Building an Ecosystem to Accelerate Data-Driven Innovation

Dr. Francine Berman
Chair, Research Data Alliance / US
Edward P. Hamilton Distinguished Professor of Computer Science, Rensselaer Polytechnic Institute
Research Data Driving Solutions to Complex Scientific and Societal Challenges

Who is most at risk to contract asthma?

How can we increase wheat yields?

How can we best address energy needs and sustain the environment?

How accurate is the Standard Model of Physics?
Data-Driven Geoscience: How Can We Respond to Large-Scale Earthquakes?

Earthquake simulations enable

- Enhanced **scientific understanding** of the physical world
- More strategic plans for bridge, building and other physical infrastructure reinforcements to **increase safety**
- Better **disaster response planning** for police, fire fighters, ER teams in high-risk areas to increase their effectiveness

Simulation courtesy of Amit Chourasia, SDSC, Table information courtesy of Southern California Earthquake Center
Earthquake simulations enable

- Enhanced **scientific understanding** of the physical world
- More strategic plans for bridge, building and other physical infrastructure reinforcements to **increase safety**
- Better **disaster response planning** for police, fire fighters, ER teams in high-risk areas to increase their effectiveness
Earthquake simulations enable

- Enhanced **scientific understanding** of the physical world
- More strategic plans for bridge, building and other physical infrastructure reinforcements to **increase safety**
- Better **disaster response planning** for police, fire fighters, ER teams in high-risk areas to increase their effectiveness

<table>
<thead>
<tr>
<th>Estimated figures for simulated 240 second period, 100 hour run-time</th>
<th>TERASCALE: Tera Shake domain (600x300x80 km^3)</th>
<th>PETASCALE: Peta Shake domain (800x400x100 km^3)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fault system interaction</strong></td>
<td><strong>NO</strong></td>
<td><strong>YES</strong></td>
</tr>
<tr>
<td><strong>Inner Scale</strong></td>
<td>200m</td>
<td>25m</td>
</tr>
<tr>
<td><strong>Resolution of terrain grid</strong></td>
<td>1.8 billion mesh points</td>
<td>2.0 trillion mesh points</td>
</tr>
<tr>
<td><strong>Magnitude of Earthquake</strong></td>
<td>7.7</td>
<td>8.1</td>
</tr>
<tr>
<td><strong>Time steps</strong> (sec/step)</td>
<td>20,000 (.012 sec/step)</td>
<td>160,000 (.0015 sec/step)</td>
</tr>
<tr>
<td><strong>Surface data</strong></td>
<td><strong>1.1 TB</strong></td>
<td><strong>1.2 PB</strong></td>
</tr>
<tr>
<td><strong>Volume data</strong></td>
<td><strong>43 TB</strong></td>
<td><strong>4.9 PB</strong></td>
</tr>
</tbody>
</table>

Simulation courtesy of Amit Chourasia, SDSC, Table information courtesy of Southern California Earthquake Center
Integrated Infrastructure Critical: Application Needs Span the Spectrum

Digital history image integration

Analysis and modeling of protein function and structures

Analysis of Data from the CERN Large Hadron Collider

Earthquake Simulation

Turbulence

Cosmology

Global Public health

Seti@home, MilkyWay@Home, BOINC

BIG DATA, SMALL COMPUTE

BIG DATA, BIG COMPUTE

SMALL DATA, SMALL COMPUTE

SMALL DATA, BIG COMPUTE

DISTRIBUTED, GRID AND CLOUD APPS, CROWD-SOURCING

DISTRIBUTED, GRID AND CLOUD APPS, CROWD-SOURCING

COMPUTE (more FLOPS)

DATA (more BYTES)

NETWORK (more BW)
Many Kinds of Technical Data Infrastructure Needed to Drive Innovation

