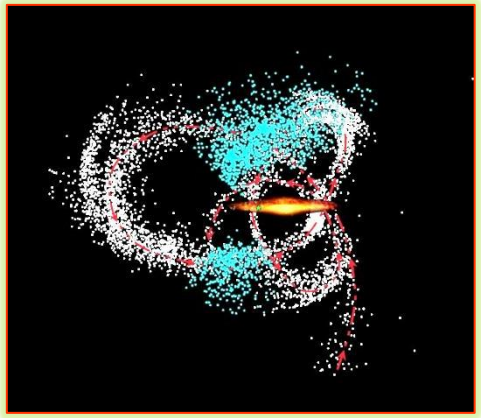
Got Data?
Building a Global Ecosystem for Data-Driven Innovation

Dr. Francine Berman

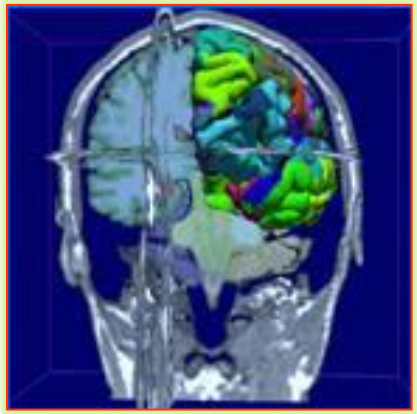
Chair, Research Data Alliance / US

Edward P. Hamilton Distinguished Professor in Computer Science, RPI

Data Drives 21st Century Innovation

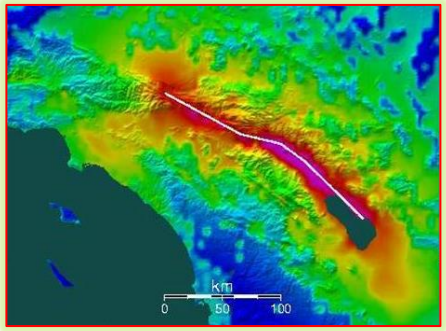


Research Insights



Transformative strategies for disease treatment and well-being

Safety and Security



amazon.com®

WebMD℠

Broad spectrum of goods and services

Efficient Physical Infrastructure

Rensselaer

Data infrastructure necessary for data-driven innovation

Data discoverability tools

Data access via portals, science gateways, etc.

Database and data collection systems

Data services to support use and re-use

Data analysis algorithms, data-driven models and simulations

Data visualization tools

Semantic frameworks

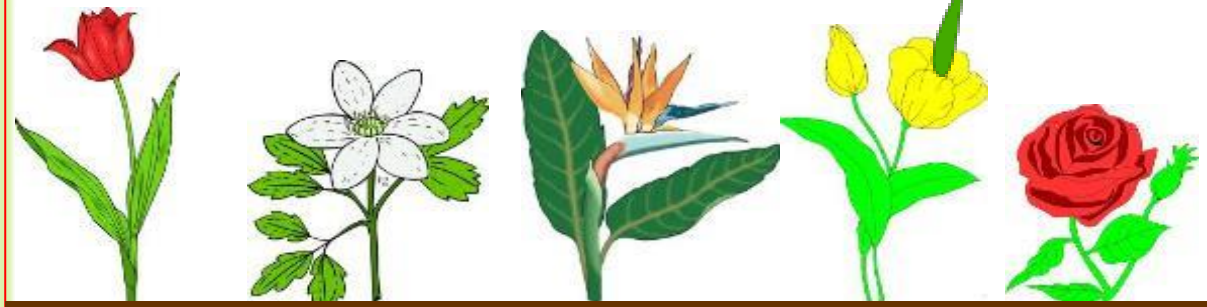
Data management systems

Data storage

Expert assistance for data-oriented applications, …

What can we do to accelerate the development of effective data infrastructure needed for 21st century innovation?

Data-driven Innovation



Data Infrastructure

Technical Infrastructure
SW and systems, Tools and algorithms, Hardware and facilities

Social Infrastructure
Policy, Practice, Standards, Rights, Community culture

Fran Berman

What is needed for effective infrastructure?

"Historical infrastructures – [transportation], electrical grids, railways, telephony and most recently the Internet – become ubiquitous, accessible, reliable, and transparent as they mature."

