Penn Libraries
University of Pennsylvania

Digital Humanists' Help Desk: Ethically Making & Sharing Data

Cynthia Heider, Public Digital Scholarship Librarian
heider@upenn.edu
with thanks to Jajwalya Karajgikar and Nicky Agate

Find these slides & other materials at https://bit.ly/data-ethics-rdds!

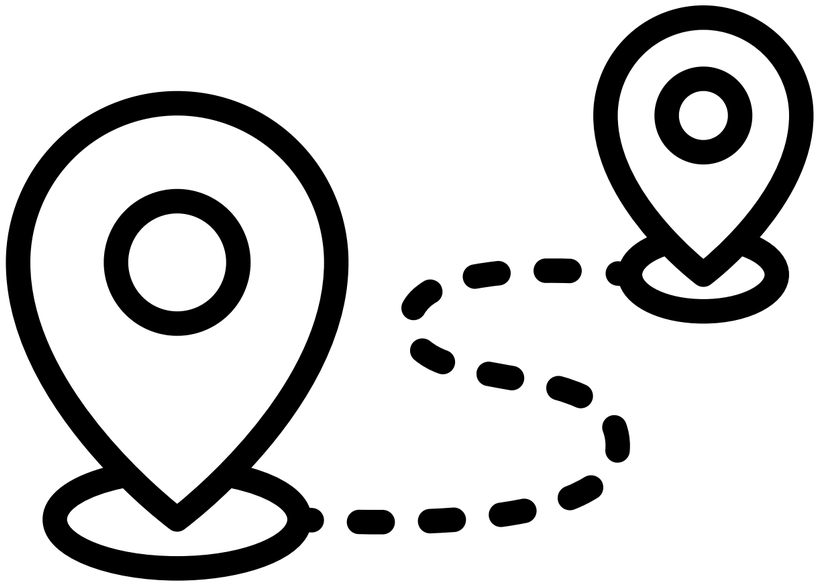
About Me

I'm **Cynthia Heider**, the Public Digital Scholarship Librarian at Penn Libraries.

I can help with:

- Digital projects for public audiences
- Digital projects involving community partners
- Choosing the right tool or platform for your digital project
- Working with data in the humanities
- Connecting with digital project resources at and beyond UPenn

Workshop Roadmap



- review data ethics principles and frameworks
- learn about the data work life cycle
- identify key actions at every step of the cycle
- put it into practice with tools and resources

