COLLECTING AND ARCHIVING MASTODON DATA

Ethical Enquiries on Decentralized Networks

Marco Wähner, Johannes Breuer,
Annika Deubel, Katrin Weller

CENTER FOR
ADVANCED
INTERNET STUDIES

THE CASE OF TWITTER AND DECENTRALIZED NETWORKS

INTRODUCTION

Before Elon Musk…

Twitter was the "model organism" for researchers (Tufekci, 2014, p.2)**:**

- Discussions of political, environmental, and other topics
- Large (English-speaking) user base
- Accessible and (mostly) free API
- Elevated access for researchers with Academic Research API
- Numerous open-source third-party tools for easy data collection

THE CASE OF ~~TWITTER~~ X AND DECENTRALIZED NETWORKS

INTRODUCTION

… After Elon Musk

X data has become inaccessible for research (Braun, 2023; La Cava et al., 2022)**:**

~~**Twitter as the "model organism" for researchers** (Tufekci, 2014, p.2)~~**:**

- Discussion of political, environmental, and other topics
- Large, but shrinking (English-speaking) user base
- ~~Accessible and (mostly) free~~ Only paid access to the API
- ~~Elevated access for researchers with Academic Research API~~
- ~~Numerous open-source third-party tools for easy data collection~~

RESEARCHING MASTODON

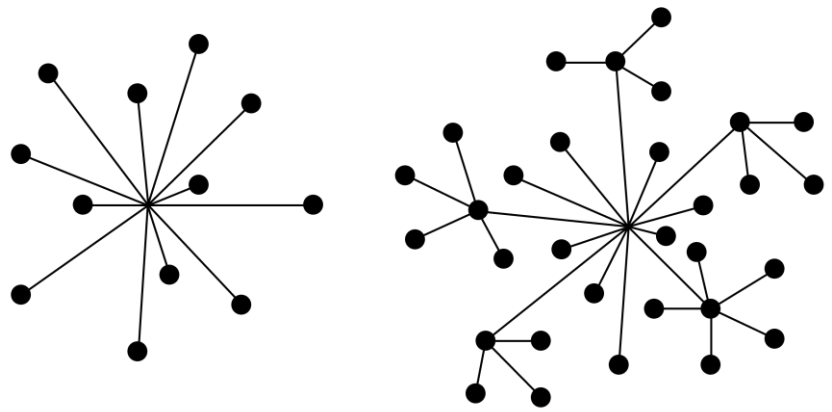
- Reasons for increased academic interest in Mastodon
  1. Increasing popularity/user numbers of the platform (including migration of “academic Twitter”)
  2. Openness of the platform (also regarding data access)
  3. New tools for research with Mastodon data, such as the R package rtoot (Schoch & Chan, 2023), the Python library mastodon-toolbox, or the new version of Communalytic (Gruzd & Mai, 2022)

THE CASE OF ~~TWITTER~~ X AND DECENTRALIZED NETWORKS

INTRODUCTION

Many researchers and users have since migrated to Mastodon, a decentralized social network

- Decentralized, open-source social networking platform
- Aims to provide a **user-friendly** and **ad-free alternative** to mainstream social media platforms
- Unlike traditional social networks, Mastodon operates on a **federated system**, meaning that it is composed of **multiple interconnected servers**, known as "instances"
- Mastodon represents the most widely adopted and recognized platform in the Fediverse (La Cava et al., 2021)
- Users can choose an instance that aligns with interests & values
- Fostering of **diverse and inclusive** online communities



