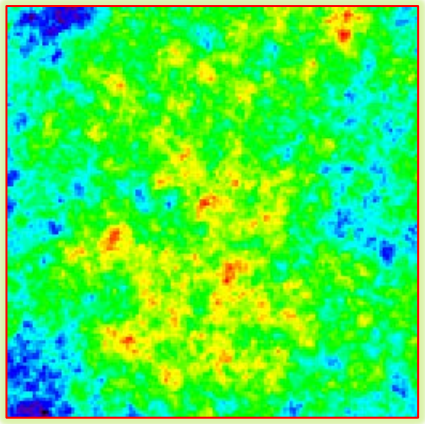
Got Data?
Sustaining the Engine for Data-Driven Innovation

Dr. Francine Berman

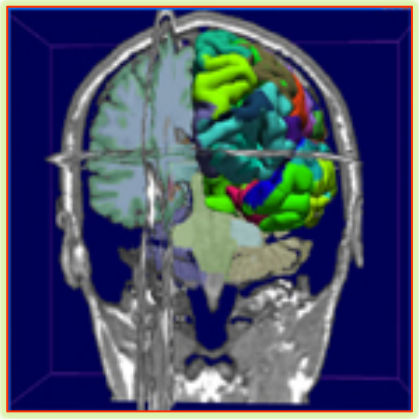
Chair, Research Data Alliance / US

Edward P. Hamilton Distinguished Professor in Computer Science, RPI

Data Drives 21st Century Innovation

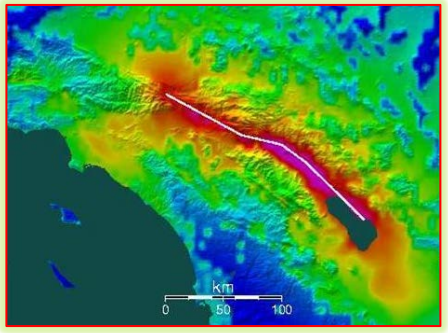


Research Insights



Transformative strategies for disease treatment and well-being

Safety and Security



amazon.com

WebMD

Broad spectrum of goods and services

Efficient Physical Infrastructure

Rensselaer

Animation of ENZO courtesy of Mike Norman

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

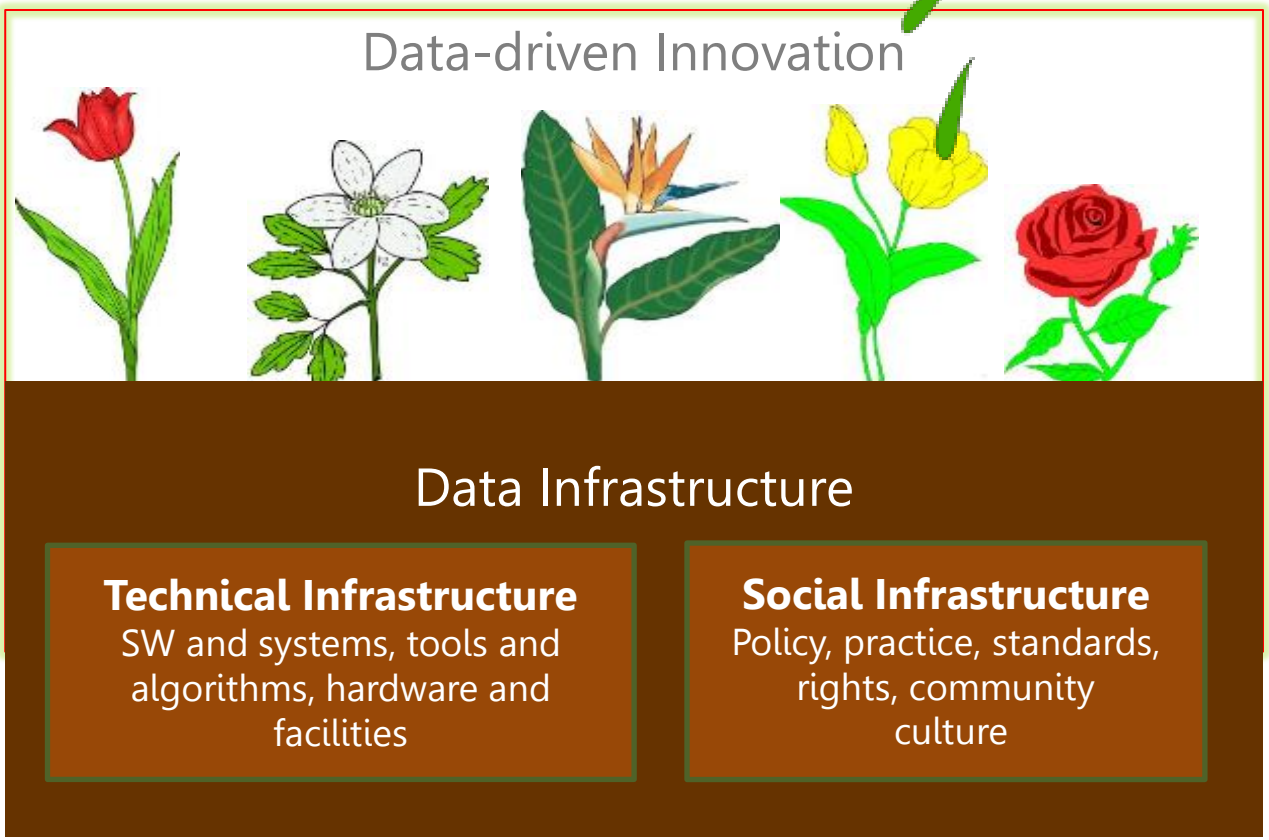
Semantic frameworks

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?



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 Research Data Infrastructure

Global Data Infrastructure

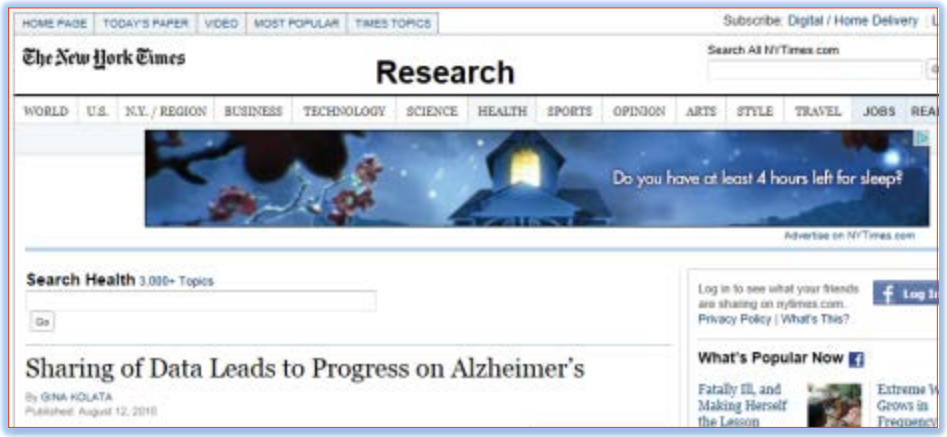
How do we accelerate open access data sharing and exchange?

National Data Infrastructure

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

Rensselaer

Data Sharing Driving New Discovery and Innovation



Rensselaer

Fran Berman

Data Sharing is a Global Challenge

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, ...

Data Scientists

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

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

Communities Efforts Can 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 **standards** that support innovation

NIST and industry development of **common standards** and **community practice** creating a more efficient and reliable energy grid.

SGIP
SMART GRID INTEROPERABILITY PANEL

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

NIH **policy** has helped create ADNI **public access** data repository accelerating insights and discovery about Alzheimer's disease

