Finding Datasets
Data Camp - 2024

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

Today's Topics

- Finding Datasets
- Evaluating Datasets
- Connection to Scholarly Literature
- Finding Scholarly Literature

Related Library Guide:

https://libguides.colorado.edu/findingdatasets/2023


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

Why would you want to find a dataset?

- Have you ever needed to find one?
- What did you look for?
- What did you do with it?
- Did you have any problems with it?

#1. Managing Expectations…

Image by Charlesdeluvio on Unsplash

1. It may not exist
2. It may not be free
3. You may not be allowed to use it
4. Mimi Onuoha's The Library of Missing Datasets

Finding Datasets

1. Define need: what topic, when, where, why
2. Get informed
3. Identify places to look
4. Search widely
5. Evaluate quality, ethics, & fit

Image by David Troeger on Unsplash

Identify Possible Sources: Who Collects This Data?

- Data sets must be created: collected, organized, stored, made accessible
- This takes time, money and effort
  - Who has the resources, responsibility/mandate, interest, in collecting this data?
    - Person? (Researcher? Scientist?)
    - Research organizations/labs?
    - Government departments and agencies? (EPA, Census Bureau?)
    - International organizations? (World Bank, World Health Organization)
    - Companies? (Facebook, Amazon, Pfizer)

Open Data, Closed Data…

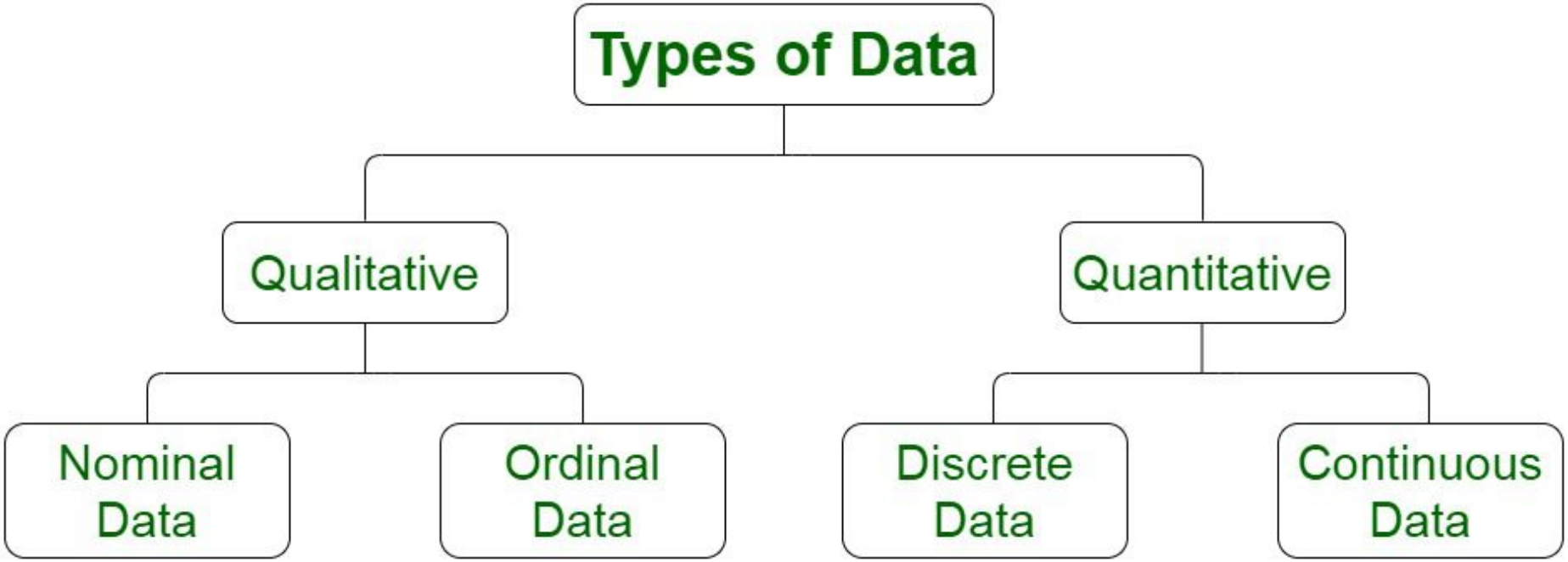
Open Data: available to everyone

- Archived in an open repository
- Data sets are often required to be openly available by grant issuing agencies
- Publicly-funded research is already required to be available in an open-access repository (see OSTP memo)

***ALSO: Some dataset owners require explanations of use to protect abuse.**

Proprietary Data: closed to public use

- Privately owned and funded; protected by copyright, patents, contracts, privacy protected
- May be related to software, business/financial information, or unpublished research (insurance data, health data, financial data, data protected by court order, recipes, designs, patterns)



