Announcements

• Data Op-Ed criteria will be discussed today

• Modification from last time:

Lecture 1 Sources (not already in text) (pdfs or links on course website)

• You are responsible for material on Lecture 1 slides and in red for Exam 1

• LI Sources


– responsible for entire report (all 5 sections)

– Previous IDC Reports: http://www.emc.com/leadership/digital-universe/index.htm#Archive

– “It’s official: 2014 was the hottest year in recorded history”, Wonkblog, Washington Post, January 16

http://www.washingtonpost.com/blogs/wonkblog/wp/2015/01/16/its-official-2014-was-the-hottest-year-in-recorded-history/ (Data Roundtable)
Cool Graphics Courtesy of Ted (and CISCO)
How you’ll be graded

Student grades are computed from:

- **2 section exams** (20 points each)
- **1 section paper** (20 points):
  - Undergrads: 6-8 page research paper on an approved Section 2 topic
  - Grads: NSF-style 10 page mini-proposal on an approved Section 2 topic
- **Class participation** (10 points)
- **3 Data Roundtable reviews / presentations** (roughly 1 per section, 10 points each)

Students can obtain up to 5 points extra credit by doing an **Op-Ed** (due any time up to May 1)
Grading Detail – Extra credit Op-Ed

• Grade distribution: 3 points on editorial content (ideas, thesis, and support), 2 points on writing (does it work as an op-ed, is it compelling, does it make sense)

• Op-eds must be turned in before the beginning of class on May 1.

• Op-eds should be in 12 pt. font and between 500 and 1000 words

• FYI: See http://www.nytimes.com/2013/10/14/opinion/op-ed-and-you.html?pagewanted=all&r=0 (pdf on course website) for an article on writing an op-ed for the NY Times.
Op-Ed Detail -- Structure

Not all Op-Eds are like this, but many good Op-Eds have this structure:

- **Lede** – *Lead-in around a news hook or personal experience*

- **Thesis** – *your position (explicit or implied)*

- **Argument** – *should be based on evidence (stats, news, reports, expert quotes, scholarship, history, experience). Arguments often presented as a series of points.*

- **Criticism pre-emption** – *take the lead in acknowledging the flaws in your argument and address potential counter-arguments*

- **Conclusion** – *circle back to lede?*

---

**Lede Options**
- Current news
- Dramatic or personal anecdote
- Reference to popular culture or twist on conventional wisdom
- Anniversary of an event
- Major new study
## Section 1: The Data Ecosystem -- Fundamentals

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
<th>Roundtable</th>
</tr>
</thead>
<tbody>
<tr>
<td>January 30</td>
<td>Class introduction; Digital data in the 21st Century (L1)</td>
<td>Data Roundtable / Fran</td>
</tr>
<tr>
<td>February 6</td>
<td>Data Stewardship and Preservation (L2)</td>
<td>L1 Data Roundtable / 5 students</td>
</tr>
<tr>
<td>February 13</td>
<td>Data and Computing (L3)</td>
<td>L2 Data Roundtable / 6 students</td>
</tr>
<tr>
<td>February 20</td>
<td>Colin Bodel, Time Inc. CTO Guest Lecture and Q&amp;A</td>
<td>L3 Data Roundtable / 5 students</td>
</tr>
</tbody>
</table>

## Section 2: Data and Innovation – How has data transformed science and society?

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
<th>Roundtable</th>
</tr>
</thead>
<tbody>
<tr>
<td>February 27</td>
<td>Section 1 Exam</td>
<td>Data and the Health Sciences (L4)</td>
</tr>
<tr>
<td>March 6</td>
<td>Paper preparation / no class</td>
<td></td>
</tr>
<tr>
<td>March 13</td>
<td>Data and Entertainment (L5)</td>
<td>L4 Data Roundtable / 5 students</td>
</tr>
<tr>
<td>March 20</td>
<td>Big Data Applications (L6)</td>
<td>L5 Data Roundtable / 5 students</td>
</tr>
</tbody>
</table>