Understanding Infrastructure: Dynamics, Tensions, and Design, January 2007

Systems Interoperability

Sustainable Economics

Policy

Community Practice

Common Standards

Rensselaer

Fran Berman

Traffic Image: Mike Gonzalez

Today's Presentation: Emerging Efforts in the Development of Effective Data Infrastructure

Global Data Infrastructure

How do we accelerate open access data sharing and exchange?

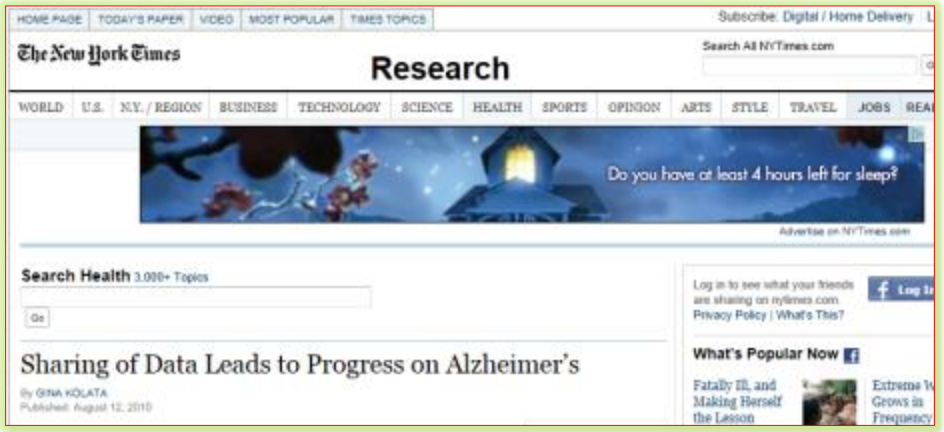
National Data Infrastructure

How do we support stewardship and preservation of publicly accessible research data?

Rensselaer

Fran Berman

Data Sharing Drives New Discovery



Rensselaer

Fran Berman

Data Sharing is 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, EPOS and EUREF, <rmanuel@di.ubi.pt>

Science, Humanities, Arts Communities

Libraries, Archives, Repositories, Museums

DATA.GOV
EMPOWERING PEOPLE

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

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

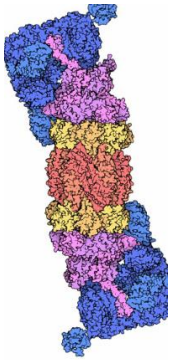
Department of Science & Technology
Ministry of science & Technology
Government of India

Data Scientists

Communities Coming Together to Build Infrastructure and Increase Impact

IETF®

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



PROTEIN DATA BANK

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

THE LANCET Neurology

Search for: All Fields

Home | Journals | Specialties | Clinical | Global Health | Audio | Conferences

The Lancet Neurology, Volume 6, Issue 6, Pages 480 - 481, June 2007
doi:10.1016/S1474-4422(07)70118-7 Cite or Link Using DOI

« Previous Arti

Alzheimer's researchers open the doors to data sharing

James Butcher

Data from the US Alzheimer's Disease Neuroimaging Initiative will be made publicly available in a central datab

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

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

Rensselaer

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*

CREATE → ADOPT → USE

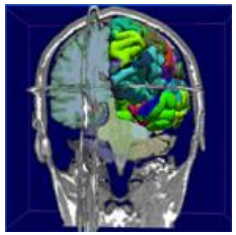
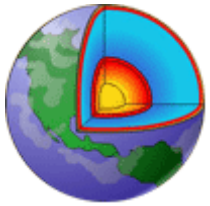
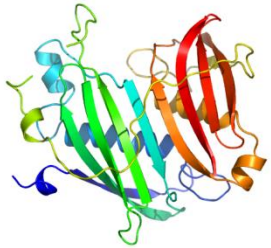
RDA Members come together as

- **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

- **Tools, policy, software 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 substantial applicability** to groups within the data community, but may not apply to everyone
- **Efforts for which working scientists and researchers can start today**

