



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



Physical Infrastructure

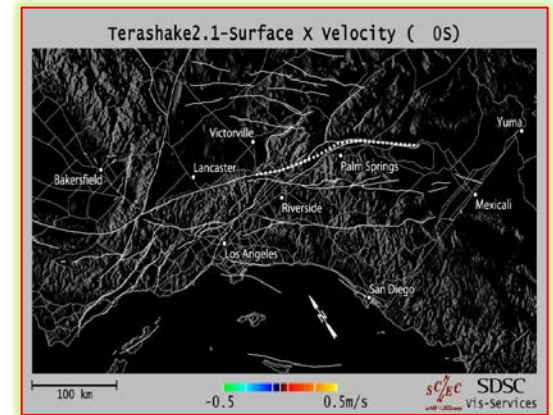
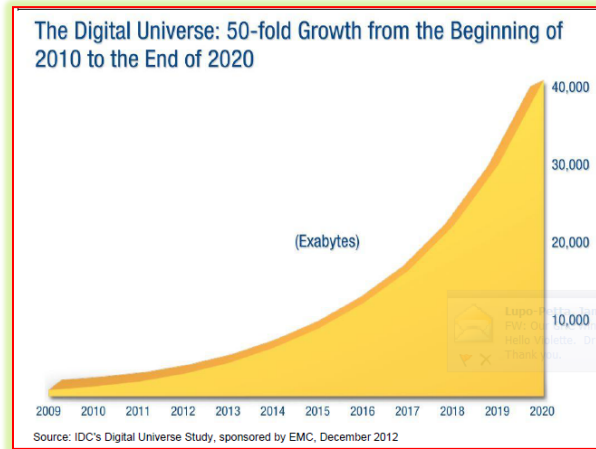
Entertainment



Commerce



Health



Research



Communication / Community

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

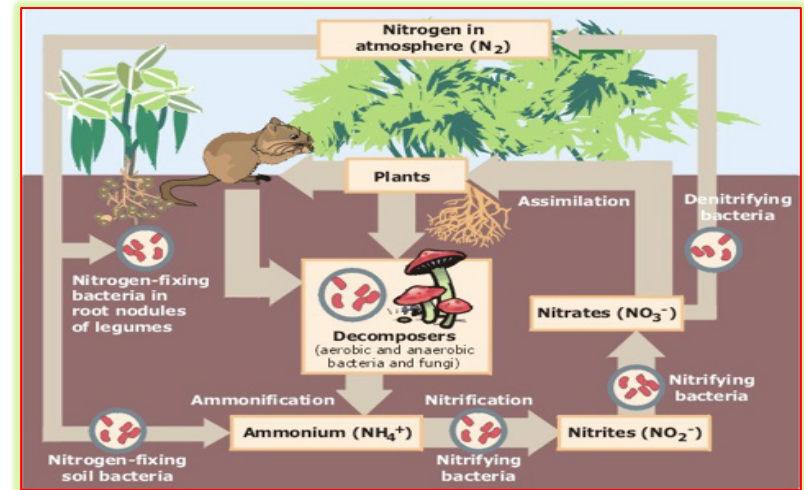


What's required to support data-driven innovation?

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

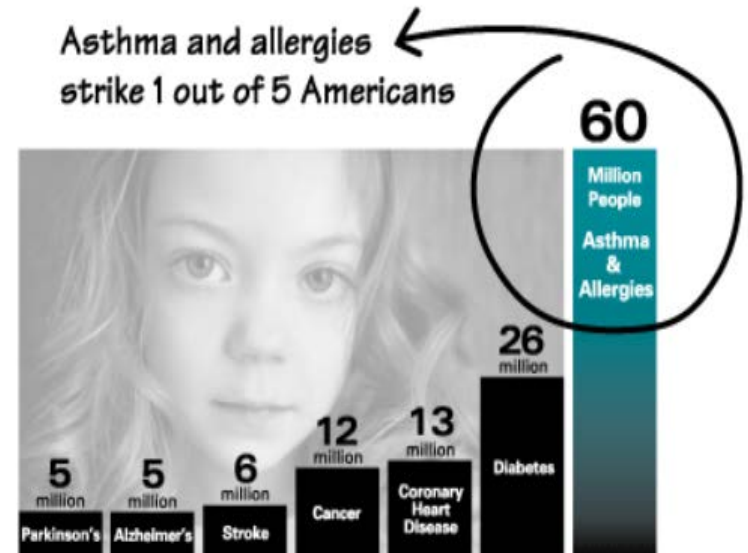


Data-Driven Health – Who is at risk for Asthma?