## Section 3: Data and Community – Social infrastructure for a data-driven world

<table>
<thead>
<tr>
<th>Date</th>
<th>Event</th>
<th>Roundtable</th>
</tr>
</thead>
<tbody>
<tr>
<td>April 3</td>
<td>Data in the Global Landscape (L7) Section 2 paper due</td>
<td>L6 Data Roundtable / 5 students</td>
</tr>
<tr>
<td>April 10</td>
<td>Bulent Yener Guest Lecture, Data Privacy / Bad guys on the Internet  (L8)</td>
<td>L7 Data Roundtable / 5 students</td>
</tr>
<tr>
<td>April 17</td>
<td>Data and the Workforce (L9)</td>
<td>L8 Data Roundtable / 5 students</td>
</tr>
<tr>
<td>April 24</td>
<td>Mike Schroepfer, Facebook CTO Guest Lecture and Q&amp;A</td>
<td></td>
</tr>
<tr>
<td>May 1</td>
<td>Data Futures (L10)</td>
<td>L9 Data Roundtable / 5 students</td>
</tr>
<tr>
<td>May 8</td>
<td>Section 3 Exam</td>
<td>L10 Data Roundtable / 5 students</td>
</tr>
</tbody>
</table>
Today (2/6/15)

• Any questions about Lecture 1?
• Lecture 2: Data Stewardship and Preservation
  – Why data stewardship and preservation matter
  – Key issues and challenges in stewardship and preservation
  – Data stewardship and preservation models and infrastructure
  – Economics of data preservation
  – Economics of research data

• Break

• Data Round Table (Karl, Amit, Charles, Sumit, Kate)
Lecture 2: 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
- Critical to sustaining that resource is the preservation of data over time
- Data stewardship and preservation provides a stable “home” for data that is a pre-requisite for data-driven applications, innovation, commerce, research
  - “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.
What is data preservation? (Wikipedia)

- “In library and archival science, digital preservation is a formal endeavor to ensure that digital information of continuing value remains accessible and usable.

- It involves planning, resource allocation, and application of preservation methods and technologies, and it combines policies, strategies and actions to ensure access to reformatted and "born-digital" content, regardless of the challenges of media failure and technological change.

- The goal of digital preservation is the accurate rendering of authenticated content over time.”
Data stewardship promotes access and use of digital data *today* and data preservation promotes the access and use of digital data *tomorrow.*

- Many aspects of data stewardship and preservation:
  - **Software systems** for managing data
  - **Data storage** and hardware
  - **Community policy and practice**
  - **Economic models**
  - **Community / professional culture** and expectation around curation, data sharing, stewardship, preservation, etc.
Good Practice in Data Stewardship and Preservation

- **Replication** – make multiple copies of data and store some off-site

- **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
Challenges in Stewardship and Preservation

• What should we preserve?
  – What materials must be “rescued”?
  – How to plan for preservation of materials by design?
  – What is the “original”?

• How should we preserve it?
  – Curation
  – Metadata
  – Storage
  – Facilities
  – Monitoring
  – Migration
  – Copies

• Who should pay?
  • Business models for “initialization”
  • Business models for “steady state”

• Who can access?
  • Who should access digital materials?
  • What tools should be provided for access?

• Who is responsible?
  • Data creators
  • Professional communities
  • Libraries and archivists
  • Government agencies

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
Which data should we save?

Digital information 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 financial data, digital photos of my kids’ graduations, etc.
The Data Pyramid: A Framework for Digital Stewardship and Preservation

Digital Data Collections
- Reference, nationally/Internationally important, irreplaceable data collections
- Key research and community data collections
- Personal data collections

Societal Scale
Increasing responsibility, increasing security, increasing stability

Community Scale

Local Scale

Increasing Societal Value

Repositories / Facilities
- National / international scale repositories, libraries, archives
- “Regional” scale libraries and targeted data archives / centers
- Private repositories – your computer, your iPhone, the cloud

The Data Pyramid
**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

---

**What do we have to save?**

<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”</td>
<td>Penalty structure unclear, likely fines?</td>
</tr>
</tbody>
</table>

---

1. “Don’t forget that email and instant messaging are business records …
2. 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”

http://searchdatamanagement.techtarget.com/news/article/0,289142,sid91_qci1186910,00.html

---

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

<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 save?**

- 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 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
  - Most 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 (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” [grantee = institution]</td>
<td>Penalty structure unclear, likely fines?</td>
</tr>
</tbody>
</table>
Regulation and organizational value often dictate level of stewardship, support

- In the **private sector**, data is a business asset and often a competitive advantage
  - Infrastructure and costs of data stewardship part of the cost of doing business

