cessda

Linking Twitter and Survey Data

Luke Sloan (Cardiff University)

Libby Bishop (GESIS)

Johannes Breuer (GESIS)

June 23rd, 2020

cessda.eu

@CESSDA_Data

Linking Twitter and Survey Data
GESIS Training Course
23rd June 2020

Luke Sloan

@DrLukeSloan

Cardiff University, UK

Libby Bishop

@LibbyBishopPhi

GESIS, Köln

Johannes Breuer

@MattEagle09

GESIS, Köln

gesis Leibniz Institute for the Social Sciences

cessda

CARDIFF UNIVERSITY PRIFYSGOL CAERDYD

SOCIAL DATA SCIENCE LAB.

Outline

1. Introductions

2. Planning
3. Collecting
4. Processing
5. Analysing
6. Archiving & Sharing

1: INTRODUCTIONS

Dr Luke Sloan

My research focuses on Twitter and how social media data can be used to understand social phenomenon on its own, or through data linkage…

- **Linking Survey and Twitter Data:** Sloan et al. 2020. Linking Survey and Twitter Data: informed consent, disclosure, security and archiving. Journal of Empirical Research on Human Reseach Ethics 15(1-2) (10.1177/1556264619853447)
- **Linking Social Media to Cohort Data:** Di Cara et al. 2020. Views on social media and its linkage to longitudinal data from two generations of a UK cohort study [version 1; peer review: 1 approved]. Wellcome Open Res 5(44) (10.12688/wellcomeopenres.15755.1)
- **Linking Survey & Twitter Data:** Al Baghal, Sloan, Jessop et al. 2019. Linking Twitter and survey data: The impact of survey mode and demographics on consent rates across three UK studies. Social Science Computer Review (10.1177/0894439319828011)
- **Who Uses Twitter?** Sloan et al. 2015. Who tweets? Deriving the demographic characteristics of age, occupation and social class from Twitter user meta-data. Plos One 10(3), article number: e0115545. (10.1371/journal.pone.0115545)
- **Who geotags?** Sloan and Morgan 2015. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. PLoS ONE 10(11), article number: e0142209. (10.1371/journal.pone.0142209)
- **Validating Proxies with Survey Data:** Sloan 2017. Who Tweets in the United Kingdom? Profiling the Twitter population using the British Social Attitudes Survey. Social Media + Society 3(1) (10.1177/2056305117698981)
- **Crime-Sensing Through Twitter:** Williams, Burnap & Sloan 2016. Crime sensing with big data: the affordances and limitations of using open source communications to estimate crime patterns. British Journal of Criminology (10.1093/bjc/azw031)

Sloan & Quan-Haase (2017)

SAGE Handbook of Social Media Research Methods

CARDIFF UNIVERSITY PRIFYSGOL CAERDYDD

SOCIAL DATA SCIENCE LAB.

Dr Libby Bishop

My research addresses ethical issues in publishing and sharing research data, most recently, social media data. Currently I am working on challenges for sharing that arise when research and data sharing span public and private boundaries.

- **Accessing digital trace data:** Breuer, J., Bishop, L., & Kinder-Kurlanda, K. Forthcoming. The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. New Media & Society.
- **Ethics and data sharing**: Corti, L. and L. Bishop. 2020. Ethical Issues in Data Sharing and Archiving, in R. Iphofen (ed.), Handbook of Research Ethics and Scientific Integrity, Springer Nature Switzerland AG https://doi.org/10.1007/978-3-030-16759-2_17.
- **Sharing research data**: Corti, Louise, Veerle Van den Eynden, Libby Bishop, and Matthew Woollard, ed. 2020. Managing and Sharing Research Data: A guide to good practice. 2nd ed. Los Angeles: Sage.
- **Ethics and sharing social media data**: Bishop, L., and D. Gray. 2017. "Ethical Challenges of Publishing and Sharing Social Media Research Data." In The Ethics of Online Research 159-188. Bingley: Emerald.
- **Sharing big data**: Bishop, L. 2017. Big data and data sharing: Ethical issues. UK Data Service, UK Data Archive. https://www.ukdataservice.ac.uk/media/604711/big-data-and-data-sharing_ethical-issues.pdf.

Dr Johannes Breuer

At GESIS my work focuses on data linking and digital trace data. My other research interests are the use and effects of digital media, computational methods, and open science.

- **Accessing digital trace data:** Breuer, J., Bishop, L., & Kinder-Kurlanda, K. (in press). The practical and ethical challenges in acquiring and sharing digital trace data: Negotiating public-private partnerships. *New Media & Society*.
- **Linking surveys and digital trace data:** Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2019). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, Advance online publication. https://doi.org/10.1177/0894439319843669
- **Using digital trace data to study online news use:** Scharkow, M., Mangold, F., Stier, S., & Breuer, J. (2020). How social network sites and other online intermediaries increase exposure to news. *Proceedings of the National Academy of Sciences*, Advance online publication. https://doi.org/10.1073/pnas.1918279117

