Data and Society
Lecture 6: Stewardship and Preservation

2/23/118
Announcements 2/23

• **Your Op-Ed final needs to be on the same topic as your Op-Ed draft.**

• **Op-ed Finals due next Friday.** Please turn in hardcopies of **the draft and the final copy** on March 2 at 9:00 a.m. If you will be doubling your draft grade, let Fran (bermaf@rpi.edu) know by March 1. No late work.

• Research Paper Instructions today at the end of class
Discussion article for March 2

<table>
<thead>
<tr>
<th>Wednesday Section</th>
<th>Friday lecture</th>
<th>First Half of Class</th>
<th>Second Half of Class</th>
<th>Assignments</th>
</tr>
</thead>
<tbody>
<tr>
<td>January 17: NO class</td>
<td>January 19</td>
<td>L1: CLASS INTRO AND LOGISTICS</td>
<td>Presentation Model / Op-Ed Instructions</td>
<td>Op-Ed instructions</td>
</tr>
<tr>
<td>January 24: NO class</td>
<td>January 26</td>
<td>L2: BIG DATA 1</td>
<td></td>
<td>4 Presentations</td>
</tr>
<tr>
<td>January 31: NO class</td>
<td>February 2</td>
<td>L3: BIG DATA 2 - IoT</td>
<td></td>
<td>4 Presentations</td>
</tr>
<tr>
<td>February 7: NO class</td>
<td>February 9</td>
<td>L4: DATA AND SCIENCE</td>
<td></td>
<td>4 Presentations</td>
</tr>
<tr>
<td>February 14: 5 Presentations</td>
<td>February 16</td>
<td>L5: DATA AND HEALTH / LESLIE McINTOSH GUEST SPEAKER</td>
<td></td>
<td>4 Presentations</td>
</tr>
<tr>
<td>February 21: 5 Presentations</td>
<td>February 23</td>
<td>L6: DATA STEWARDSHIP AND PRESERVATION</td>
<td>5 Presentations</td>
<td>Research Paper instructions</td>
</tr>
<tr>
<td>February 28: 5 Presentations</td>
<td>March 2</td>
<td>L7: DATA INFRASTRUCTURE</td>
<td></td>
<td>4 Presentations</td>
</tr>
<tr>
<td>March 7: 4 Presentations</td>
<td>March 9</td>
<td>NO CLASS / PAPER PREPARATION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 14: Spring Break</td>
<td>March 16</td>
<td>SPRING BREAK</td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 21: NO class</td>
<td>March 23</td>
<td>NO CLASS / PAPER PREPARATION</td>
<td></td>
<td></td>
</tr>
<tr>
<td>March 28: 5 Presentations</td>
<td>March 30</td>
<td>L8: DATA RIGHTS, POLICY, REGULATION</td>
<td></td>
<td>4 Presentations</td>
</tr>
<tr>
<td>April 4: NO class</td>
<td>April 6</td>
<td>L9: DATA AND ETHICS</td>
<td></td>
<td>4 Presentations</td>
</tr>
<tr>
<td>April 11: 5 Presentations</td>
<td>April 13</td>
<td>L10: DATA AND COMMUNICATION</td>
<td></td>
<td>4 Presentations</td>
</tr>
<tr>
<td>April 18: 5 Presentations</td>
<td>April 20</td>
<td>L10: DATA FUTURES</td>
<td></td>
<td>4 Presentations</td>
</tr>
<tr>
<td>April 25: 5 Presentations</td>
<td>April 27</td>
<td>L11: HOT TOPICS / TBD</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Today (2/23/18)

• Lecture 6: Stewardship and Preservation
• Research Paper instructions
• Discussion Paper
• Break
• 5 Student Presentations
Lecture 6: Data Stewardship and Preservation
Why should we care about digital stewardship and preservation?

• In the data ecosystem, data provides a “natural resource”

• Critical to *accessing and utilizing* that resource is the *stewardship* of data today

• Critical to *sustaining* that resource is the *preservation* of data over time

• Data stewardship and preservation ensure a stable “home” for data used for data-driven applications, innovation, commerce, research, etc.
  
  – “Homeless” data ceases to exist ...
Why data stewardship and preservation matter

http://youtu.be/N2zK3sAtr-4
What is data stewardship?

• *Environmental stewardship* refers to responsible use and protection of the natural environment through conservation and sustainable practices. *Wikipedia*

• We can define *data stewardship* similarly as the responsible use and protection of digital assets through management, infrastructure support, and sustainable practices.
“When” is data stewardship?