Broad spectrum of infrastructure needed to support data sharing

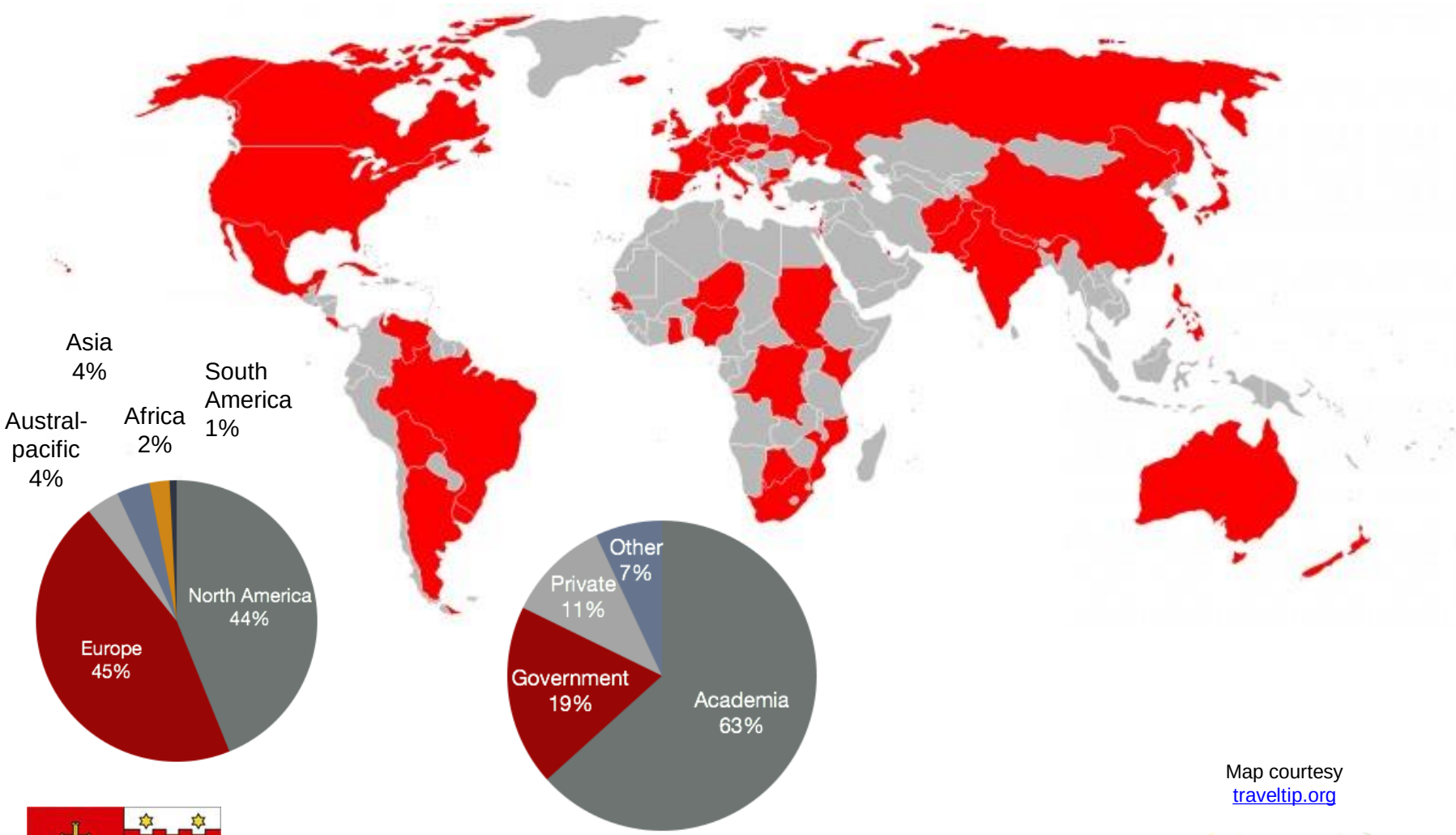




Technical and Social Data Infrastructure

- Data Services
- Data Analytics
- Data Sharing Practice
- Data Management Systems
- Digital Object Identifiers
- Common Metadata Standards
- Data Citation Standards
- Data Access and Distribution Policy
- Data Discoverability Tools
- Preservation Policy
- Data Storage
- Sustainable Economic Model

Images: NASA, SDSC / UCSD

The RDA Community (as of 2/14): Over 1300 members from 64+ countries



Map courtesy
traveltip.org

Fran Berman

RDA Plenaries: Venue for community building and WG / IG progress

RDA Plenary 1 / Launch

- March 2013 in Gothenburg, Sweden
- 240 participants
- 3 WG, 9 IG

RDA Plenary 2

- September 2013 in Washington, DC
- 380 participants
- 6 WG, 17 IG, 5 BOF
- Data Citation Summit co-located in RDA "neutral space"
- First Organizational Assembly meet-up