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

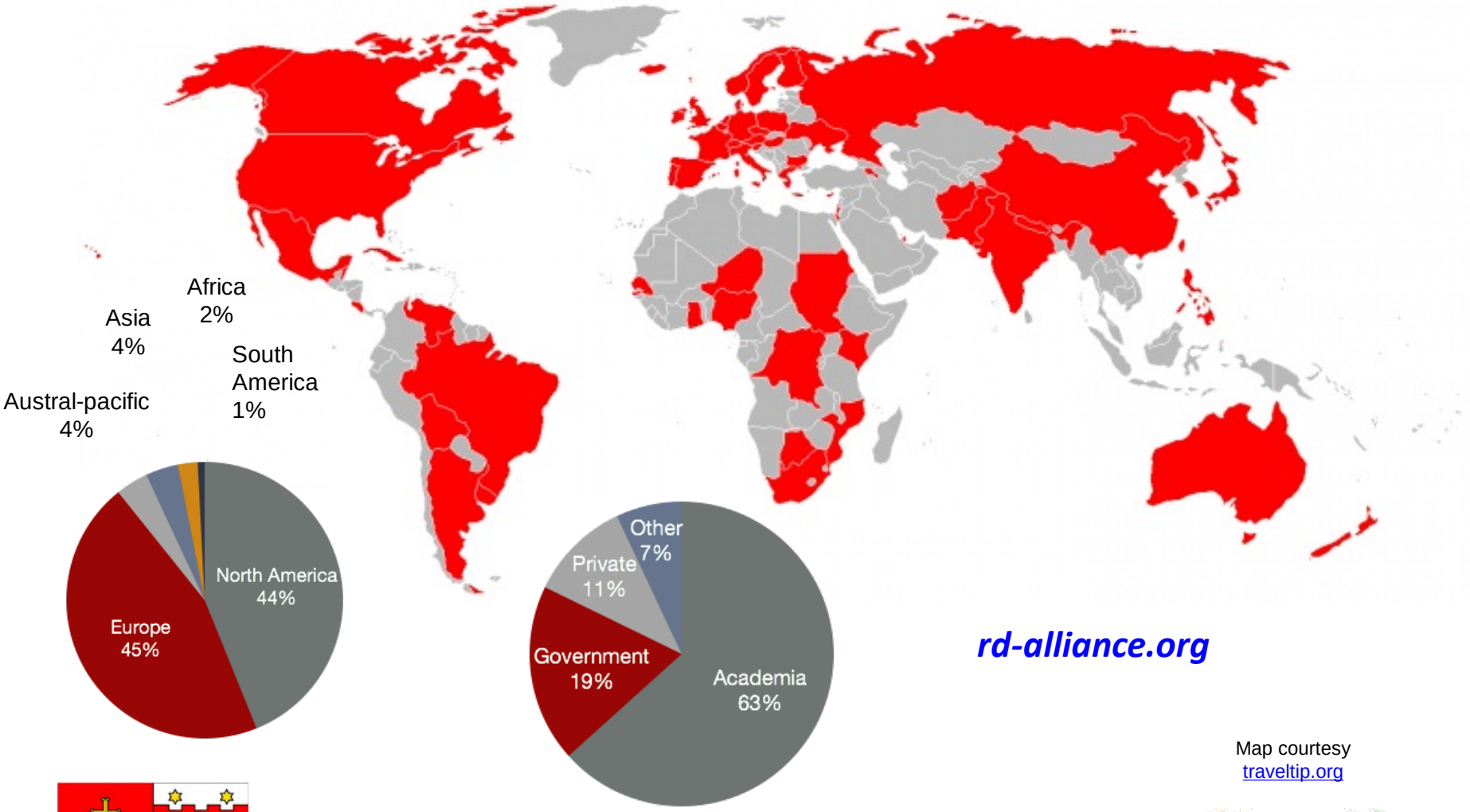
- **Working Groups** – 12-18 month efforts to build, adopt, and use specific pieces of infrastructure
- **Interest Groups** – longer-lived discussion forums that spawn Working Groups as specific pieces of needed infrastructure are identified.

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

- **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**

The RDA Community:

Over 1300 members from 64+ countries



rd-alliance.org

Map courtesy
traveltip.org

Fran Berman

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)

Rensselaer

Fran Berman

Community-Driven RDA Groups by Focus (as of 12/13, more by 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

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

Global Data Science Challenge: 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; New York Times Health Section September 23, 2013

