Data and Society
Data Stewardship, Preservation and Cyberinfrastructure – Lecture 15

3/25/21

Today's Class

- **Personal Essay (Face Recognition) due on April 1 / Instructions in Lecture 1**
- Lecture / Discussion
- Student Presentations

Read before 3/29

- **The Wired Guide to the Internet of Things**, Wired
- https://www.wired.com/story/wired-guide-internet-of-things/?code=3HIcRlnhQIU--k_iXDHCMnir6TWv6sknycysnX6wUQr&state=%7B%22redirectURL%22%3A%22https%3A%2F%2Fwww.wired.com%2Fstory%2Fwired-guide-internet-of-things%2F%3Futm_source%3DWIR_REG_GATE%22%7D&utm_source=WIR_REG_GATE

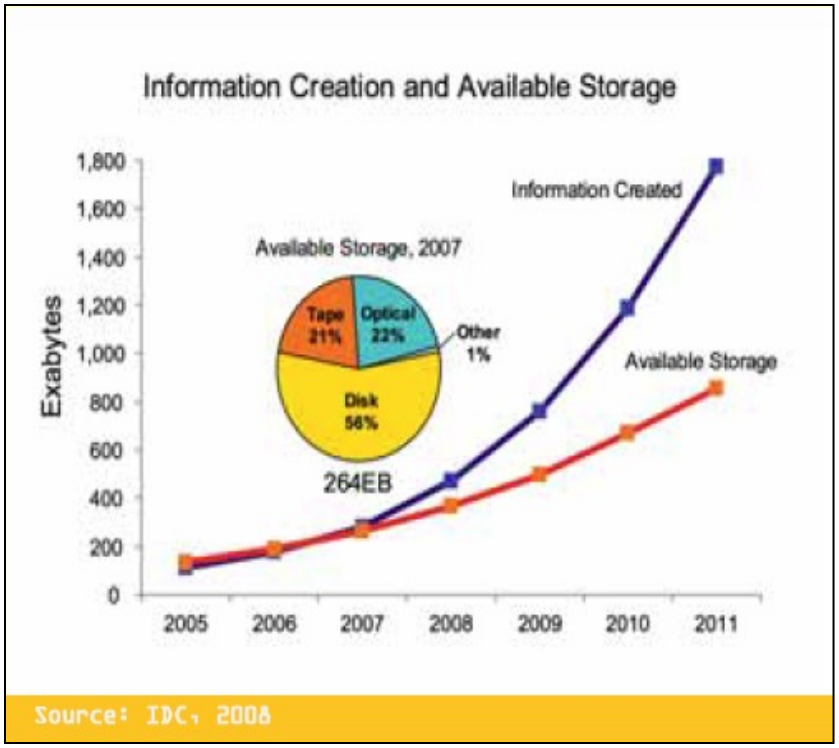
Date	Topic	Speaker	Date	Topic	Speaker
1-25	Introduction	Fran	1-28	The Data-driven World	Fran
2-1	Data and COVID-19	Fran	2-4	Data and Privacy -- Intro	Fran
2-8	Data and Privacy – Differential Privacy	Fran	2-11	Data and Privacy – Anonymity / Briefing Instructions	Fran
2-15	NO CLASS / PRESIDENT'S DAY		2-18	NO CLASS	
2-22	Legal Protections	Ben Wizner	2-25	Data and Discrimination 1	Fran
3-1	Data and Discrimination 2	Fran	3-4	Data and Elections 1	Fran
3-8	Data and Elections 2	Fran	3-11	NO CLASS / WRITING DAY	
3-15	Data and Astronomy (Op-Ed due)	Alyssa Goodman	3-18	Data Science	Fran
3-22	Digital Humanities	Brett Bobley	3-25	Data Stewardship and Preservation	Fran
3-29	Data and the IoT	Fran	4-1	Data and Smart Farms	Rich Wolski
4-5	Data and Self-Driving Cars	Fran	4-8	Data and Ethics 1	Fran
4-12	Data and Ethics 2	Fran	4-15	Cybersecurity	Bruce Schneier
4-19	Data and Dating	Fran	4-22	Digital Rights in the EU and China	Fran
4-26	Tech in the News	Fran	4-29	NO CLASS	Fran
5-3	Wrap-up / Discussion				

Lecture

- **Data Stewardship and Preservation**
- **The Internet Archive**

Data stewardship promotes access and use of digital data ***today*** and **data preservation** promotes the access and use of digital data ***tomorrow***.

