Data Collection

2022-02-17

Bin Chen
CME, Propaganda Vertical, Data & Democracy Project

Jo Lukito
CME, Propaganda Vertical, Data & Democracy Project

The University of Texas at Austin
Center for Media Engagement
Moody College of Communication

Workshop Goals

Our goal is to provide you with information about where to find digital data.

But your projects should **not** be driven by what data you can collect: ***your projects should be driven by your research questions.***

Sources of Data

Static Sources of Data

The dataset collected is "complete." There is a defined time frame

Streaming Sources of Data

The data collected are ongoing.

Regardless of your **source**, the dataset you will need to collect will always end up being static, because you will (at some point) need to stop collecting data so you can analyze it and write your paper.

Sources of Data

1. Archives with Dashboard *(streaming)*
    a. Public Archives
    b. Social Listening Tools (e.g., Brandwatch)
2. APIs *(streaming)*
    a. Common for social media
3. Data Scraping *(“static”)*

You may need to combine these strategies.

(e.g., collecting URLs from an archive and then scraping it)

Data Policies and Ethics

Data Policies

Different platforms will have different policies.

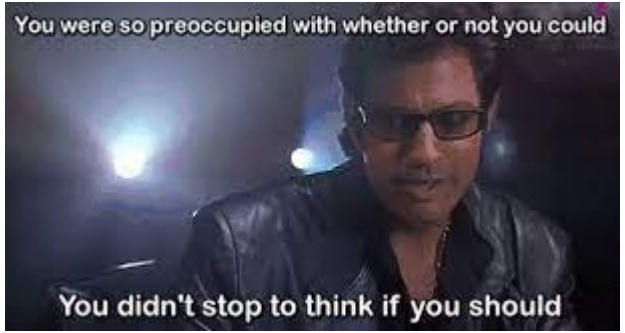
Consider:

- Social Media Terms of Service
- News Copyright

Data Ethics

Policies != Ethics

What are you data collection AND your data analysis ethics?



Steps to Collecting Digital Data

1. (Develop your RQ)
2. Research your data source
    a. Contact the source of the data to know what's available
    b. Contact IRB if necessary (UT has a secondary data IRB)
3. Plan out your data collection
    a. Time frame, source, keywords
4. Do the data collection
5. Check and Store the Data
    a. Consider the size of the data
6. (Analyze the data)

Data Collections We'll Cover Today

1. CrowdTangle (Facebook/Instagram)
2. 4CAT (Social media, especially alternative platforms)
3. MediaCloud (News)
4. Twitter API (Twitter)

We intentionally organized this from "no coding" sources to "some coding" sources, but all of these sources are accessible to beginners programmers.

CrowdTangle in a nutshell

A company-provided archive (CrowdTangle is owned by Meta)

We'll cover:

- How to get access as a researchers
- How to query content
- How to export content
- Exporting limits (rate limit issues)

Case: Beijing 2022

4CAT in a nutshell

A multi-platform data archive built by and for academic researchers.

We'll cover:

- How to get access as a researchers
- What social media data are accessible
- How to collect data
- How to download data

Case: Beijing 2022

MediaCloud in a nutshell

A great source for digital news

We'll cover:

- How to sign up for an account
- What sources are available (Source Tab)
- How to collect data (Explore Tab)
- How to download data (and what can be collected)
- Next steps: scraping

Case: Immigration (first 3 months of Trump's Presidency)

Twitter API in a nutshell

Twitter as a gateway to research with APIs

We'll cover:

- How to get access as a researchers (Twitter 2.0 Academic Track)
- AcademicTwitteR (R package accessing Twitter API)
  - How to input your bearer token
  - How to query the data
- How to download data

Case: Beijing 2022

And now, demos!

The University of Texas at Austin
Center for Media Engagement
Moody College of Communication

CrowdTangle
4CAT
MediaCloud
Twitter API