Categories & labels: Nationality, blood-type, zip code

Categories & labels with a distinct ranking: level of education, income level,

Can be counted and has a specific answer: # of employees, # of new customers, # items on shelf
CAN BE COUNTED

Changes over time, has infinite range between values: temp, weight
CAN BE MEASURED

What is 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. **Require minimal cleaning**
3. **Good metadata / documentation**
    a. Explains data collection
    b. Clear labels: variables, column headers
    c. Clear about conflict of interest, source of funding
4. **License Information**
5. **Ethical & Protects privacy**
6. **Usable Format**

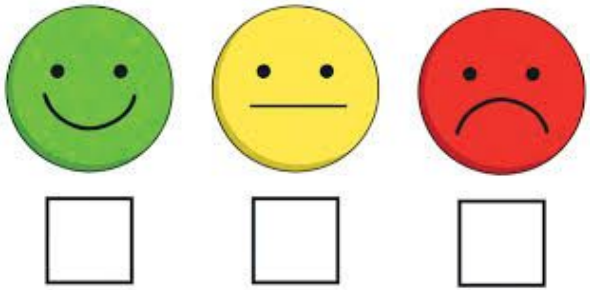
Image by Richard Brutyo

Bad Datasets

1. Incomplete or have errors
2. Formatting inconsistencies, require lots of cleaning
3. Outdated
4. No or poor documentation
    a. No info about source
    b. Poor labeling/metadata
5. Unethical/biased
6. Hard to use

Evaluate Datasets

1. **Is the dataset:**
    a. Usable: readable, well-documented, and available to all
    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

- Data.gov
- Google Dataset Search
- 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 can give you advice and assistance**
- **Ask the researchers of a project for their data**
  - **They might - or might not - be willing to share**
    - **In a recent article, researchers gave reasons for not sharing data:**
      - **lack of time to find their data (29.2%)**
      - **loss of data (27.7%)**
      - **privacy or legal concerns (23.1%)**
  - **You may be asked about your intentions**

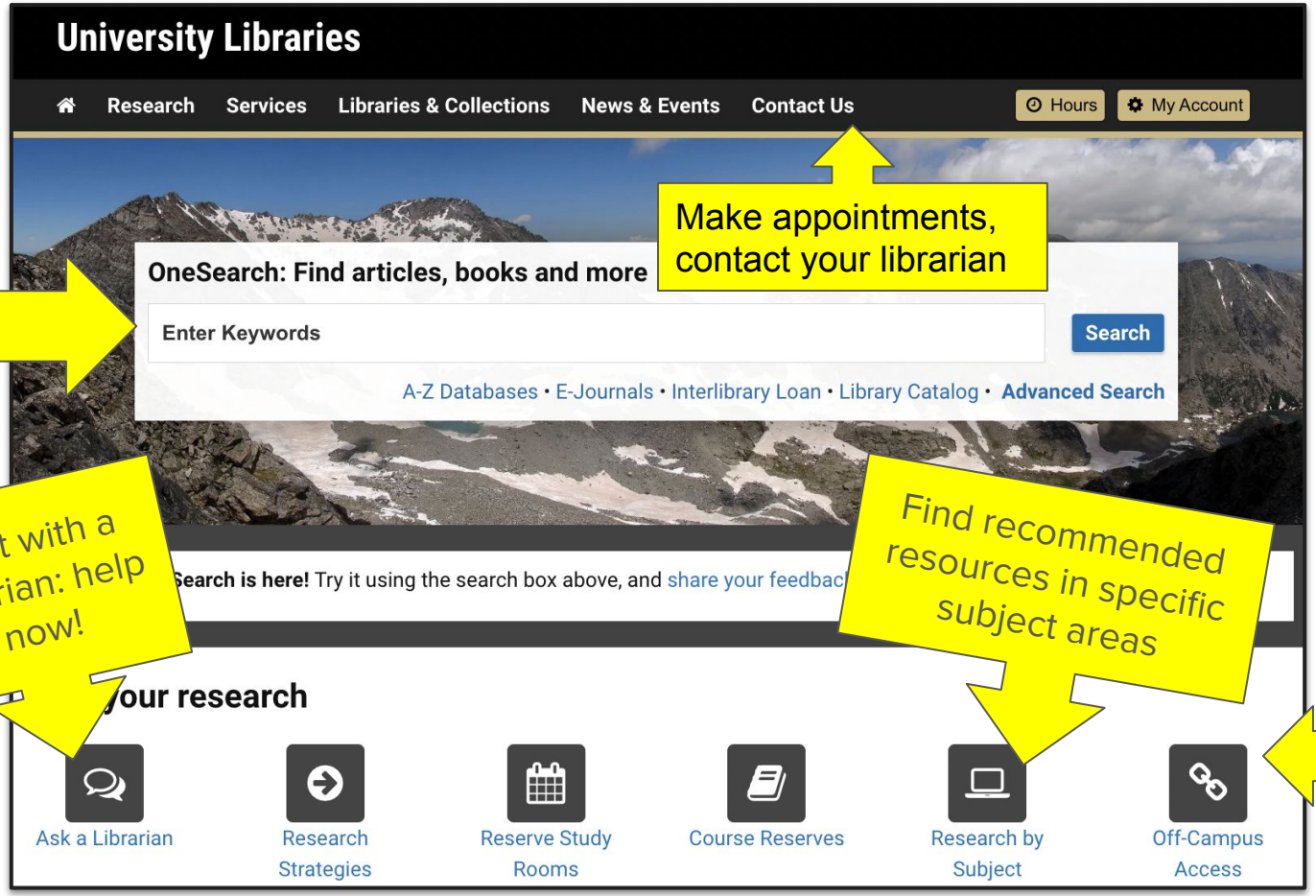
Academic Literature: Why Bother?