More information: https://www.johannesbreuer.com/

gesis
Leibniz Institute
for the Social Sciences

Objectives

- The learning objectives of this workshop are to…

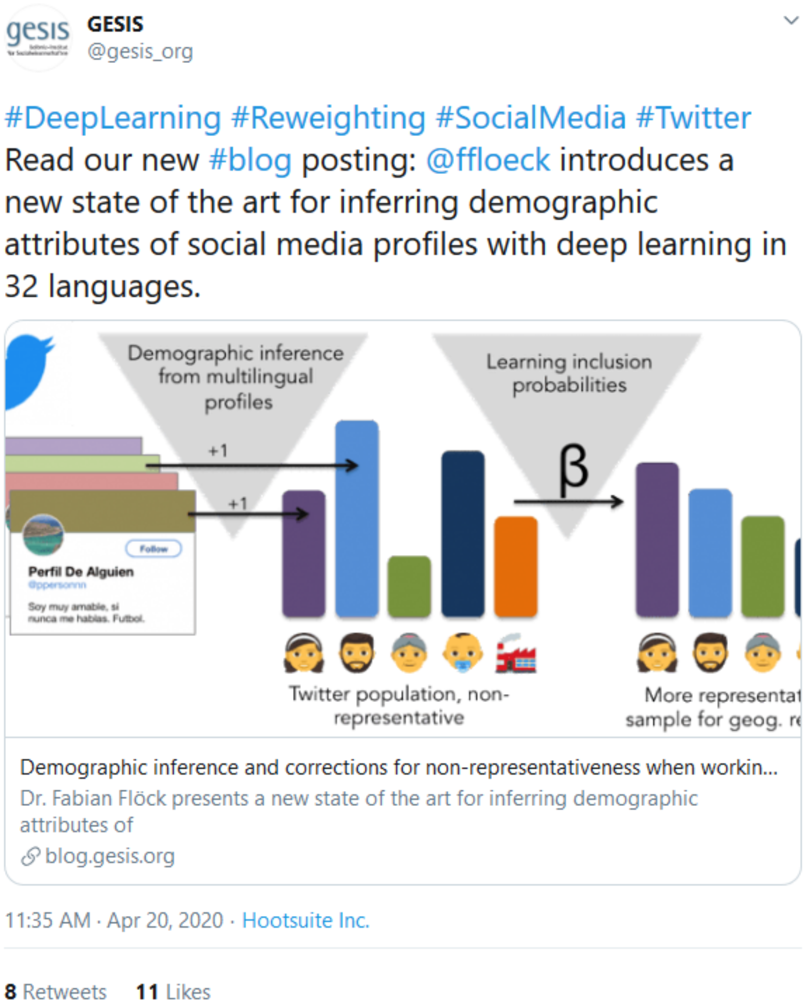
1. Understand why and how to link survey and Twitter data
2. Be aware of the key practical and ethical challenges in linking survey and Twitter data
3. Be familiar with the types of disclosure risks associated with linked survey and Twitter data
4. Know strategies for minimising risk in linked survey and Twitter data projects

2: PLANNING

What is Twitter Data?

- There are different types of Twitter data
  - Textual data: Tweets, retweets, replies
    - + metadata: reactions, time, location, language, …
  - Network data: Followers (directed networks)
  - User data: # of followers, profile information, …
- Which type(s) you need depends on your specific research question: Are you, e.g., interested in user activity, user interactions/networks or exposure (to specific content)?

Example: Data for one Tweet



GESIS
@gesis_org

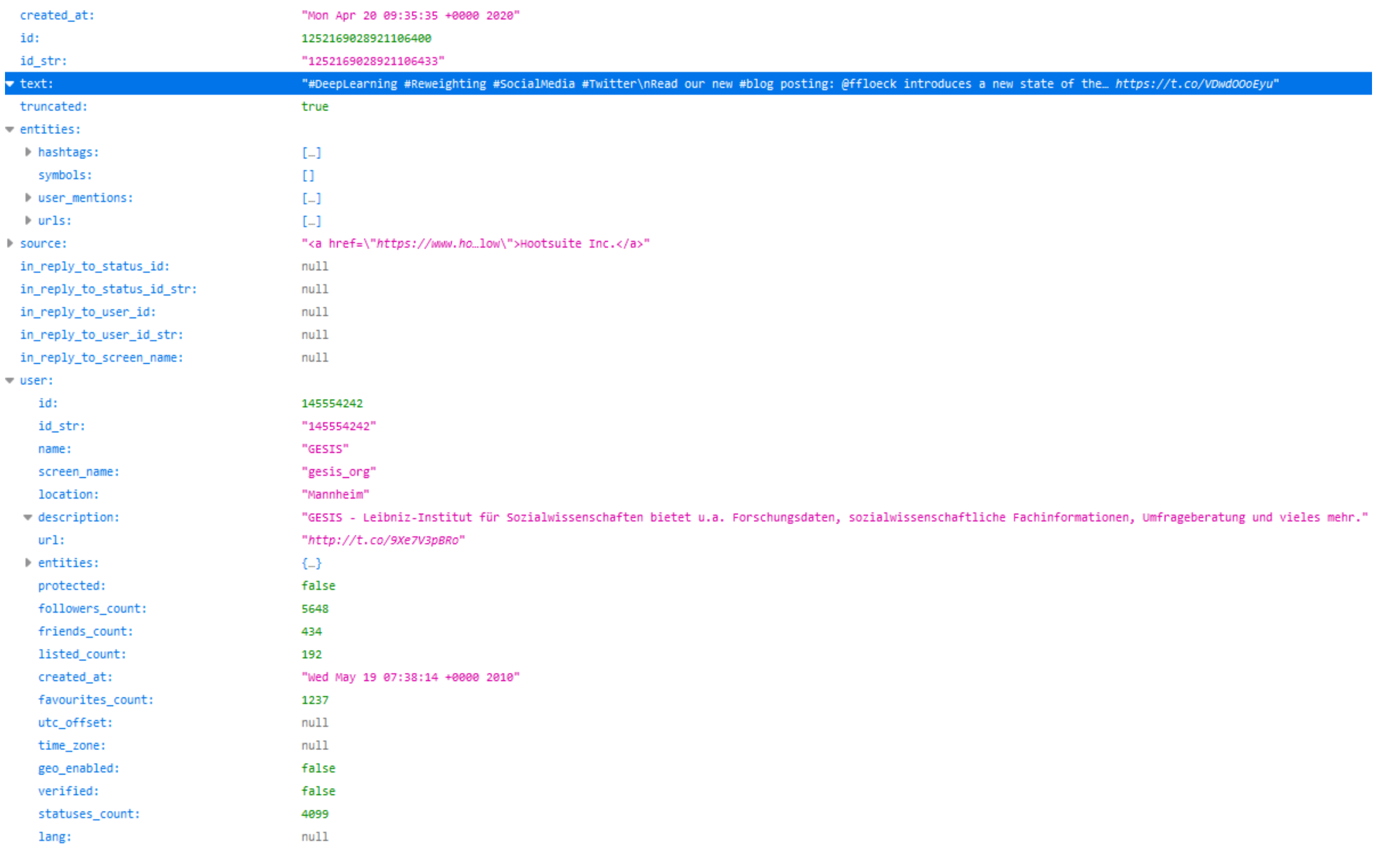
#DeepLearning #Reweighting #SocialMedia #Twitter
Read our new #blog posting: @ffloeck introduces a new state of the art for inferring demographic attributes of social media profiles with deep learning in 32 languages.