RDA Plenary 3

- March 26-28, 2014 in Dublin, Ireland
- Hosted by Australia and Ireland

RDA Plenary 4

- September 2014 in The Netherlands

RDA Plenaries Emerging as a Data Community "Town Square"

Emerging Plenary Format:

- **Working sessions:** Face-to-face opportunities for global Interest Groups, Working Groups, and BOFs to meet and advance their agendas
- **All-hands sessions:** Place for community networking and exchange of information (funding agencies, data organizations, key stakeholders)
- **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)

Community-Driven RDA Groups by Focus (as of 12/13, more at Plenary 3)

Domain Science - focused

- Toxicogenomics Interoperability IG
- Structural Biology IG
- Biodiversity Data Integration IG
- Agricultural Data Interoperability IG
- Digital History and Ethnography IG
- Defining Urban Data Exchange for Science IG
- Marine Data Harmonization IG
- Materials Data Management IG

Community Needs - focused

- Community Capability Model IG
- Engagement IG
- Clouds in Developing Countries IG

Reference and Sharing - focused

- Data Citation IG
- Data Categories and Codes WG
- Legal Interoperability IG

Data Stewardship - focused

- Research Data Provenance IG
- Certification of Digital Repositories IG
- Preservation e-infrastructure
- Long-tail of Research Data IG
- Publishing Data IG
- Domain Repositories IG
- Global Registry of Trusted Data Repositories and Services IG

Base Infrastructure - focused

- Data Foundations and Terminology WG
- Metadata Standards WG
- Practical Policy WG
- PID Information Types WG
- Data Type Registries WG
- Metadata IG
- Big Data Analytics IG
- Data Brokering IG

First RDA Infrastructure Deliverables

Scheduled to Complete Summer 2014

Data Type Registries WG

- **Deliverables:** System of data type registries, formal model for describing types, working model of a registry.
- **Initial Adopters and Users:** CNRI, International DOI Foundation, Deep Carbon Observatory

Practical Code Policies

- **Deliverables:** Survey of policies in production use, testbed of machine actionable policies, deployment of 5 policy sets, policy starter kits
- **Initial Adopters and Users:** RENCI, DataNet Federation Consortium, CESNET, Odum Institute

Persistent Identifier Information Types

- **Deliverables:** Minimal set of PID types, API
- **Initial Adopters and Users:** Data Conservancy, DKRZ

Scheduled to Complete Fall 2014

Language Codes

- **Deliverables:** Operationalization of ISO language categories for repositories.
- **Initial Adopters and Users:** Language Archive, Paradisec

Data Foundations and Terminology

- **Deliverables:** Common vocabulary for data terms, formal definitions and open registry for data terms
- **Initial Adopters and Users:** EUDAT, DKRZ, Deep Carbon Observatory, CLARIN, EPOS

Metadata Standards

- **Deliverables:** Use cases and prototype director of current metadata standards starting from DCC directory
- **Initial Adopters and Users:** JISC, DataOne

RDA Medium-term (3-5 years) Strategic Goals

- **Create a pipeline of data sharing infrastructure efforts**
  - that are adopted and used by communities during their development
  - that increase their impact through greater adoption over time
- **Build and expand the research data community for effective impact**
  - globally, regionally, and within constituent groups
- **Evolve as a useful, relevant, and agile organization**
  - that helps the community capitalize on opportunity and respond to challenges within the data community

RDA/US: Collaborate Globally, Contribute Locally

RDA/US Goals:

- ⬅ **Contribute to RDA "international" efforts and leadership**
- ⬅ **Bring US efforts to broader RDA community**
- ➡ **Build the RDA community within the US**
- ➡ **Leverage and implement RDA deliverables in the US to amplify impact**
- ↔ **Collaborate closely with other RDA "regions" on key programs and initiatives**

Open Positions

http://www.cs.rpi.edu/~bermaf/

1) RDA/US Managing Director
2) Communications Specialist

NSF-supported RDA/US initiatives:

- Outreach (RDA → RDA/US)
- RDA Deliverables Amplification
- Student / Early Career Engagement

RDA/US Steering Committee

- Fran Berman, RPI
- Larry Lannom, CNRI
- Mark Parsons, RPI
- Beth Plale, IU

Rensselaer

Fran Berman

(U.S.) National Infrastructure for Data-Driven Innovation

Global Data Infrastructure

How do we accelerate open access data sharing and exchange?