Image by Jongsun Lee on Unsplash

- What is known/has been done
- Emerging research
- Methods, instruments
- Datasets

Library Resources

CU Libraries Website: colorado.edu/libraries



Main search bar for library resources. Good for topic searching

Make appointments, contact your librarian

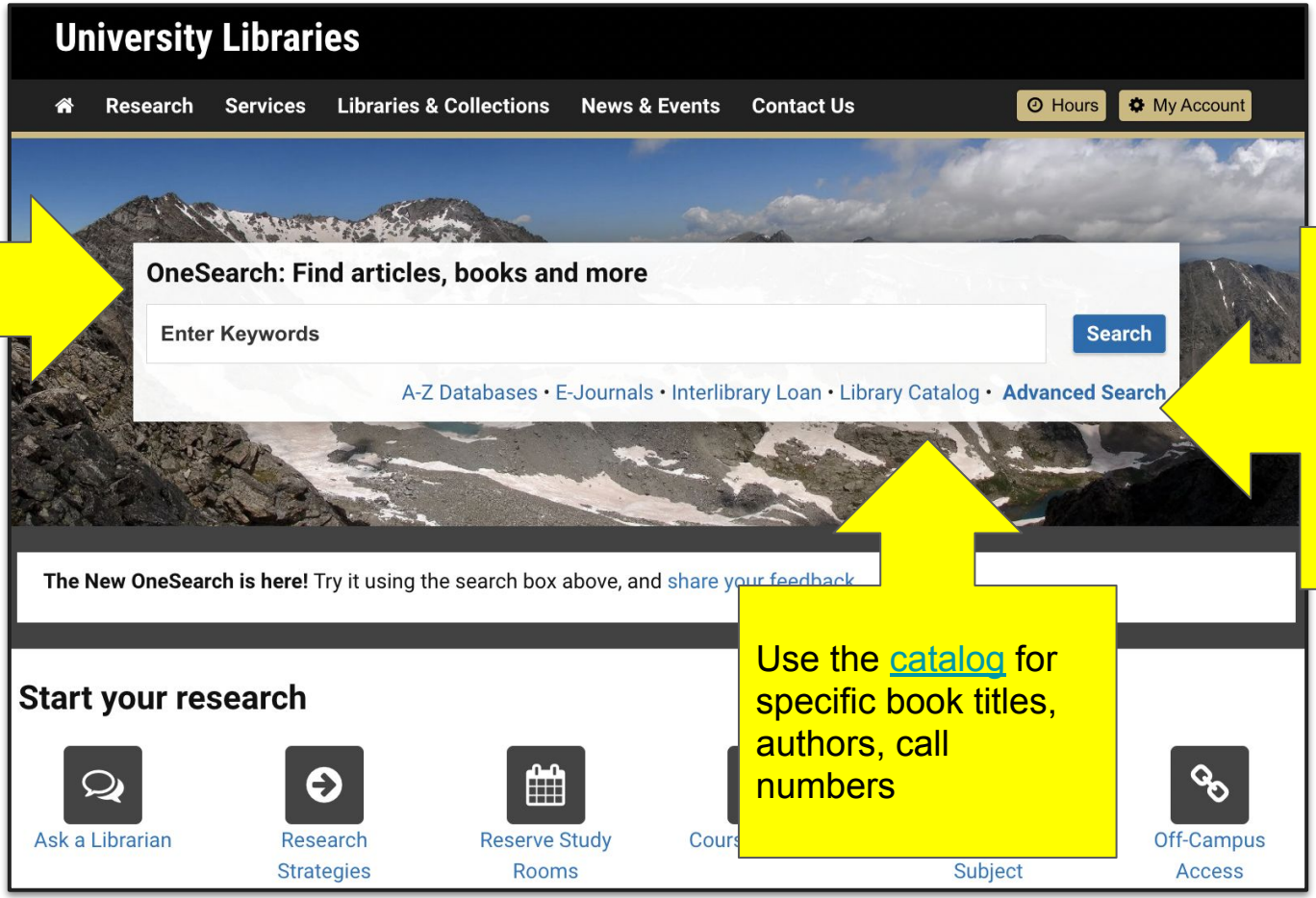
Chat with a librarian: help now!