Themes

Data Integrity

vs

Bias & Potential for Misuse

Security, Privacy & Surveillance

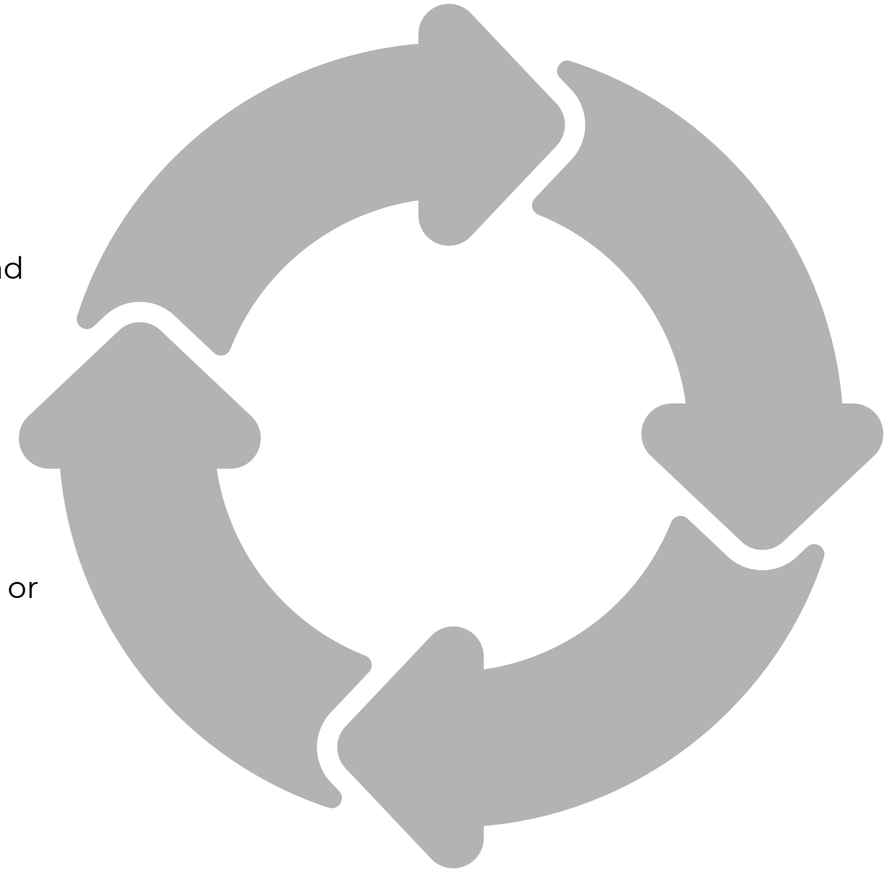
vs

Openness

The Data Work Life Cycle

1. Acquisition

The stage at which people decide which data to collect and why **(conception)**, determine how to collect them (**instrumentation)**, and take action to obtain them **(collection)**.



2. Processing & Analysis

The stage at which people decide how to **tabulate** and **interpret** data. This includes determining which data to include or exclude in analyses, how to process the data to create new variables or measures like indexes, and which people will be involved in the process to decide what the data mean.

3. Dissemination

The stage at which people **publish** and **share** data and/or **report findings** from their analysis.

4. Disposition

The stage at which people **destroy** or **archive** data, either completely or partially.

Principles & Precedents

Centering Ethics in Data Work

Principles

An Ethical Approach to Research:
"The Belmont Report" on Ethical Principles and Guidelines for the Protection of Human Subjects of Research (1979)

Beneficence

The commitment to maximize benefits and avoid causing harm to the extent possible, even if it is not a formal or legal requirement

Centering Ethics in Data Work

Principles

An Ethical Approach to Research:
"The Belmont Report" on Ethical Principles and Guidelines for the Protection of Human Subjects of Research (1974)

Respect for Persons

The responsibility to uphold people's power to make decisions that are in their best interest and to protect people who do not have that power

Centering Ethics in Data Work

Principles

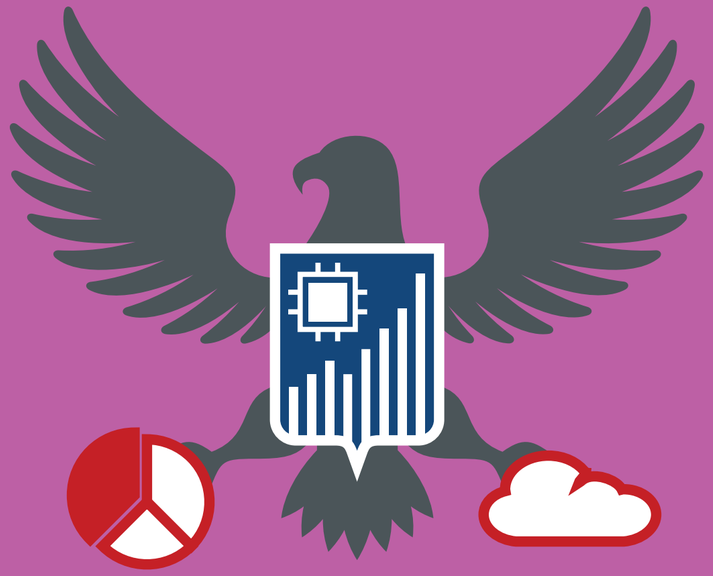
An Ethical Approach to Research:
"The Belmont Report" on Ethical Principles and Guidelines for the Protection of Human Subjects of Research (1974)

Justice

The commitment to the fair distribution of burdens and benefits among people

Centering Ethics in Data Work

Frameworks



An Ethical Approach to Data:

Federal Data Strategy "Data Ethics Framework" (2020)

"Data Ethics are the norms of behavior that promote appropriate judgments and accountability when acquiring, managing, or using data, with the goals of protecting civil liberties, minimizing risks to individuals and society, and maximizing the public good." [Ref]

Uphold applicable statutes, regulations, professional practices, and ethical standards.

Respect the public, individuals, and communities.

