Hill
Kerstin Lehnert, LDEO, Columbia University
Carol Meyer, Foundation for Earth Science
William Michener, University Libraries, University of New Mexico
Erin Robinson, Foundation for Earth Science
Jennifer Schopf, International Networks, Indiana University
Kaitlin Thaney, Mozilla Foundation
Andrew Turner, Esri R&D Lab
Paul Uhlir, BRDI, U.S. National Academies
Anne Wilson, LASP, University of Colorado at Boulder

Appendix 2: Workshop Agenda

8:30 - Welcome and goals of workshop

8:45 - Introduction

9:00 - Session 1: Scientific Data Infrastructure Challenges
"What are three of the "grand challenges in scientific data infrastructure" from either your personal perspective or that of a "stakeholder" of your choosing (e.g., funder, decision-maker, researcher, student)? "

10:30 - Break

10:45 - Session 2: Potential Outcomes of a Community Study
"What are two to three key recommendations or actions that you would like to see emerge from a study of the "grand challenges in scientific data infrastructure"? (The idea here is to establish study context/need: Why is a study of "grand challenges in scientific data infrastructure" needed now?)

12:15 - 1 pm - Lunch

1:00 - Session 3: Identification of Community Stakeholders
"Who are the stakeholders that should be invited to participate in a study of "grand challenges in scientific data infrastructure". Each participant identifies up to 10 stakeholders…"

2:15 - Break

2:30 - Session 4: Engaging the Community in a National Study
"How would your team design an effective study of "grand challenges in scientific data infrastructure" that could be carried out within a two-year period (e.g., "structure of the meetings to maximize participation and productivity", "number and size of meetings within the two-year period", "location(s) of meeting(s) (physical places? virtual? hybrid?)", "pre-meeting preparation")?"

4:00 - Session 5: Wrap-up and the Road Ahead
"Each group … is given 10 minutes to explain its design of a study, followed by 20 minutes and Q&A and a 15-minutes high level summary and statement summarizing plans for moving forward."

5:15 Adjourn

Appendix 3: Workshop Artifacts

Session 1: Scientific Data Infrastructure Challenges
"What are three of the "grand challenges in scientific data infrastructure" from either your personal perspective or that of a "stakeholder" of your choosing (e.g., funder, decision-maker, researcher, student)? "

Top ranked grand challenges:
https://docs.google.com/document/d/1GhgLk8WieezQ5rrz2rR12LJyq05Xr0FPLUgPQ4_wuz4/edit?usp=sharing
All grand challenges:
https://docs.google.com/a/esipfed.org/document/d/1n1bJbmx_PTsi5QgZzsQ-zPwp5iK8-YpSBS8603N9c5E/edit?usp=sharing

Session 2: Potential Outcomes of a Community Study
"What are two to three key recommendations or actions that you would like to see emerge from a study of the "grand challenges in scientific data infrastructure"? (The idea here is to establish study context/need: Why is a study of "grand challenges in scientific data infrastructure" needed now?)

Worksheet content:
https://docs.google.com/spreadsheet/ccc?key=0AoyYlkf4MJfSdDFPdVdoLTZoS1VJSUFQZmNQQjQzQ2c&usp=sharing
Scanned worksheets:
https://drive.google.com/file/d/0B4yYlkf4MJfSQVhxTTYtWmNmZHc/edit?usp=sharing

Session 3: Identification of Community Stakeholders
"Who are the stakeholders that should be invited to participate in a study of "grand challenges in scientific data infrastructure". Each participant identifies up to 10 stakeholders…"

Stakeholders, grouped:
https://docs.google.com/spreadsheet/ccc?key=0AoyYlkf4MJfSdEpqNXp2NlYyNHo2aUl3Wm53WlZBckE&usp=sharing
Individual stakeholder categories and instances: https://docs.google.com/document/d/19idl-yQpSVUezJ11N2IkeLiZdsfcNDgMx1ljpP4pZ7U/edit?usp=sharing

Session 5, Wrap Up and the road ahead, raw notes:
"Each group … is given 10 minutes to explain its design of a study, followed by 20 minutes and Q&A and a 15-minutes high level summary and statement summarizing plans for moving forward."
https://docs.google.com/document/d/1Mq-_ycB_77LGKfu76-Xg9U6dXlPnb4Xl2nok086oyDQ/edit?usp=sharing

Appendix 4: Partial List of Prior Studies around Data

National Research Council. Preserving Scientific Data on Our Physical Universe: A New Strategy for Archiving the Nation's Scientific Information Resources. Washington, DC: The National Academies Press, 1995.

National Research Council. Geoscience Data and Collections: National Resources in Peril. Washington, DC: The National Academies Press, 2002.

*"Priority Area Assessment on Scientific Data and Information", International Council for Science, 2004,*http://www.icsu.org/publications/reports-and-reviews/priority-area-assessment-on-scientific-data-and-information-2004

*"Report from the ad hoc Strategic Committee on Information and Data", International Council for Science, June 2008,*http://www.icsu.org/publications/reports-and-reviews/scid-report

*"GEO Vision", National Science Foundation Advisory Committee for Geosciences GEO Vision Working Group, 2009,*http://www.nsf.gov/geo/acgeo/geovision/start.jsp

"Harnessing the Power of Digital Data for Science and Society", Report of the Interagency Working Group on Digital Data to the Committee on Science of the National Science and Technology Council, January 2009, http://www.nitrd.gov/About/Harnessing_Power_Web.pdf.

"Sustainable Economics for a Digital Planet: Ensuring Long-term Access to Digital Information", Blue Ribbon Task Force on Sustainable Digital Preservation and Access Final Report, Online Computer Library Center, March 2010:http://www.oclc.org/news/releases/2010/201011.htm

"Riding the Wave", European Commission's High-Level Group on Scientific Data, June 2010:http://ec.europa.eu/information_society/newsroom/cf/itemlongdetail.cfm?item_id=6204

"Harnessing the Power of Digital Data: Taking the Next Step", Scientific Data Management (SDM) for Government Agencies: Report from the Workshop to Improve SDM, July 2010, http://www.cendi.gov/publications/CENDI-2011-1.pdf

"Report of the Ad-hoc Strategic Coordinating Committee on Information and Data (SCCID Report)", International Council for Science, April 2011, http://www.icsu.org/publications/reports-and-reviews/strategic-coordinating-committee-on-information-and-data-report/ad-hoc-strategic-coordinating-committee-on-information-and-data

"Data sharing in neuroimaging research", Poline, Breeze, Ghosh, et al, *Frontiers in Neuroinformatics*, April 2012,http://www.frontiersin.org/Neuroinformatics/10.3389/fninf.2012.00009/abstract

"Strategic Frameworks for Education & Diversity, Facilities, International Activities, and Data and Informatics in the Geosciences", National Science Foundation Advisory Committee for Geosciences (AC GEO), 2012, http://www.nsf.gov/geo/acgeo/geovision/start.jsp

References

[Altmetrics] altmetrics.org.

[Jetzek 2013] Jetzek, Thorhildur; Avital, Michel; and BjÃ¸rn-Andersen, Niels, "The Generative Mechanisms Of Open Government Data" (2013),
ECIS 2013 Completed Research. Paper 156. http://aisel.aisnet.org/ecis2013_cr/156.

[Jisc 2014] The Value and Impact of Data Sharing and Curation, A synthesis of three recent studies of UK research data centres", March 2014, http://repository.jisc.ac.uk/5568/1/iDF308_-_Digital_Infrastructure_Directions_Report%2C_Jan14_v1-04.pdf.