- Which data should we preserve?
- Who maintains and preserves it?
- Who preserves the Internet?



Who is preserving data?

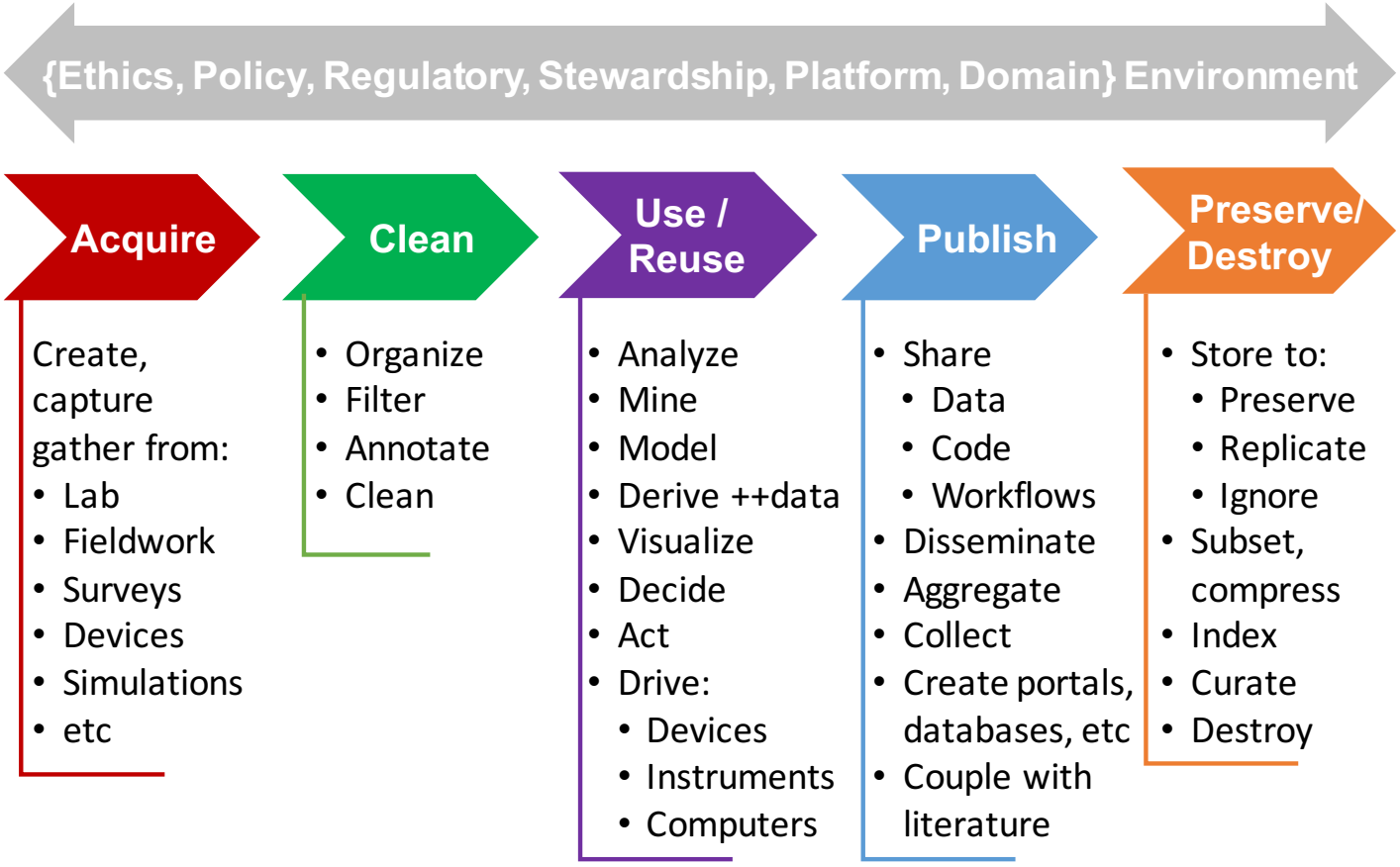
- **Personal data you want to keep**: You are preserving your data (on your own gear or via a service). You are responsible for ensuring that data is sustained over time (through fees, hardware migration, etc.)
- **Business data**: Companies determine what is valuable to them and include data preservation as part of their 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 valuable (e.g. through NARA, the Library of Congress, GAO, agencies, NSA, etc.). You do not have access to all of it.
- **Research data**: Researchers preserve their data at their discretion if it is valuable, or as required by funding sponsors, their institutions, or publications. Where that data goes and who is responsible for it is often left up to the researcher.