- **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, ...



Asthma and allergies strike 1 out of 5 Americans



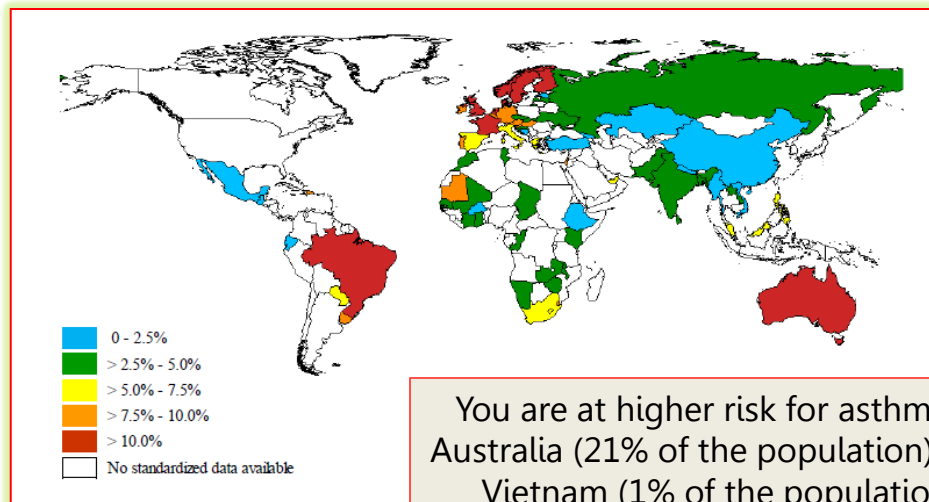
* Annual U.S. Prevalence Statistics for Chronic Diseases



Image, Kim Fortun, RPI; Graph: Asthma and Allergy Foundation of America,
<http://www.aafa.org/display.cfm?id=8&sub=42>

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 journal
WORLD ALLERGY ORGANIZATION

Search WAO Journal for

Home Articles Authors Reviewers About this journal My WAO Journal

Top
Abstract
Introduction
Methods
Results
Discussion

Original research **Open Access**

The Impact of a Program for Control of Asthma in a Low-Income Setting

Alvaro A Cruz^{1,*}, Adelmir Souza-Machado², Rosana Franco³, Carolina Souza-Machado², Eduardo V Ponte², Pablo Moura Santos² and Maurício L Barreto²

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%.



- **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



InformationWeek Healthcare Digital Bundle

Software Security Cloud Mobility Social Business Big Data Windows Global CIO Government Healthcare Education Financial SMB More

Electronic Medical Records Mobile & Wireless Clinical Information Systems Security & Privacy CPOE The Patient Leadership More Healthcare

HP ProLiant Gen8 servers powered by AMD Opteron® 6300 Series processors. The power of HP Converged Infrastructure is here. Now you can afford them. Watch video

NEWS
Sharing Psychiatry EHR Data Cuts Readmission Rates

Get InformationWeek Daily

Don't miss each day's hottest technology news, sent directly to your inbox, including occasional breaking news alerts.



OFFICE OF JUSTICE PROGRAMS

NATIONAL INSTITUTE OF JUSTICE
Research • Development • Evaluation

HOME | FUNDING & AWARDS | PUBLICATIONS & MULTIMEDIA | EVENTS | TRAINING | TOPICS

NIJ Home Page > NIJ Journal > NIJ Journal No. 267

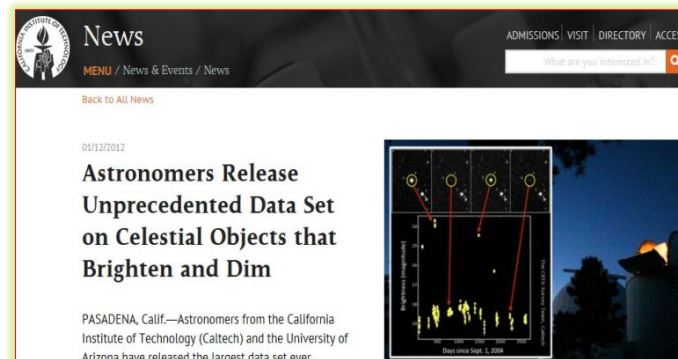
NIJ JOURNAL NO. 267

In Brief: Expanding Research by Sharing Data
by NIJ staff

Director's Message
NIJ makes data available for future research.