Find recommended resources in specific subject areas

Get VPN here!

CU Libraries Website: colorado.edu/libraries

OneSearch searches almost everything we have; not great to look for a SPECIFIC book; good if you are searching a topic



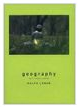
Use the catalog for specific book titles, authors, call numbers

Use "advanced search" for known titles/ authors

Firefox File Edit View History Bookmarks Tools Window Help
Tue Aug 29 3:18 PM
https://research-ebsco-com.colorado.idm.oclc.org/c/3czfwv/results?q=geography&facetFilter=sourceTypes%3ANDMwQks%3D&limiters=FT1%3AY
Import bookmarks... University Libraries ... LibraryH3lp (1) LibApps Login @ Un... Libraries Hub - Home Request Materials f... Reservations - Clas... American Library As... ProQuest OASIS ALTEC Spanish Inter... ProQuest OASIS Service Points Guid... Log in - FOLIO

OneSearch
University Libraries

Geography : art, race, exile / Ralph Lemon ; performance text by Tracie Morris ; afterword by Ann Daly.



Subjects: Geography (Choreographic work : Lemon); Modern dance; African American dancers -- Biography; Choreographers -- United States -- Biography; Lemon, Ralph -- Diaries

Published in: 2000, Library Catalog

By: Lemon, Ralph

Status:	Location:	Call number:
Available	Norlin Library - Stacks	GV1785.L45 A3 2000

Norlin

Access options | View details

Microfiche

Geography [microform] : Concepts, Maps, and Activities. Update: Geography Education Program, No. 9, Fall 1987.

Summary: This issue of Update contains four separate **geography** lesson plans on the topics of: (1) the mobility and interactions of people, goods, and ideas; (2) creating thematic maps and comparing the relative wealth of South...

Subjects: Elementary Secondary Education; Geographic Concepts; Geography; Geography Instruction; Human Geography; Instructional Materials; +8 more

Published in: 1987, Library Catalog

By: National Geographic Society (U.S.)

Status:	Location:	Call number:
Available	PASCAL Offsite	ED293742

Off-site

My dashboard
Overview
Projects
Saved
Searches
Viewed

Research tools
General search
Publications
Concept map
Supplemental sources

Library Links
Library Home
Complete our 2-question survey
Search Tips

Related resources

javascript:void(0)
222%

Google Scholar

- To connect with your library:
  - “Settings”:
    - “Library Links”
    - “Account”
- Do NOT pay for articles!
- If you can’t get full text on GS:
  - get citation and search in your library catalog

The “Good”

- Uses natural language
- Familiar/easy
- Finds much of what databases find
- Can connect to institutional databases to give you access

The “Bad”

- Fewer filters to narrow results
- Not full-text
- Algorithm is unknown
- Pulls from across internet; not all sources are reliable

Disciplinary Databases

- Specific focus
- Limited number of journals they pull from
- Example: Web of Science, Engineering Village

vs. General Databases

- Contain articles from many disciplines
- Good for broad, interdisciplinary searching

Citation Mining

1. Find a “good” article on your topic
2. Go forward in the research by seeing who has cited it
    a. (you can find this on Web of Science and Google Scholar)
3. Go backwards in the research by seeing what papers the researchers cited
4. Search the authors of this paper or any of the authors they cite
5. Look at articles in the journal of publication
    a. In Web of Science, you can see who funded their research and some other information

Use Citation Management Software!

- Zotero
- EndNote
- Mendeley
- EasyBib
- RefWorks

Interesting Data-Related Websites:

- StackOverflow
- Data Colada
- InsideAINews
- Data Science Central
- Diversity in Tech: 40 Resources to Promote Equity and Representation for People of Color

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

https://libguides.colorado.edu/findingdatasets/2023