The Federation of Earth Science Information Partners
Fostering connections to make data matter

Establishing Long Term Data Management Research Priorities via a NRC Data Study

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

Data and Science Goals

Data should be

Discoverable

Understandable

Accessible

Interoperable

Preserved, usable in the future

Science should be

Verifiable

Repeatable

Transparent

Inclusive

Attributable

Cross Domain/Organizational

Impacts of Technology on Data and Science

- Changes in the data landscape are changing the performance and culture of science Visionary science: executable publications

- Rate of change is analogous to Moore's Law

"The number of transistors on integrated circuits doubles approximately every two years"

A new world with new implications

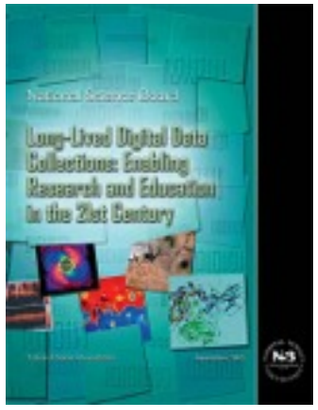
"It is exceedingly rare that fundamentally new approaches to research and education arise. Information technology has ushered in such a fundamental change. Digital data collections are at the heart of this change. ... **Through their very size and complexity, such digital collections provide new phenomena for study.** "

[NSB Report: "Long-lived Data Collections: Enabling Research and Education in the 21st Century," Sept 2005]

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

We've been talking about 'the data problem' for years!

"Priority Area Assessment on Scientific Data and Information", International Council for Science, 2004



NSB Report: "Long-lived Data Collections: Enabling Research and Education in the 21st Century," Sept 2005

"We have a shared responsibility to create and implement strategies to realize the full potential of digital information for present and future generations." –eGY Declaration, 2007

"Report from the ad hoc Strategic Committee on Information and Data", International Council for Science, June 2008

"Sustainable economics for a digital planet: Ensuring long term access to digital information", The Blue Ribbon Task Force on Sustainable Digital Preservation and Access, 2010

"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

NSF: "Changing the Conduct of Science in the Information Age" 2011

"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

"Report of the Ad-hoc Strategic Coordinating Committee on Information and Data (SCCID Report)", International Council for Science, April 2011

We're working on it:
Current Declarations, Efforts, and Reports

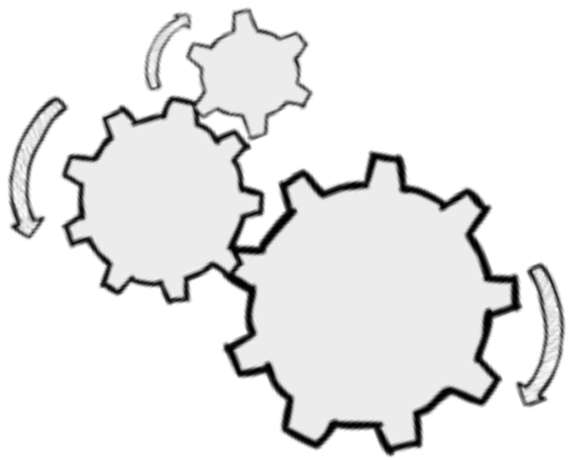
- The Blue Ribbon Task Force on Sustainable Digital Preservation and Access
- Data.gov
- EarthCube
- ESIP
- eGy Declaration
- National Research Council
- NASA's Earth Science Data System Working Groups
- National Consortium for Data Science
- NOAA's Environmental Data Management Committee
- OSTP declaration
- Research Data Alliance
- Sustainable Digital Data Preservation and Access Network Partner (DataNet) and its funded projects, such as DataONE and Terra Populus
- USGS Community for Data Integration

and others!

With all this effort, why are we here?
Why are we worried?

Stakeholders in the Science Data Enterprise

Scientific
Technical
Organizational/Institutional
Economic/financial
Policy/legal
Socio-cultural
Publishing
Information science
Library science



These groups are addressing the problems **piecemeal** and from **limited perspectives**, which can negatively impact coherence and effectiveness.

No Sustainable Economic Model

"There is no general agreement, however, about *who is responsible* and *who should pay* for the access to, preservation of, valuable present and future digital information. Creating sustainable economic models for digital access and preservation is a major challenge for all all sectors, ..."