CENTRALIZED DECENTRALIZED

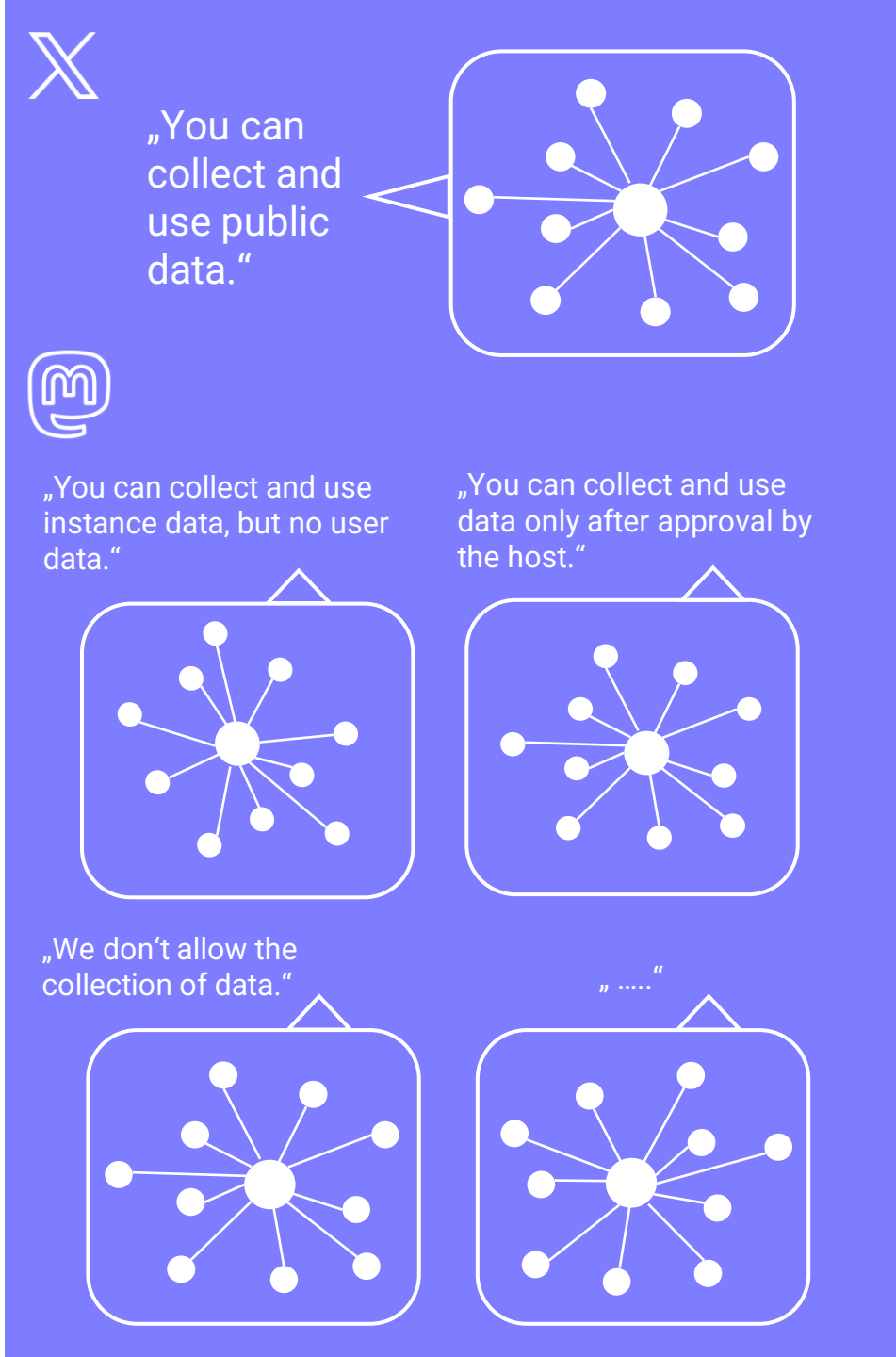
DATA PRIVACY AND MASTODON

- Mastodon emphasizes user privacy and control
- User base is more aware of their privacy and less accepting of the usage of their data

  At least one research article had to be retracted due to unlawful use and publication of user data

  (Statement of Removal, 2022)

- Important: Mastodon does not have universal Terms of Service (Gehl & Zulli, 2022)
- Instances have their own rules and regulations that may (or may not) address collecting and using data for research purposes



