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

Executive Summary

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 uncoordinated fashion. The problem continues to grow as data volumes, complexities, and sources increase exponentially.

In order to be well positioned to address these challenges, our nation needs an overarching, unifying strategy for collaboratively addressing the management of 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 fullest potential of our data. Over the last year, the ESIP Data Study Working Group has investigated the idea of initiating a National Research Council study on science data infrastructure.

This workshop, "Planning for a Community Study of Scientific Data Infrastructure", held on January 7, 2014 in Washington, D.C. was designed to frame a community study of challenges and opportunities associated with scientific data infrastructure. The workshop sought to:

- 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 participants envisioned a sustained Science Data Infrastructure (SDI) and associated technical and cultural changes that better enabled science in the face of major challenges now and into the future. Participants agreed that the government should provide funding for a SDI in some sustainable manner. SDI stakeholders were identified, and ideas for executing the study and engaging the community were discussed.

The challenges that the workshop identified are opportunities to achieve progress in science, innovation, the economy, and broader society. ***To capture the fullest value of the investment in our data, the consensus of the workshop participants was that a study is needed to investigate the costs and benefits of providing sustainable infrastructure for the long-term management and stewardship of scientific data ….. further, the consensus of the workshop participants was that the National Research Council (NRC) is the logical entity to oversee such an effort in order to ensure an authoritative and unbiased assessment of requisite strategic, 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.

The challenges that the participants identified are also opportunities to achieve progress in science, innovation, the economy, and broader society. The consensus of the workshop participants was that, in order to capture the fullest return from our investments in data and to realize important potential societal benefits, authoritative guidance is needed in important aspects of scientific data, 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. Also, participants felt that the the National Research Council (NRC), being the authoritative and unbiased body for science in the United States, would be best suited to provide a consensus study to investigate these areas and produce a report with impact. This report would inform and guide decision makers in the government, academia, and industry in helping to improve their practices and priorities for managing scientific data. It was strongly felt that better usage of scientific data could give the U.S. an enhanced competitive advantage not just scientifically, but also economically and socially.

This workshop was made possible with generous support from the Gordon & Betty Moore Foundation, the National Consortium for Data Science and the ESIP Federation. We are also grateful to Bill Michener for facilitating the workshop and for the support of the ESIP workshop steering committee.

Workshop Overview

This workshop, "Planning for a Community Study of Scientific Data Infrastructure", held on January 7, 2014 in Washington, D.C. was hosted by the Foundation for Earth Science on behalf of the Federation for Earth Science information Partners' Data Study Work Group. The workshop brought together twenty-three participants, consisting of individual experts from academic institutions and research institutions and was facilitated by Dr. William Michener, University of New Mexico. A list of the attendees is appended in Appendix 1 of this report.

The workshop was designed to frame a community study of challenges and opportunities associated with scientific data infrastructure. The workshop sought to:

- 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 Workshop agenda is presented in Appendix 2 of this report. The one-day program started with a brief welcome, round of introductions and goals of the workshop. The group was then led through a series of questions, both in full group discussion and in small groups, facilitated by Dr. Michener. A summary of these discussions follows.

Identify Scientific Data Infrastructure Challenges

The first workshop session was a group discussion around the questions: 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.

It is noteworthy that early in the discussion an acronym emerged for Science Data Infrastructure (SDI). The group seemed unanimous in thinking in terms of a ubiquitous, reliable, easy to use system for publishing, finding, understanding, using, and accrediting scientific data. This term was used throughout the workshop.

Another phrase used was the “science data enterprise”, to refer to all relevant management and stewardship aspects of scientific data throughout the data lifecycle.

An analysis of this session and the following session 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 two economic challenges identified by the group were

- Developing an economic model for sustained infrastructure without competing with research dollars
- 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.

To date, the economic value and impact of scientific data has not been adequately studied, but interest in economic models, cost models, and return on investment of research data is growing. Efforts are underway to begin to measure impact and consider generative potential. Altmetrics, which attempts to measure the impact of academic literature from the web downloads, views, storage, links, bookmarks, electronic conversation and other novel ‘traces’, was mentioned as an example of new ways to measure value [Altmetrics]. [1]

[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.