Police Use of Force: The Impact of Less-Lethal Weapons and Tactics

Toward a Better Way to Interview Child Victims of Sexual Abuse



News

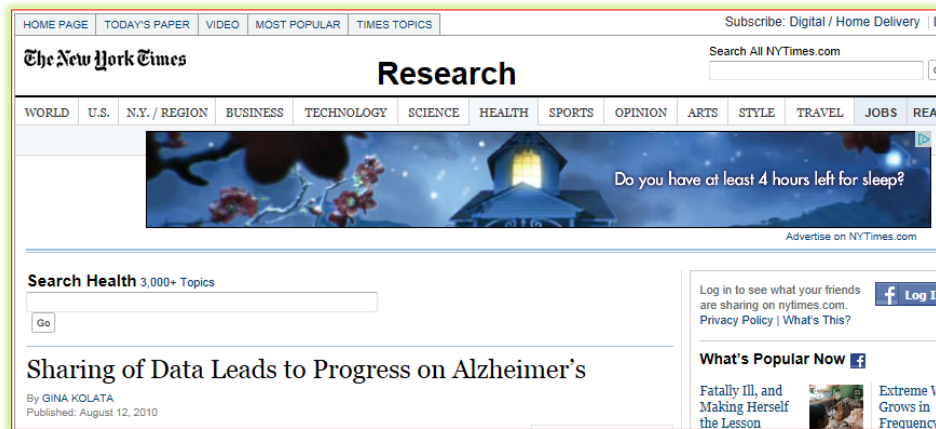
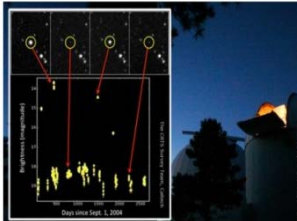
MENU / News & Events / News

Back to All News

01/12/2012

Astronomers Release Unprecedented Data Set on Celestial Objects that Brighten and Dim

PASADENA, Calif.—Astronomers from the California Institute of Technology (Caltech) and the University of Arizona have released the largest data set ever



HOME PAGE TODAY'S PAPER VIDEO MOST POPULAR TIMES TOPICS

Subscribe: Digital / Home Delivery

The New York Times

Research

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION ARTS STYLE TRAVEL JOBS REAL ESTATE

Do you have at least 4 hours left for sleep?

Search Health 3,000+ Topics

Log in to see what your friends are sharing on nytimes.com. Privacy Policy | What's This?

What's Popular Now

Fatally Ill, and Making Herself the Lesson

Extreme Weather Grows in Frequency

Sharing of Data Leads to Progress on Alzheimer's

By GINA KOLATA
Published: August 12, 2010



nature medicine

nature.com | journal home | archive | issue | news | abstract

ARTICLE PREVIEW

view full access options

NATURE MEDICINE | NEWS

日本語要約

The delay in sharing research data is costing lives

Josh Sommer

Research Data Sharing Ecosystem – Technical, social and organizational infrastructure

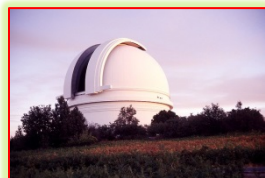


Data from birth to death / immortality: The Digital Research Data Life Cycle

Create

Data creation / capture / gathering from

- laboratory experiments
- fieldwork
- surveys
- devices
- simulation output ...