Demographic inference from multilingual profiles
Learning inclusion probabilities
+1
+1
β
Follow
Perfil De Alguien
@ppersonnn
Soy muy amable, si nunca me hablas. Futbol.
Twitter population, non-representative
More representat
sample for geog. re

Demographic inference and corrections for non-representativeness when workin...
Dr. Fabian Flöck presents a new state of the art for inferring demographic attributes of
blog.gesis.org

11:35 AM · Apr 20, 2020 · Hootsuite Inc.

8 Retweets **11** Likes



```
  created_at:                  "Mon Apr 20 09:35:35 +0000 2020"
  id:                          1252169028921106400
  id_str:                      "1252169028921106433"
▾ text:                        "#DeepLearning #Reweighting #SocialMedia #Twitter\nRead our new #blog posting: @ffloeck introduces a new state of the… https://t.co/VDwdOOoEyu"
  truncated:                   true
▾ entities:
  ▸ hashtags:                  […]
    symbols:                   []
  ▸ user_mentions:             […]
  ▸ urls:                      […]
▸ source:                      "<a href=\"https://www.ho…low\">Hootsuite Inc.</a>"
  in_reply_to_status_id:       null
  in_reply_to_status_id_str:   null
  in_reply_to_user_id:         null
  in_reply_to_user_id_str:     null
  in_reply_to_screen_name:     null
▾ user:
    id:                        145554242
    id_str:                    "145554242"
    name:                      "GESIS"
    screen_name:               "gesis_org"
    location:                  "Mannheim"
  ▾ description:               "GESIS - Leibniz-Institut für Sozialwissenschaften bietet u.a. Forschungsdaten, sozialwissenschaftliche Fachinformationen, Umfrageberatung und vieles mehr."
    url:                       "http://t.co/9Xe7V3pBRo"
  ▸ entities:                  {…}
    protected:                 false
    followers_count:           5648
    friends_count:             434
    listed_count:              192
    created_at:                "Wed May 19 07:38:14 +0000 2010"
    favourites_count:          1237
    utc_offset:                null
    time_zone:                 null
    geo_enabled:               false
    verified:                  false
    statuses_count:            4099
    lang:                      null
```

Why is Twitter data valuable for social research?

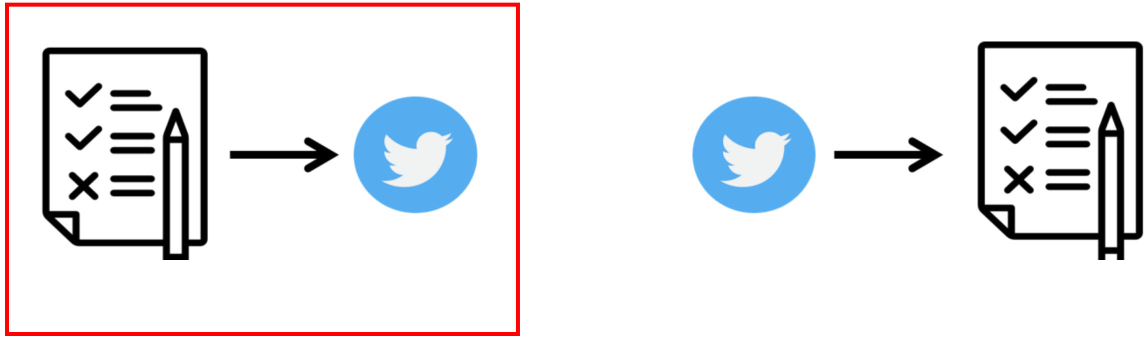
- Self-report data (from surveys) are often biased
  - Social desirability
  - Problems with recollection
- Twitter (or other social media) can provide behavioral data (posts, comments, reactions)
- If researchers are interested in studying the use of Twitter (or social media), using data the platform generates is much more reliable than self-reports
- However, such data can also be used to study the formation and expression of opinions and attitudes

Why combine survey and Twitter data?