Sustainable Stewardship to Support Data-Driven Innovation

Global Data Infrastructure

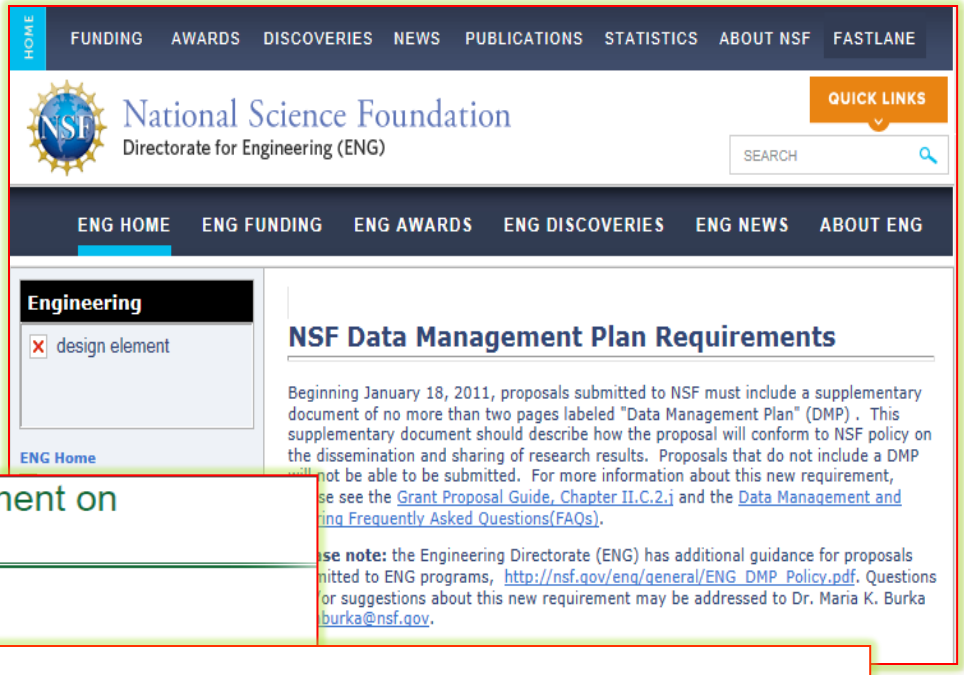
How do we accelerate open access data sharing and exchange?

National Data Infrastructure

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

Fran Berman

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



Summary: Office of Science Statement on Digital Data Management

Requirements (1 of 3)

- To integrate data
  proposals subm
  required to inclu
  pages that descri
  research will be s
  preservation are
  DMPs must descr
  of results, or how
  preserved.

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY GUIDELINES, INFORMATION QUALITY STANDARDS, AND ADMINISTRATIVE MECHANISM

PART I: BACKGROUND, MISSION, DEFINITIONS, AND SCOPE

BACKGROUND

Section 515 of the Treasury and General Government Appropriations Act for Fiscal Year 2001 (Public Law 106-554), hereinafter "Section 515," directs the Office of Management and Budget (OMB) to issue government-wide guidelines that "provide policy and procedural guidance to Federal agencies for ensuring and maximizing the quality, objectivity, utility, and integrity of information (including statistical information) disseminated by Federal agencies." OMB complied by issuing guidelines which direct each Federal agency to (A) issue its own guidelines ensuring and maximizing the quality, objectivity, utility, and

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

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

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"

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.

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

Publicly Accessible Data has to Live Somewhere

Public Access, Use, and Re-Use of Data Now and in the Future Presupposes Sustainable Stewardship *Today*

