Liz Novosel
Computer Science, Mathematics,
& Social Sciences Librarian
elizabeth.novosel@colorado.edu

Chris Pusateri
E-Resources Acquisitions & Licensing
Librarian
christopher.pusateri@colorado.edu

Finding Datasets & Data Licensing

Today's Topics

Finding Datasets
Evaluating Datasets
Finding articles
Licensing issues with datasets


CAUTION: BAD DATA
BAD DATA QUALITY MAY RESULT IN FRUSTRATION AND LEAD TO DROP KICKING YOUR COMPUTER

AI & Data

Obviously, datasets are necessary for training of AI models.
AI can help create test datasets and synthetic data.
Are there any concerns about AI generated data?

- **Bias and misrepresentation of groups**
- **Lack of accuracy (it is made-up, after all)**
- **Does not reflect "new" information j**ust extrapolates from data it has already
- **Privacy/Safety:** ask yourself: where is the LLM/AI getting this data? Could there be privacy or safety concerns that could impact people?

Can AI make scientific data more equitable?. *Nat Rev Bioeng* **2**, 981 (2024). https://doi.org/10.1038/s44222-024-00263-5
Try googling "privacy issues AI" – it's fun.

How Have You Used Datasets?

- Have you ever worked with a dataset?
- How did you find it? Why did you use it or select it?
- What did you do with it?
- Did you have any problems working with the dataset you selected?

https://libguides.colorado.edu/findingdata/2025

To get to guide:

- **Google: “CU Boulder Libaries”**
- **Type: “Finding Data” in One Search Box**

Finding Datasets

Step 1:
Manage Your Expectations

Image by Charlesdeluvio on Unsplash

The perfect dataset…

1. may not exist
2. may not be free
3. may not be “good” or useable
4. may need significant cleaning
5. you may not be allowed to use it

The Library of Missing Datasets by Mimi Onuoha

"Missing data sets" are the blank spots that exist in spaces that are otherwise data-saturated…something does not exist, but it should…That which we ignore reveals more than what we give our attention to…Spots that we've left blank reveal our hidden social biases and indifferences.

—Mimi Onuoha

(File being grabbed is titled: “Publicly available gun trace data”)

Step 2:
Define Your Project

Image by Charlesdeluvio on Unsplash

1. What is your research question?
2. Who or what are you studying?
3. Which discipline(s) are related?
4. Time period & location?
5. How much data do you need?
6. How much data can you handle?
7. What software is available to you?
8. Where will you store your data?
9. Has anyone else done a similar project? (Look at the literature, more later)

Step 3:
Who Collects This Data?

1. **Data sets are collected, organized, stored, & made accessible (Time, money and effort)**
2. **Who has the resources and interest to collect this data?**
    a. **Specific researchers / labs?**
    b. **Government departments and agencies? (EPA, Census Bureau?)**
    c. **International organizations? (World Bank, World Health Organization)**
    d. **Companies? (Facebook, Amazon, Pfizer, Academic Publishers)**
3. **What relevant data repositories exist?**

Image by Toshi on Unsplash

Can you get the dataset?