Some data must be preserved by law

Regulations	Type of data	Retention Requirement	Penalty
Sarbanes-Oxley	**Business data** for U.S. public company boards, management, and public accounting firms	Auditors must retain relevant data for at least 7 years	Fines to $5M and 20 years in prison
HIPAA	**Health data** created or maintained by health care providers	Retain patient data for 6 years	$250K fine and up to 10 years in prison
OMB Circular A-110 / CFR Part 215 (applies to federally funded research data)	**Federally funded research data** – including supporting documentation, scientific notebooks, financial records, etc. be maintained by the grantee (typically institution)	"a three year period is the minimum amount of time that research data should be kept by the grantee"	Penalty structure unclear, likely fines?

Data stewardship and preservation should be planned from the start

- Data stewardship and preservation important focus all throughout the **"research data life cycle"**



Research Data

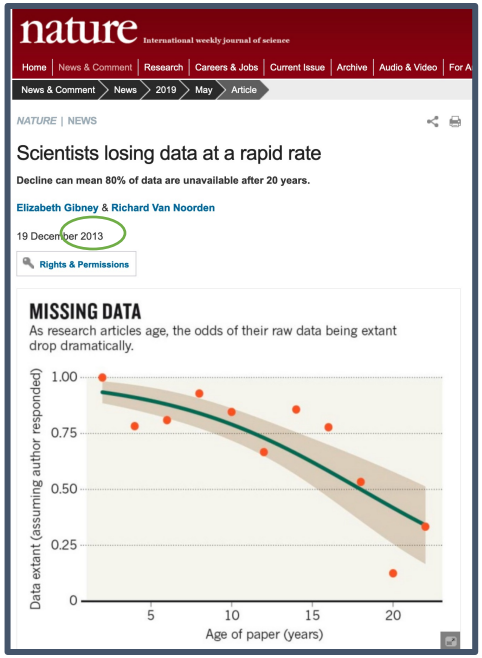
Why data stewardship and preservation matter
(4:40 min)

http://youtu.be/N2zK3sAtr-4

In research world, stewardship and preservation is uneven

The Dataverse® Project

Dataverse provides open source research data repository software, which can be used by researchers, journals, institutions and developers. Many institutions (e.g. Harvard) provide a local Dataverse used for stewardship and preservation of eligible data.



tair

The Arapidopsis Information Resource is a community database for plant biology. Originally funded by NSF for many years, the group launched a non-profit (Phoenix Bioinformatics) and now offers this DB and others through subscription.

Best 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** – create 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

Professionals 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.

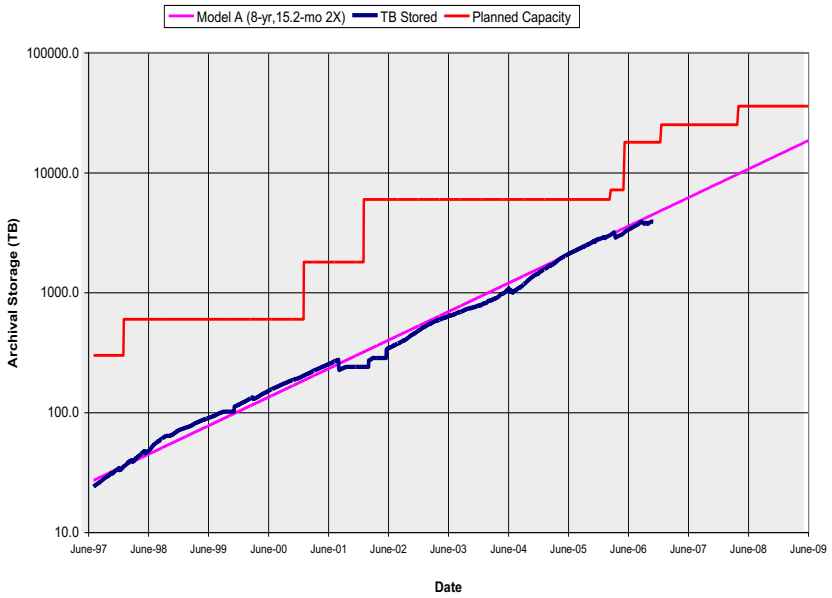
- **Traditional professional skills expanded with key areas from information science:**
  - Knowledge of information architecture and information management systems
  - Markup languages, metadata formats, file types
  - Digitization, database management
  - Standards, policy and regulation
  - Data integrity,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

