The Social Media Archive at ICPSR (SOMAR)

ICWSM 2023 Tutorial: Collecting and Sharing Twitter Data for Academic Research

Marley Kalt, Inter-university Consortium for Political and Social Research

Anmol Panda, University of Michigan School of Information

Libby Hemphill, University of Michigan School of Information & ICPSR

ICPSR

Agenda

- Web scraping: an API alternative
- Introduction to SOMAR
- Demo: Deposit data to SOMAR

ICPSR

Web Scraping

Web Scraping vs APIs

Web Scraping

- Retrieving data from a website's HTML
- Publicly-available content, what humans would see while browsing
- Programmatically extract content from websites that may not provide structured data through other means

Application Programming Interfaces

- Retrieving structured data through a web platform's provided interface
- Data often provided in JSON or XML
- May require authentication or agreement to provider's Terms of Service

Web Scraping is Controversial

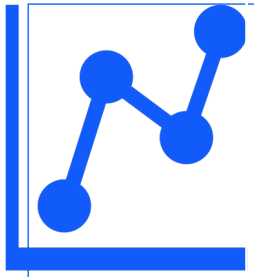
Ethics

No formal authentication or consent

Some websites prohibit web scraping

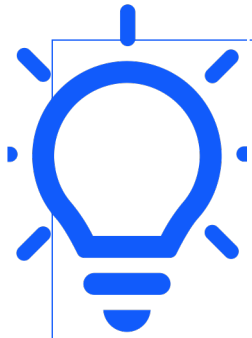
Privacy

Users have not consented to the data use case



Website Performance

Web crawlers can overburden and slow down websites



Intellectual Property

Who owns the website content?

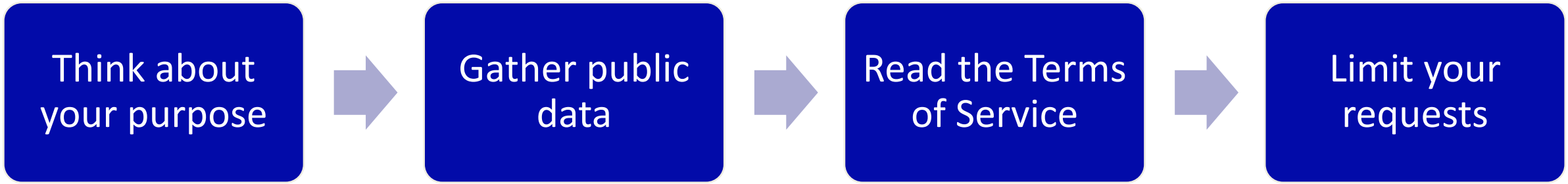
Web Scraping in the US

- Courts have ruled both ways on web scraping
  - Facebook, Inc. v Power Ventures, Inc.
    - Power Ventures aggregated user data from several social media sites, with user permission. Used Facebook data for promotional campaigns.
    - Court ruled Power Ventures violated the Computer Fraud and Abuse Act and California Penal Code section 502.
  - hiQ Labs, Inc. v LinkedIn Corp.
    - Data analytics company that scrapes public data from LinkedIn.
    - Court ruled web scraping of public data does not violate Computer Fraud and Abuse Act.
- Social media platforms try to prevent web scraping

Web Scraping in the EU

- French Data Protection Authority (CNIL) published guidelines on web scraping of publicly-available data
  - Publicly available data is still personal data
  - Focused on commercial uses
  - Recommended guidelines: obtain consent, only collect the data that you need, allow individuals to opt out, complete a Data Protection Impact Assessment

Web Scraping Tips



Web Scraping Resources

- Web Scraping Code of Conduct (from Library Carpentry workshop on web scraping)
- Platforms vs. PhDs: How tech giants court and crush the people who study them (Protocol)
- Victory! Ruling in hiQ v. Linkedin Protects Scraping of Public Data (Electronic Frontier Foundation)
- Data Protection Impact Assessment guidelines (CNIL)

SOMAR

socialmediaarchive.org

What is SOMAR?

- Data archive for research data from and about social media
- Accepts open access and restricted data
- Accepts data from individual researchers and social media platforms
- Launched on TIND platform in January 2023

ICPSR

Where is SOMAR?

- Part of the Inter-university Consortium for Political and Social Research (ICPSR)
  - Consortium of over 800 academic institutions and research organizations
  - Provides leadership and training in data access, curation, and methods of analysis for the social science research community
  - Maintains largest social science data archive
- ICPSR is a unit of the Institute for Social Research at the University of Michigan
- More about ICPSR: icpsr.umich.edu

Mission:
Democratize access to social media data

> *...making social media data accessible to researchers who lack the technical or computing resources to capture data independently.*"

> *…enabling researchers to share and access data from social media while respecting platforms' terms of service and users' privacy expectations.*"

ICPSR

Why use SOMAR?