CAIS RESEARCH FOR THE DIGITAL AGE

DOING ETHICALLY SOUND RESEARCH ON SOCIAL MEDIA

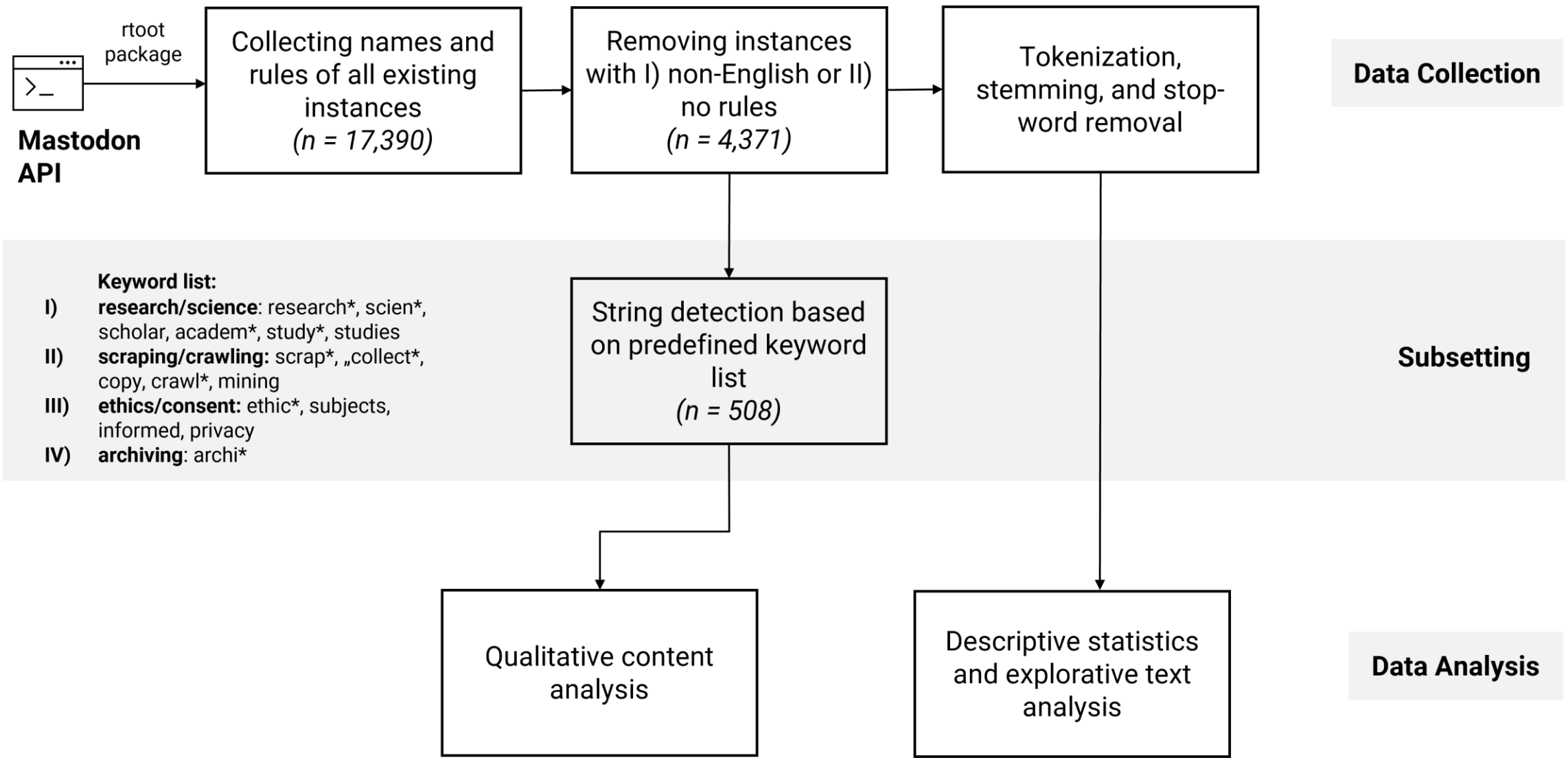
- Information privacy as the “ability (i.e. capacity) of the individual to control personally information about oneself” (Stone, et al., 1983, p. 460)
- Prevalent “but-the-data-is-already-public” attitudes in early years of social media research (Zimmer, 2010; Salganik, 2017)
- Call for a privacy violation framework to focus on ethical deficiencies, e.g., related to informed consent and (unauthorized) secondary data use