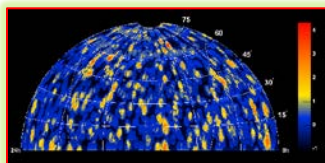
Edit

- Organize
- Annotate
- Clean
- Filter



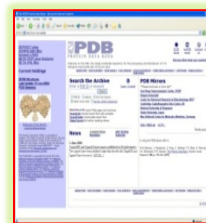
Use / Reuse

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



Publish, Disseminate

- Disseminate
- Create portals / data collections / databases
- Couple with literature

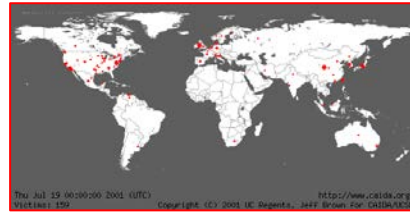
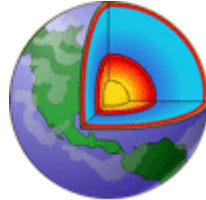
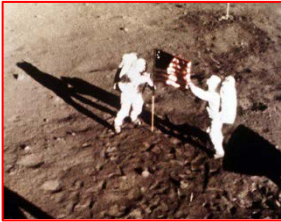


Preserve / Destroy

- Store / preserve
- Store / replicate / preserve
- Store / ignore
- Destroy



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**

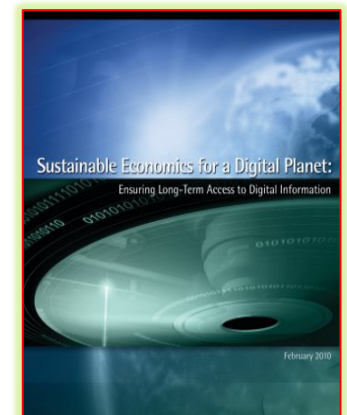
**Data
Preservation**

**Data
Storage**

**Sustainable
Economic Model**

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 Snively, Mike Norman et al.), ... etc.

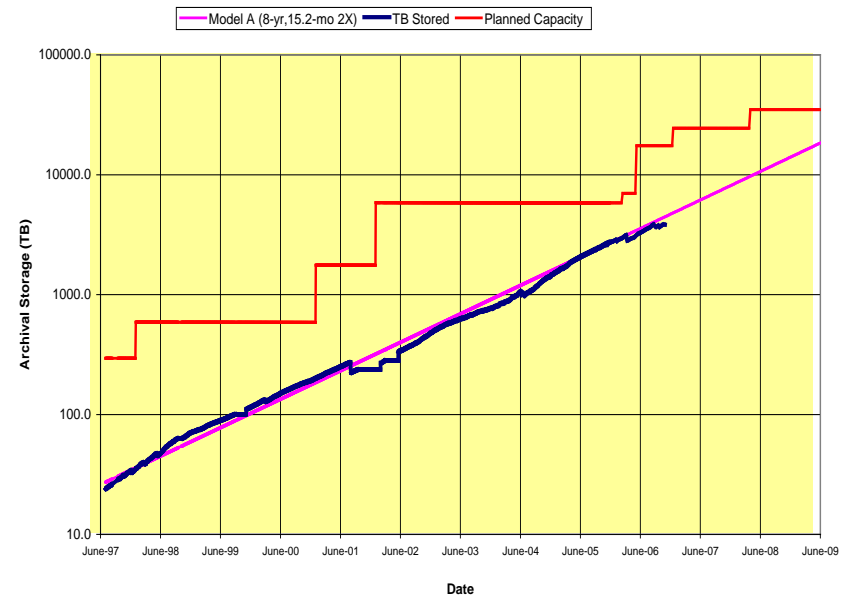


Data Infrastructure Pre-Supposes Viable Data Stewardship

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. ...