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

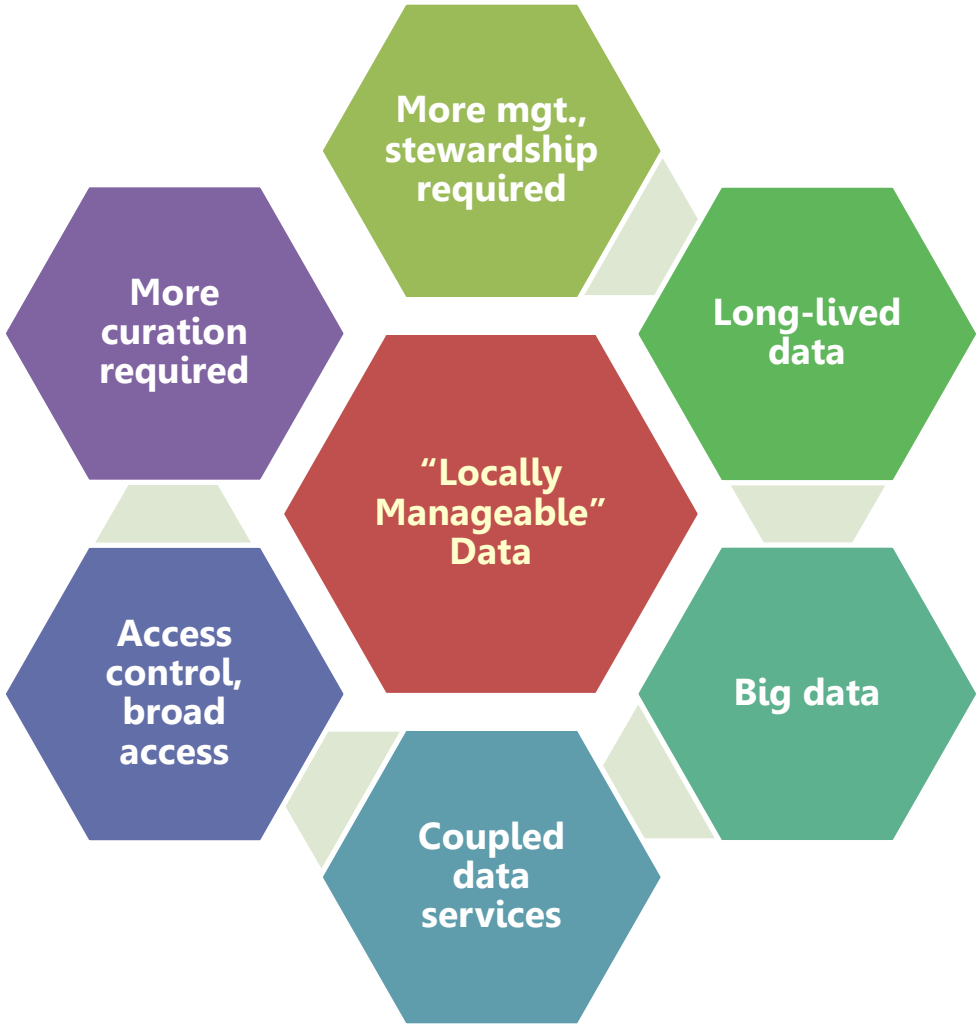
- Economically sustainable **data infrastructure** necessary to support
  - Federally mandated data management plans
  - Public access to research data
  - Use and re-use
  - Reproducibility

- **The "bigger", more long-term, more complex, or more valuable the data is, the greater the importance of sustainable data stewardship and infrastructure**

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

It's not just about the cost of storage. Data infrastructure costs increase with usage, stewardship and access requirements, perceived value

Greater costs beyond the center (including "big" data)



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]

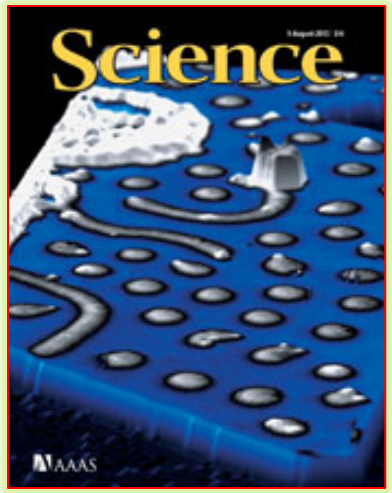
On 22 February, the U.S. Office of Science and Technology Policy (OSTP) released a memo calling for public access for publications and data resulting from federally sponsored research grants (*1*). The memo directed federal agencies with more than $100 million R&D expenditures to "develop a plan to support 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 h… up to take on the problem alone… unrealistic to expect as much. In … sector, federal R&D agencies ar… to allocate enough resources to s… federally funded research data. Th…



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

Op-Ed Recommendations: Distribute the Preservation and Stewardship Responsibilities Across 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