RESEARCH QUESTIONS

RQ 01 How do Mastodon instances **address the collection, use and archival of data** for research purposes?

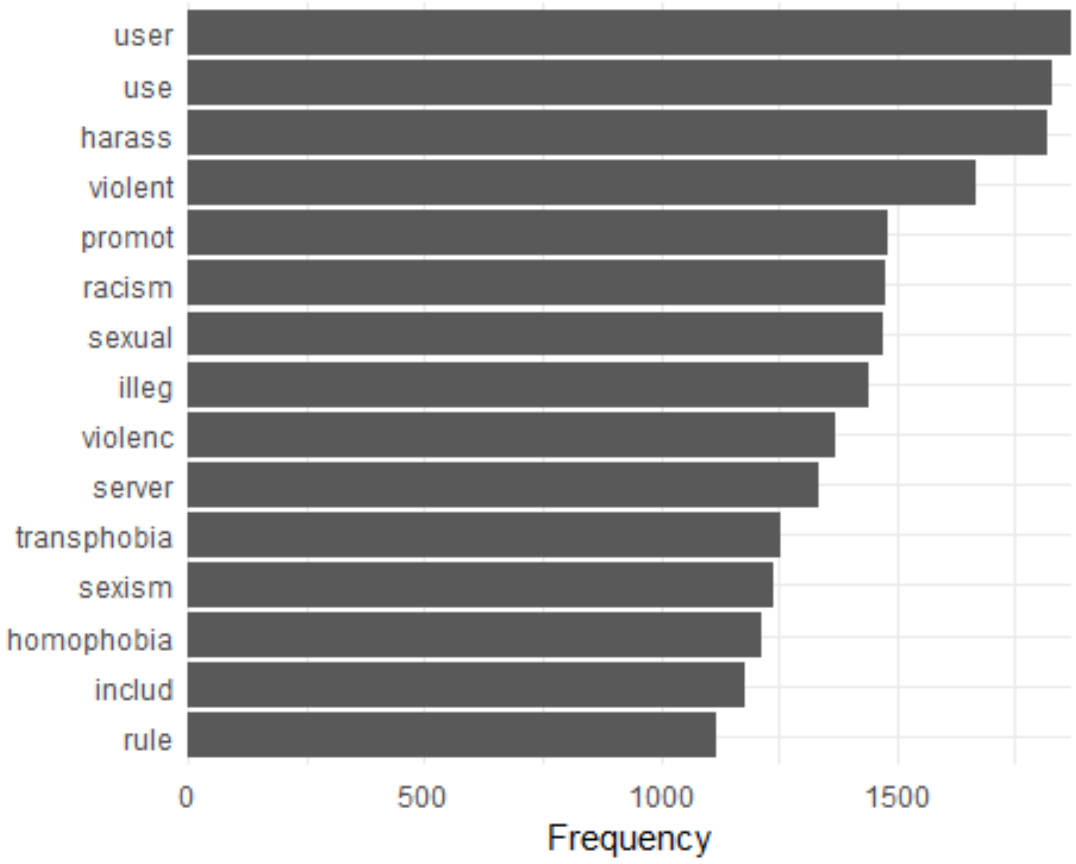
RQ 02 How can researchers use and archive Mastodon data in a manner that is **ethically sound and compliant with user expectations** in a decentralized social network?