- While self-report data can be biased to due social desirability or problems with recollection, social media data also have specific limitations
- Although there are tools for inferring user attributes from social media profiles (such as M3 for Twitter), the information about the users is typically limited in social media data
- In addition, relevant outcome variables (e.g., voting intention) are often missing from social media data
- Combining data from surveys and Twitter can help to overcome some of the respective limitations of the two data types (Stier et al., 2019)

How can survey and Twitter data be combined?

Two possible sequences of data collection



Each option is associated with specific sampling biases (also see the *Total Error Framework for Digital Trace Data* by Sen et al., 2019)[1]

- Data can be (linked) on the **individual level** or aggregated (e.g., for geographic regions or specific periods of time)
- Data can be collected together for the same people and period of time (ex-ante linking) or combined from different sources at a later point in time (ex-post linking: e.g., existing data from large survey programs and Twitter data collections)

[1] Two of the authors of this paper will also offer the GESIS workshop “Using Social Media Data for Research: Potentials and Pitfalls” on Nov 9-10, 2020

Practicalities of linking survey and Twitter data

- You need a unique identifier for linking survey and Twitter data
- A participant’s Twitter screen name/handle is the obvious candidate, however…
  - For privacy reasons the data should be stored separately, so you need a process for splitting and combining the data sets (more on that later)
  - People might also not know/remember their screen name (you can log in to your Twitter account using your e-mail address)
  - User names can be changed (User IDs, however, remain the same)
  - People may provide incorrect screen names in the survey (intentionally or unintentionally)
    - ➢ Potential solution to the issues related to providing a screen name: have people follow your account (an account for the project/study) and/or send a direct message for verification
    - ➢ Also bear in mind that you cannot track/get data from protected accounts

Case Study: Linking in practice

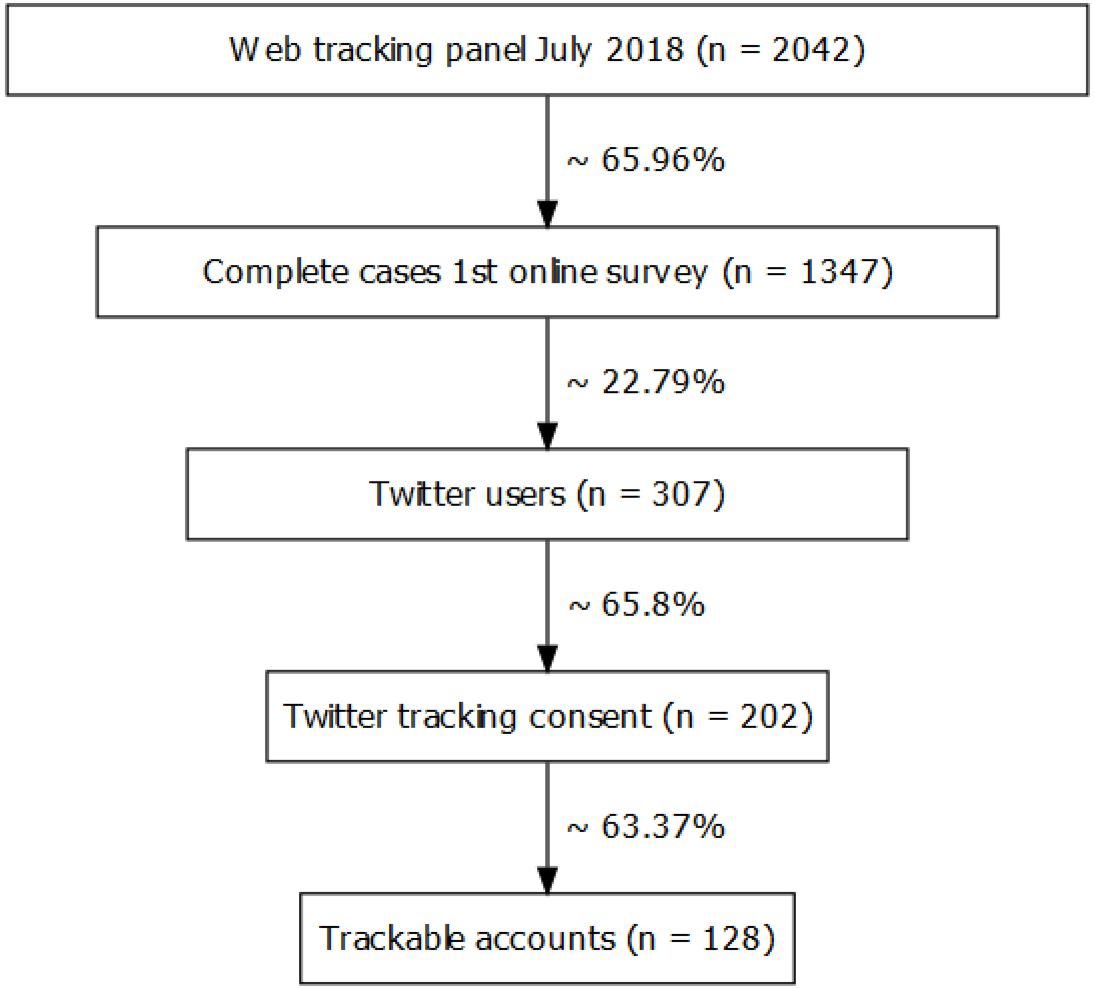
- Project by GESIS

- Web tracking panel by market research company with N ~ 2000 participants per month: data from June 2018 to May 2019