Data Stewardship and Preservation is not free

Costs of stewardship and preservation 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

Who preserves data for the public interest?

- **Library of Congress** preserves digital materials related to American history and culture. Digital items include ”born digital” materials (audio, video, films, photographs, tweets, etc.) as well as digitized materials. LoC has stringent selection process to decide what it will and will not preserve within its collections.

- The **National Archives** preserves records of the U.S. including the Constitution, Bill or Rights, military records, etc. Digital holdings include digitized and born digital records including presidential emails and other materials.

- Average lifespan of a **website** is ~2.6 months. Who preserves the Internet?

Preserving the Internet: The Internet Archive

- Internet Archive is a digital library whose mission is “universal access to all knowledge”
  - Non-profit
  - Started by Brewster Kahle
- Free public access to collections of digitized materials, including websites.
- Internet Archive currently holds > 48 PB of materials including 20+M books and texts, 6+M movies and videos, 600K SW programs, 15M audio files and **480+B web pages in the Wayback Machine.**

How does the Internet Archive preserve the web?

- Web crawlers work to preserve as much of the public web as possible. Webpages stored in the **Wayback Machine**.
- Users can view archived webpages.
  - Provides public access to code, images, source code from websites that may no longer exist or have been updated
  - About half of website hyperlinks included
- Not everything is crawled, only public Internet. Website owners can opt out.
- Frequency of website capture also varies per website, based on which crawl list(s) it's on

Expanding content

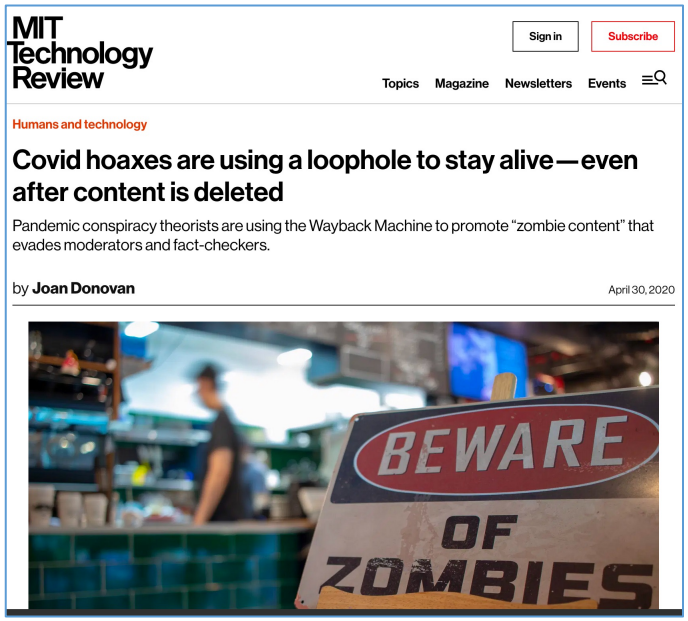
The Internet Archive is now preserving Flash games and animations, The Verge, https://www.theverge.com/2020/11/19/21578616/internet-archive-preservation-flash-animations-games-adobe

- "The Internet Archive — the non-profit digital library known for the Wayback Machine — announced that it will now preserve Flash animations and games, ahead of Adobe's planned demise for the defunct web software at the end of 2020. The Archive will emulate the content so it plays as it used to, preserving critical elements of early internet culture for browsers that can no longer run them.