National Data Infrastructure

How do we support stewardship and preservation of valuable research data?

Fran Berman

Increasing U.S. Federal R&D Agency Requirements for Data Management

HOME | FUNDING | AWARDS | DISCOVERIES | NEWS | PUBLICATIONS | STATISTICS | ABOUT NSF | FASTLANE

National Science Foundation
Directorate for Engineering (ENG)

QUICK LINKS

SEARCH

ENG HOME | ENG FUNDING | ENG AWARDS | ENG DISCOVERIES | ENG NEWS | ABOUT ENG

Engineering

design element

ENG Home

NSF Data Management Plan Requirements

Beginning January 18, 2011, proposals submitted to NSF must include a supplementary document of no more than two pages labeled "Data Management Plan" (DMP) . This supplementary document should describe how the proposal will conform to NSF policy on the dissemination and sharing of research results. Proposals that do not include a DMP will not be able to be submitted. For more information about this new requirement, see the Grant Proposal Guide, Chapter II.C.2.j and the Data Management and Frequently Asked Questions(FAQs).

Note: the Engineering Directorate (ENG) has additional guidance for proposals submitted to ENG programs, http://nsf.gov/eng/general/ENG_DMP_Policy.pdf. Questions or suggestions about this new requirement may be addressed to Dr. Maria K. Burka mburka@nsf.gov.

Summary: Office of Science Statement on Digital Data Management

Requirements (1 of 3)

- To integrate data management planning into the overall research plan, **all proposals submitted to the Office of Science for research funding are required to include a Data Management Plan (DMP)** of no more than two pages that describes how data generated through the course of the proposed research will be shared and preserved or explains why data sharing and/or preservation are not possible or scientifically appropriate. At a minimum, DMPs must describe how data sharing and preservation will enable validation of results, or how results could be validated if data are not shared or preserved.

In NIH's view, all data should be considered for data sharing. **Data should be made as widely and freely available as possible while safeguarding the privacy of participants, and protecting confidential and proprietary data.** To facilitate data sharing, investigators submitting a research application requesting $500,000 or more of direct costs in any single year to NIH on or after October 1, 2003 are expected to include a plan for sharing final research data for research purposes, or state why data sharing is not possible.

U.S. DEPARTMENT OF ENERGY | Office of Science

6

Fran Berman

Management, Access, and Use of Data Now and in the Future *Presupposes* Adequate Data Stewardship and Preservation Today

- **Stewardship and Preservation fundamentall:**
  "Homeless" data ceases to exist

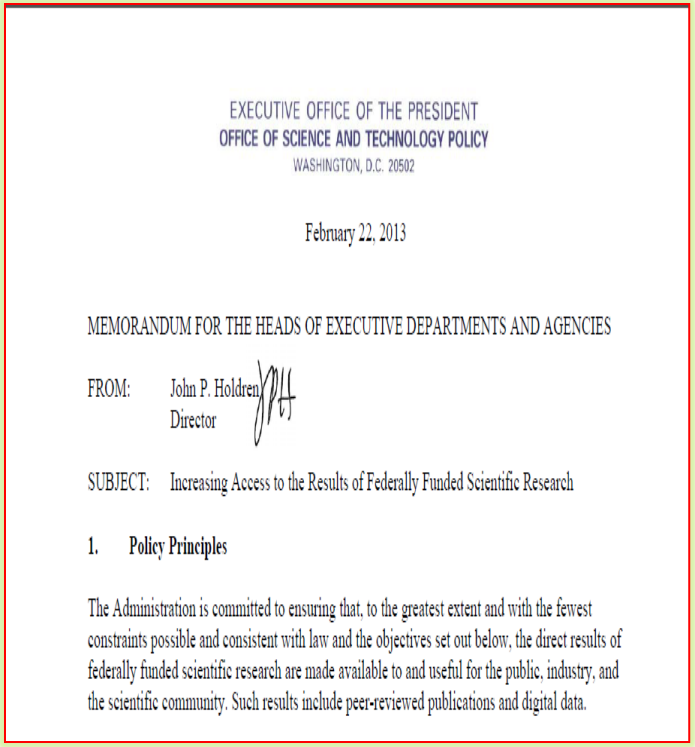
- Economically sustainable data infrastructure necessary to support
  - Federally mandated data management plans
  - Public access to research data
  - Use and re-use
  - Reproducibility