- Online surveys among panelists
  - First one (July 2018) included questions about use of Twitter
  - Twitter users in the sample were asked whether they would consent to having their Twitter activities tracked (via the API)
    - Additional incentive (5 €) for consenting to Twitter tracking

- Short informed consent in the questionnaire + extended privacy information on GESIS website linked in the short informed consent (texts were adapted from the study by Sloan et al., 2020 – will be presented/discussed in detail later)

Case Study: Linking in practice



Web tracking panel July 2018 (n = 2042)

~ 65.96%

Complete cases 1st online survey (n = 1347)

~ 22.79%

Twitter users (n = 307)

~ 65.8%

Twitter tracking consent (n = 202)

~ 63.37%

Trackable accounts (n = 128)

Ethical Issues: Social Media Data

- Data are naturally occurring, not produced (like surveys) for research
- The data collection was not subject to any formal ethical review process, e.g., research ethics committees
  - Protections applied when data are collected (e.g., informed consent) and processed (e.g., de-identification), often not implemented
- Using the data for research is different from its original purpose (e.g., user sharing with own "network")
- "Context collapse": original purpose, research use, archiving and sharing are all distinct (consider a rose bud). (boyd, 2002 & 2008)
- Data are often "personal", or worse, hard to assess how personal

Ethical Issues: Whose Ethics?

- Questions above presume a framework of social sciences…
- Other researchers (computing, linguistics) have different, frameworks. Different starting place… (Halford, 2017)
  - Is it human subjects data? (no intervention, public, not “identifiable”)
  - Is the “setting” public or private? (who, intention, platform norms)
  - Does “public” mean “anything goes”? (any use permitted)
- Essence of ethics: reasoned debate on conflicting moral claims (duties, rights, harms, etc.)
  - Often complex, rarely black/white
  - Few absolute rules, much depends on situation specifics
  - Can be frustrating…

Ethical Issues: Informed Consent & Linking

- If you start with survey data, you can (and must) get informed consent for collecting ***and linking*** social media data from your respondents
  - Probably required by the survey, and
  - “Off-Twitter matching” requires opt-in consent OR info given by user or public
- Researchers need to inform participants about…
  - What data they collect
  - For what purpose(s) they collect it
  - How the data is stored and who can access it
- Informed consent needs to adhere to legal regulations (GDPR in Europe) and satisfy ethical standards (as defined by Institutional Review Boards, etc.)
- Practical challenge: Finding the right balance between properly informing participants and overwhelming them with information and (technical) details

Ethical Issues: Beyond Consent

- Data is not usually a discrete collections, the value of big data lies in the capacity to accumulate, pool and link many sources

- When consent is not possible (who decides what is possible?)
  - Scale – cannot reach millions for direct consent
  - Problems with direct contact using platforms (cannot private message on Twitter unless mutual following…)

- Main point with linking - disclosure risk increases, but not easy to measure
- Must assess specific situation in light of basic principles
  - What were users' intentions? (**Respect** for persons, **Autonomy**)
  - What harms are possible, direct and indirect? (**Beneficence**)
  - Who benefits from this research? Who can access the data? (**Justice**)

Case Study: Informed Consent

- Understanding Society Innovation Panel 2017
- Experiments with survey design in longitudinal context
- This project looked at the feasibilities and practicalities of linking social media (in particular Twitter) and survey data in a longitudinal context, and how they can be combined to improve the quality of both
- Full details: Al Baghal et al. (2019) https://journals.sagepub.com/doi/10.1177/0894439319828011

Case Study: Informed Consent

- Designing appropriate questions to gain informed consent

- Three questions…
  - Do you have a personal Twitter account? [Yes/No]
  - *Question for consent to data linkage* (complicated!)
  - What is your Twitter username?

- Inclusion of four detailed help screens…
  - What information will you collect from my Twitter account?
  - What will the information be used for?
  - Who will be able to access the information?
  - What will you do to keep my information safe?

Case Study: Informed Consent

Q1 [Ask All]

Do you have a personal Twitter account?

1. Yes
2. No

Q2 [IF Q1 = Yes]

We would like to know who uses Twitter, and how people use it. We are also interested in being able to add people's answers to this survey to publically available information from your Twitter account such as your profile information, tweet content, and information about how you use your account.

Your Twitter information will be treated as confidential and given the same protections as your interview data. Your Twitter username, and any information that would allow you to be identified, will not be published without your explicit permission.

Case Study: Informed Consent

HELP SCREEN: What information will you collect from my Twitter account?

We will only collect information from your Twitter account that is publically available. This will include information from your account (such as your profile description, who you follow, and who follows you), the content of your tweets (including text, images, videos and web links), and background information about your tweets (such as when you tweeted, what type of device you tweeted from, and the location the tweet was sent from).

We will collect information from your past tweets (up to the last 3,000) and will update this with information from more recent tweets on a regular basis. This information will be collected and stored for as long as they are useful for research purposes, or you contact us to withdraw your permission. You can do this at any time, and do not have to give a reason.

Case Study: Informed Consent

HELP SCREEN: What will the information be used for?