[Blue Ribbon Task Force on Sustainable Digital Preservation and Access Interim Report, 2008]

Agencies are being asked to do more and better data management with less funding.

The needs are greater than the scope of any single agency or organization

Agencies are not tasked to address:

Economics: *who is responsible* and *who should pay* [BRTS Interim, 2008]

Connecting publications to data

Legal issues: e.g., "Develop an Open Research License (ORL) to resolve conflicts between reproducibility and copyright law [CCSIA]

General data science issues: e.g., "Develop pilot projects that compare different technical solutions for developing and maintaining open data platforms, fostering the replication of research, and ensuring attribution for intellectual contributions." [CCSIA]

Creation of a system of persistent ids for both researchers and their outputs, usable by publishers. "This involves creating taxonomies of data and other entities, incentives for changing
encourage the development, implementation, and use of standardized identification systems to facilitate attribution by requiring system registration as a prerequisite for agency funding." [CCSIA]

Inconsistent, short term funding

Inadequate incentives

The role of the private sector

Workforce training

The issue of attribution

Cross organizational infrastructure: e.g. "Establish federally funded platforms for data and code sharing that are independent of institutions and individual researchers" [CCSIA]

"Confusion, lack of alignment between stakeholders, roles, and responsibilities" [BRTF]

Uneven appreciation of issues, goals: "the urgency of the matter is not uniformly appreciated" [BRTF]

The fear that it's all too big to take on [BRTF]

Spearheading an NRC Data Study

Working in partnership with the community, the BRDI, and with support from the Moore Foundation and the National Consortium for Data Science, ESIP is leading efforts to define and scope **an NRC data study** to provide a consensus view on **research priorities for the science data enterprise**

BOARD ON RESEARCH DATA AND INFORMATION

BOARD ON RESEARCH DATA AND INFORMATION
Policy and Global Affairs

Under Policy and Global Affairs, rather than one of the science Boards

ESIP is cross agency, community-led, neutral

Study initiation is a community-led effort

Organizations, agencies would fund the study

What a NRC Data Study Could Accomplish

Overarching, independent, authoritative, expert-derived scientific advice from the highest U.S. authority

Establishment of an economic model

Planning that **spans Administrations and Congressional turnover**

Influence, the possibility to impact public policy, including Congress, who controls the purse

Guiding principles and approaches that can help inform organizations that fund research, scientific research organizations, and publishing houses. [CCSIA]

Advice on technical constructs: what are the most important digital technologies that could be used to facilitate data and knowledge access? To what extent is progress already being made and how can progress be accelerated? [CCSIA]

Long term visioning for future technological changes, trends

Inclusion of the private sector: What role might the private sector play in bringing about change? [CCSIA]

Advice on social constructs: What incentives are necessary to engage scientists in making data accessible and shared? What are the appropriate business models necessary to promote connecting publications to data? How might the private sector be engaged? What are the social barriers to adopting and using [unique identifiers for researchers]? [CCSIA]

A coherent vision across all stakeholder roles

Integration of disparate experience: What lessons have we learned already? What can we learn from libraries and publication efforts? What can we learn from domain-specific successes? [CCSIA]

The possibility for **ongoing advice, guidance**

[CCSIA] Changing the Conduct of Science in the Information Age, June 2011 [http://www.nsf.gov/pubs/2011/oise11003/index.jsp]

Why now?

"The problem is urgent!
Access to data tomorrow requires
decisions today." [BRTF, 2008]

Data Study Risks

- Producing a costly, ignored report
- Very broad topic, how to constrain?
  - Either too broad or too narrow is not helpful
- What scope? Earth Science or beyond?

Next steps

ESIP plans to write recommendations to the NRC regarding a data study to be initiated in FY 2015

Invited workshop was held yesterday to address scope, etc.

We seek input from community at Data Study Community Forum: ESIP wiki, http://bit.ly/Jf5Idx

"Data Study Community Forum Input Form"

Please comment!

Write the report.

But, we've been doing that! Yet another study

- What if we hadn't made those efforts?
- What if we don't make this effort?

Info

me: anne.wilson@lasp.colorado.edu

ESIP wiki: http://wiki.esipfed.org

Community Forum: http://bit.ly/Jf5Idx

"Data Study Community Forum Input Form"

Thank you to the ESIP Data Study Working Group for assistance and support in this project!

Thank you, LASP, for providing support for this effort!