Resources and Resource Refresh



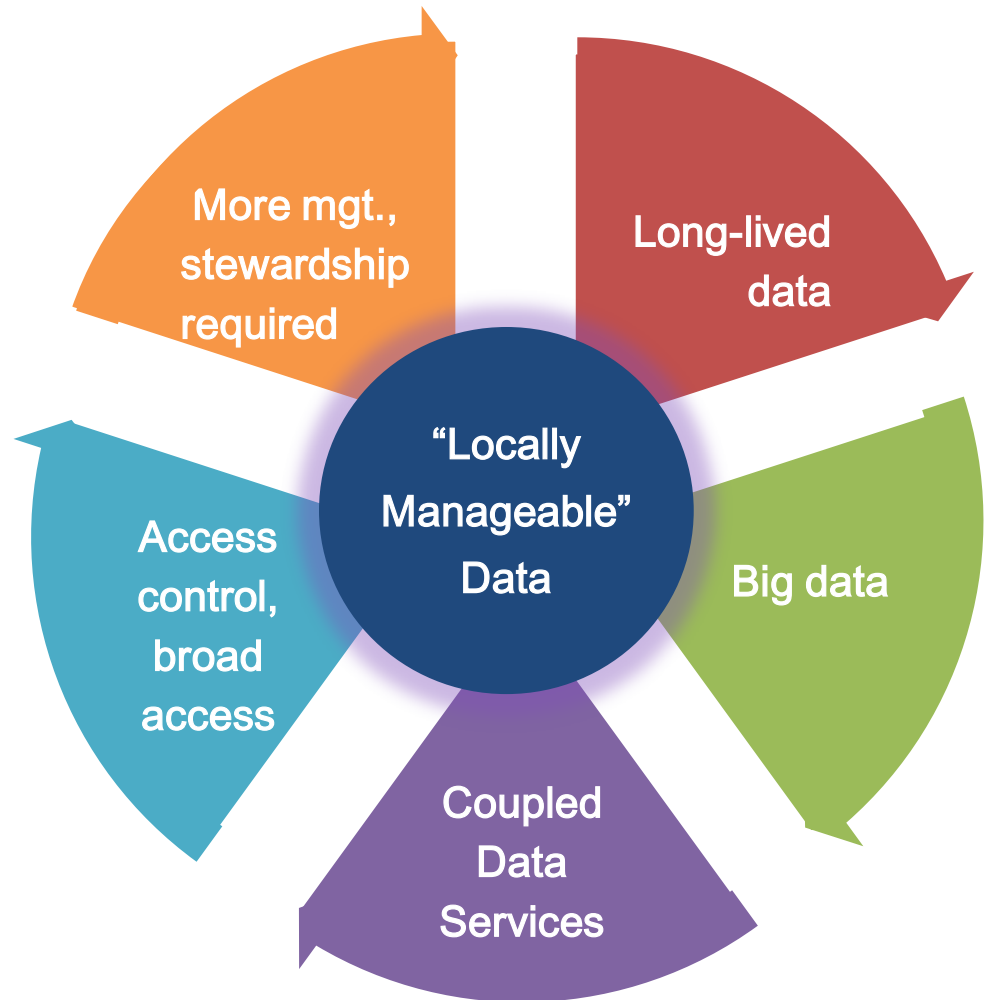
SDSC Data Storage Growth '97-'09

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

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 *JPH*
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 that 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

The New York Times

The Opinion Pages

WORLD U.S. N.Y. / REGION BUSINESS TECHNOLOGY SCIENCE HEALTH SPORTS OPINION

WE'RE READY TO WORK FOR YOU.

EDITORIAL

We Paid for the Research, So Let's See It

Published: February 25, 2013

The Obama administration is right to direct federal agencies to make public, without charge, all scientific papers reporting on research financed by the government. In a memorandum issued on Friday, John Holdren, the president's science adviser, directed federal agencies with more than \$100 million in annual research and development expenditures to develop plans for making the published results of almost all the research freely available to everyone within one year of publication.

Connect With Us on Twitter

For Op-Ed, follow @nytopinion and to hear from the editorial page editor, Andrew Rosenthal, follow @andyrNYT.



The agencies must submit plans to the [White House Office of Science and Technology Policy](#) within the next six months that will apply to both peer-reviewed scientific papers and digital manuscripts and supporting data.

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

FACEBOOK

TWITTER

GOOGLE+

SAVE

E-MAIL

SHARE

PRINT

REPRINTS

THE WAY BACK
WATCH TRAILER

Data Economics: Who Pays the Bill?

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

February

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: John P. Holdren *JPH*
Director

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

1. Policy Principles

The Administration is committed to ensuring that the constraints possible and consistent with law and policy on the use of federally funded scientific research are made available to the scientific community. Such results include publications, data, and other information.