The information will be used for social research purposes only. Adding your Twitter information and your survey answers will allow researchers from universities, charities and government to better understand your experiences and opinions.

For example, using extra information from your Twitter account, researchers can start to:

- Understand who uses Twitter and how they use it
- See what Twitter information can tell us about people, and how accurate it is
- Know what people in the UK are saying about things we don’t ask in our survey
- Look at additional information related to questions asked in the survey

Case Study: Informed Consent

HELP SCREEN: Who will be able to access the information?

Matched data which includes both your survey answers and Twitter information will be made available for social research purposes only. Researchers who want to use your detailed Twitter information must apply to access it and present a strong scientific case to ensure that the information is used responsibly and safely.

Statistical information from your Twitter account which you cannot be identified from (e.g. how often you Tweet, or whether you follow any politicians) will have the same access controls as your other survey answers.

At no point will any information that would allow you to be identified be made available to the public

Case Study: Informed Consent

HELP SCREEN: What will you do to keep my information safe?

All information we collect will be held in accordance with the Data Protection Act 1998.

Because Twitter information is public data that anyone can search, it is impossible to anonymise completely. To keep your information safe, researchers will only be able to access the matched survey answers and detailed Twitter information in a secure environment set up to protect this type of data. Only approved researchers who have gone through special training may access this information, and they will have to apply to do so.

Statistical information from your Twitter account which you cannot be identified from (e.g. how often you Tweet, or whether you follow any politicians) will have the same level of protection as your other survey answers.

Case Study: Informed Consent

Are you willing to tell me your personal Twitter username and for your Twitter information to be added to your answers to this survey?

1. Yes
2. No

SOFTCHECK: If answer does not begin with ‘@’ or contains any spaces: ‘Please check and amend. Twitter usernames should begin with an @ character and should not contain any spaces

Q3 [IF Q2 = Yes]

What is your Twitter username?

3: COLLECTING

Accessing Social Media Data

Three options for collecting social media data:

1. Collect data yourself
   a) API
   b) Web scraping
2. Cooperate with companies that produce or hold these data to gain privileged access (e.g., as embedded researcher)
3. Purchase the data from data resellers (or market research companies)

(see Breuer, Bishop, Kinder-Kurlanda, in press)

+ possibility to reuse already collected data (examples for Twitter data: GESIS Social Media Monitoring, TweetsKB, DocNow, TweetSets, Geotagged Tweets from the US)

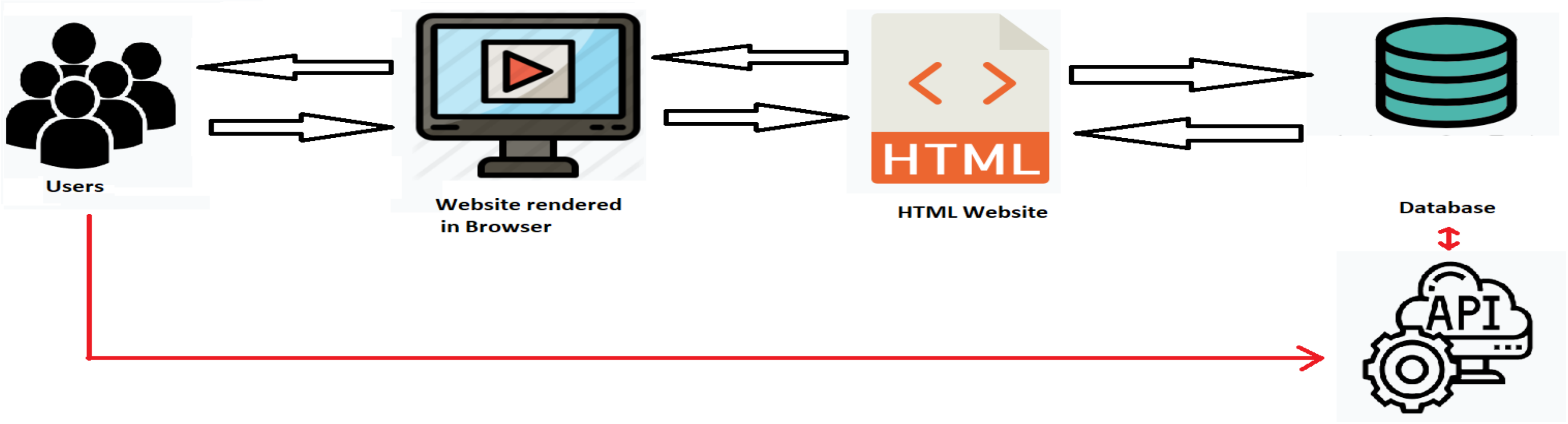
Accessing Social Media Data

	API/web scraping/own tools	Direct cooperation (e.g. embedded researcher)	Purchase data from market research or data reseller
Monetary costs	potential costs for recruitment/incentives or hardware	normally no additional costs	high costs
Required effort and skills	substantial amount of time and technical skills required	depends on the agreement, but typically less than in a full DIY approach	recruitment and/or data collection are taken care of
Control over data collection	depends on available options or documentation	depends on the agreement, but possible conflicts of interest	researchers have to buy data "as is" but many data resellers, for example, offer options for creating bespoke data collections
Comprehensiveness and depth of the data	depends on sample and/or API limitations	potentially richest source of data	depends on how data are collected
Data sharing	subject to Terms of Service (ToS)	subject to contractual agreements, but typically restricted	subject to contractual agreements, but typically restricted
Independence	only limited by options of the API or tool	companies might want to have a say in what is/can be studied	researchers can choose what data to purchase based on their research interests