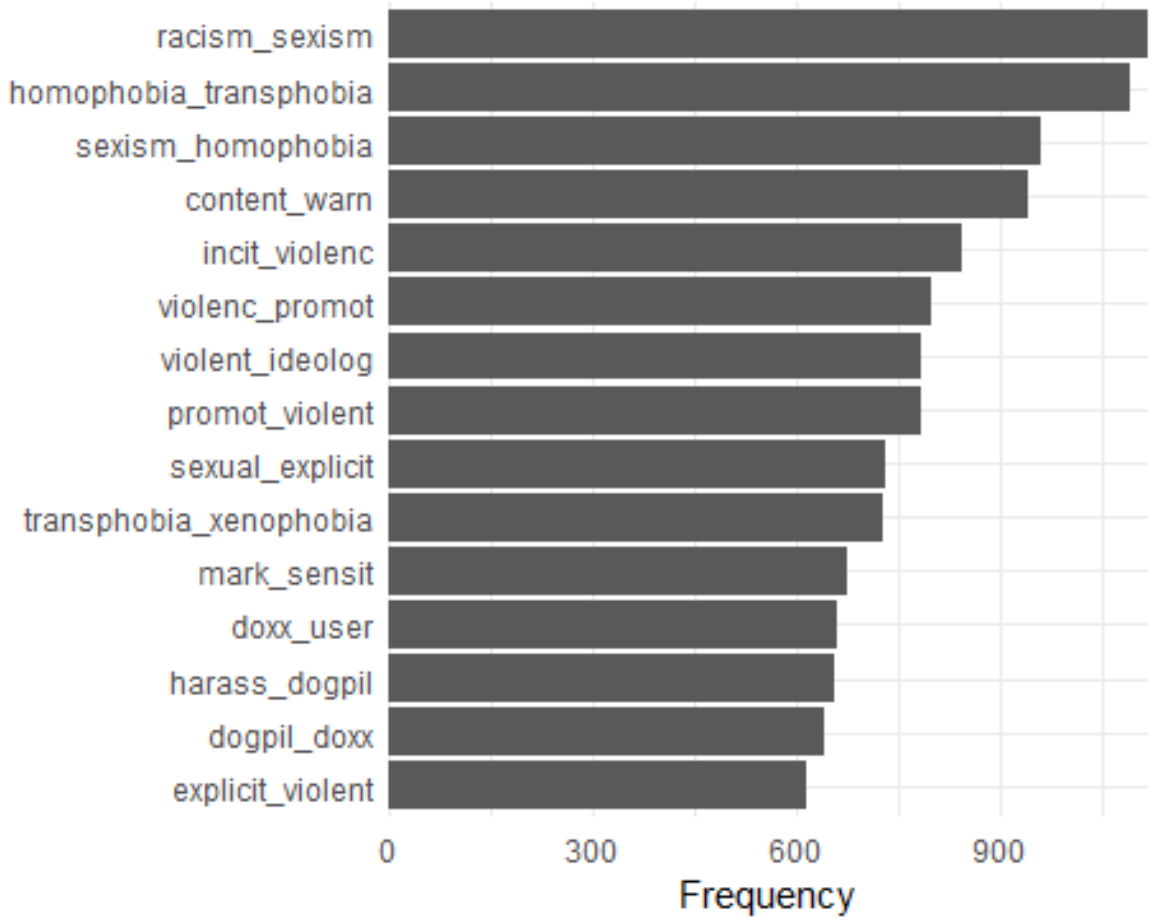
METHOD


rtoot package
Mastodon API
Collecting names and rules of all existing instances (n = 17,390)
Removing instances with I) non-English or II) no rules (n = 4,371)
Tokenization, stemming, and stop-word removal
Data Collection
Keyword list:
I) research/science: research*, scien*, scholar, academ*, study*, studies
II) scraping/crawling: scrap*, „collect*, copy, crawl*, mining
III) ethics/consent: ethic*, subjects, informed, privacy
IV) archiving: archi*
String detection based on predefined keyword list (n = 508)
Subsetting
Qualitative content analysis
Descriptive statistics and explorative text analysis
Data Analysis