Respect privacy and confidentiality.

Act with honesty, integrity, and humility.

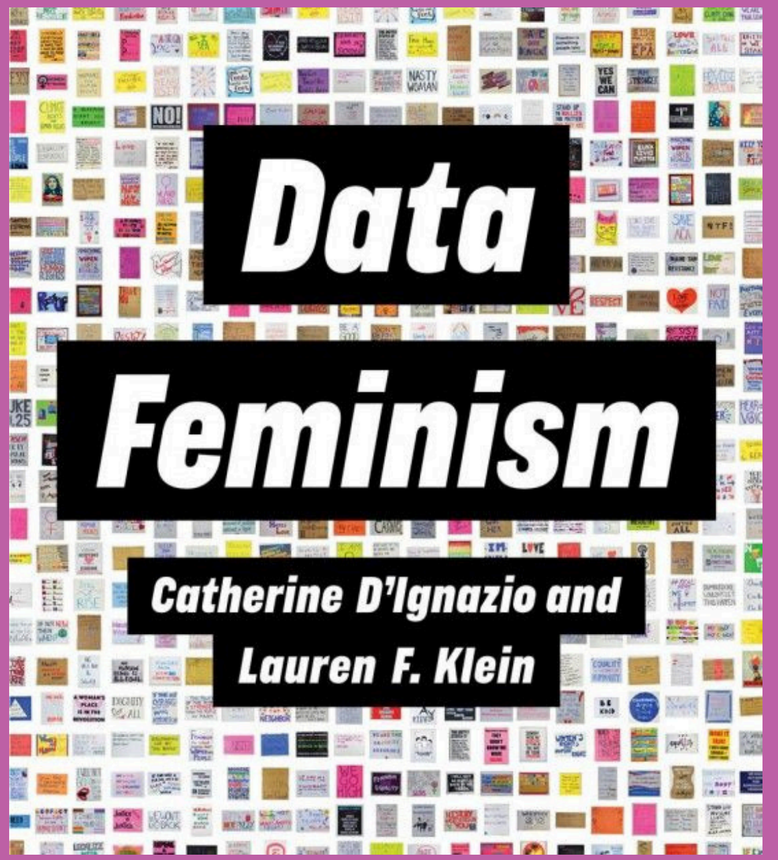
Hold oneself and others accountable.

Promote transparency.

Stay informed of developments in the fields of data management and data science.

Centering Ethics in Data Work

Frameworks



A Feminist Approach to Data:

Data Feminism by Catherine D'Ignazio and Lauren Klein (2020)

"We derived these principles from the major ideas that have emerged in the past several decades of intersectional feminist activism and critical thought. At the same time, we welcome the notion that there are many other possible starting points that share the end goal of using data (or refusing data) in order to end oppression." [Ref]

Examine Power

Challenge Power

Elevate Emotion & Embodiment

Rethink Binaries & Hierarchies

Embrace Pluralism

Consider Context

Make Labor Visible

Centering Ethics in Data Work

Frameworks

An Ethical Approach to Data Science:
Academic Data Science Alliance's Data Science Ethos (2022)

"The Data Science Ethos highlights what we believe to be data science's true spirit - namely to provide insight about the world around us while centering ethical and responsible services, outcomes, and technologies... Data scientists have an ethos, too, and we believe this ethos includes considering the potential impacts of our work on the world around us." [Ref]

Positionality: Diversity of Human Experience

Sociotechnical Systems: Technology Interacting with Society

Power: Asymmetries in Agency

Narratives: Dominant Discourse

Putting It Into Practice

A note about AI

- Recent advancements in artificial intelligence technology, especially generative AI and machine learning, have introduced new aspects of consideration (and concern) for data ethics
- These considerations may occur at any stage of the data work life cycle
- When storing and sharing data, be aware of the data sharing policies of your tool - it may be harvesting your data for AI/ML model training
- When working with AI, think about how the Belmont report principles might guide your work:

 **Beneficence**

 **Respect for Persons**

 **Justice**

Example scenario

A class action lawsuit filed in 2024 alleged Anthropic AI used the contents of millions of digitized copyrighted books to train the large language models behind their chatbot, Claude, including at least two works by each plaintiff. The company also bought some hard copy books and scanned them before ingesting them into its model. The company has admitted to doing as much, a fact that the plaintiffs raise in their complaint.