- In the **public sector**, data is both a business and a public asset
  - Agencies include infrastructure for business data stewardship as part of the cost of doing business
  - Agencies may provide data as a public asset for the broader public

- In the **academic sector**, organizational data is a business asset; research data is a critical driver of new innovation
  - Institutions include infrastructure for data stewardship as part of organizational infrastructure
  - Research data stewardship often left up to researcher’s discretion for dissemination, stewardship, support
Key Players in 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 Glossary

• **Metadata** – Documentation relating to data content, structure, provenance (history), and context

• **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
DCC Life Cycle and Data Creators

• It is a common misconception that data is created or captured and then passed on to someone else to curate. Much of the crucial information required for effective long-term curation and reuse must be captured at the conceptualization and collection stages.

• Key considerations for data creators.
  – How meaningful the data is to other users
  – Whether it can be accessed, shared, and re-used in the short or long-term
  – Whether the data may or may not be selected for ingest into an archive (i.e., does the data conform to archival standards, can the data be stored and preserved?)
  – Which transformations can be performed on the data (e.g. migration to new file formats)
  – How easily other researchers can find and understand the data for reuse
  – Whether the data can be proven to be authentic and have integrity (i.e. is what it purports to be and has not been changed or tampered with since creation, a crucial characteristic of scientific data)
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
More About OAIS

• OAIS is an archive that promotes preservation for a designated community.
  – “Open” in OAIS: model developed in open forum (not that all information in OAIS archive is unrestricted)

• Mandatory responsibilities for OAIS archives. **OAIS must**
  – Negotiate for and **accept appropriate information** from Information Producers
  – **Obtain sufficient control** of the information provided to the level needed to ensure long-term preservation
  – Determine which communities should become the “**Designated Community**” (who will understand the information provided)
  – Ensure that the information to be preserved is **independently understandable** to the Designated Community
  – Follow documented policies and procedures which **ensure that the information is preserved against all reasonable contingencies**, and which **enable the information to be disseminated** as authenticated copies of the original, or as traceable to the original
  – Make the preserved information **available** to the Designated Community

Fran Berman, Data and Society, CSCI 4967/6963
Preservation Infrastructure: **LOCKSS**

- **LOCKSS (Lots of Copies Keep Stuff Safe)** is an open-source, library-led digital preservation system.
  - Publishers participate at no cost, libraries pay between 2K-12K per year, depending on size

- LOCKSS allows librarians at each institution to take custody of and preserve access to the e-content to which they subscribe (like print purchase model).

- **LOCKSS model provides libraries’ the ability to build and preserve “local” e-collections.**
  - Libraries acquire digital content in their local LOCKSS box.
  - When publisher’s web site unavailable, content served from the library’s LOCKSS box, guaranteeing continuous user access.

- LOCKSS preserves the original publisher’s copy of each item, ensuring that the most authoritative version persists.

- LOCKSS box performs 5 main functions:
  - **Ingests** content from target websites using a web crawler similar to those used by search engines.
  - **Preserves** content by continually comparing the content it has collected with the same content collected by other LOCKSS Boxes, and repairing any differences.
  - **Delivers** authoritative content to readers by acting as a web proxy, cache or via Metadata resolvers when the publisher’s website is not available.
  - **Manages** content through a web interface that allows librarians to select new content for preservation, monitor the content being preserved and control access to the preserved content.
  - **Migrates content** dynamically to new formats as needed for display.
Preservation Infrastructure: Digital Preservation Network

- The Digital Preservation Network (DPN) formed to ensure that the scholarly record is preserved for future generations.

- DPN uses a **federated approach** to preservation. By replicating multiple dark copies of these collections in diverse nodes, **DPN protects against the risk of catastrophic loss due to technology, organizational or natural disasters.** Diversity of the DPN nodes mitigates the risk of a single point of failure.

- **Implementation**
  - Local repositories become contributing nodes, which ingest new forms of scholarship and new collections.
  - DPN creates federated, replicating nodes, which are digital repositories for the contributing nodes with a specific focus on long-term preservation.
    - Replicating nodes multiple preservation repositories (at least 3) to reduce single point of failure
    - Replicating nodes contain redundant, dark copies of all deposits that can be “brightened” in cases of catastrophic loss.
    - Objects and metadata are replicated across nodes that embody organizational, technical, physical, and political diversity. A single point of failure unlikely to jeopardize centuries of scholarship.