- The "bigger", longer-term, more complex, or more valuable the data is, the greater the importance of sustainable data infrastructure

February 2013 OSTP Memo: New Policies for Public Access to Research Publications *and Data*

Agency Plan Requirements:

- Strategy for **capitalizing on what exists** and **fostering public-private partnerships** with scientific journals
- Strategy for increasing / enhancing **discoverability, access, dissemination, stewardship, preservation**
- Approach for **measuring and enforcing compliance**
- **No new money:** "Identification of resources within the existing agency budget to implement the plan"



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.

Rensselaer

Fran Berman

The New York Times

The Opinion Pages

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

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.

FACEBOOK
TWITTER
GOOGLE+
SAVE
E-MAIL
SHARE
PRINT
REPRINTS

THE WAY WAY BACK
WATCH TRAILER

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

Opinion

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 year for a subscription and $30 or more for an individual copy. That is simply too much for many people and small businesses to afford.

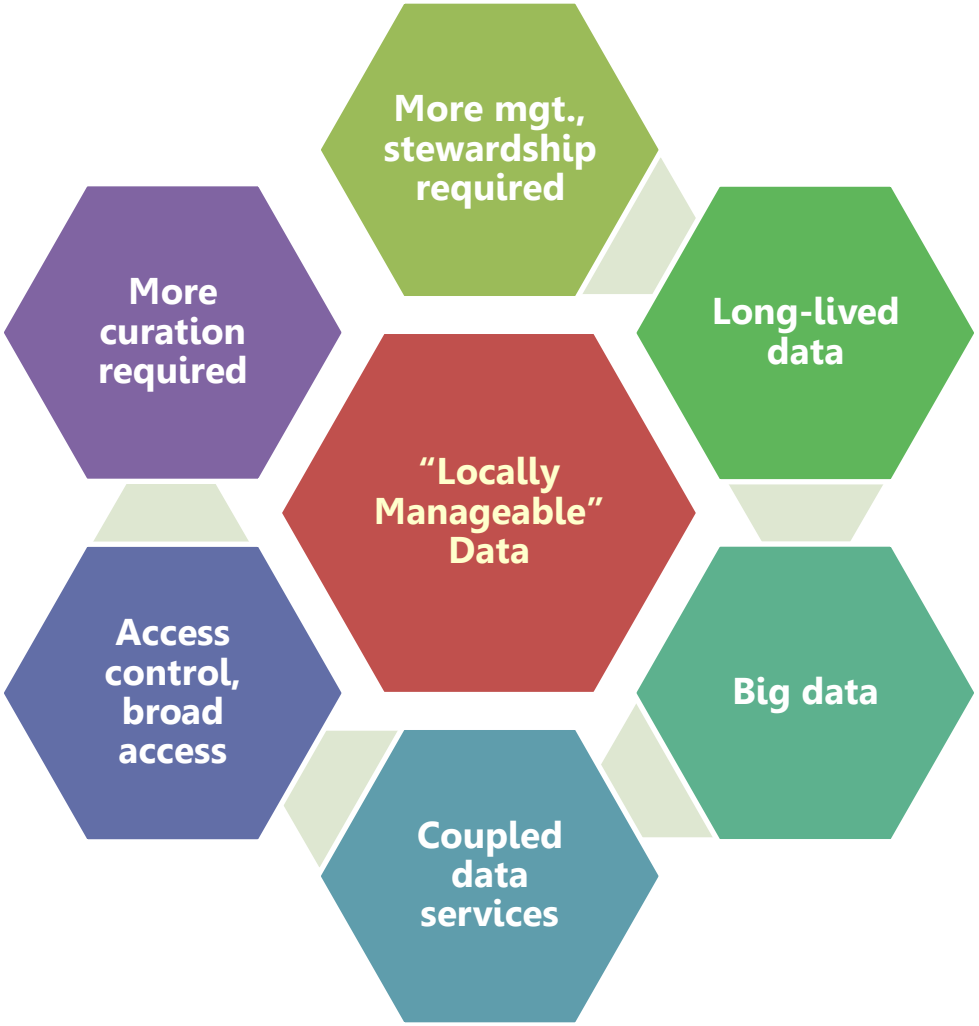
Rensselaer

Fran Berman

Public Access to data requires a *viable business model* for sustaining its underlying infrastructure