Scientific research supported by the Federal Government drives our economy. The results of that research are used for progress in areas such as health, energy, the environment, and national security.

POLICYFORUM

SCIENCE PRIORITIES

Who Will Pay for Public Access to Research Data?

Francine Berman¹ and Vint Cerf²

On 22 February, the U.S. Office of Science and Technology Policy (OSTP) released a memo calling for public access for publications and data resulting from federally sponsored research grants (1). The memo directed federal agencies with more than \$100 million R&D expenditures to “develop a plan to support increased public access to the results of research funded by the Federal Government.” Perhaps even more succinctly, a subsequent *New York Times* opinion page sported the headline “We Paid for the Research, So Let’s See It” (2). So who pays for data infrastructure?

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, federal R&D agencies are working hard to foster approaches



Research data of community value are supported today in a variety of ways. Some of them, like those in the Protein Data Bank (PDB) (3)—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

When economic incentives are not in place, federally funded research data are “at risk.”

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 not just for the duration of the project but for the next decade or more to collaborators and a broader community of researchers. 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.

Up to this point, no one sector has stepped

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



www.sciencemag.org on August 29, 2013



Article: *Science Magazine*, August 9, 2013. Free public access link at <http://www.cs.rpi.edu/~bermaf/>




F#\$*#ing brilliant - article by Cerf & Berman "Who Will Pay for Public Access to Research Data?" costs \$20 to read <http://t.co/UOZwGuiXr4>

@phylogenomics 9 hours ago 13 Follow @phylogenomics

93 retweets | 10 replies

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.)



The Charleston Ballet 2010-2011
REFLECTIONS

**SPONSOR PERFORMING ART
RECEIVE A WEST VIRGINIA TAX CREDIT**

Sponsor the Charleston Ballet, receive tickets, program ad and a West Virginia Tax Credit, too!
You or your business can receive a West Virginia Neighborhood Investment Program (NIP) tax credit equal to one-half of your sponsorship of \$500 or more. This includes show tickets (October and March) and program advertising, if you choose it.
Program press deadlines are fast approaching, so call the Charleston Ballet Office at 304.342.6541 for details and to secure your sponsorship. Tax Credits are limited and available on a first-come, first-served basis.



GRACE HOPPER CELEBRATION OF WOMEN IN COMPUTING
Presented by the Anita Borg Institute for Women and Technology and the Association for Computing Machinery

CONFERENCE * PARTICIPATE * SPONSORSHIP * K-12 COMPUTING TEACHERS WORKSHOP * COMMUNITY

Current Sponsors

PLATINUM CORPORATE SPONSORS

amazon BROADCOM ca technologies
CISCO facebook Google hp
IBM intel intuit
LOCKHEED MARTIN Microsoft
NetApp



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

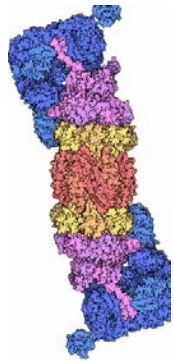
The screenshot shows the University of Maryland Medical Center's website with a search bar at the top right. The main navigation bar includes links for Patients & Visitors, Centers & Services, Health Information, Research & Clinical Trials, For Health Professionals, and News & Events. The page title is "Drug Interaction Tool". The tool interface includes a search box for drug names, a "Search" button, and an "Advanced" checkbox. Below the search box, there are two sections: "Drugs searched for" and "Drugs to check interactions", each with a list of drug names and "Add" and "Remove" buttons. A "Check Interactions" button is located at the bottom of the tool. The page also contains a disclaimer and copyright information.

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



PROTEIN DATA BANK

Creation / adoption of **data sharing policies** have accelerated research innovation

Development and adoption of **parallel communication protocols** through the **MPI Forum** drove a generation of advances



why not change the world? SM



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



I E T F[®]

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

*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

Falk Amelung, University of Miami, USA (GEO task lead)
 Craig Dobson, NASA and Committee of Earth Observation Satellites (CEOS)
 Rui Fernandes, EROS and ELREE, <rmanuel@di.ubi.pt>

Science, Humanities, Arts
 Communities



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



Libraries, Archives,
 Repositories, Museums



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



Department of Science & Technology
 Ministry of science & Technology
 Government of India

Fran Berman

Data
 Scientists



why not change the world? SM

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, ...



