Got Data?
Building a Sustainable Environment for Data-Driven Innovation

Dr. Francine Berman
Chair, Research Data Alliance / US
Hamilton Distinguished Professor in Computer Science, Rensselaer Polytechnic Institute
It’s a Data-Driven World

Entertainment

Physical Infrastructure

Commerce

Health

Communication / Community

Research

Fran Berman
What is the likely impact of a large-scale earthquake?

How does disease spread in large urban populations?

How can we increase wheat yields?

Data and Research: Digital Research Data Driving Solutions to Complex Science and Societal Challenges

What’s required to support data-driven innovation?
Data-Driven Innovation Requires an Ecosystem

Impact is dependent on effective development and integration of all components

• **“Natural Resources”**
  - Data
  - Data collections and databases

• **Infrastructure**
  - **Software and hardware:** tools, systems, storage, data centers
  - **Social and organizational:** policy, practice, people, standards

• **Resource management**
  (cross-cutting)
  - Stewardship
  - Sustainability
  - Economic support

---

Figure: Wikipedia

Rensselaer

why not change the world?™

Fran Berman
Data-Driven Health – Who is at risk for Asthma?

- **Asthma** is a major cause of disability, health resource utilization, and poor quality of life worldwide.
  - Most common chronic disease among children and young adults
  - Expected that by 2025, 400M people world-wide will have asthma

- **Asthma is a socio-cultural-health issue**
  - **Relevant data:** Health records, biological data, environmental data, location of physical infrastructure and hospitals, population data, ...
Data-Driven Asthma Results Advance Treatments and Outcomes, Accurate Disease Models, Development of Public Policy

• **Asthma risk factors include:**
  – Pollution
  – Smoking
  – Age, race, economic status
  – Allergy to cockroaches and dust mites
  – Exposure to formaldehyde
  – Use of antibiotics in early life, etc.

You are at higher risk for asthma in Australia (21% of the population) than Vietnam (1% of the population) [World Health Organization study]

WAO article: [http://www.waojournal.org/content/3/4/167](http://www.waojournal.org/content/3/4/167); WHO article: [http://www.biomedcentral.com/1471-2458/12/204](http://www.biomedcentral.com/1471-2458/12/204)
The Data “back story”

- **Results summary:** Program for Control of Asthma (ProAR) was able to decrease costs for asthma treatment in low-income families and asthma hospital admissions in Salvador City, Brazil by 82%.

  WAO article: http://www.waojournal.org/content/3/4/167 ; Global alliance against chronic respiratory disease image: http://www.who.int/gard/countries/demonstration_project_bahia/en/

- **Data coordinated and used for analysis:**
  - Public health data
  - Patient personal data – age, gender
  - Asthma family costs
  - Asthma quality of life factors
  - Patient information
  - Pharmacy data on provided medications
  - Statistical database of hospital admissions and stays, etc.
Data Sharing and Interoperability – key driver for innovation
Research Data Sharing Ecosystem –
Technical, social and organizational infrastructure
Data from birth to death / immortality:
The Digital Research Data Life Cycle

Create
- Data creation / capture / gathering from
  - laboratory experiments
  - fieldwork
  - surveys
  - devices
  - simulation output...

Edit
- Organize
- Annotate
- Clean
- Filter ....

Use / Reuse
- Analyze
- Mine
- Model
- Derive additional data
- Visualize
- Input to instruments / computers / devices ....

Publish, Disseminate
- Disseminate
- Create portals / data collections / databases
- Couple with literature ....

Preserve / Destroy
- Store / preserve
- Store / replicate / preserve
- Store / ignore
- Destroy ....

Information adapted from Chris Rusbridge and Liz Lyon; images: wikipedia, coneslayer at en.wikipedia, SDSC
Data Infrastructure: Enabler for Data-Driven Research

Data Access

Data Sharing

Data Visualization

Data Analysis

Data Services

Data Mining

Data Sharing Practice

Data Management

Digital Object Identifiers

Common Metadata Standards

Data Citation Standards

Data Access and Distribution Policy

Tools and infrastructure that promote Discoverability

Data Preservation

Data Storage

Sustainable Economic Model

Images: NASA, SDSC / UCSD
**SDSC / UCSD** – history of leadership in data infrastructure