The group felt that reliable, long term funding as well as changes in financial incentives and rewards are needed. Data management expenses should 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 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

Several top challenges were about changing cultural values, such as

- The need for new academic values around data management and stewardship
- Cultural incentive and reward structures do not reflect the importance of quality data management
- Providing attractive, rewarding career paths for system builders in science
- Integrating data science into ways we teach science

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.

Cultural issues are involved in coordinating the data-related efforts of disparate communities, organizations, and domains. Today's societal needs require research at the intersection of once disparate domains. New research questions can be studied by integrating scientific data that have been shared and used across different disciplines. For example, disaster response and planning for climate change require earth science data as well as perhaps biological, GIS, population, and sociological data.

Another cultural issue ~~problem~~ is that policy makers, academic institutions, scientists, and the public do not fully understand the importance of data infrastructure (or the lack thereof) to their work. The group felt it was necessary to raise awareness of the importance and potential value of managing scientific data, and to be clear about the risks of not doing so. It is necessary for stakeholders in the science data enterprise to understand the value of scientific data so that resources will be allocated for their management and stewardship.

Finally there is a cultural divide between the domain scientists and the data scientists who develop and work with systems that manage data. Domain scientists need to appreciate data management and stewardship advice, while data scientists need to better understand scientific needs and goals in designing and building tools and infrastructure.

Data Science Research, Goals

The term 'science data enterprise' may seem new due to the unheralded efforts that have evolved to become unified data management and stewardship practices. While the importance of data interoperability has been recognized across scientific disciplines, as it should be, a successful science data enterprise that truly maximizes the generative value of the data needs to consider much more than just data interoperability.

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

Data science related questions 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?

The importance of metadata came up many times throughout the workshop. The challenge of achieving “magic metadata” was identified - not too much, not too little, but just enough of the right information. What metadata fields are needed? At what resolution? How can potential users discover data that meets their needs? The answer is difficult because it depends on what unintended uses of the data arise, with new potential uses occurring over time.

As often mentioned, technical challenges continue to include making data interoperable across domains, groups, and organizations and tools. Are there frameworks or mathematical solutions that could help? How can high performance computing be leveraged? Another challenge: the paradoxical situation of having both too much and not enough data at the same time, finding a needle in a haystack. Also often cited is the 80/20 rule of working with data - a scientist oftens spends 80% of their resources getting data in a usable format and 20% doing actual science. Finally, networks are far from being used optimally and major gains could be achieved if they were simply managed better.

What coordinated practices are essential to increasing the generative value of our data? The group held the view that the nature of scientific review is changing, with the traditional publishing process being replaced with more ‘crowd’ and social networking types of review. For example, a mechanism was recommended for unintended users of a data set to give feedback to data providers and tool developers (who would presumably have resources enabling them to respond).
student development and life-long professional development for the current workforce. The Next Generation

Education Challenges

As eScience is changing the very nature of doing science, we face major challenges in workforce development issues, starting in K-12 with science data in the classroom and extending through graduate Science Standards call for data in the classroom and there are still major challenges for getting data to teachers in ways that they can easily use in lessons.

Legal, Ethical & Policy Challenges

Science progresses through opportunities for repeatability and verification. Challenges exist around the ethical use of data, data licensing, data transparency, and data verifiability. The true costs and benefits of good data management, as well as the trade-offs, need to be understood in order to make informed decisions for science infrastructure.

While science is evolving to be cross disciplinary, current funding vehicles contribute to disparate, potentially duplicative, possibly non interoperable efforts around data. The vision and funding for a science data infrastructure needs to be considered on a national scale rather than only on organizational scales. Perhaps a new entity is needed, e.g., the National Science Infrastructure Foundation? This could be a way to accomplish sustained funding. At the very least, organizations that are part of the scientific enterprise need to prioritize their funding and related actions to better support data management and stewardship for sustainable science.

Workshop participants unanimously agreed that overcoming these challenges to form a more coordinated science data infrastructure could not only enable new science that could not have been achieved previously, such as interdisciplinary science and geographically distributed collaborations, from which other societal benefits could be achieved. While some of that is starting to occur, we identified huge challenges that remain. A truly coordinated effort could provide a substantial increase in the generative potential of data along with associated scientific and societal benefits.