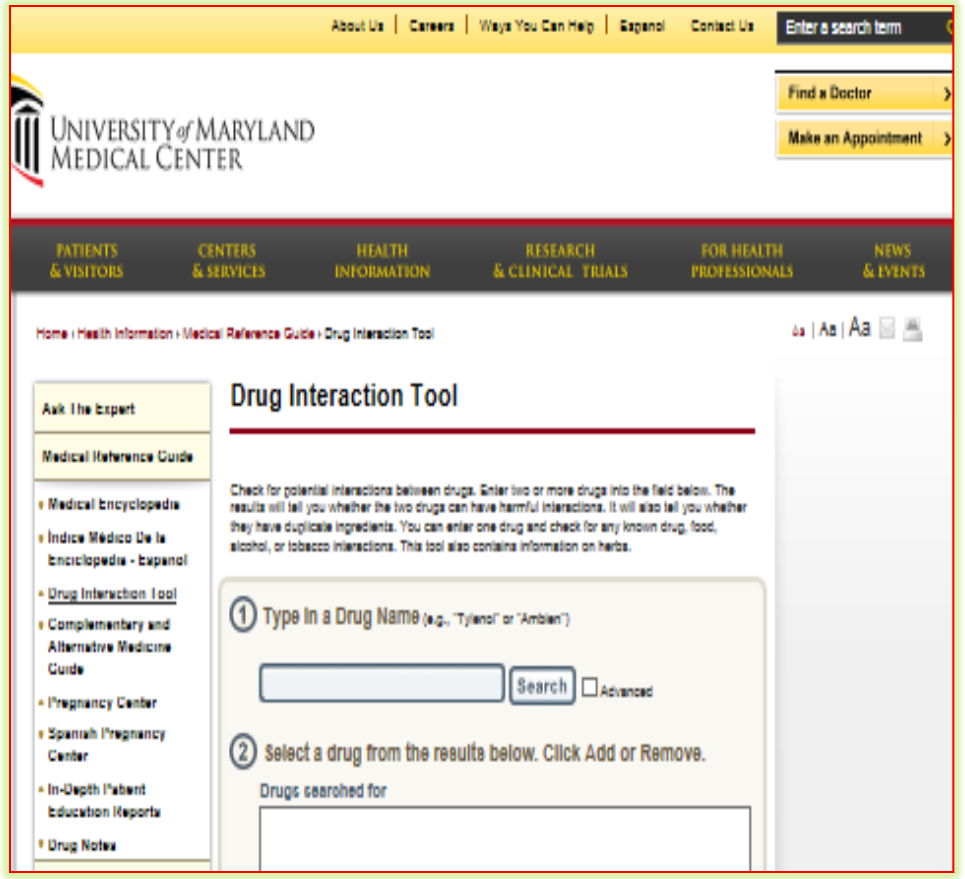
3. **ACADEMIC SECTOR:** Use public sector investment to jumpstart sustainable university library / community repository stewardship solutions

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.)

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

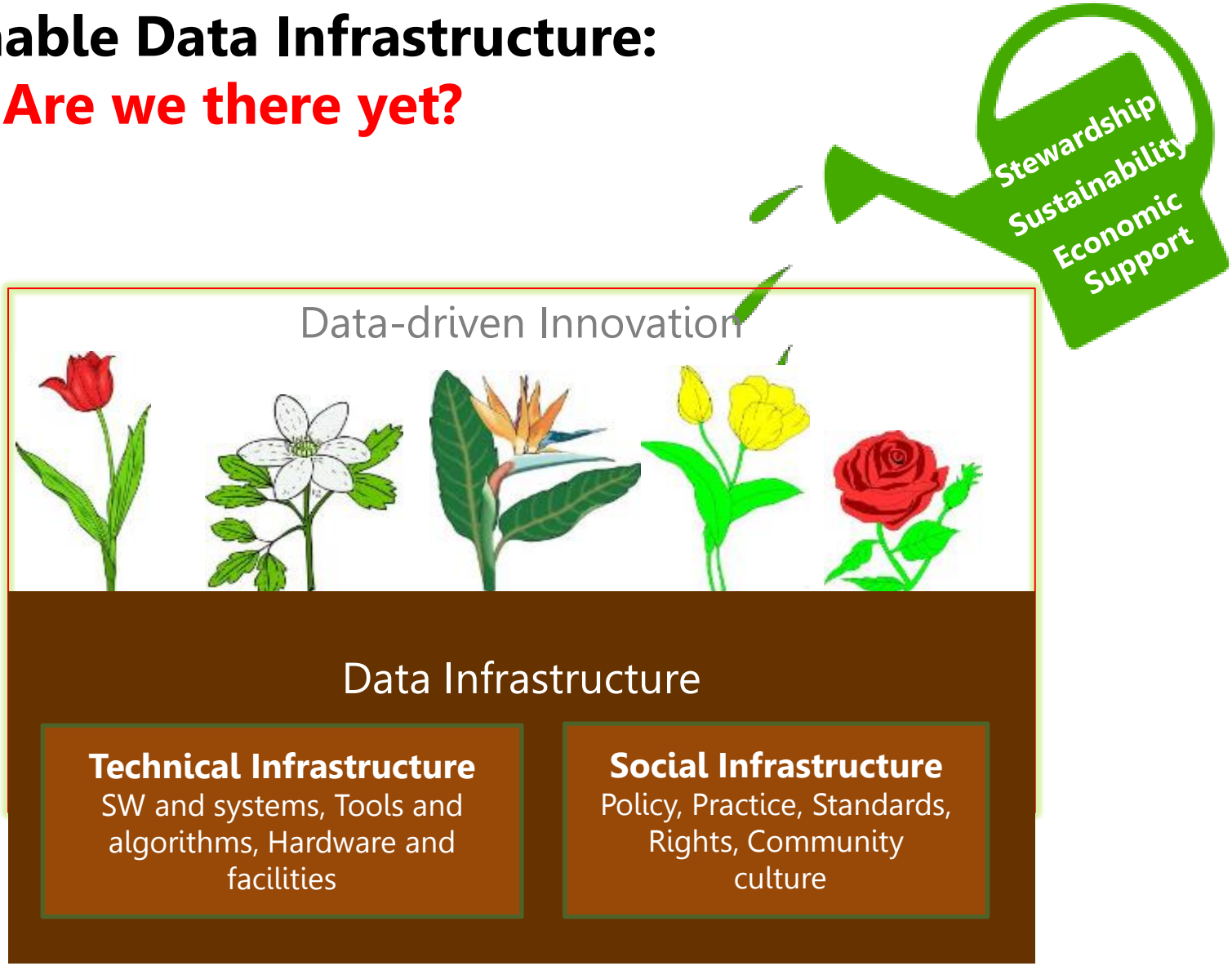
Data Science Challenge: Improving the Usefulness of Research Data Infrastructure