- **SRB / IRODS** (Reagan Moore et al.) – collection and policy management
- **GEON, NEES, Safe&Well**, etc. (Chaitan Baru et al.) – data-driven community infrastructure
- **Protein Data Bank** hosting (Phil Bourne et al.) – support for invaluable community data collection and services
- **Data Central** (Natasha Balac et al.) – community data hosting and services
- **Blue Ribbon Task Force on Sustainable Digital Preservation and Access** (Fran Berman et al.) – data sustainability economics
- **Chronopolis** (Brian Schottlaender / UCSD Libraries, David Minor / SDSC et al.) – national-scale preservation grid
- **Data-intensive supercomputing** (Phil Andrews, Richard Moore, Wayne Pfeiffer, Alan Snavely, Mike Norman et al.), ... etc.
Data Infrastructure Pre-Supposes Viable Data Stewardship

Costs / components of Data infrastructure include

- Maintenance and upkeep
- Software tools and packages
- Utilities (power, cooling)
- Space
- Networking
- Security and failover systems
- People (expertise, help, infrastructure management, development)
- Training, documentation
- Monitoring, auditing
- Reporting costs
- Costs of compliance with regulation, policy, etc. ...

SDSC Data Storage Growth ‘97–’09

- Most valuable data replicated
- As research collections increase, storage capacity must stay ahead of demand

Information courtesy of Richard Moore, SDSC
Economics of Data Stewardship

It’s not just about size ...

Data costs increase with usage, management requirements, perceived value
Data Stewardship rising as a National Priority --

New Federal Policies for Access to Publicly Funded Data

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

February 22, 2013

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren
Director

SUBJECT: Increasing Access to the Results of Federally Funded Scientific Research

1. Policy Principles

The Administration is committed to ensuring that, to the greatest extent and with the fewest constraints possible and consistent with law and the objectives set out below, the direct results of federally funded scientific research are made available to and useful for the public, industry, and the scientific community. Such results include peer-reviewed publications and digital data.

Scientific research supported by the Federal Government catalyzes innovative breakthroughs to drive our economy. The results of that research become the grist for new insights and are assets for progress in areas such as health, energy, the environment, agriculture, and national security.

No new money: Agencies asked to identify resources within existing agency budgets to implement public access plans.
Data Economics: Who Pays the Bill?

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

MEMORANDUM FOR THE HEADS OF EXECUTIVE AGENCIES

FROM: John P. Holdren
Director

SUBJECT: Increasing Access to the Results of Federal Science and Technology

1. Policy Principles

The Administration is committed to ensuring the maximum possible public access to the results of federally funded scientific research and development (R&D). The Administration is also committed to supporting the scientific community in its efforts to make research results available to the public. Such efforts include providing guidance to federal agencies on the release of research results, and ensuring that the results of federally funded research are accessible to the public.

Scientific research supported by the Federal Government is a vital component of our economy. The results of such research provide the underpinnings for progress in areas such as health, energy, the environment, and national security. The Administration is committed to ensuring that such research results are made available to the public in a timely and meaningful manner.

The OSTP memo requested agencies to provide plans by September 2013 that describe their strategies for providing public access to both research publications and research data. Plans are expected to be implemented using “resources within the existing agency budget,” i.e., no new money should be expected. Currently, no federal R&D agencies are working hard to foster approaches to public access to research data.

Research data of community value are supported today in a variety of ways. Some of them, like those in the Protein Data Bank (PDB) (1)—a database of protein structure information used heavily by the life sciences community—are supported by the public sector. In particular, U.S. funding from the National Science Foundation (NSF), the National Institutes of Health (NIH), and the Department of Energy (DOE) have played a key role in creating and maintaining the PDB.

What happens to valuable data when project funding ends? Consider, for example, a 3-year research project in which valuable sensor data are collected from an environmentally sensitive area. Those data may be useful for a period of time, but subsequently the data may be of little importance. For the first 3 years, the costs of stewardship (including development of a database that supports analysis, access to the data for the community through a portal, adequate storage and management of the data collection, and so on) may be paid for by the grant. But who pays for subsequent support? In such cases, research data may become more valuable just as the economics of stewardship become less viable.

Upto this point, no one sector has stepped up to the plate to say that they will continue to provide necessary stewardship for these important datasets. Federal agencies have so far been reticent to fund the collection and stewardship of this data, even though, in the case of the PDB, publicly available data were collected with public funds. Clearly, we need a new strategy to ensure that research data are made available to the public.

