Got Data?  The Role of Digital Information in Shaping 21st Century Research

University of Buffalo

Dr. Francine Berman
Vice President for Research
Professor of Computer Science
Rensselaer Polytechnic Institute
The World We Live In
The Expanding Impact of Digital Information

The new Egypt, where the PM resigns on Facebook

Prime Minister Ahmed Shafiq’s resignation was announced today not on state TV nor in a press release, but on Facebook — a key tool protesters used to overthrow Mubarak.

The New York Times
Science

Arms of Expensive Lawyers, Replaced by Cheaper Software

Cyberbullying an epidemic
Jennifer Dudley-Nicholson | National Features | March 01, 2011 11:30pm

ONE in four school children is being bullied using modern technology, and that frightening ratio could go even higher as more children get access to computers and phones.
What is the potential impact of Global Warming?

How will natural disasters effect urban centers?

How do we ensure a future pipeline of innovation?

How do we harness the power of community?

How can we create a more sustainable environment?

What therapies can be used to cure or control cancer?

Digital information also shaping our ability to address our most challenging science and societal challenges

Fran Berman
Today’s Presentation

• Data-enabled Research and Innovation

• Supporting Digital Research Data – Enabling Infrastructure

• Preserving Digital Data Over the Long Term

• Who Will Pay? Economics and Digital Preservation
Data-Driven Astronomy
How was the Milky Way Galaxy formed?

Data-Driven Discovery: Create physical and dynamic model of the Milky Way Galaxy using data from all-sky surveys

- Trace the density of stars from large imaging surveys (SDSS)
- Constrain the dynamics of the stars using large spectroscopic surveys (Sloan Extension for Galactic Understanding and Exploration, Chinese LAMOST survey of galactic stars, etc.)

- Use a model of the detailed motions of Galactic stars to recover the gravitational potential of the Galaxy, which is dominated by dark matter

New results from data-driven research include discovery of the Sagittarius Dwarf Tidal Stream, shedding light on the formation of the Milky Way.
Data-Driven Geoscience

Where are the areas of most impact during a large-scale earthquake?

- **Input for supercomputer simulation:**
  - Model of 7.7 earthquake targeted to lower San Andreas fault
  - 10 years of sensor data on southern CA terrain

- **Post simulation**
  - Additional computation (80,000+ CPU hours) used for visualization of seismic wave propagation and analysis
  - Derived data products (velocity magnitude, displacement vector field, cumulative peak maps, statistics, etc.) included in SCEC digital library (>168 TB).

- **Results used to create safer building codes, plan disaster response**

Visualization courtesy of Amit Chourasia, SDSC; Model from SCEC
Data in the Life Sciences
Enabling Community Research

The Protein Data Bank
- worldwide repository for the processing and distribution of 3-D structure data of large molecules of proteins and nucleic acids.

- PDB represents $80 billion + investment in research resulting in PDB structures
- PDB supported by funds from NSF, NIGMS, DOE, NLM, NCI, NCRR, NIBIB, NINDS, NIDDK, etc.

![Yearly Growth of Total Structures in PDB](image)

Information courtesy of Phil Bourne, SDSC/UCSD and Helen Berman, Rutgers
Beyond Science
Data-driven Innovation

• Dance in the Digital World
  – Today’s choreographers using information technologies for
    • Fundraising
    • Publicity and outreach
    • Movement capture
On the Horizon (Information Technology) …

- Un-structured data, multi-modal data
- Real-time analytics
- Data \( \rightarrow \) Information \( \rightarrow \) Decisions

Augmented Reality: “X-analytics”

Low-Barrier User interfaces: “There’s an App for that”

Crowd-sourcing

Exascale Supercomputing and Commodity Petascale Computing
Supporting Digital Research Data -- Enabling 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
- Create portals / data collections / databases
- Couple with literature ....

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

Information adapted from Chris Rusbridge and Liz Lyon
Digital Research Data
Digital Research Data
It’s Not Just About Size …

- **RETENTION TIMEFRAME:**
  Short-term (few months, years) to long-term (decades, centuries, …)

- **PREPARATION:**
  Well-tended (sufficient metadata, cleaned and filtered) to poorly tended (flat files, insufficient metadata)

- **POLICY / REGULATION RESTRICTIONS:**
  Subject to more restrictive policy and regulation (HIPAA) vs. subject to less restrictive policy and regulation (NSF)

- **LIFE CYCLE PLANNING:**
  Has a data management and sustainability plan (PDB, PSID, NVO) vs. ad hoc approach

- **STANDARDS**
  Uses community standards vs. creates own standards (or uses no standards at all)

- **COMMUNITY CULTURE**
  Shared broadly with community vs. kept private for competitive advantage

- **SIZE / SCALE:**
  Small-scale (GBs) to large-scale (PBs +)
Cyberinfrastructure Access and services

Needed Support for Data-driven Research

Interoperability reduces the infrastructure burden for researchers

File systems, Database systems, Collection Management
Data Integration, etc.