(source: Breuer, Bishop, Kinder-Kurlanda, in press)

APIs

- An [Application Programming Interface](#)...
  - is a system built for developers
  - directly communicates with the database of a service
  - has a defined vocabulary of queries
  - controls what information is accessible, to whom, and in which quantities



APIs

- “An API is very much the same thing as a UI [user interface], except that it is geared for consumption by software instead of humans” (David Berlind, ProgrammebleWeb)
- However, APIs can also be used by researchers for collecting data
- ProgrammableWeb provides a good overview of available APIs, including the Twitter APIs

Words of caution about APIs

- APIs are services offered by the providers of specific platforms
- They may change or completely close off APIs as they wish and at any time
  - Facebook's essential lockdown of its Graph API in the wake of the Cambridge Analytica scandal is a good example
  - Freelon (2018) has, hence, argued that computational research is entering a "post-API age" and Bruns (2019) writes of an "APIcalypse"
- APIs typically have rate limits that regulate how much and how often you can make (certain) requests (and these also change)
- Like other services, APIs have specific Terms of Service (ToS) that users need to adhere to (more on that later)

A (Brief!) Introduction to JSON

- A very common format for the structured data provided by APIs is **JSON**
- “**J**ava**S**cript **O**bject **N**otation is an open standard file format, and data interchange format, that uses human-readable text to store and transmit data objects consisting of **attribute–value** pairs and array data types (or any other serializable value)” (Wikipedia)
- You can open and edit JSON files with text editors like Notepad++ or Atom
- You can also use browser extensions for Firefox or Chrome to explore JSON files (RStudio offers some nice options for exploring JSON files as well)
- Twitter provides a detailed explanation of the JSON data for Tweets
- Many tools for collecting social media data can convert JSON to other formats (like CSV)

The Twitter APIs

- Twitter provides extensive documentation for its APIs
- Twitter also provides information and resources specifically for academic researchers
- Twitter has different APIs
  - The **REST API** can be used to collect information about user accounts (e.g., their profile information or followers) as well as a limited number of historical tweets (currently up to 3200 per user)
  - The **Streaming API** allows the collection of tweets in real time
    - The free version of the Streaming API allows the collection of up to 1% of all tweets produced within 10 milliseconds of a request
    - Twitter promises that this sample is random, but some researchers have found reason to doubt this (Pfeffer et al., 2018)
    - If you limit your collection by specifying filter parameters like user accounts, geographic regions or keywords, it is possible to collect all relevant tweets (if the tweets matching your defined criteria represent less than 1% of all tweets posted within 10 milliseconds of your request)

Alternatives to APIs

	Pros	Cons
Web scraping	• Flexible • Independent of API limitations	• Unstructured data • Methodologically challenging • Not allowed by (ToS of) most social media platforms
Data donation from users (see, e.g., Halavais, 2019): Users can export their Twitter archive and share it with researchers	• Direct involvement of participants • Can be more transparent for participants • Independent of API ToS	• Effort for participants • Solutions for receiving and processing the data required

Tools for Collecting Twitter Data

- There are dozens of free (and open source) tools for collecting Twitter data
- The Social Media Lab at Ryerson University curates the Social Media Research Toolkit that provides a good overview
- The available tools differ in many regards
  - Do they offer a graphical user interface (GUI)?
  - Do they require programming skills?
  - Do they require API keys/a developer account?
  - What type of data do they collect/provide?
  - Can they also be used for analysis?
  - …

Tools for Collecting Twitter Data

	Description	GUI	Programming skills required?	API Key required?
TAGS	TAGS is a free Google Sheet template which lets you setup and run automated collection of search results from Twitter	Yes (Google Sheets)	No	Yes
COSMOS	COSMOS Open Data Analytics software provides ethical access to social media data for social science researchers	Yes	No	No
Chorus	Chorus is a free, evolving, data harvesting and visual analytics suite designed to facilitate and enable social science research using Twitter data	Yes	No	Yes
Facepager	Facepager was made for fetching public available data from YouTube, Twitter and other websites on the basis of APIs and webscraping	Yes	No, but requires a good understanding of the API	Yes
rtweet	R client for accessing Twitter's REST and stream APIs	No	R	Yes
Tweepy	An easy-to-use Python library for accessing the Twitter API	No	Python	Yes
GetOldTweets3	A project written in Python to get old tweets, it bypass some limitations of Twitter Official API	No	Python/Bash	No
TWINT	Twint is an advanced Twitter scraping tool written in Python that allows for scraping Tweets from Twitter profiles without using Twitter's API	Not yet	Python/Bash	No
Twitter Scraper	Python library for the collection of Twitter data without using the API	No	Python	No

Try out some of the tools

You can try out the R and Python libraries with interactive notebooks which you can access via this GitHub repository (without the need to install anything on your computer)

4: PROCESSING

The Disclosive Nature of Twitter Data

- We need to think very carefully about how we collect and link survey and Twitter data

- Surveys promise anonymity, and this needs to be maintained

- Even a list of Twitter usernames will identify who is in the sample

- Twitter data is highly disclosive, and not just for the reasons you might think…

