Data Study Working Session

The Workshop

Supporters:

- The ESIP Federation
- Moore Foundation
- LASP
- DataONE
- the Board on Research Data and Information
- National Consortium for Data Science

The workshop seeks 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
(and their dream vacation spots)

- Bill Michener, University Libraries, UNM (Australia)
- Bryan Heidorn, SLIS, University of Arizona (Wherever his son gets a job)
- Jim Frew, UCSB/Bren (Scotland)
- Chris Lenhardt, Renci (Scotland)
- Karl Benedict, EDAC/Libraries/UNM ()
- Jennifer Schopf, UI International Networks (home!)
- Roberta Johnson, Albany/NESTA (Patagonia)
- Bob Cook, ORNL (Pilot)
- Juliana Friere, NYU ()
- Bob Downs, CIESIN (South Florida)
- Kaitlin Thaney, Mozilla Foundation (anywhere not near her inbox)
- Lee Allison, AZGS (anywhere not there before)
- Stan Ahalt, UNC/Renci (Spain)
- Kerstin Lehnert, LDEO/Columbia (New Zealand & Iceland)
- Anne Wilson, LASP (Tierra del Fuego)
- Sara Graves, UAH/ITSC (Israel)
- Steve Diggs, Scripps (Fiji)
- Paul Uhlir, BRDI (home)
- Steve Gustafson, GE Global Research (Hawaii)

Session 1: What are three of the "grand challenges in scientific data infrastructure" from either your personal perspective or that of a "stakeholder" of your choosing?

13 6. Sustaining infrastructure information: Developing economic model competing with research and developing an economic model
13 36. New academic values with respect to data management and stewardship – “reward structures”
10 42. Supporting o enticing career paths for system builders in science
10 57. Integrating data science into ways we teach science
9 12. Integration and portability of Diversity, enabling interpretability
9 27. Effective data sharing and selection by people who don’t know what they are looking for or how to find it.
8 18. Usability of today’s date for future users
8 65. The Internet of Samples: Physical objects (maps) need to be in the mix
8 66. Commoditizing SDI (engineering, architect, aspects, etc)
7 7. Balancing Need for broad solutions vs. domain specific ones
7 26. Distributed Big Data; moving data to computation; "location agnostic"
7 28. Infrastructure for interactive data exploration that is scalable (size and complexity)
7 32. Magic Metadata: tools for binding metadata to data
7 44. Uncertainty, data, quality, trust
6 19. Support for reproducibility, sharing of software data and experiments,
6 41. Educate scientific workforce: SDI without widespread use does not equal funding
5 1. How SDI enables high value of secondary use of data
5 4. Updating distributed data sources “global R”
5 17. Close interaction between science to informatics “sociopoidial”
5 23. No data formats
5 29. Mapping to desired changes in behavior: Understanding user needs
5 45. Cybersecurity of SDI
5 49. Overcoming data sharing barrier
5 53. Extra temporal ‘curation’ data valuation “looking forward”
4 2. Integration with web
4 9. Online clearinghouse of data infrastructure activities
4 34. Maximizes utility of legacy data (data rescue)

4 39. Full SDI transparency and verifiability

- Facilitating future access to currently available data
- Portability of data to operate like the web?
- Support for reproducibility sharing of solution, data with experiments

4 48. How to value data (planning for throwing away

4 67. Digital Sheepskin; making digital data preservable for centurie

3 8. Decision Makers understanding the importance of data

3 14.Adoption of Common Open Usage Licensing

3 20.“Usability” Low barriers to accessibility

3 21.All the data you use is visible with your name space

3 37. Better network use (shipping data)

3 38. Intervention of coordination of public sector policies

3 51.Well-socialized ethical data use standards

3 60. Changing understanding of values ‘societal understanding’

3 64. SDI Governance?

2 55. Define data science

2 54. Business model for secondary use ‘data market’

2 24. Ubiquitous scientific data cloud (an implementation); portability of data to operate like the public web

2 33. Interdisciplinary science that crosses agencies, organizations to public sector

2 35.Technologies preventing unethical use

1 3.Too little/much data (sparse data)

1 5. Elimination of multiple similar efforts

1 11.Too many problems that are not grand challenges

1 13. How to achieve purposeful use of data by educators

1 61. Leveraging (partnering) with commercial enterprises for SDI

0 40. Open feedback and collaboration channels

- Commoditizin
- Getting publishers out of the solution

Session 2: 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"?

Stan Ahalt:
Establish a new agency that focuses on establishing a national data infrastructure.

Kaitlin Thaney:
Openness as a default – repositories, techonlogoes

Bob Downs
Revise science education at all levels to include scientific data management across data life cycle.

Juliana Friere

Bob Cook
Envisions a data system that has data across domains and that enables discovery access

Roberta Johnson
Provide professional development to engage effectively with education community

Jim Frew
Reliable funding streams need to be made to institutions that are committed to curating data for the long term. ROI? Value

Chris Lenhardt
Create the new discipline of data science to work on the grand challenges and have an applied element.

Karl Benedict
Differentiate between common data characteristics v. domain specific ones.
Common data models where they exist.

Jen Schopf
Do we really need to do all the things we talk about? What about ROI?

Kerstin Lehnert
Pull out ACI from under CISE and have it be a stand alone under the top of NSF. Need actionable recommendations.

Anne Wilson
Explore a domain independent data model using mathematics.