Under current procedures, much of the federally financed research is published in scientific and medical journals that can cost thousands of dollars.

Multiple Approaches Can Provide Stewardship Options in All Sectors

1. **PRIVATE SECTOR:** Create federal and state incentives to facilitate private sector stewardship of public access research data

2. **PUBLIC SECTOR:** Create and clarify public sector stewardship commitments: articulate what data will and what won’t be supported

3. **ACADEMIC SECTOR:** Use public sector investment to jumpstart sustainable university library / community repository stewardship solutions

4. **RESEARCH COMMUNITY:** Encourage research culture change to take advantage of what works in the private sector (e.g. subscription, advertising, low-barrier-to-access fees, etc.)
Frontier Challenges: Research Data Sharing Infrastructure

• **Data Quality** – How do you know if your data is credible / clean / accurate?

• **Data Compatibility** -- How can we ensure that data from distinct sources can be combined?

• **Data → Information Literacy** – What does data mean in context? Is this data relevant evidence for this issue?

• **Data Discoverability** – How do we find relevant data?
  - “Whole Earth Catalog” or Advanced Search Tools
  - Discoverable by people and machines
  - Targeted discovery for specific kinds of uses

http://umm.edu/health/medical/drug-interaction-tool
Research Data Sharing Ecosystem – Community Engagement / Efforts
The Power of Many: Community organizations initiated to accelerate impact beyond individuals / projects / institutions

• “Just do it” -- Focused efforts help communities drive tangible progress

Now 25 years old, the Internet Engineering Task Force’s mission “to make the Internet work better” has resulted in key specifications of Internet community standards that support innovation

Development of a public access to shared data collection enabling new results for Alzheimer’s

MPI Forum photo by Erez Heba, PDB molecule of the month at http://www.rcsb.org/pdb/home/home.do
Data Sharing and Public Access a Global Issue

A Europe-Japan-United States GNSS data-sharing pilot project for the Geohazard Supersites and Natural Laboratories

Fran Berman

Data Scientists

Libraries, Archives, Repositories, Museums

D ATA .gov

Cyberinfrastructure professionals, data analysts, data center staff, ...

National Data Sharing and Accessibility Policy-2012 (NDSAP-2012)

Fran Berman
The Research Data Alliance (RDA)

- Global community-driven organization launched in March 2013 to accelerate data-driven innovation

- RDA focus is on building the **social, organizational and technical infrastructure** to
  - reduce barriers to data sharing and exchange
  - accelerate the development of coordinated global data infrastructure
Goal of RDA Infrastructure: Support Data Sharing and Interoperability Across Cultures, Scales, Technologies

- Common metadata types for data Interoperability
- Persistent identifiers
- Domain-focused portals
- Harmonized standards
- Digital object identifiers
- Data access and preservation policy and practice
- Tools for data discoverability, ...
CREATE → ADOPT → USE

RDA Members come together as

• **Working Groups** – 12-18 month efforts to build, adopt, and use specific pieces of infrastructure

• **Interest Groups** – longer-lived discussion forums that spawn Working Groups as specific pieces of needed infrastructure are identified.

**Working Group efforts focus on the development and use of data sharing infrastructure**

• **Code, policy, infrastructure, standards, or best practices that are adopted and used** by communities to enable data sharing

• **“Harvestable” efforts** for which 12-18 months of work can eliminate a roadblock

• **Efforts that have substantive applicability** to groups within the data community, but may not apply to everyone

• **Efforts for which working scientists and researchers can start today**
What RDA Groups are Working On --
Groups that Met at RDA Plenary 2 in DC

- **Birds-of-a-Feather**
  - Linked Data
  - Chemical Safety Data
  - Education and Skills Development in Data Intensive Science
  - Libraries and Research Data
  - Cloud Computing and Data Analysis Training for the Developing World

- **Working Groups**
  - Data Type Registries
  - Metadata Standards
  - Practical Policy
  - Persistent Identifier Types
  - Data Foundations and Terminology
  - Data Categories and Codes

- **Interest Groups**
  - Agricultural Data
  - Big Data Analytics
  - Data Brokering
  - Certification of Trusted Repositories (joint with ICSU-WDS)
  - Long tail of Research Data
  - Marine Data Harmonization
  - Community Capability Model
  - Data Publishing (joint with WDS)
  - Toxicogenomics Interoperability
  - Research Data Provenance
  - Data Citation
  - Metadata

- **Data Citation Harmonization Summit**
  - DataCite, FORCE11, CODATA/ICST, ESIP, DCC, etc.