- The Internet Archive says you can already browse over 1,000 games and animations that it's saved, including classics like "Peanut Butter Jelly Time" and "All your base are belong to us". The organization says emulation is made possible by an in-development Flash emulator called Ruffle that it's incorporated into its system. While Ruffle's developers say it isn't currently compatible with a majority of Flash projects made after 2013, having any amount of access to the culture that defined many people's adolescence and young adulthood is a win for preservation."

Challenges for the Internet Archive: Misinformation

- In October, 2020, the Internet Archive began to provide fact checks and context for Wayback Machine webpages
  - Idea is not to store misinformation without labeling it as such



We would like to acknowledge the hard work of the organizations we are building upon in order to provide context for archived web pages: FactCheck.org, Check Your Fact, Lead Stories, Politifact, Washington Post Fact-Checker, AP News Fact Check, USA Today Fact Check, Graphika, Stanford Internet Observatory, and Our.news.

Challenges for the Internet Archive: Copyright and rights

- **What's Happening:** Publishing companies filed a lawsuit against the Internet Archive's "Emergency Library" which allows readers to "check out" the same digital copy of books more than onece. IA's "Open Library" allows one reader at a time to check out books in the public domain and books under copyright.

- **IA perspective:** Digitized books are owned and lent in the same way books are lent from a library. Emergency Library goes back to usual waitlist when emergency is over.

- **Publishers' perspective:** Lending of digitized books books without appropriately licensing books and compensating authors is piracy. Emergency Library exacerbates copyright problems.

- **Current status:** Suit focuses on 127 books, for which damages would be $19M and forced destruction of 1.4M e-books. Not to court yet. At state is the goal of an "open internet"

Discussion

- What do you preserve? How?
- What do you think should be preserved? Who should pay for it?
- Who should make decisions about preservation and access?

Lecture Sources

- **A lawsuit is threatening the Internet Archive, but it's not as dire as you think**, Vox, https://www.vox.com/2020/6/23/21293875/internet-archive-website-lawsuit-open-library-wayback-machine-controversy-copyright
- **Wayback Machine**, https://en.wikipedia.org/wiki/Wayback_Machine
- **Internet Archive**, https://en.wikipedia.org/wiki/Internet_Archive
- **Internet Archive Website**, https://archive.org/

Presentations

Upcoming Presentations

March 29

- "**Animal Planet**", New York Times, https://www.nytimes.com/interactive/2021/01/12/magazine/animal-tracking-icarus.html?referringSource=articleShare (Julian C.)
- "**Ring and Nest helped normalize American surveillance and turned us into a nation of voyeurs**", Washington Post, (Hannah L.) https://www.washingtonpost.com/technology/2020/02/18/ring-nest-surveillance-doorbell-camera/?utm_campaign=wp_post_most&utm_medium=email&utm_source=newsletter&wpisrc=nl_most

April 1

- "**Smart farms are hackable farms**," IEEE Spectrum, https://spectrum.ieee.org/riskfactor/telecom/security/cybersecurity-report-how-smart-farming-can-be-hacked
- "**Farms are going to need different kinds of robots**," BBC News, https://www.bbc.com/news/business-56195288

April 5

- "**Your self driving car isn't ready. Smarter roads might change that**," CNN Business, https://www.cnn.com/2021/03/05/cars/cavnue-self-driving-lanes/index.html
- "**Waymo simulated real world crashes to prove its self-driving cars can prevent deaths**", The Verge, https://www.theverge.com/2021/3/8/22315361/waymo-autonomous-vehicle-simulation-car-crash-deaths

Need Volunteers – 4/8

- "**Vaccine passports pose ethical thicket for Biden Administration**," Politico, https://www.politico.com/news/2021/03/17/vaccine-passports-ethics-biden-administration-476384 (Davis E.)

- "'**This is bigger than just Tinmit': How Google tried to silence a critic and ignited a movement**". Fast Company, https://www.fastcompany.com/90608471/timnit-gebru-google-ai-ethics-equitable-tech-movement (Grant B.)

Today's Presentations

March 25

- "**How Scientists scrambled to stop Donald Trump's EPA from wiping out climate data**", The Verge, https://www.theverge.com/22313763/scientists-climate-change-data-rescue-donald-trump (Justin O.)

- "**More than 100 scientific journals have disappeared from the Internet**", Nature, https://www.nature.com/articles/d41586-020-02610-z (Liam M.)