Harmonized standards

Policy and Practice



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 --

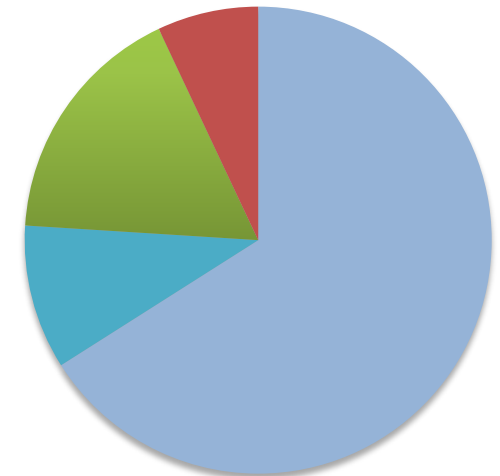
Groups that Met at RDA Plenary 2 in DC

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

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%)

Fran Berman

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)

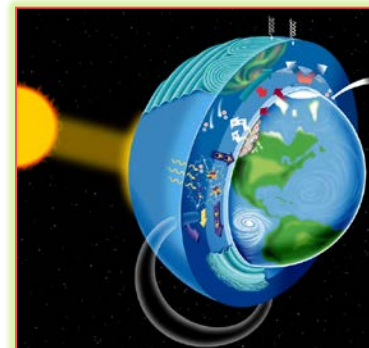
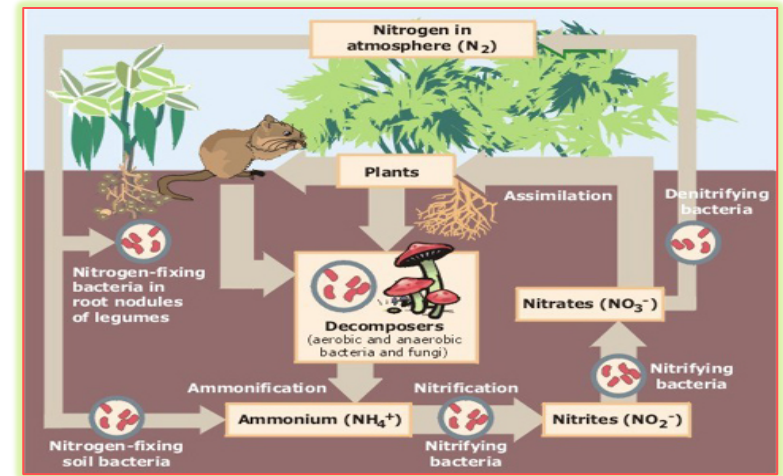


Ultimate Goal: Accelerate data sharing world-wide through targeted community infrastructure development

- **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.)

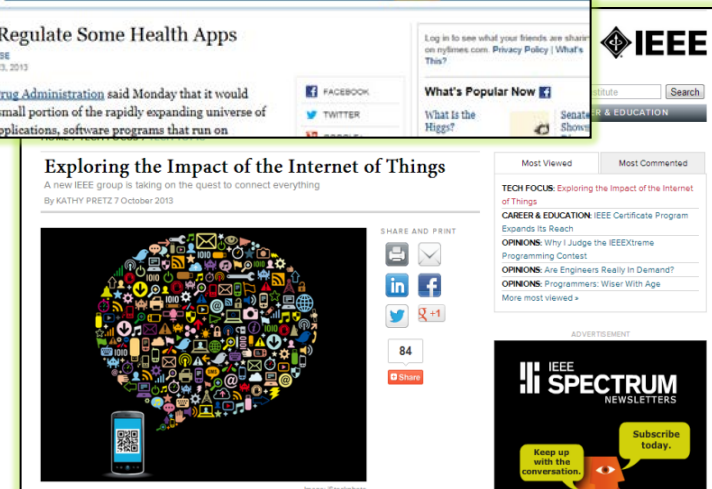
Data Ecosystems

- “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



Articles: Huffington Post Tech section, June 26,, 2013; IEEE The Institute, October 7, 2013;
New York Times Health Section September 23, 2013

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!