In one of the largest copyright settlements involving generative artificial intelligence, Anthropic AI, a leading company in the generative AI space, agreed to pay $1.5 billion to settle a copyright infringement lawsuit brought by a group of authors, in 2025.

From NPR

Penn Libraries
University of Pennsylvania

Centering Ethics in Data Collection Work

Implementing principles in the Acquisition phase of the data life cycle

 **Beneficence**

 **Respect for Persons**

 **Justice**

- Involve stakeholders from the start
- Collect only the minimum viable data
- Require informed consent and "opt-in"
- Provide reciprocal value in exchange for data
- Anticipate potential security risks, such as re-identification

Example scenario

A team designing an in-home security system was charged with making it easier to use, while also offering a greater sense of protection. Users wanted the system to monitor activity in and around the home but also preserve privacy.

To understand the line between protection and privacy, the team mapped out all human interactions that occur daily at home. This allowed them to pinpoint times when visual recording was critical for protection and when it wasn't. Inspired by analog camera shutters, the team designed a "privacy shutter" to open and close automatically at specific points.

From IDEO's "AI Ethics Cards"

Penn Libraries
University of Pennsylvania

Centering Ethics in Data Analysis Work

Implementing principles in the Processing & Analysis phase of the data life cycle

 **Beneficence**

 **Respect for Persons**

 **Justice**

- Forefront the human element inherent in data, rather than approach it as an abstraction
- Be transparent about the data's context and limitations
- Seek out and incorporate stakeholders' interpretation of the data
- Identify encoded bias and consider mitigation actions

Example scenario

A large equipment company wanted to explore the tradeoffs between usage-based repair and time-based.

Rather than present graphs and charts to show the difference between the two approaches, the team created two contrasting stories demonstrating what happened in each case. One told of an operator who overspent on parts that were replaced before new ones were needed. In the other scenario, the operator had to send his entire team home due to a part failure. Adding these simple, relatable stories helped everyone in the room understand the real-world implications of a data-driven solution.

From IDEO's "AI Ethics Cards"

Penn Libraries
University of Pennsylvania

Centering Ethics in Data Sharing

Implementing principles in the Dissemination phase of the data life cycle

Beneficence **Respect for Persons** **Justice**

- Incorporate quality control frameworks such as FAIR, CARE, and domain-specific specifications
- Account for how publication may reinforce inequities or close disparities
- Share to reduce the burden of duplicate data collection
- Return data and research results to stakeholders and contributors in a form they can use

Example scenario

In 2018, the City of Chicago launched CityKey, a single card that could serve as an ID, library card, and public transit farecard. The card was specifically designed for residents of Chicago who might have a hard time acquiring a driver's license or state ID—including undocumented immigrants.

In designing the CityKey, the City Clerk's office wanted to avoid replicating the experience New York City had faced with its own IDNYC card. New York City had to go to federal court to protect the personal information contained in its applicant database, which the Trump administration sought to use for immigration enforcement. Unlike New York's program, Chicago's CityKey database does not retain images of any documents, nor any personal information about CityKey holders.

From IDEO's "AI Ethics Cards"

Penn Libraries
University of Pennsylvania

Centering Ethics in Data Afterlives

Implementing principles in the Disposition phase of the data life cycle

Beneficence **Respect for Persons** **Justice**

- Empower individuals to order the destruction of their data
- Be transparent about the plans for the data after project conclusion
- Put in place accountability mechanisms to provide redress for harms that may arise from data misuse

Example scenario

In 2010 Mario Costeja González filed a complaint with the Spanish Data Protection Agency (AEDP) against a local newspaper and Google Spain for claims relating to auction notices mentioning González published in 1998. The notices concerned real estate auctions held to secure repayment of González's social security debts. González contended that these pages were no longer necessary because "the attachment proceedings concerning him had been fully resolved for a number of years and that reference to them was now entirely irrelevant." He sought to have the local newspaper remove the pages or alter them so his personal information was no longer displayed. He also sought for Google Inc. to remove the links to the articles in question so that the information no longer appeared in Google Search results.