Fran Berman, Data and Society, CSCI 4967/6963
Data economics: Responsible data stewardship requires a viable business model for sustaining its underlying infrastructure

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

Greater costs at the extremes (including “big” data) …
Data Stewardship and Preservation Incurs Real Costs

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
Who should pay for data?

• 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.
Current Economic Support Models for Digital Data

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

Fran Berman, Data and Society, CSCI 4967/6963
Deep Dive Exploration of Data Economics:
Blue Ribbon Task Force on Sustainable Digital Preservation and Access (BRTF)

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

Web page and materials at brtf.sdsc.edu

Fran Berman, Data and Society, CSCI 4967/6963
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

*From Blue Ribbon Task Force Interim Report*
Organizational challenges to economically sustainable stewardship
(from the Blue Ribbon Task Force Interim Report [at brtf.sdsc.edu])

• **Poor alignment between stakeholders** in the digital preservation and access world and their roles, responsibilities and support models

• There is a **lack of institutional, enterprise, and/or community incentives** to support the collaboration needed to enforce sustainable economic models

• **Complacency that current practices are “good enough”** and / or the problem is not urgent. Both “carrots” (in the form of recognition that access to information is an investment in current and future success) and “sticks” (in the form of penalties for non-compliance, accounting of explicit opportunity costs, or costs of lost information) are needed

• Fear that digital access and **preservation is too big to take on**

• **Prioritization of resources for pressing problems over infrastructure**
The Stakeholder Problem

• 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
Commerically Owned Cultural Content

- **Stakeholders who benefit:** the general public, cultural historians
- **Stakeholders who select:** Studios, third-party organizations
- **Stakeholders who own / have rights:** Studios, third-party organizations
- **Stakeholders who preserve:** Institutional and individual repositories, third-party organizations, etc.
- **Stakeholders who pay:** Studios, professional organizations, private owners, custodial organizations, etc.

**Needed actions involve**

- Alignment of requirements for copyright deposit with the requirements of digital preservation and access
- Development and involvement of organizations that can ensure secure handoffs of cultural materials from private owners to economically viable public preservers
Collectively-produced Web Content

- **Stakeholders who benefit:** the general public, public, private, academic sectors
- **Stakeholders who select:** Often the entities that preserve the data
- **Stakeholders who own / have rights:** Often unclear
- **Stakeholders who preserve:** Private and public organizations for their own or public use
- **Stakeholders who pay:** Private, public, preservation-focused and other organizations

**Needed actions involve**
- the development of appropriate licensing and regulations that permit third-parties to preserve web content
- the development of incentives for host sites or third parties to preserve

Findings from Blue Ribbon Task Force Final Report
Scholarly Discourse

- **Stakeholders who benefit:** the greater research and learning community
- **Stakeholders who select:** Publishers, based on community review
- **Stakeholders who own / have rights:** Publishers generally own rights
- **Stakeholders who preserve:** Publishers and third-party entities
- **Stakeholders who pay:** Publishers, third party organizations

**Needed actions involve**
- Clarification (with respect to licensing, ownership, rights, etc.) of the responsibilities of publishers, third-party archives, and scholars
- Granting of non-exclusive rights to content by scholars to enable decentralization of publishing and preservation

*Findings from Blue Ribbon Task Force Final Report*
Research Data

- **Stakeholders who benefit:** the greater research community, broader public, private sector
- **Stakeholders who select:** Often the data creators
- **Stakeholders who own / have rights:** Often the Data creators and/or their institutions
- **Stakeholders who preserve:** Often the data creators, their institutions, or some third party
- **Stakeholders who pay:** Often Federal agencies, institutions

**Needed actions involve**
- the development of federal agency policies that mandate the stewardship of important research data
- the identification of viable support options for third-party archives (e.g. university libraries) to host valuable research data
Increasing federal expectations for data management and preservation of research data and the scholarly record
New expectations from U.S. Office of Science and Technology Policy

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.
OSTP Memo applies to both scholarly publications and digital research data

• Memo targeted to Federal agencies with over $100M in research and development expenditures.
  – Requested a plan by September 2013