RESULTS

MOST FREQUENT TERMS





RESULTS

STRING DETECTION

Keyword list:

I) **research/science**: research*, scien*, scholar, academ*, study*, studies
II) **Scraping/crawling:** scrap*, „collect*, copy, crawl*, mining
III) **ethics/consent:** ethic*, subjects, informed, privacy
IV) **archiving**: archi*

Our keyword-in-context-search returned **508 (11.62%) unique instances** with **584 rules** that match our keyword topics I-IV.

	Number of Rules	Positives	False Positives
		MANUAL ANALYSIS	
RESEARCH/SCIENCE	**69** (11.81%)	**6** (9.5%)	**63** (90.5%)
SCRAPING/CRAWLING	**197** (33.73%)	**30** (15.23%)	**167** (84.77%)
ETHICS/CONSENT	**298** (51%)	**19** (6.81%)	**279** (84.77%)
ARCHIVING	**20** (3.42%)	**3** (15%)	**17** (85%)

RESULTS

QUALITATIVE CONTENT ANALYSIS

Prohibiting research and data collection in general

*"Don't index us, **don't research us**."*
(Instance 723)

*"Toots from or carried by this instance **should not be scraped or indexed** / the instance's privacy policy gives details of some of the anti-scraper measures which are and can be used […]"*

(Instance 330)

Requirements for research ethics

*"If you are doing research […] you are likely performing Human Subjects Research and expected to abide by **highest standards of ethics** and professional practice. Meeting this expectation is your responsibility. We consider **engaging in unethical research […] to be a cause for bans** or other sanctions […]."*
(Instance 395)

CAIS RESEARCH FOR THE DIGITAL AGE

The data is public

*"**This forum is NOT confidential**. All messages are archived on the web and can be widely read. Do not say anything that would compromise anyone's confidentiality or violate your professional ethics in any way […]"*
(Instance 166)

RESULTS

QUALITATIVE CONTENT ANALYSIS – DATA ARCHIVING

- Only three instance rules explicitly mention the archiving of data
- In all cases, the archiving of data is also related to the scraping of the respective instance

2x

*"**You will not conduct bulk recording or archiving of material from this instance** without specific written consent from all users and the administration and moderation team i e do NOT scrape this instance's data"*

1x

*"no scraping or **archiving**"*

CAIS RESEARCH FOR THE DIGITAL AGE

IMPLICATIONS

With regard to RQ1…

- Although the Mastodon API is almost as accessible as the deprecated Twitter API, using Mastodon data is more challenging on a practical as well as an ethical level due to the decentralized nature of the platform.
- Our results suggest that only a minimal amount of Mastodon instances address the collection and use of data. Those that do mention it tend to be rather restrictive.
- Instance rules are primarily used to set regulations for user interactions and posting (content).

IMPLICATIONS

With regard to RQ2…

- Balance between making general recommendations and leaving room for project-specific decisions

- General recommendations include closely considering instance rules as well as the option of informing users about data collection
  - *put yourself in everyone else's shoes*; and *think of research ethics as continuous, not discrete* (Salganik, 2017)
  - Researchers should take instance rules into consideration
  - option specific to Mastodon: include instance hosts in the process (e.g., for getting consent or informing users)

- Weighing between benefit of research and potential harm: interest of users (privacy) vs. public interest (in research results/topic) & tasks/obligations as researchers (also regarding transparency, open science, etc.)