SDSC Data e-infrastructure

DATA PRESERVATION
DATA STORAGE
DATA SERVICES
DATA ANALYTICS
DATA VISUALIZATION
DATA MANAGEMENT
DATA PORTALS
HPC DATA

Fran Berman

Many Kinds of Technical Data Infrastructure

2001-2009

DATA PORTALS

DATA MANAGEMENT

DATA VISUALIZATION

DATA SERVICES

DATA PRESERVATION

DATA STORAGE
Social, Organizational, and Human Infrastructure
Equally Important

Human Infrastructure / Workforce

Data Scientists

Data-focused Curriculum and Training

Community Practice

Sustainable Economics

Policy

Common Standards

McKinsey Global Institute 2011 Report,
Traffic Image: Mike Gonzalez
Today’s Presentation:
Two perspectives on Research Data Infrastructure

• **Opportunity:** Maximizing data-driven innovation through data sharing and exchange
  – Efforts to build a coordinated global Infrastructure to support data access, sharing and use

• **Challenge:** Prioritizing the Development, Implementation and Sustainable Support of Data Infrastructure
  – Strategies to accelerate efforts within organizations, communities and sectors
Opportunity: Maximizing Data-Driven Innovation Through Data Sharing and Exchange
Data-Sharing Driving innovation Across Sectors and Communities
World-wide Efforts Focusing on Infrastructure to Support Research Data Sharing, Access, Use

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 Fernandez, EOS and EURIS - rmanuel@di.ubi.pt

Science, Humanities, Arts Communities

Libraries, Archives, Repositories, Museums

E-Infrastructure 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

Fran Berman
Many Infrastructure Building Blocks Needed to Accelerate Progress

- Research Dissemination and Reproducibility
- Data Use and Re-use
- Data Access (now) and Preservation (later)
- Data Discovery and Data Sharing

- Data Access and Distribution Policy
- Institutional Data Sharing Practice
- Data Discovery Tools

- Digital Object Identifiers
- Common Metadata Standards
- Data Citation Standards

- Data Preservation Practice
- Data Analytics Algorithms
- Data Scientists and Expert Support

- Curation Practice and Policy
- Sustainable Economic Models
- Auditing, Certification and Reporting Practice
Research Data Alliance Created to Accelerate Development of Research Data Sharing Infrastructure Worldwide

- RDA is an emerging, global community-driven organization created to **accelerate the development of research data-sharing infrastructure** world-wide.

- RDA community efforts focus on building **social, organizational and technical infrastructure** to
  - **reduce barriers to data sharing and exchange**
  - **accelerate the development of coordinated global data infrastructure**
RDA Approach:

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 Today: Over 1600 members from 70+ countries (as of 15/3/14)
Precipitous Growth

First RDA organizational telecon: August 2012

Global Data Planning Meeting: October 2012

RDA Launch / First Plenary

First Working Groups and Interest Groups
240 participants

RDA Second Plenary

First “neutral space” community meeting (Data Citation Summit)

RDA Third Plenary

First Org. Partner Meet-up

14 BOF, 12 Working Groups, 22 Interest Groups

497 participants

RDA Fourth Plenary

First BOFs

380 participants from 22 countries

First Working Group exchange meeting

March 2013  September 2013  March 2014  September 2014

Fran Berman
Professional Title | Total | %
--- | --- | ---
Advisor/Consultant | 22 | 4%
CEO / Managing Director / Chief Executive | 35 | 7%
CTO / IT Director | 20 | 4%
IT Specialist / IT Architect | 53 | 11%
Journalist / Editor / Copywriter | 6 | 1%
Librarian | 27 | 5%
Other | 93 | 19%
Student | 38 | 8%
Policy Development Manager / Policy Consultant | 12 | 2%
Professor | 42 | 8%
Programme Manager / Project Manager | 62 | 12%
Researcher | 87 | 18%
Total | 497 | 100%

Organizational Assembly Meeting: 20+ new members

6 co-located data-focused events
### 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

### 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

### Community Needs - focused
- Community Capability Model IG
- Engagement IG
- Clouds in Developing Countries 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

---

HPC Members welcome!
**First RDA Infrastructure Deliverables coming this Fall**

### 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, EUDAT

### Persistent Identifier Information Types
- **Deliverables:** Minimal set of PID types, API
- **Initial Adopters and Users:** Data Conservancy, DKRZ

### 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 directory of current metadata standards starting from DCC directory
- **Initial Adopters and Users:** JISC, DataOne
Next Steps for the RDA