Data infrastructure costs increase with usage, stewardship and access requirements, perceived value

Greater costs away from the center (including "big" data) …



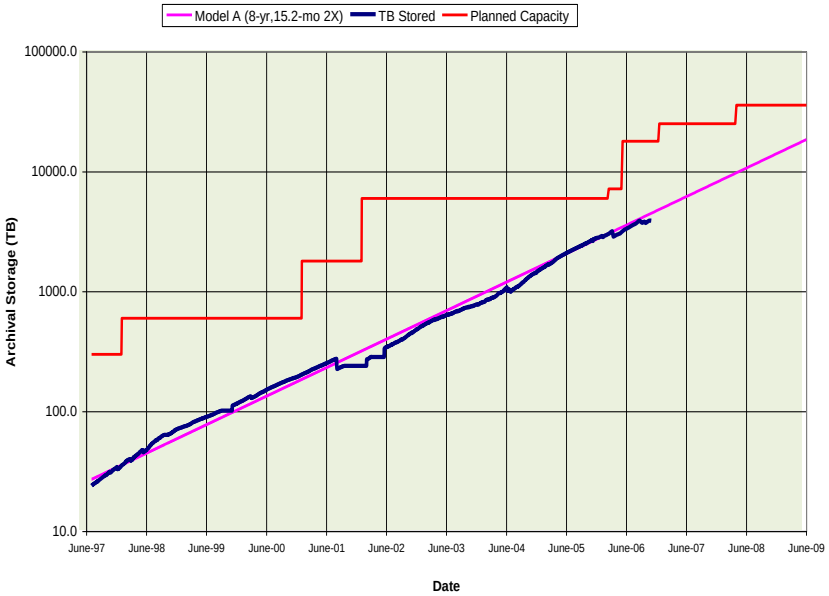
Fran Berman

It's not just about the cost of storage

Data Infrastructure costs 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*

Information courtesy of Richard Moore, SDSC

Research data infrastructure has a Stakeholder Problem

- **Many Stakeholders in digital access and preservation**
  - **Stakeholders who benefit** from use of the preserved asset
  - **Stakeholders who generate / create / select** what to preserve
  - **Stakeholders who own / have rights** to the asset
  - **Stakeholders who manage / host / preserve** the asset
  - **Stakeholders who pay**

- **The greater the alignment between key stakeholder groups, the better the prospect for sustainable preservation**

Economics of Public Access: Who Pays the Data Bill?

POLICYFORUM

SCIENCE PRIORITIES

Who Will Pay for Public Access to Research Data?

When economic models and infrastructure are not in place to ensure access and preservation, federally funded research data are "at risk."

Francine Berman[1] and Vint Cerf[2]

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 to public access, to assess needs for supporting partnerships and enabling infrastructure, and to develop timetables and approaches for implementation. We focus here on the research data portion of the OSTP memo.

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 National Institutes of Health (NIH), and the U.S. Department of Energy for the Research Collaboratory for Structural Bioinformatics (RCSB) PDB is $6.3 million annually.) Other data, as from the Longitudinal Study

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 up to take on the problem alone. unrealistic to expect as much. In t sector, federal R&D agencies are to allocate enough resources to s federally funded research data. Th

Digital Repository
@dri_ireland

Follow

Berman and Cerf "Who will pay for public access" behind paywall :(
m.sciencemag.org/content/341/61...
#ipres2013 #irony