Sara Graves
Transparency of the data & software supporting science. Need better incentives at all levels for data usage and rewards.

Steve Diggs
Change in how the funding agencies value data curation, data access. A plan for public access memo – use it to frame actions.

Paul Uhlir
Global Research Foundation (EC & NSF created) should form a permanent council from international science agencies

Peter Fox
Producers will use the same SDI as consumers.

Stefan Falke
Advancing frameworks, tools and policies that encourage orgs (private, academia, government) to assess what is truly proprietary.

Steve Gustafson
Focus commercial innovation on solving some of the challenges discussed here.

Bryan Heidorn
Develop a model for analyzing the cost of running a long-term data center to enable appropriate allocation of resources. These need to be more transparent. Develop a tool (to be applied across disciplines) to assess real costs.

Andrew Turner
How does citation & reuse of data metrics align with metrics for other publications.

Session 3: 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 as part of a story-boarding exercise in order to identify stakeholders and categories. (clustering exercise) https://drive.google.com/folderview?id=0Bw1bvGCz9K3_aV8tdWthMGxOaE0&usp=sharing&usp=sharing#

Session5: Wrap-up and the Road Ahead

- SDI, Science Data Infrastructure
  - An agency for this?
    - Issue of an entity that can receive funding (Brian Wee)
    - National Data Infrastructure Foundation
- One group not convinced that this study is needed (group 1, Jen Schopf)
  - Have there been benefits to prior studies? What if they had not been done?
  - What is the risk of NOT doing this study?
- rigorous pre study
  - synthesis, analysis of prior work, reports, followed by gap analysis
- The messenger
  - Very important (Stan Ahalt)
    - VPs, CEOs
    - Stan recommended someone in particular

- Leverage education initiative PCAST STEM Education Report (Roberta Johnson)
- Data, the last frontier
- Better alternatives than an NRC study?
  - Keep trying, wait for window of opportunity to open (Stan)
- Timing
  - It is everything
  - It may be ripe now

Data Study Dialog
Breakout Session

Attendees: Anne Wilson, Bob Downs, Kevin Ashley, Chris Lenhardt, Jane Greenberg, Reid Boehm, Tiffany Mathews, Sharon Hays, Adam Steckel, Brian Wee, Steve Aulenbach, Lee Allsion, Jeff Walter, Rama Ramapriyan, Mark Parsons, Paul Uhlir, and Martha Maiden.

Problem: no clear target to fund

- US Department of Data?
- NSDI?
- Failure of National Climate Center to launch

Economic value, ROI

- Kevin Ashley, 400% - 1200% return on data centers
  - There is an method to calculate this that probably could be applied
  - DCC Digital Curation Center, UK
- How much more valuable would the data/data centers be if they were interoperable?

A general collaboration is asking too much. Just get agencies to act in similar ways (Kevin Ashley). What if agencies don’t play?

- NRC doesn’t tell how, but what.
  - paul - unless asked

Needs depend on perspective (Sharon Hays, Congress, OSTP, now CSC).
Must help NRC understand all of them.
Must identify the key questions we are asking.
SBA (Brian), what is the data good for? it's purpose
Wilbanks spoke of the "generative value" of data (Parsons)
 4 axis of measurement: accessible, adaptable, ease of mastery, ?
 How are our agencies maximizing the generative value?
Frame the problem - how to enable value of [better data] to economic impact
Another axis besides economics - health
 in India and other poor countries climate change will have a big impact
 http://cal-adapt.org/ a site on impact of sea level rising
 what science do we need to enable to be able to generate sites like that?
 What are the gaps?
 A different team for each societal benefit area?
Stay focused on Earth Science (Sharon)
SDI as a research area (Bob)
"Harnessing the Power..." OSTP was good because it was not prescriptive, 2009
 Interagency Working Group on Digital Data

OSTP's Science and Policy Institute might do gap analysis

scipi ("stpi") sci and tech policy inst

https://www.ida.org/stpi/about%20stpi.php

ROSES E.2 will pay for a workshop. Could survey and synthesis be done as a workshop??

DataONE has done a synthesis within [its purview?] (Steve)

Hiring a consultant can free agencies from being on the hook

Final strategy: do gap analysis and get right messengers, esp from private sector

Outcome from Working Session:
Next steps

1) Synthesize stuff from workshop and breakout
Update and agree on list of prior work to survey.
Do synthesis of prior work

- graduate student work?
- a seminar?
- new ESIP fellow?
- shop around a statement of work
  - ucsb center for info tech and society

ESIP for $? an esip fellow is $.

2) Do gap analysis, including interviewing people re: impact of prior reports
   a) gap analysis of what report recommended vs what got done
   b) what are common recommendations across studies? Were they implemented? If implemented, did they work?   If not, why not?
Why must we keep asking? Wrong q's?? political issue?

Make sure ESIP agency reps part of the process: Martha, Jeff dlb, EPA sontag?, USGS Kevin Gallagher? tailor for their perspective

Keep Earth Sci focus

Broad publicity all along.

3) Produce ESIP recommendation to agencies accompanied by a draft charge letter and supporting material

Study:
Gap analysis of what is still needed, e.g. to enable sites like cal-adapt.org. What is needed for supporting various societal benefit areas? This would be one outcome of NRC study. This is part of the charge letter.

4) agencies issues charge letter to NRC to do study