Announcements

• Office hours today are 1:15-2:00 (AE 218)
• Data Op-Ed criteria will be discussed today
• Session 1 exam: February 26 (20 points)
How you’ll be graded

Student grades are computed from:

• **2 section exams** (20 points each)

• **1 section paper** (20 points):
  – Undergrads: 7 page research paper on an approved topic
  – Grads: 10 page research paper on an approved topic

• **Class participation** (10 points)

• **1 Op-Ed** (10 points, due before April 29)

• **2 Data Roundtable reviews / presentations** (roughly 1 per section, 10 points each)

Time permitting, one “do-over” op-ed or data roundtable may be accepted. More information on whether that is possible will be provided after Section 1.
Grading Detail – Op-Ed (do one of these)

• Grade distribution: 5 points on editorial content (ideas, thesis, and support), 5 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 April 29.

• 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.
Writing op-eds

• Op-eds can have tremendous influence on community and stakeholders
  – Can establish you as an expert
  – Can get your point of view into the public discourse
  – Can be useful to your company, project or community

• Who is your audience: General public

• What is your purpose: Persuasively get your point of view across
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
Op-Ed Tips

• Write it in a way that smart people can relate to, even if they are not in your discipline. Don’t use buzzwords or talk “inside baseball” without explaining things.

• Pay attention to publication word count – op-eds are usually quite short

• *If you do this for real:*
  – The final version may be reviewed and/or edited – what you send in may not be the final draft
  – Do your homework – everyone will read this
  – Be prepared for feedback – blogs, tweets, etc.
| Section 1: The Data Ecosystem -- Fundamentals | January 29 | Class introduction; Digital data in the 21st Century (L1) | Data Roundtable / Fran
| | February 5 | Data Stewardship and Preservation (L2) | L1 Data Roundtable / 5 students
| | February 12 | Data-driven Science (L3) | L2 Data Roundtable / 5 students
| | February 19 | Future infrastructure – Internet of Things (L4) | L3 Data Roundtable / 5 students
| | February 26 | Section 1 Exam | L4 Data Roundtable / 5 students
| Section 2: Data and Innovation – How has data transformed science and society? | March 4 | Paper assignment description | Section 1 Data Roundtable / 5 students
| | March 11 | Data and Health: Phil Bourne guest lecture (L5) | Section 2 Data Roundtable / 5 students
| | March 18 | Spring Break / no class | |
| | March 25 | Data and Entertainment (L6) | L5 Data Roundtable / 5 students
| | April 1 | Big Data Applications (L7) | L6 Data Roundtable / 5 students
| Section 3: Data and Community – Social infrastructure for a data-driven world | April 8 | Data in the Global Landscape (L8) | L7 Data Roundtable / 5 students
| | April 15 | Digital Rights (L9) | L8 Data Roundtable / 5 students
| | April 22 | Bulent Yener Guest Lecture, Data Security (L10) | L9 Data Roundtable / 5 students
| | April 29 | Digital Governance and Ethics (L11) | L10 Data Roundtable / 5 students
Today (2/5/16)

• Any questions about Lecture 1?
• Lecture 2: Data Stewardship and Preservation
  – Why data stewardship and preservation matter
  – Stewardship and preservation of valuable data
    • What should we preserve?
    • How should we preserve it?
    • Who should access it?
    • Who is responsible for stewardship?
    • How do we pay for it?

• Break

• Data Round Table
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 today
- 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*.

**Key Questions:**

• What should we preserve?
• How should we preserve it?
• Who should access 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 21st 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
The Data Pyramid: A Framework for Value and Stewardship

Digital Data Collections
- Reference, nationally/ internationally important, irreplaceable data collections
- Key research and community data collections
- Commercial data; personal data collections

Increasing social or community value
Increasing expectation of public / open access

Societal Scale
- Increasing responsibility, increasing need for security and policy
- Increasing need for economic stability

Repositories / Facilities
- National / international scale repositories, libraries, archives
- Community scale or institutional libraries and targeted data archives / centers
- Commercial stewardship, private repositories – your computer, your iPhone, the cloud

The Data Pyramid
• **Administrative / business data is an organizational utility**
  – Maintained as part of the business systems of the organization
  – Ensures that the organization complies with relevant policy and regulation concerning data
  – Works with internal “users” to develop and maintain effective systems

• **Customer / client data can provide a competitive advantage**
  – Companies may keep data assets private and not engage in community data sharing efforts
  – Some data sets (e.g. Twitter) released to the public for analysis and research at scale
  – Company policy highly dependent on competitive landscape and value of data

• **Research data often at risk**
  – No clear designation of responsibility among stakeholders (more about this later)
**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”
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 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</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]
- 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.**

Fran Berman, Data and Society, CSCI 4370/6370
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 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</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>
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?
- **How should we preserve it?**
- Who should access it?
- Who is responsible for stewardship?
- How do we pay for it?