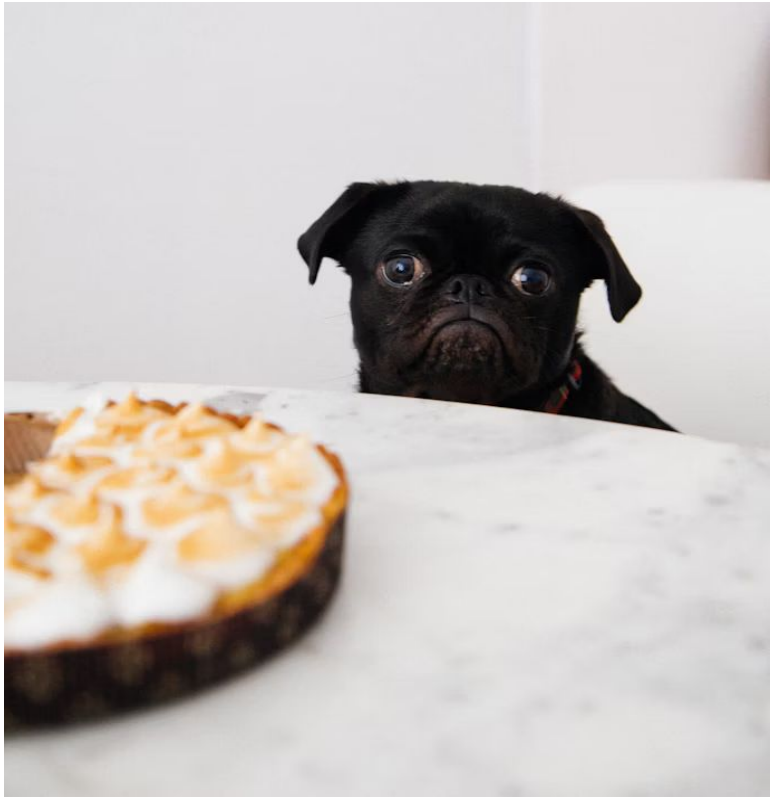


Image by Charlesdeluvio on Unsplash

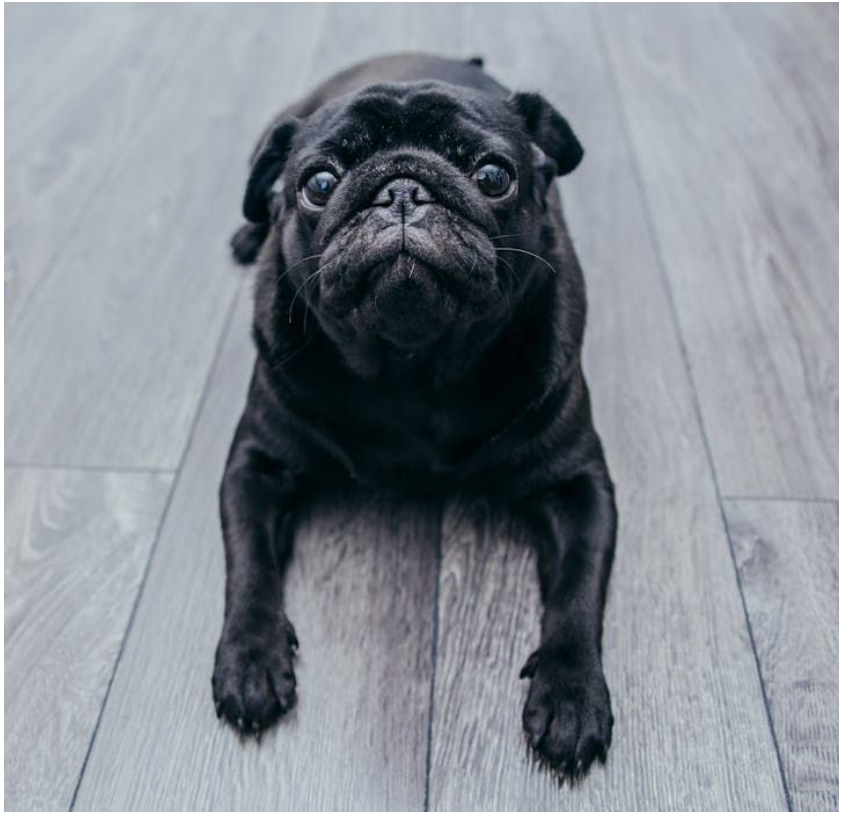
- **Open Data: publicly available**
  - Grants often require open data
  - Office of Science and Technology Policy (OSTP) requires open data for publicly funded research (or used to?)
- **Closed data: no public access*****
  - Protected by copyright, patents, contracts
  - Privacy protected for medical or other sensitive information
  - Software,
  - Business/financial information,
  - Unpublished research (insurance data, health data, financial data, data protected by court order, recipes, designs, patterns)

***Sometimes access can be purchased. Check with your librarian & department for options.

Step 4: Start the search.

Places to look:

1. Internet Search (regular and Google Data Search)
2. Data repositories (subject specific vs. general)
3. Government websites
4. Academic library collections
5. Scholarly articles



What is “good” data?

Image by Ivan Babydov on Pexels

Ethical Considerations

Photo credit: meanmugpug / Instagram

- As discussed by the Handbook of the Modern Development Specialist, it is easy to forget that data is about people, places, animals, etc.; this contributes to unethical practices.
-
- Who collected, funded, compiled, and published the dataset?
  - Why? How?
- Is there personal information in the dataset? Was it anonymized? Who could be harmed by the data?
- Did participants give consent?