- **More Infrastructure**: Continuing pipeline of infrastructure deliverables adopted and used to accelerate data sharing
- **Effective Community**: Increasing coordination of infrastructure
- **Synergistic Programs**: Increasing cross-boundary collaborations between domains, sectors, organizations
- **Focus on Industry**: International and regional programs focusing on workforce, outreach, expansion of infrastructure impact
- **New partners in the Organizational Assembly**: Focused strategy to support development of industry infrastructure for data sharing
Challenge: Supporting and Sustaining Research Data Infrastructure
Research Data and Data Sharing Key to Innovation. Research Data an Accelerator for All Sectors.

- National governments increasingly calling for public access to research data.
  - Valuable to all sectors as a driver of current and future innovation
- Research data infrastructure is necessary to support
  - Use and re-use
  - Research reproducibility
  - Federally mandated data management plans
  - Public access to research data
Yet Research Data Infrastructure is Difficult to Sustain. Why?
Sustainable Data Infrastructure Starts with a Sustainable Economic Model

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

Greater costs beyond the center (including “big” data)
Data Economics: Data Management, Stewardship, and Use Incur Continuing Infrastructure Costs

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, etc.

Resources and Resource Refresh

- Model A (8-yr, 15.2-mo 2X)

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

Fran Berman
In the Public Sector, Research and Infrastructure often compete for limited funding. Infrastructure Investment a hard sell …

<table>
<thead>
<tr>
<th></th>
<th>Research</th>
<th>Infrastructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>What is Newsworthy?</td>
<td>New discoveries and breakthroughs</td>
<td>Failure of systems</td>
</tr>
<tr>
<td>What is the value proposition?</td>
<td>Domain and national leadership and competitiveness</td>
<td>Enabler of innovation</td>
</tr>
<tr>
<td>Funding Model</td>
<td>Fixed-term funding</td>
<td>Continuous long-term support</td>
</tr>
</tbody>
</table>

Crisp photos of moon landing are missing
Spectacular images of day were stored, forgotten -- and lost
Marc Kaufman, Washington Post
Sunday, February 4, 2007

Higgs boson scientists win Nobel prize in physics
By James Morgan, Science reporter, BBC News

Scientists losing data at a rapid rate
Decline can mean 80% of data are unavailable after 20 years.
Elizabeth Gibney & Richard van Noorden
Data Infrastructure particularly important in light of increasing National R&D Agency Requirements for Data Access and Management.
Economics of Public Access: Who Pays the Data Bill?

When economic models and infrastructure are not in place to ensure access and preservation, 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 been willing or able to take on the problem alone. It is unrealistic to expect as much. In the sector, federal R&D agencies are not allocated enough resources to support federally funded 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) (/)—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 Studies...
Op-Ed Recommendations: Partner Across Sectors to Distribute the Preservation and Stewardship Responsibilities

- Facilitate private sector stewardship of public access research data as a public good
- Create sustainable university library and repository stewardship solutions
- Clarify public sector stewardship commitments: articulate what data will / won’t be supported
- Evolve research culture to take advantage of what works in the private sector
- Evolve research culture to take advantage of what works in the private sector

Fran Berman
Next Steps Towards Sustainable Stewardship

• Identify and evolve an expanding network of repositories for publicly accessible research data

• Create useful metrics for successful stewardship and economic stability that can be used to support the development of effective organizational support

• Create a plan and actionable recommendations for strategic investment
Last Words: Information Infrastructure is necessary for 21st Century Innovation

• **Value Proposition:**
  - Virtually all fields becoming data driven
  - **Adequate and sustainable data infrastructure is critical to drive innovation and HPC applications**

• **What can we do:**
  - Include data stewardship, management, use, access and preservation as part of project planning, budget and efforts
  - Recognize and publish the data contributions of our work as well as the research contributions

**Small steps** (things to do on Monday morning):

1. If you don’t have one, create a data management plan for your current project for a reasonable fixed term of time
2. Make your data available to the community (as appropriate) by curating it and ingesting it into a publicly accessible repository
3. Cite and publish your data as appropriate when you write about your results
Thank You!