- Data stewardship and preservation important focus all throughout the “data life cycle”

{Ethics, Policy, Regulatory, Stewardship, Platform, Domain} Environment

**Acquire**
- Create, capture, gather from:
  - Lab
  - Fieldwork
  - Surveys
  - Devices
  - Simulations
  - etc

**Clean**
- Organize
- Filter
- Annotate
- Clean

**Use / Reuse**
- Analyze
- Mine
- Model
- Derive ++data
- Visualize
- Decide
- Act
- Drive:
  - Devices
  - Instruments
  - Computers

**Publish**
- Share
- Data
- Code
- Workflows
- Disseminate
- Aggregate
- Collect
- Create portals, databases, etc
- Couple with literature

**Preserve / Destroy**
- Store to:
  - Preserve
  - Replicate
  - Ignore
  - Subset, compress
  - Index
  - Curate
  - Destroy

Fran Berman, Data and Society, CSCI 4370/6370
Data stewardship promotes access and use of digital data today and data preservation promotes the access and use of digital data tomorrow.

**Key Questions:**

- What should we preserve?
- Who should access it?
- How should we preserve it?
- Who is responsible for stewardship?
- How do we pay for it?
What should we preserve?

Data that is commonly of value ...

• Administrative / “business” data
  – Increasing business automation in 21\textsuperscript{st} century. Stewardship part of the cost of doing business
  – Regulation and policy often mandates preservation

• Client / customer data
  – Competitive advantage in the private sector. Stewardship and preservation required to capitalize on this asset

• Research / public data
  – Access increasingly expected to support research, innovation, public information
Value is in the eye of the beholder ...

Digital information we* want to keep over the long-term:

**We = “Society” / Public Sector**
- Official and historically valuable data (Census information, presidential emails, Shoah Collection, etc.)

**We = Private Sector**
- Administrative / business data (organizational utility); Client/customer data (competitive advantage)

**We = Research Community**
- Data important to projects, colleagues, domains, current and future efforts

**We = Me**
- My financial data, digital photos of my kids’ graduations, etc.

---

- Many kinds of valued research data
  - Data that is in demand by researchers for replication or reuse
  - Data that is mandated to be preserved by policy or regulation
  - Data that is expected to be preserved as part of good scholarly practice
  - Data that is highly cited
  - Data for which value accrues over time
  - Data that underlies assessment reports
  - Data that is costly to reproduce or cannot be reproduced
  - Data that is timely, costly or difficult to obtain, etc.
### Sarbanes-Oxley (Public Accounting Reform and Investor Protection Act of 2002)

*Applies to all U.S. public company boards, management, and public accounting firms*

*Includes electronic records* (correspondence, work papers, memoranda, etc.) that are created, sent, or received in connection with an audit or a review

<table>
<thead>
<tr>
<th>Regulations</th>
<th>Retention Requirement</th>
<th>Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarbanes-Oxley</td>
<td>Auditors must retain relevant data for at least 7 years</td>
<td>Fines to $5M and 20 years in prison</td>
</tr>
<tr>
<td>HIPAA</td>
<td>Retain patient data for 6 years</td>
<td>$250K fine and up to 10 years in prison</td>
</tr>
<tr>
<td>Gramm-Leach-Baily</td>
<td>Ensure confidentiality of customer financial information</td>
<td>Up to $500K and 10 years in prison</td>
</tr>
<tr>
<td>SEC 17a</td>
<td>Broker data retention for 3-6 years. Some require longer retention</td>
<td>Variable based on violation</td>
</tr>
<tr>
<td>OMB Circular A-110 / CFR Part 215 (applies to federally funded research data)</td>
<td>“a three year period is the minimum amount of time that research data should be kept by the grantee”</td>
<td>Penalty structure unclear, likely fines?</td>
</tr>
</tbody>
</table>

1. “Don’t forget that email and instant messaging are business records …
4. Don’t assume that the retention requirement … is … 7 years. ... most lawyers that understand information retention agree that business records need to be kept indefinitely.

*Kevin Beaver, “Thirteen Data Retention Mistakes to Avoid”*  

---

Table information partly based on “Data Retention – More Value, Less Filling”, John Murphy, [http://www.tdan.com/view-articles/5222](http://www.tdan.com/view-articles/5222)
What do we *have* to preserve?