Potential Outcomes of a Community Study

The second session focused on the questions, "What are the 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 intent was to solicit from participants their expectations of what a study of "grand challenges in scientific data infrastructure" could achieve and why it is needed now. The content of the resulting worksheets and their categorization is available in a spreadsheet format in Appendix 3.

Not surprisingly, these goals tended to fall under the categories there were also identified as challenges. From this exercise, the following more specific themes were identified.

A Ubiquitous Science Data Infrastructure (SDI) is needed

Economic Goals

- don't take funding away from science research for infrastructure
- sustained funding
- Economic models, cost models, ROI
- measures of progress/sharability/impact
- a way to evaluate the value of a data set , e.g. to decide to what extent to fund maintenance
    - e.g., altmetrics.org
- financial incentives and rewards

Cultural Issues/Changes

- Improved cultural incentives, rewards
- Elevated awareness of importance of data

Research in Data Science, Data Science Goals

- common data model
- transparency
- interoperability

Improved Education

- education: all ages (k - grave), all domains, training current researchers
- Elevated awareness of importance of data

Policy

- Establishment, movement of an agency, council, or office

General Concerns

- Need for specifics
- dealing with commonalities vs domain specific needs
- alignment of solutions for US or global network

Identification of Community Stakeholders

In this session the group was asked: “Who are the stakeholders that should be invited to participate in a study of “grand challenges in scientific data infrastructure”. Each participant identified up to 10 stakeholders. Then each subgroup identified common stakeholders and clustered their ten stakeholders into categories (raw notes are in Appendix 3).

The result of this group work spanned multiple dimensions - data lifecycle, economic sector and organizational type and then into these categories:

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

Wrap-up and the Road Ahead

The group broke into small groups again and envisioned study process, time frame, and methods for community engagement, then reconvened and shared the ideas that emerged from their discussions. The raw notes are in Appendix 3.

The question was asked, "Do we really need to do all the things we talk about? Is the return on investment worth it?" The answer is unknown because we do not have a good measure.

Nonetheless, the group agreed that the main goal for this effort is to convince the Federal government to fund science data infrastructure and for science data infrastructure to get the same level of treatment that science gets at a cabinet level. This high level exposure would provide overarching coordination of funding and implementation.

The groups generally agreed that an innovative new approach is needed and a business as usual series of meetings and workshops is not adequate because the field is changing too quickly. That said, the groups seemed to converge on a two year timeline to complete a study, hopefully with ongoing assessment and evaluation after that. The first task, done in six months or less, would be to identify trends and perform a gap analysis of what is needed. There was also strong consensus that a review and meta-analysis of prior studies should occur to identify past successes and failures. Following this analysis, if it is determined that a Community Study is warranted, has support, and is feasible, the group would meet to identify overarching themes and chapters for the proposed study.

Call to Action

We live in a world rich with data. But due to insufficient planning, management, and resources, the potential benefits of all these data often go unrealized. 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. These problems will grow as data volumes and sources increase.

The challenges that the workshop identified are also 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) calls upon the National Research Council (NRC) to offer an authoritative and unbiased assessment for strategic scientific investments. This assessment 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.

The assessment should:

- Synthesize and analyze prior work in science data management and infrastructure, such as, what was successful, what was not successful, and why have past efforts 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.

Next Steps and Acknowledgement

The workshop report will be presented to the NRC's Board on Research Data and Information and made available broadly throughout the Earth and environmental science community, and beyond. There are several complementary efforts in US GEO Earth Observatory Assessment, NSF's CIF21 initiatives and NASA's ESDSWG recently proposed work group for Vision 2020 that we will reach out to in an effort to coordinate around a shared agenda.

This workshop was made possible with generous support from the Gordon & Betty Moore Foundation, the National Consortium for Data Science and the ESIP Federation. We are also grateful to Bill Michener for facilitating the workshop and for the support of the workshop steering committee members: Robert Downs, Chris Lenhardt, Anne Wilson, Carol Meyer and Erin Robinson.

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-yQpSVUezJ11N2IkeLiZdsfcNDgMx1IjpP4pZ7U/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-Xg9U6dXIPnb4XI2nok086oyDQ/edit?usp=sharing

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.