Many Data Sources

modeling

analysis

simulation

visualization

 computers

computers

Data Use

Data Access

Data Management

Data Storage

Data Preservation

Services make data more usable

• Data Analytics
• Data visualization
• Collection Management
• Data mining
• Hosting and Preservation Services
• Statistical packages
• Domain-specific tools
  • Biology Workbench
  • Montage (astronomy mosaicking)
  • Workflow management
• Data de-identification, etc.
Cyber-security – Digital “Achilles Heel”

• **What / who should you trust?**
  – Your hardware?
  – Your system / software?
  – Your respondent?

• Technical, social, policy, and regulatory solutions needed to address cyber-security challenge
  – Most valued data must be replicated
    (Byzantine General’s Problem …)

In 2003, the **Slammer computer virus** exploited a weakness in SQL server software to launch a “denial of service” attack that

• Shut down over 13,000 Bank of America ATMs
• Caused difficulties in Continental Airline’s electronic reservation and ticketing systems, causing cancellation of some regional flights
Data Cyberinfrastructure Must be Reliable and Easy to Use

Key Characteristics:
- Usability
- Scalability
- Interoperability
- Reliability
- Capability
- Predictability
- Accessibility
- Sustainability
- Cost-Effectiveness

### Reliability:
How to minimize the risk of data loss or damage?

### Accessibility:
Will the information be there when you need it?

<table>
<thead>
<tr>
<th>Entity at risk</th>
<th>What can go wrong</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>File</td>
<td>Corrupted media, disk failure</td>
<td>1 year</td>
</tr>
<tr>
<td>Tape</td>
<td>+ Simultaneous failure of 2 copies</td>
<td>5 years</td>
</tr>
<tr>
<td>System</td>
<td>+ Systemic errors in vendor SW, or Malicious user, or Operator error that deletes multiple copies</td>
<td>15 years</td>
</tr>
<tr>
<td>Archive</td>
<td>+ Natural disaster, obsolescence of standards</td>
<td>50-100 years</td>
</tr>
</tbody>
</table>
Reliable Data Cyberinfrastructure Incurs Real 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 regulations, …

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
Data Stewardship and Preservation: Technical Challenges / Costs Increase at the Extremes
New Opportunities for Synergy between Research Needs and Library / Museum Strengths

- Research community characterized by culture of innovation
  - Periodic new starts
  - Experimentation
  - Customized solutions to ill-defined problems
  - Collaboration and competition

- Researchers need help with things
  Libraries / Museums are good at
  - Developing reliable management, preservation and use environments
  - Proper curation and annotation
  - Navigating policy, regulation, intellectual property
  - Sustainability
The “Local” Digital Research Data Repository: Emerging Role for University Libraries

- Researchers are increasingly required to retain the digital products of their research.
- University libraries can play a new role as local stewards of digital research data.
- “Preservation Jumpstart” funding needed to make this realistically viable on a broad scale.
Preserving Digital Research Data Over the Long Term
Access to Digital Information Tomorrow Requires Preservation Today

- Digital Access and Preservation is a technical, management, policy, regulatory, social, and economic problem

- Key issues to resolve:
  1. What should we preserve?
  2. Who is responsible for digital information?
  3. Who pays for digital information and its supporting cyberinfrastructure?
What Should We Save?
Saving Everything Isn’t An Option …

2007 was the “crossover year” where the amount of digital information exceeded the amount of available storage (~264 exabytes)

By 2023, the amount of digital data will exceed Avogadro’s number. (6.02 X 10^23, the number of atoms in 12 grams of carbon).

U.S. Library of Congress manages over 300 TB of digital data

World Data Center for Climate Database = 220 TB

1 novel = 1 MB

IBM Watson’s Reference Data = 15+ TB

<table>
<thead>
<tr>
<th>Kilo</th>
<th>10^3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mega</td>
<td>10^6</td>
</tr>
<tr>
<td>Giga</td>
<td>10^9</td>
</tr>
<tr>
<td>Tera</td>
<td>10^12</td>
</tr>
<tr>
<td>Peta</td>
<td>10^15</td>
</tr>
<tr>
<td>Exa</td>
<td>10^18</td>
</tr>
<tr>
<td>Zetta</td>
<td>10^21</td>
</tr>
</tbody>
</table>

What do We Want to Save?

Data we* want to keep over the long-term:

– We = “Society”
  • Official and historically valuable data (Census information, presidential emails, Shoah Collection, etc.)