<table>
<thead>
<tr>
<th>Regulations</th>
<th>Retention Requirement</th>
<th>Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarbanes-Oxley</td>
<td>Auditors must retain relevant data for at least 7 years</td>
<td>Fines to $5M and 20 years in prison</td>
</tr>
<tr>
<td>HIPAA</td>
<td>Retain patient data for 6 years</td>
<td>$250K fine and up to 10 years in prison</td>
</tr>
<tr>
<td>Gramm-Leach-Baily</td>
<td>Ensure confidentiality of customer financial information</td>
<td>Up to $500K and 10 years in prison</td>
</tr>
<tr>
<td>SEC 17a</td>
<td>Broker data retention for 3-6 years. Some require longer retention</td>
<td>Variable based on violation</td>
</tr>
<tr>
<td>OMB Circular A-110 / CFR Part 215 (applies to federally funded research data)</td>
<td>“a three year period is the minimum amount of time that research data should be kept by the grantee”</td>
<td>Penalty structure unclear, likely fines?</td>
</tr>
</tbody>
</table>

**HIPAA (Health Insurance Portability and Accountability Act)**

- **Applies to health information created or maintained by health care providers “who engage in certain** electronic transactions, health plans, and health care clearinghouses”[www.hipaa.org](http://www.hipaa.org)
- **Title II:** Requires HHS to create rules and standards for the use and dissemination of health care information
- Healthcare providers must retain healthcare records for a period of **not less than 6 years.**
What do we have to preserve?

- The U.S. Office of Management and Budget requires that **federally funded research data**, supporting documentation, scientific notebooks, financial records, etc. **be maintained by the grantee (typically institution) for 3+ years**

- Sponsored research grants: research data typically owned by the grantee (institution).

- Some federal and private sector contracts require that sponsors be granted ownership or some/all rights to data
  - Many institutions retain rights to data for research and education

<table>
<thead>
<tr>
<th>Regulations</th>
<th>Retention Requirement</th>
<th>Penalty</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sarbanes-Oxley</td>
<td>Auditors must retain relevant data for at least 7 years</td>
<td>Fines to $5M and 20 years in prison</td>
</tr>
<tr>
<td>HIPAA</td>
<td>Retain patient data for 6 years</td>
<td>$250K fine and up to 10 years in prison</td>
</tr>
<tr>
<td>Gramm-Leach-Baily</td>
<td>Ensure confidentiality of customer financial information</td>
<td>Up to $500K and 10 years in prison</td>
</tr>
<tr>
<td>SEC 17a</td>
<td>Broker data retention for 3-6 years. Some require longer retention</td>
<td>Variable based on violation</td>
</tr>
<tr>
<td>OMB Circular A-110 / CFR Part 215</td>
<td>“a three year period is the minimum amount of time that research data should be kept by the grantee” [grantee = institution]</td>
<td>Penalty structure unclear, likely fines?</td>
</tr>
</tbody>
</table>
Public access expectations for sponsored research data and publications

- February 2013 OSTP Memo focus: Access to data and publications from federally funded research
- Federal R&D agencies asked to develop
  - 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
Challenge: Economics of providing public access

Holdren memo aims to

• **Publications:** “Ensure that the public can read, download, and analyze in digital form final peer reviewed manuscripts or final published documents within a timeframe that is appropriate for each type of research conducted or sponsored by the agency.

  – Specifically, each agency: i) shall use a twelve-month post-publication embargo period as a guideline for making research papers publicly available …”

• **Data:** “Maximize access, by the general public and without charge, to digitally formatted scientific data created with Federal funds, while

  I. protecting confidentiality and personal privacy,

  II. recognizing proprietary interests, business confidential information, and intellectual property rights and avoiding significant negative impact on intellectual property rights, innovation, and U.S. competitiveness, and

  III. preserving the balance between the relative value of long-term preservation and access and the associated cost and administrative burden;”

Fran Berman, Data and Society, CSCI 4370/6370
Economic issues for publishers

- Business models of most publishers built on restricting access to publications to paying customers (subscriptions, article fees, etc.).
- Access to publications for the general public without charge breaks traditional economic models.
- Rise of “open science” publications, such as the PLOS (Public Library of Science) journals provide free access to research.
  - PLOS business model is “author pays” (around $1500 to publish). Journal provides free access to all readers.
Data stewardship promotes access and use of digital data *today* and data preservation promotes the access and use of digital data *tomorrow*.

**Key Questions:**
- What should we preserve?
- Who should access it?
- **How should we preserve it?**
- Who is responsible for stewardship?
- How do we pay for it?