METADATA

- **Different types:**
  - Licensing information (who can use the data and for what)
  - Technical requirements for using a dataset (how to use it)
  - The who, what, where, when, why and how the data was created
- **Where to find it:**
  - Readme file, data dictionary, codebook, attached file, repository page
- **Why is it important?**
  - Helps you use and understand a dataset

Examples of Metadata Standards

- Astronomy Visualization Metadata
- Darwin Core
- Data Documentation Initiative (DDI) to document numeric data files
- Dublin Core, a general purpose metadata standard
- ISO 19115 or FGDC's Content Standard for Digital Geospatial Metadata for geospatial data
- Ecological Metadata Language

Good Datasets

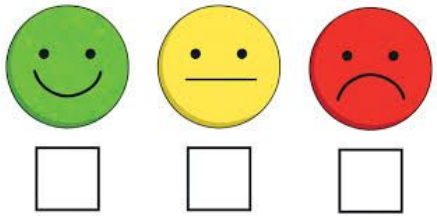
1. Complete
2. Requires minimal cleaning
3. Explains data collection methods
4. Clear labels (i.e., variables, column headers)
5. Declares bias
    a. Conflicts of interest
    b. Source of funding
6. License Information
7. Ethical & Protects privacy
8. Usable Format

“Bad” Datasets

1. Incomplete / errors
2. Require cleaning
3. Outdated
4. No / poor documentation
    a. No info about context or methods
    b. Poor metadata
5. Unethical / biased
6. Too much data
7. Incompatible with software

Evaluation Checklist:

1. **Is the dataset:**
    a. Usable: readable, well-documented, and available (to you)
    b. Functional format for software/analysis
    c. Complete, has good metadata (readme file!)
    d. Minimal “cleaning” or “wrangling” needed
    e. Data is current



2. **Does it follow a Metadata Standard?**
3. **How was the data set created and why?**
4. **What kinds of bias or issues exist in the dataset?**
5. **Could the use of the dataset be harmful in some way?**
6. **How has the dataset been used? How could it be used?**

Examples of Data Repositories

- Kaggle
- Data.gov
- Earthdata.nasa.gov
- Microsoft Research Open Data
- Reddit Datasets
- ICPSR (Inter-university Consortium for Political and Social Research)
- World Bank Open Data World Health Organization Data
- Dryad
- Amazon Web Services (AWS) Data Exchange
- Data.europa.eu
- Figshare
- Zenodo
- CU Scholar

Dataset Search Tools

a. Google Data Search
b. Re3data.org
c. Open Access Directory's List of Open Repositories
d. Nature's List of Scientific Data Repositories
e. NIH Guide to Finding Datasets and Repositories

What if you can’t find a dataset you need?

- **Ask advisor, instructor, research team, and subject librarian**
- **Ask researchers who have used similar data for theirs**
  - ***Note: You may be asked about your intentions***

Other Library Resources

Academic Literature: Why Bother?

Image by Jongsun Lee on Unsplash

Existing research is helpful:

- What is known/has been done on your topic
- Emerging research
- Methods, instruments
- **Datasets**

Use Citation Management Software!

- Zotero
- EndNote
- Mendeley
- EasyBib
- RefWorks

For help on citations, contact your librarian or consult Purdue's OWL (Online Writing Lab)

Find databases recommended by subject librarians:

Library Guides

Please feel free to contact me:
elizabeth.novosel@colorado.edu

https://libguides.colorado.edu/findingdata/2025

Using Datasets: Licensing Issues

Chris Pusateri

E-Resources Acquisitions & Licensing Librarian

Data & Licenses

Photo by Alotrobo on Pexels

1. How do I intend to use the content?
2. Who owns the rights to the content?
3. What rights do I have to use the content?

Common Licensing Terms

- Authorized Users
- Print or Digital Copies
- Reproduction of Data
- Data Privacy
- Text and Data Mining

Photo by Blogtrepreneur on Flickr

Text and Data Mining (TDM)

Photo by NeedPix

Questions

1. Can I mine the resource?
2. How much content do I want to harvest?
3. How do I plan to harvest it?
4. Are there other restrictions or conditions that I need to follow?
5. How do I determine the answers to these questions?

Text and Data Mining (TDM)

Photo by NeedPix

Answers

- Contact libraries@colorado.edu for specific questions about the use of licensed content in Libraries' electronic collections.
- Look for relevant licensing terms in Libraries discovery systems (e.g. OneSearch).