The AEDP dismissed the plaintiff's claims against the newspaper, but allowed those against Google.

From EPIC

Tools & Resources

The ODI Data Ethics Canvas

Improve project planning

Grow impact and trust

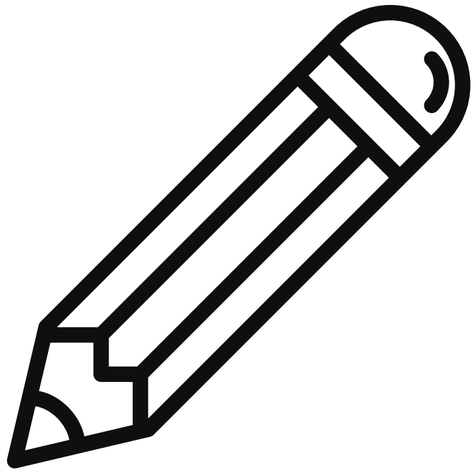
Complement other ethics guidelines

Manage data ethics in the long-term

Interactive tool:
https://theodi.github.io/interactive-data-ethics-canvas/

Graphic: bit.ly/dataethicscanvas

Google Doc with user and facilitator guides:
bit.ly/ethicscanvasdoc



The ODI Data Ethics Canvas

Digital Humanists' Help Desk: Ethically Making & Sharing Data
https://bit.ly/data-ethics-rdds

Open Data Institute #DataEthicsCanvas

theodi.org/tools

Data Hazards Labels

Teach students about ethics

Communicate potential harms about new research

Prompt discussion, reflection, and though

Include in Data Protection Impact Assessments (DPIA)

Labels:
https://datahazards.com/labels.html

Self assessment template: https://datahazards.com/materials/self-assessment.html

PERVADE Data Ethics Tool

Become familiar with factors that matter most to ethical data use

Find relevant resources to guide research

Increase reflexivity of data scientists towards data ethics

Interactive tool:
https://umdsurvey.umd.edu/jfe/form/SV_3wLh5p8hGWbCEya

Other resources

Data Ethics Decision Aid (DEDA) from Utrecht University

Data Ethics in the Participatory Sciences Toolkit from the Association for Advancing Participatory Sciences

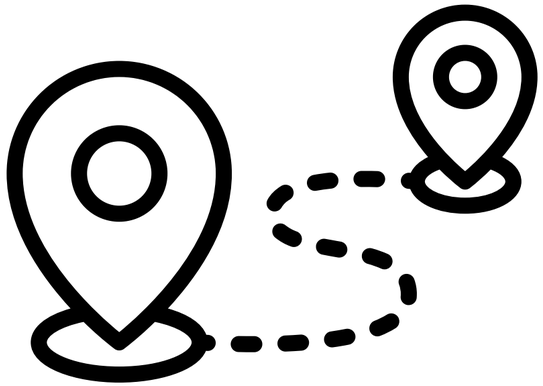
CARE Principles for Indigenous Data Governance and **Traditional Knowledge (TK) Labels**

Dataset Nutrition Labels from Data Nutrition Project

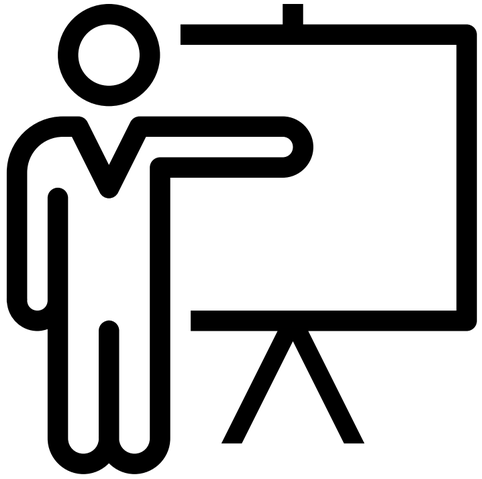
"Datasheets for Datasets" and **The Data Cards Playbook** from Google

- **Datasheets for Digital Cultural Heritage Datasets** by the Europeana Foundation
- **Datasheets for Earth Science Datasets** from Colorado State University scientists

Penn Resources



[Data Management Resources Libguide]



RDDS [Data Workshops and Consultations]

[Office of Research Services]