Print media provides easy access for long periods of time but is hard to data-mine

Digital media is easier to data-mine but requires management of evolution of media and resource planning over time
Good Practice in Data Stewardship and Preservation

- **Replication** – make multiple copies of data and store some off-site
- **Refreshing** – transfer of data between “old” versions of the same storage to new versions of the same storage to reduce bitrot and alteration of data
- **Integrity assurance** – incorporate sufficient metadata, provenance information, checksums and other techniques to ensure the integrity of data systems, content, and context
- **Forward planning / migration** – pro-actively plan and transition data to ensure sustainability across multiple technology generations
- **Sustainable economic support** – develop business model to stably support data preservation efforts, technologies, and staffing over time
- **Compliance** – Ensure that preservation systems comply with current regulations, policies, and penalties that pertain to data
- **Security and disaster planning** – ensure appropriate levels of system security to demonstrate good practice and plan ahead for recovery from disaster scenarios

**Why are 3 copies used as best practice?**

- Approach comes from Lamport, Shostak, and Pease’s solution to the *Byzantine General’s Problem*
  - Method for agreement on a battle plan for a group of Byzantine generals communicating only by messenger
  - Analogous to reliable computer systems with malfunctioning components
- Solution: When generals can send unforgeable signed messages to one another, the minimum number required for agreement is 3.
- *Paper in lecture references*
Key Players in research and public data stewardship and preservation: Librarians and Archivists

- **Archives** are the non-current records of individuals, groups, institutions, and governments that contain information of enduring value. The primary task of the **archivist** is to establish and maintain control, both physical and intellectual, over records of enduring value and ensure their content accessible for posterity.

- A **library** is an organized collection of sources of information and similar resources, made accessible to a defined community for reference or borrowing. The primary task of the **librarian** is to manage the information for discovery and use, and assist individuals in accessing and using library information.

- **Traditional professional skills expanded with key areas from information science:**
  - Knowledge of information architecture and information management systems
  - Markup languages, metadata formats, file types
  - Digitization, database management
  - Standards, policy and regulation
  - Data integrity, security, etc.
Data Stewardship and Preservation Glossary

- **Metadata** – Documentation relating to data content, structure, provenance (history), and context, “data about the data”
- **Identifier** – unique label used to reference an object or record
- **Curation** – maintaining and adding value to a trusted body of digital information for current and future use
- **Appraisal** – evaluation and selection of digital material for long-term curation and preservation
- **Authentication** – security measure designed to establish the validity of a transmission, message, or originator, or a means of verifying an individual’s authority
- **Ingest** – Controlled or secure transfer of material to an archive, repository, data center, or other custodial environment
- **Integrity** – Condition when data is unchanged from its source and has not been accidently or maliciously modified, altered, or destroyed
- **Digital Rights Management** – use of technologies to control how digital content is used and re-used
The Digital Curation Center’s Digital Data Life Cycle: Digital curation and preservation stages

Image: http://www.dcc.ac.uk/resources/curation-lifecycle-model
Archivist’s Perspective: **Open Archival Information system (OAIS) Reference Model**

- **AIP**: Archive Information Packages
- **DIP**: Dissemination Information Packages
- **SIP**: Submission Information Packages

Figure: OAIS functional entities, Wikipedia
One size doesn’t fit all: Spectrum of research data libraries, archives, domain repositories

Fran Berman, Data and Society, CSCI 4370/6370
Data stewardship promotes access and use of digital data *today* and data preservation promotes the access and use of digital data *tomorrow*.

**Key Questions:**

- What should we preserve?
- Who should access it?
- How should we preserve it?
- **Who is responsible for stewardship?**
- How do we pay for it?
Who is responsible for data stewardship and preservation?

• The “Free Rider” non-solution:
  
  – (Economics) **Free rider** refers to someone who benefits from resources, goods, or services without paying for the cost of the benefit.
  
  – Free riding may be considered as a **free rider problem** when it leads to under-provision of goods or services, or when it leads to overuse or degradation of a common property resource. [Wikipedia]

• **Free Rider problem for data**: someone else (Google, the Govt., libraries, my institution, data creators, etc.) should pay for data stewardship / preservation, but not me
Aligning the Stakeholders

• Many Stakeholders in digital access and preservation
  – Stakeholders who benefit from use of the preserved asset
  – Stakeholders who select what to preserve
  – Stakeholders who own / 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
Who is preserving your data? Common approaches to digital preservation

- **Personal data you want to keep**: You are preserving your data (on your own gear or via a service). You are responsible for ensuring that data is sustained over time (through fees, hardware migration, etc.)