- **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?
- **Metrics of Success** – What is good infrastructure?
  - How do we measure it?
  - What are appropriate paths for improvement / evolution?
  - How do we create agile systems that can support unintended new uses?



http://umm.edu/health/medical/drug-interaction-tool

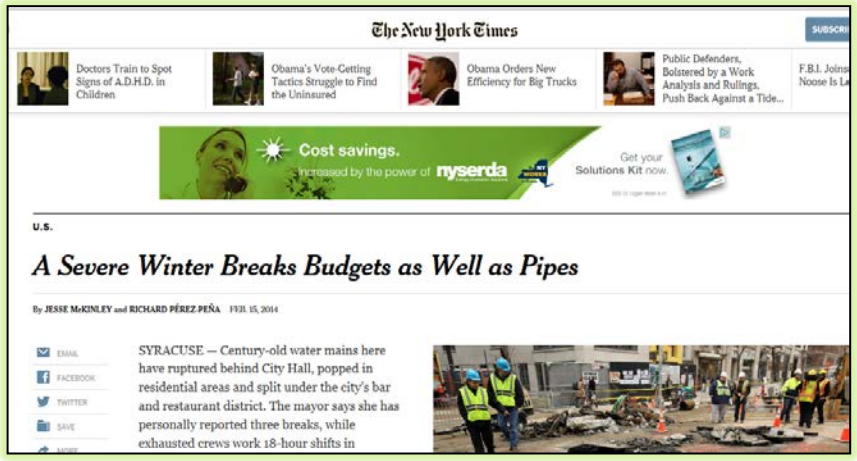
Sustainable Data Infrastructure:
Are we there yet?



Fran Berman

Sustainable Data Infrastructure:
Are we there yet?

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



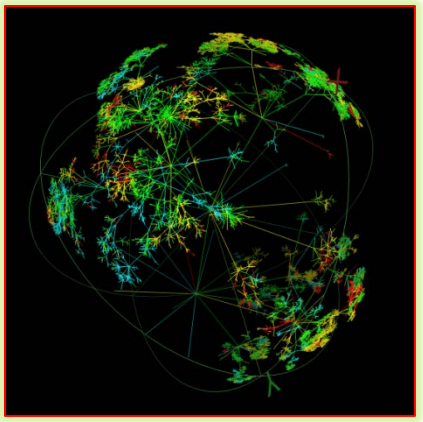
Stephanie A. Miner, the Syracuse mayor, said [infrastructure is] too often overlooked when politicians want to spend money on economic development. ***"You don't cut ribbons for new water mains, but that's really what matters."***

NY Times, Feburary 15, 2014

Rensselaer

Fran Berman

Value Proposition:
Data Investments Do and Will Pay Off



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

Image: CAIDA Internet visualization;
Article: HBR October 2012

Thank You!

Rensselaer

Fran Berman