- Economic Models and Infrastructure for Federated Materials Data Management
- Engagement
- Preservation e-Infrastructure
- Legal Interoperability (joint with CODATA)
- Global Registry of Trusted Data Repositories and Services
- Digital Practices in History and Ethnography
The RDA Community: ~1300 participants from 50+ countries and a broad spectrum of data cohorts

1. Albania
2. Australia
3. Austria
4. Bangladesh
5. Belgium
6. Bolivia
7. Botswana
8. Brazil
9. Bulgaria
10. Canada
11. China
12. Congo {Dem. Rep.}
13. Costa Rica
14. Czech Republic
15. Denmark
16. Estonia
17. Finland
18. France
19. Germany
20. Greece
21. Iceland
22. India
23. Iran
24. Ireland
25. Ireland (Rep.)
26. Italy
27. Japan
28. Kyrgyzstan
29. Kuwait
30. Mexico
31. Netherlands
32. New Zealand
33. Norway
34. Palestine
35. Poland
36. Portugal
37. Russian Federation
38. Rwanda
39. Serbia
40. Singapore
41. Slovenia
42. South Africa
43. South Korea
44. Spain
45. Sweden
46. Switzerland
47. Taiwan
48. Turkey
49. United Arab Emirates
50. United Kingdom
51. United States
52. Vatican City
53. Venezuela

RDA by Sector

- Academics (66%)
- Private Sector (10%)
- Public Sector (17%)
- Unknown (7%)
RDA Plenaries as a Data Community “Town Square”

Emerging Plenary Format:

• **All-hands sessions:** Place for community networking and exchange of information (funding agencies, data organizations, key stakeholders)

• **Working sessions:** Face-to-face opportunities for global Interest Groups, Working Groups, and BOFs to meet and advance their agendas

• **Neutral meeting place:** Place for multiple groups to meet and form a common agenda and action plan (e.g. Plenary 2 Data Citation Harmonization Summit)

2014 RDA Plenaries:

• *Plenary 3* – Ireland / March 2014

• *Plenary 4* – Netherlands / September 2014
On the Horizon for the RDA (rd-alliance.org)

Ultimate Goal: Accelerate data sharing world-wide through targeted community infrastructure development

- Create / expand a pipeline of adopted infrastructure used by the community to increase data sharing and exchange
- Build “regional” communities and strength to address national issues (e.g. RDA / US engagement in public access and big data priorities in US)
- Build an effective organization that supports coordination and impact across the broader data community (e.g. data summits, engagement with G8 + 5, etc.)
Data Ecosystems

• **“Natural Resources”**
  – Data
  – Data collections and databases

• **Infrastructure**
  – **Software and hardware:** tools, systems, storage, data centers
  – **Social and organizational:** policy, practice, standards

• **Resource management** *(cross-cutting)*
  – Stewardship
  – Sustainability
  – Economic support
Frontier Challenges: Data “Governance”

- Many data ecosystems -- domain, sector, data cohort, national, etc. Each has a “data culture” and is subject to multiple interacting rules and influences
  - Community reward and collaboration / competition structure
  - Community policy and practice (ethics, rights, privacy)
  - National / international regulation, etc.
- How can we interoperate / harmonize between the cultures of distinct data communities?
  - How should we approach conflict resolution, ecosystem management, coordination of distinct cultures

Building a Sustainable Data Environment for Data-Driven Innovation

*Sustainable development:* "development that meets the needs of the present without compromising the ability of future generations to meet their own needs."

*Our Common Future, U.N. Brundtland Commission*

- **Key components**
  - Ecological sustainability
  - Cultural / institutional sustainability
  - Economic sustainability
  - Political sustainability

“We call for a common endeavor and for **new norms of behaviour at all levels** and in the interests of all. The **changes in attitudes and aspirations** that the report urges will depend on vast campaigns of **education, debate and public participation. ...**”

*Gro Harlem Brundtland*
Thank You
Happy 25th Anniversary to UCSD CSE!