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

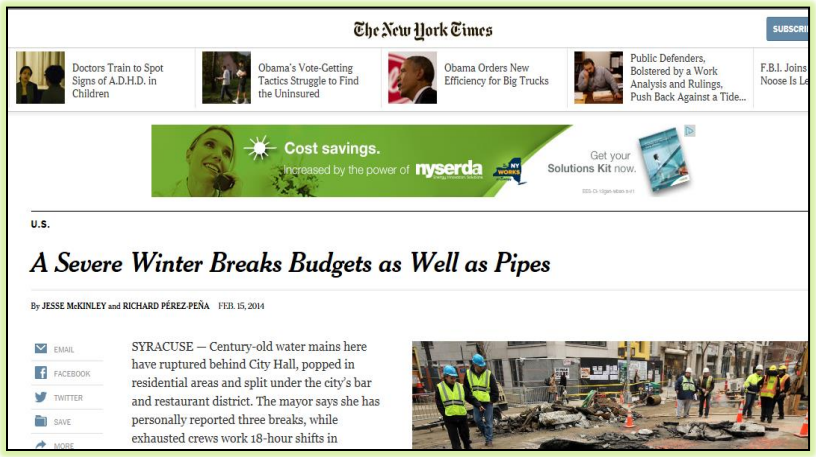
Op-Ed Recommendations: Cultivate Preservation and Stewardship Options in Every Sector

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.)

Charleston Ballet blog: http://allianceblog.org/tag/charleston-ballet/ ; iTunes gift card

Why aren't we there yet?

- **Infrastructure investments are often a hard sell**
  - Business model must be sustainable and address infrastructure refresh and evolution
  - Quantifying ROI a challenge
  - Hard to "market" compared to more newsworthy competing priorities

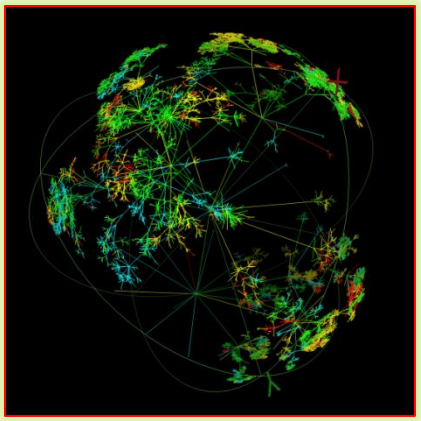


Stephanie A. Miner, the Syracuse mayor, said [infrastructure is] too often overlooked when politicians want to spend money on economic development. ***"You don't cut ribbons for new water mains, but that's really what matters."***

NY Times, Feburary 15, 2014

Making the Case for Sustainable Stewardship is More Important Than Ever

The Research landscape is changing



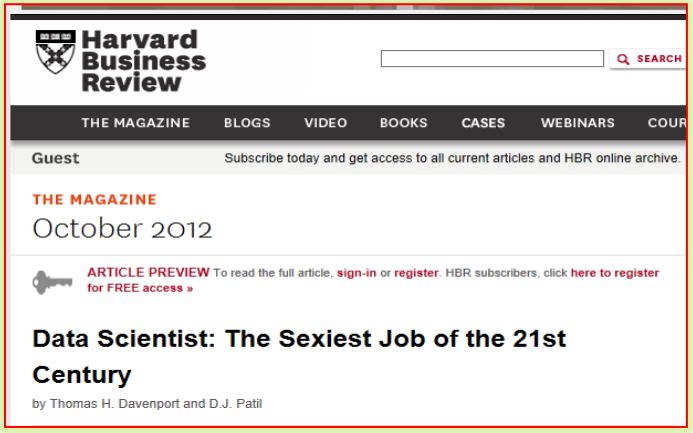
- Data is accelerating new innovation and discovery
- Greater need for access, ease-of-use, interoperability of data
- Traditional modes of research recognition evolving: new approaches to collaboration / competition, publication, citation, analysis all involve digital data

The Educational landscape is changing

- University curricula becoming more data-driven
- Increasing integration of on-line / on-site options supported by data infrastructure
- More digital monitoring, tracking, accountability needed; more policy and regulation involving digital data

The Workforce is changing

- More data literacy required from everyone
- More data science embedded in everything
- Data scientists increasingly critical for competitiveness and leadership



Harvard Business Review

SEARCH

THE MAGAZINE BLOGS VIDEO BOOKS CASES WEBINARS COUR

Guest Subscribe today and get access to all current articles and HBR online archive.

THE MAGAZINE

October 2012

ARTICLE PREVIEW To read the full article, sign-in or register. HBR subscribers, click here to register for FREE access »

Data Scientist: The Sexiest Job of the 21st Century

by Thomas H. Davenport and D.J. Patil

Image: CAIDA Internet visualization;
Article: HBR October 2012

Thank you!

Fran Berman