IMPLICATIONS FOR ARCHIVING

- Option of contacting instance admins to ask for permission
- Possible practical approach: First collecting data from all instances but only using and/or archiving data from instances that do not explicitly disallow this
- restricted data access (on-site and ideally also remote) or "remote code execution" as potential solutions

POINTS OF DISCUSSION

Dynamic use case:

- Changes in instances and their rules
- Other new alternatives for Twitter users (e.g., BlueSky or Threads)

How to consider **user-level „rules“** like #nosearch, #nobots, or #noindex in profile information?

Changes brought about by the **Digital Services Act** (DSA), Art. 40?

THANK YOU FOR YOUR ATTENTION!

We‘re looking forward to your questions & comments.

REFERENCES

Braun, J. (2023). Journalism, Media Research, and Mastodon: Notes on the Future. *Digital Journalism*, *Advance online publication*, 1–8. https://doi.org/10.1080/21670811.2023.2208619

Gehl, R. W., & Zulli, D. (2022). The digital covenant: non-centralized platform governance on the mastodon social network. *Information, Communication & Society*, 1-17.

Gruzd, A., & Mai, P. (2022). *Communalytic: A Research Tool For Studying Online Communities and Online Discourse*. Available at https://Communalytic.com

La Cava, L., Greco, S., & Tagarelli, A. (2022). Information consumption and boundary spanning in Decentralized Online Social Networks: The case of Mastodon users. *Online Social Networks and Media*, *30*, 100220. https://doi.org/10.1016/j.osnem.2022.100220

Salganik, M. J. (2019). *Bit by bit: Social research in the digital age*. Princeton University Press.

Schoch, D., & Chan, C. (2023). rtoot: Collecting and Analyzing Mastodon Data. *Mobile Media & Communication*. Advance online publication. https://doi.org/10.1177/20501579231176678

Statement of Removal. (2022). *Proceedings of the International AAAI Conference on Web and Social Media*, *13*(01). https://doi.org/10.1609/icwsm.v13i01.22003

Stone, E. F., Gueutal, H. G., Gardner, D. G., & McClure, S. (1983). A field experiment comparing information-privacy values, beliefs, and attitudes across several types of organizations. *Journal of Applied Psychology*, *68*(3), 459.

Tufekci, Z. (2014, May). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. In *Proceedings of the international AAAI conference on web and social media* (Vol. 8, No. 1, pp. 505-514).

Zimmer, M. (2010). “But the data is already public”: On the ethics of research in Facebook. Ethics and Information Technology, 12(4), 313–325. https://doi.org/10.1007/s10676-010-9227-5

ADDITIONAL INFORMATION

CAIS
CENTER FOR ADVANCED INTERNET STUDIES

FALSE POSITIVES

RESEARCH/SCIENCE

research*, scien*, scholar, academ*, study*, studies

- Prohibiting content that imply or address the **denial of science**

> *„**Malicious or misleading information** (such as but not limited to **anti-science** fake news) are bannable offences This also includes the encouragement and normalising of said information“*

SCRAPING/CRAWLING

Scrap*, collect*, copy, crawl*, mining

- **Word stem „copy“** referring to copyright

> *“Do not **post content that breaches copyright** defamation or libel laws”*

ETHICS/CONSENT

ethic*, subjects, informed, privacy

- Rules that refer to a **privacy policy** or prompt users to respect the **privacy of others**

> *„By signing up and using the service […] you are agreeing to **the terms of service outlined in the privacy policy**“*

> *„No person […] shall use the account for harassment threats or intimidation nor to **compromise the safety privacy or well-being** of any other person“*

ARCHIVING

archi*

- **Archival of user accounts** after a certain time of inactivity

> *„60 Days of Inactivity will result in account being warned No response within 30 days of warning will result in archival of user“*

- Words belonging to the **word stem other than archiving/archival**; e.g. „anarchist“, „non-hierarchical“, „monarchists“, etc.

> *„We believe in cyberanarchism; and like any good old-fashioned anarchist collective we look out for each other“*

- Instances about museums or **archives**

> *„[Instance] is a space for productive conversations about libraries archives museums or academia“*

CAIS RESEARCH FOR THE DIGITAL AGE