• Stated goal of memo:
  – “to the greatest extent and with the fewest constraints possible ... the direct results of federally funded scientific research are made available to and useful for the public, industry and scientific community. Such results include peer-reviewed publications and digital data.”
  – “Federal agencies investing in research and development must have clear and coordinated policies for increasing [access to federally funded published research and digital scientific data]”
OSTP Requested Agency Plan Elements

• **Strategy for capitalizing on what exists:**
  – Strategy for leveraging existing archives and fostering public-private partnerships with scientific journals

• **Strategy for increasing / enhancing discoverability, access, dissemination, stewardship, preservation:**
  – Strategy for improving the public’s ability to discover and access digital data resulting from federally funded research
  – Approach for optimizing search, archival, and dissemination features that encourage innovation in accessibility and interoperability, while ensuring long-term stewardship

• **Approach for Accountability:**
  – Plan for notifying awardees and federally funded scientific researchers of their obligations
  – Agency strategy for measuring and enforcing compliance with its plan
  – Timeline for implementation

• **No new money:**
  – Identification of resources within the existing agency budget to implement the plan
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.

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.
Economics of OSTP Memo Public Access: Who Pays the Data Bill?

Who Will Pay for Public Access to Research Data?

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 timelines 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

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 stepped up to take on the problem alone. It is unrealistic to expect as much. In any sector, federal R&D agencies are expected to allocate enough resources to support research data from federally funded research data. The...
Op-ed recommendations: **Cultivate / coordinate preservation and stewardship options in every sector**

- Private Sector
- Public Sector
- Academia
- Individuals
Op-ed recommendations: Cultivate / coordinate preservation and stewardship options in every sector

Private Sector

• Facilitate private sector stewardship of public access research data as a public good

Public Sector

• Clarify public sector stewardship commitments: articulate what data will / won’t be supported

Charleston Ballet blog: http://allianceblog.org/tag/charleston-ballet/; corporate and collection logos
Op-ed recommendations: Cultivate / coordinate preservation and stewardship options in every sector

Academic Sector
• Create sustainable university library and repository stewardship solutions

Individuals
• Evolve research culture to take advantage of what works in the private sector
No magic economic bullet. Coordination between approaches can provide even more robust options for stewardship.
Individual Action: What individuals can do to promote data stewardship and preservation

• **Cite and publish your data** when you write about your results. Work with your professional societies and conferences to include “data sessions” (*idea from Sibel Adali*)

• Create and implement a **data management and stewardship plan** for your project for a reasonable fixed term of time.
  – **Budget realistically** for the costs of data stewardship and preservation
  – **Prioritize the “data bill”** at the same level as other critical infrastructure.

• **Contribute /create a local / community culture of data sharing**
  – **Make your data available** to the community (as appropriate) by curating it and ingesting it into a publicly accessible repository
  – **Adopt / support policy and practice** that enables the development and continued maintenance of sustainable stewardship, data sharing, and broad access
Lecture 2 Sources (links on website)

- You are responsible for material on Lecture 2 slides and in red for Exam 1
  - Digital Curation Center Data Life Cycle [http://www.dcc.ac.uk/resources/curation-lifecycle-model](http://www.dcc.ac.uk/resources/curation-lifecycle-model)
  - OAIS [http://public.ccsds.org/publications/archive/650x0m2.pdf](http://public.ccsds.org/publications/archive/650x0m2.pdf)
  - “Sustainable Economics for a Digital Planet” Blue Ribbon Task Force on Sustainable Preservation and Access Final Report, [brtf.sdsc.edu](http://brtf.sdsc.edu)

Fran Berman, Data and Society, CSCI 4967/6963
Break
Data Round Table
Next Week: Lecture 2 Data Roundtable (February 13)


- Princeton Single-Pay Storage model (2 students):
  - “DataSpace: A Funding and Operational Model for LongTerm Preservation and Sharing of Research Data” White Paper (August, 2010), [http://dataspace.princeton.edu/jspui/bitstream/88435/dsp01w6634361k/1/DataSpaceFundingModel_20100827.pdf](http://dataspace.princeton.edu/jspui/bitstream/88435/dsp01w6634361k/1/DataSpaceFundingModel_20100827.pdf) (Lars Olsson)
Two weeks: L3 Data Roundtable for February 20 (Need volunteers)


Today: Lecture 1 Data Roundtable (today)