- **Administrative / “business” data**
  - Increasing business automation in 21st 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**
  - Increasingly access expected to support research, innovation, public information
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
More value → more infrastructure, more cost

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

Information courtesy of Richard Moore, SDSC
Many challenges in Research / Public Data Stewardship and Preservation

**What should we preserve?**
- What materials must be “rescued”?
- What is valuable to the community?
- How do we plan for preservation of materials by design?
- What is the “original”?

**How should we preserve it?**
- Curation
- Metadata
- Storage
- Facilities
- Monitoring
- Migration
- Copies
- Disaster mitigation

**Who can access and use?**
- Who should access digital materials?
- What tools should be provided for access?
- What tools are needed for use?

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

**Who should pay?**
- Business models for “initialization”
- Business models for “steady state”
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 accidentally 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 archives 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 4370/6370
Certification / Evaluation Approaches for Preservation Environments

- **TRAC** [Trustworthy Repositories Audit and Certification: Criteria and Checklist] – document allowing digital repositories to assess their capability to reliably store, migrate and provide access to digital content

- **DRA MBORA** [Digital Repository Audit Method Based on Risk Assessment] – 6 stage methodology and toolkit for digital repository self-assessment based on risk and risk management

- **Data Seal of Approval** – 16 guidelines to assess digital repositories to ensure that data can be discovered, understood and used in the future

- **Audit and Certification of Trustworthy Digital Repositories (ISO 16363)** – approved national standard based on OASIS and TRAC checklist. Provides a detailed specification of criteria against which the trustworthiness of a digital repository can be evaluated.
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.
Multiple Approaches to Stewardship and Preservation Infrastructure

- **Dryad** – curated resource that makes the data from underlying scientific publications discoverable, freely reusable and citable. Non-profit membership organization.

- **DataVerse** -- Open source web application to share, preserve, cite, explore and analyze research data. Code available on GitHub. Harvard Dataverse repository hosts research data. Funded by Harvard, Sloan, NSF, etc.

- **DurasSpace** – open source repository software package used for creating open access repositories. Merged from Fedora Commons and Dspace technologies. Serves a specific need as a digital archives system.

- **DuraCloud** – open source technology project for preserving and archiving digital content. Service provided by DuraSpace. Uses heterogeneous cloud technologies to support data replication.

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

- **Administrative / “business” data**
  - Increasing business automation in 21st 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**
  - Increasingly access expected to support research, innovation, public information
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) ...
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 Research Data

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

Fran Berman, Data and Society, CSCI 4370/6370
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
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*

**Digital sustainability** concentrates less on the solution and technology and more on building an infrastructure and approach that is flexible with an emphasis on interoperability, continued maintenance and continuous development.

**Digital sustainability** incorporates activities in the present that will facilitate access and availability in the future. [Wikipedia]
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.
EDITORIAL

We Paid for the Research, So Let’s See It

Published: February 25, 2013

The Obama administration is right to direct federal agencies to make public, without charge, all scientific papers reporting on research financed by the government. In a memorandum issued on Friday, John Holdren, the president’s science adviser, directed federal agencies with more than $100 million in annual research and development expenditures to develop plans for making the published results of almost all the research freely available to everyone within one year of publication.

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 funded 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.
Public Access for Research Data: Who Pays the Data Bill?

On 22 February, the U.S. Office of Science and Technology Policy (OSTP) released a memo calling for public access for publications and data resulting from federally sponsored research grants (1). The memo directed federal agencies with more than $100 million R&D expenditures to “develop a plan to support increased public access to the results of research funded by the Federal Government.” Perhaps even more succinctly, a subsequent New York Times opinion page sported the headline “We Paid for the Research, So Let’s See It” (2). So who pays for data infrastructure?

The OSTP memo requested agencies to provide plans by September 2013 that describe their strategies for providing public access to both research publications and research data. Plans are expected to be implemented using “resources within the existing agency budget,” i.e., no new money should be expected. Currently, federal R&D agencies are working hard to foster approaches to public access, to assess needs for supporting partnerships and enabling infrastructure, and to develop timetables and approaches for implementation. We focus here on the research data portion of the OSTP memo.

Research data of community value are supported today in a variety of ways. Some of them, like those in the Protein Data Bank (PDB) (3)—a database of protein structure information used heavily by the life sciences community—are supported by the public sector. (In particular, U.S. funding from the National Science Foundation (NSF), the National Institutes of Health (NIH), and the U.S. Department of Energy for the Research Collaboratory for Structural Bioinformatics (RCSB) PDB is $6.3 million annually.) Other data, as from the Longitudinal Study...

Digital Repository
@dnireland
Berman and Cerf "Who will pay for public access" behind paywall :( m.science mag.org/content/341/61...
#ipres2013 #irony

Op-ed recommendations: Cultivate / coordinate preservation and stewardship options in every sector
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.
Lecture 2 Sources

- 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)
Break
Two weeks: L3 Data Roundtable for February 19


Next week: **Lecture 2 Data Roundtable (February 12)**


<table>
<thead>
<tr>
<th>Data Round Table</th>
</tr>
</thead>
</table>

Fran Berman, Data and Society, CSCI 4370/6370
Today: Lecture 1 Data Roundtable