– We = Research Community
  • Protein Data Bank, National Virtual Observatory, etc.

– We = Me
  • My medical record, my Quicken data, digital family photos, etc.

---

* We = "Society"
**What do We Have to Save?**

- **HIPAA** applies to health information created or maintained by health care providers.

- **Sarbanes-Oxley** regulations apply to all U.S. public company boards, management, and public accounting firms.

- **OMB** regulations apply to federally funded research data (NIH, NSF, DOE, etc.)

<table>
<thead>
<tr>
<th>Crime and Punishment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regulations</td>
</tr>
<tr>
<td>HIPAA</td>
</tr>
<tr>
<td>Sarbanes-Oxley</td>
</tr>
<tr>
<td>Gramm-Leach-Baily</td>
</tr>
<tr>
<td>SEC 17a</td>
</tr>
<tr>
<td>OMB Circular A-110 / CFR Part 215 (applies to federally funded research data)</td>
</tr>
</tbody>
</table>

*Table information partly based on “Data Retention – More Value, Less Filling”, John Murphy, http://www.tdan.com/view-articles/5222*
Data Economics: The Blue Ribbon Task Force on Sustainable Digital Preservation and Access

BRTF Charge:

1. Conduct a comprehensive **analysis** of sustainable digital preservation
2. Identify and evaluate **best practices**
3. Make specific **recommendations for action**
4. Articulate **next steps** for further work

Sponsoring Agencies / Institutions:

- National Science Foundation
- Library of Congress
- Mellon Foundation
- National Archives and Records Administration
- Council on Library and Information Resources
- National Coordination Office for Networking and Information Technology Research and Development
- Joint Information Systems Committee, U.K.
- Member institutions
What is required to support digital information over the long term?

Economic sustainability for digital information requires:

- Recognition of the benefits of long-term access and preservation
- Incentives for decision-makers to act
- Means of selecting “valued” information for long-term preservation
- Mechanisms to support ongoing, efficient allocation of resources
- Appropriate organization and governance of preservation and access activities

BRTF Interim Report
- Scoping of problem space
- Survey of current practices
- Identification of systemic challenges
- Over 43,000 downloads to date
- Available at brtf.sdsc.edu
Who’s Paying the Data Bill?

• The “free rider” non-solution: “Let X do it”
  where X is:
  – The Government
  – The Libraries
  – The Archivists
  – Google, Microsoft, etc.
  – Data users
  – Data owners
  – Data creators, etc.
There is no magic bullet

- Many economic models for digital information currently used, none of them are “free”

Federal grants
Pay per service
Advertisements
Donations, etc.

“Tax”
The Stakeholder Problem

- Many Stakeholders in digital preservation
  - Stakeholders who benefit from use of the preserved asset
  - Stakeholders who select what to preserve
  - Stakeholders who have rights to the asset
  - Stakeholders who preserve the asset
  - Stakeholders who pay

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

• Many Stakeholders in digital preservation
  – Stakeholders who benefit from use of the preserved asset
  – Stakeholders who select what to preserve
  – Stakeholders who have rights to the asset
  – Stakeholders who preserve the asset
  – Stakeholders who pay

• The greater the alignment between key stakeholder groups, the better the prospect for sustainable preservation
Towards a Comprehensive Approach for Data Preservation and Access: Task Force Recommendations 1

- **INVEST IN / SEED PRESERVATION INFRASTRUCTURE**
  - Invest in building / seeding stewardship capacity and capability throughout the system.
  - Create financial incentives to encourage organizations to preserve digital materials on the public behalf.
  - Fund the modeling and prototyping of preservation infrastructure approaches

- **CREATE SUSTAINABILITY-FRIENDLY POLICY, REGULATION, AGREEMENTS**
  - Create preservation mandates as appropriate
  - Reform national and international copyright legislation to address and support digital preservation needs
Towards a Comprehensive Approach for Data Preservation and Access:
Task Force Recommendations 2

- **CONVENE PRESERVATION-AWARE COMMUNITIES**
  - Create public-private partnerships to align distinct stakeholder groups
  - Convene expert communities to address the selection and preservation needs of valuable materials

- **TAKE INDIVIDUAL RESPONSIBILITY**
  - Provide nonexclusive rights to preserve and distribute your content
  - Partner with preservation experts throughout the data lifecycle to ensure long-term sustainability
Raise Awareness
Make the Case for Digital Preservation and Access

• To Decision Makers:
  • What are liabilities and the opportunity costs of *not acting*?
  • What specific actions need to be made a priority now?

• To the General Public:
  • Does your dry cleaner know what digital preservation is?
Thank You

Blue Ribbon Task Force Final Report

- Economic analysis
- Findings and recommendations
- Priorities for near-term action
- Over 80,000 downloads to date