Free and easy to deposit

- Self-deposit model
- Accept data of any format & size (5GB direct upload, contact team for larger files)
- Accept data and code, provide integration with GitHub

Professional archiving services

- Provide persistent identifiers (DOIs)
- Quality assurance through professional curator review
- Long-term preservation through ICPSR

Secure access to data

- Restricted data application process
- Virtual data enclave for working with restricted data

Using SOMAR: Public Data

Public Download

- Files anyone can download
- No login required
- Typically documentation files, associated publications

Login Required

- Data files with no restrictions
- Logins help us obtain user statistics, contact users as needed
- Users must agree to ICPSR Terms of Use
- Privacy Policy

Using SOMAR: Restricted Data

Restricted and Controlled Access

- Restricted data provided in a secure virtual enclave
- Closed environment for data access and analysis, any output must be approved
- Data may be identifiable or otherwise sensitive content
- Requires an approved Restricted Data Application: https://somar.inforeadv4.com/

Controlled Access, not Restricted

- Data may not have sensitive or identifiable content, but may have restrictions on use
- Data may be provided directly for researcher download
- Requires an approved Restricted Data Application to ensure compliance with terms of use

Deposit Process

Self-Deposit Process

Create SOMAR account

Complete the submission form

Upload data and documentation

Files under 5GB

- Upload through website interface

Files over 5GB

- Contact somar-help@umich.edu

SOMAR review

- Ensure metadata is complete
- Check files and documentation

Data are published!

ICPSR

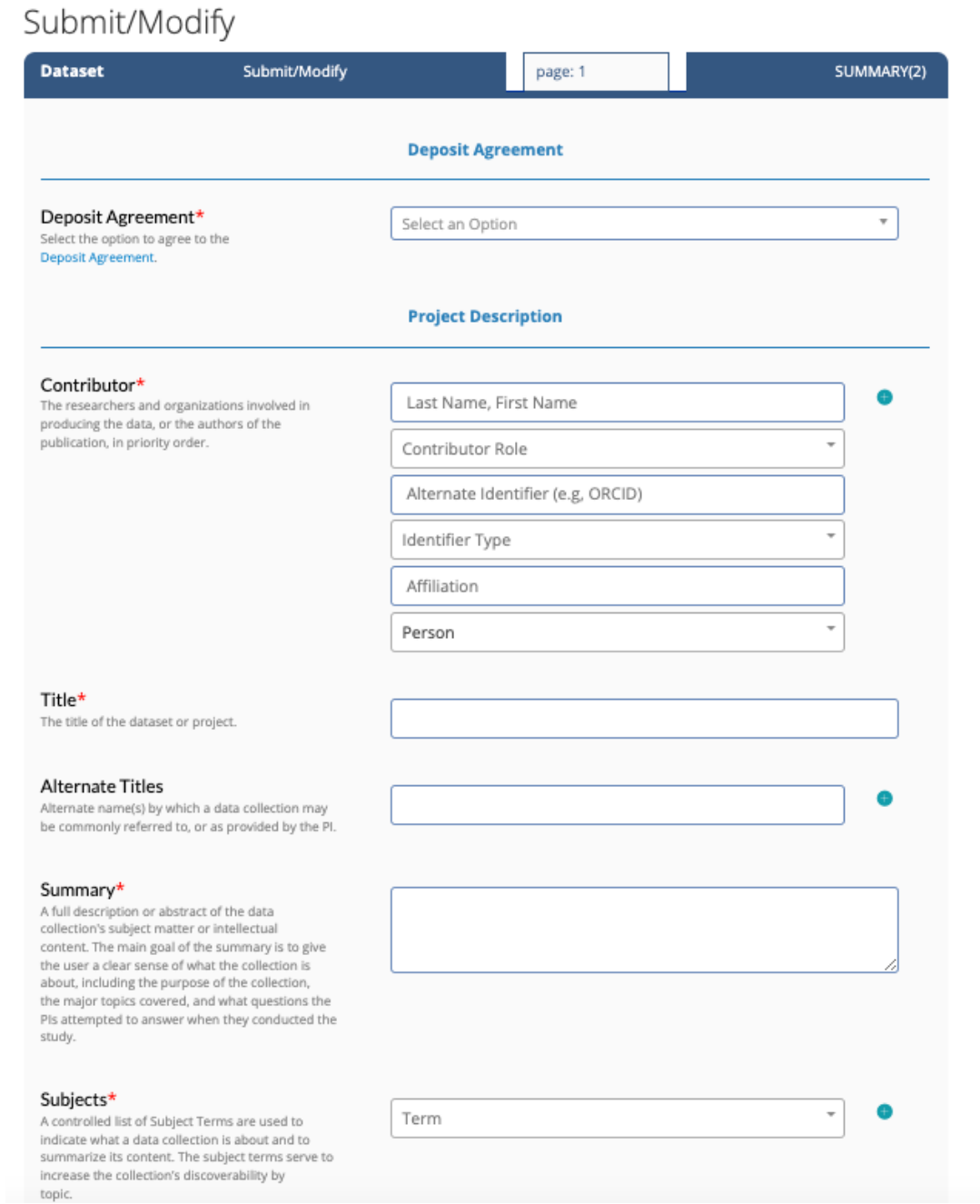
Dataset Submission Form

https://socialmediaarchive.org/

Click “Submit” in the top navigation bar

Required fields:

- Deposit Agreement
- Contributor(s)
- Title
- Summary
- Subjects



ICPSR

SOMAR Deposit Demo

Questions?

Thank You!

Social Media Archive at ICPSR (SOMAR)

Website: https://socialmediaarchive.org/

Email: somar-help@umich.edu

ICPSR