- **Business data**: Companies determine what is valuable to them and include data preservation as part of their own infrastructure. Choices are made based on business priorities and regulation on what to retain and what to discard.

- **Government data**: The government is required to preserve many different kinds of data based on what is considered value (e.g. through NARA, the Library of Congress, GAO, agencies, etc.). You do not have access to all of it.

- **Research data**: Researchers preserve their data at their discretion if it is valuable, or required by funding sponsors, their institutions or publication. Where that data goes and who is responsible for it is often left up to the researcher.
The Stewardship Gap in Research Data

• **NIH estimates** for 2011 PubMed Central publications:
  - 12% of publication data sets deposited in recognized repositories, 88% of the data sets were invisible
  - Estimated approximately **200,000-235,000 invisible data sets** generated NIH work published in 2011
  - 87% of the invisible data sets are new, 13% reflect data re-use
  - More than 50% of the datasets were derived from live human or animal subjects

• **Lack of comprehensive understanding about the broader sustainable stewardship gap hampers evidence-based discussion, prioritization and potential strategic investments.**

* From PLOS ONE
Fran Bern  http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0132735
The Resource Gap – Data economics

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

Greater costs at the extremes (including “big” data) ...
What are we paying for?

Data infrastructure components may 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
Current Economic Support Models for Digital Research Data

- Subscription
- Crowd-sourcing, philanthropy
- Federal grants
- Fee for service
- Advertisement

Fran Berman, Data and Society, CSCI 4370/6370
Stewardship and Preservation for Research Data: Work in Progress

• Active community discussions around
  – “open science” and current publication models
  – Role of universities, libraries, repositories, government, etc. with respect to research data stewardship and preservation
  – Reproducibility of results
  – Business models for stewardship and preservation – who pays?
  – Role of private sector with respect to research data, etc.
Lecture 6 Sources


• Digital Curation Center Data Life Cycle http://www.dcc.ac.uk/resources/curation-lifecycle-model


• OAIS http://public.ccsds.org/publications/archive/650x0m2.pdf

• “Who will pay for Public Access to Research Data?”, Science Magazine, August 9 (on http://www.cs.rpi.edu/~bermaf/)


• “Sizing the Problem of Improving Discovery and Access to NIH-Funded Data: A Preliminary Study”, PLOS ONE http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0132735
Research Paper Assignment
Research Paper is 25% of your grade

**Paper Assignment**

- **Undergrads: 5-7 page research paper on a data-related topic (does not include references)**
- **Grads: 7-8 page research paper on a data-related topic (does not include references)**

**Paper due March 28 before class. Bring hardcopy to class**
Research Paper Structure

Specs

• Paper: 5-7 pages (undergrad) or 7-8 pages (grad), 1.5 spaced, 12 pt font

• Focus of paper should be an area of science or society that has been transformed by the availability of digital data

• General outline:
  – Description of the area/problem and the role of data in transforming it
  – Specifics on how data has made the transformation possible
  – Specifics on the kind of data tools, skills, policy, infrastructure, etc. that is needed to support this transformation
  – Conclusion and thoughts about the potential of data in this area in the future

• Paper should include adequate references and bibliography (not included in the page count).
  – If you use material from a source, reference it in the bibliography.
  – If you copy material from a source, put it in quotes and reference it in the bibliography.
Paper Grading Metrics (25 points total)

Content (13 points):

- Does the paper content provide adequate depth and evidence to describe the transformation of an area by digital data?
- Does the paper include appropriate specifics on innovation and infrastructure?
- Are the references reasonable and adequate?

Writing (12 points):

- Is the paper well-organized, readable by non-specialists, and credible to specialists?
- Does the writing tell a story? Is it articulate and compelling?
- Is the paper well-structured with the main points backed up by evidence?
- Are all the areas covered – description, specifics, infrastructure, conclusion?

Fran Berman, Data and Society, CSCI 4370/6370

- If you want to check your topic or approach, please come to office hours on March 2 (Friday 1-2, AE 218) or make an appt. with Fran before break (bermaf@rpi.edu)
- This can save you time and help you meet the expectations of the assignment
Discussion Article for February 23

• “How the data that Internet companies collect can be used for the public good”, Harvard Business Review, https://hbr.org/2018/01/how-the-data-that-internet-companies-collect-can-be-used-for-the-public-good
Break
Presentations
Presentation Articles for February 28


• “Librarians Saving the Internet”, Science Friday, https://apps.sciencefriday.com/data/librarians.html [Wei P]


Presentation Articles for March 2


Presentation Articles for March 7


Presentation Articles for Today