Understanding Our Data

- Do we understand what Twitter data actually is?
- Do we know how the API works?
- Do we understand what is in the JSON?
- A single tweet can come with over 150 associated ‘attributes’!
- Consider the tweet, the user, and the geography

Relating to:	Attribute:	Description:	Nature of Risk:	Risk of Identifying an Individual:
Tweet	text	The actual text of the tweet	If not a retweet, then unique content and directly identifiable	HIGH
Tweet	retweet_count	The number of times a tweet has been retweeted	Changeable and dynamic, unlikely to be unique * unless extreme	LOW*
Tweet	favorite_count	The approximate number of times a tweet has been liked by other users	Changeable and dynamic, unlikely to be unique * unless extreme	LOW*
Tweet	favorited	Indicates whether a user has favourited the tweet	Binary categorical variable, common practice to 'favourite' a tweet	NEGLIGIBLE
Tweet	truncated	Whether a tweet text has been truncated (greater than 140 characters)	Binary categorical variable, truncation common with new 280 character tweet limit	NEGLIGIBLE
Tweet	id_str	The numeric (string) version of the unique identifier for this tweet	Unique content, directly identifiable - often deposited to allow other researchers to 'rehydrate' Twitter datasets	HIGH
Tweet	in_reply_to_screen_name	If the tweet is a reply to another tweet, this is the name of the original tweet's author	Evidence of Twitter correspondence with another unique user, may or may not represent someone in their network, often used for responding to public individuals (e.g. politicians) but could also be used to respond to users who are closely connected	VARIABLE
Tweet	source	The utility used to post the tweet (e.g. Tweets posted from the Twitter website have a source of 'web')	Unlikely to pose a risk as alternative Twitter posting tools are in widespread use	NEGLIGIBLE
Tweet	retweeted	Indicates whether the tweet has been retweeted by the user	Binary categorical variable, common practice to retweet	NEGLIGIBLE
Tweet	created_at	Creation date and time of the tweet to the second (in UTC) e.g. Tue Nov 23 12:46:54 +0000 2018	On average there are 6,000 tweets created every second (http://www.internetlivestats.com/twitter-statistics/), and difficult (if at all possible) to acquire all historic tweets made in a given second without access to the firehose (100% feed). Note that offset ('+0000') could be used to determine time zone (but see later comment on GDPR)	LOW
Tweet	in_reply_to_status_id_str	If the tweet is a reply to another tweet, this is the ID of the original tweet	Represents part of a conversation that the user is partaking in, could be used to identify an individual if number of responses to original tweet	VARIABLE

			are small	
Tweet	in_reply_to_user_id_str	If the tweet is a reply to another tweet, this is the ID of the original tweet's author	Evidence of Twitter correspondence with another unique user, may or may not represent someone in their network, often used for responding to public individuals (e.g. politicians) but could also be used to respond to users who are closely connected	VARIABLE
Tweet	lang	The language of the tweet text (machine-detected)	Machine detection will allocate to one language or mark as 'undetected', will only identify a single language, might well not be the same as language of interface, can change with every tweet (dynamic) * but might result in 'low cell count problem' for minority languages	NEGLIGIBLE*
Tweet	expanded_url	Full (expanded) version of a URL included in the tweet	Depends where the URL points to, often to generic content (e.g. BBC News story) but could be to personal website or blog	VARIABLE
Tweet	url	Wrapped URL corresponding to the value directly embedded into the raw tweet text	Depends where the URL points to, often to generic content (e.g. BBC News story) but could be to personal website or blog	VARIABLE
User	listed_count	The number of public lists that the user is a member of	Unlikely to be unique * unless extreme	LOW*
User	verified	Whether account has been verified (account of 'public interest')	Binary categorical variable, not unusual and could include actors, musicians, journalists, politicians, organisations etc	NEGLIGIBLE
User	location	The location defined by the user	May or may not represent where the user lives or works, but potentially could place user in a low level spatial unit	VARIABLE
User	user_id_str	The numeric (string) version of the unique identifier for this user	Unique identifier, directly identifies the user	HIGH
User	description	User-defined description of their account, often used as a 'bio'	Regardless of what the user writes, this is likely to unique to the individual	HIGH
User	geo_enabled	User has enabled the possibility of geotagging their tweets	Simply enables geotagging, does not enforce it. Binary categorical variable - research suggests that 41.6% of users have this setting enabled (Sloan & Morgan 2015)	NEGLIGIBLE
User	user_created_at	Creation date and time of the user account to the second (in UTC) e.g. Tue Nov 23 12:46:54 +0000 2018	Potentially unique to the individual due to high level of temporal granularity, note that offset ('+0000') can be used to determine time zone (but see later comment on GDPR)	HIGH
User	statuses_count	The number of tweets and retweets posted by the user	Changeable and dynamic, unlikely to be unique * unless extreme	LOW*
User	followers_count	The number of followers the user account currently has	Changeable and dynamic, unlikely to be unique * unless extreme	LOW*

User	favourites_count	The number of tweets the user has favourited since the account was created	Changeable and dynamic, unlikely to be unique * unless extreme	LOW*
User	protected	Whether account is protected (tweets only visible to followers)	Binary categorical variable, not unusual practice	NEGLIGIBLE
User	user_url	A URL given by the user, normally a link to a personal/organisational website	Not necessarily unique, but will be in some cases, not unusual for users to direct to personal websites	HIGH
User	name	The self-defined name of the user	Not necessarily the name of a person, but often is	HIGH
User	time_zone	The time zone of the user	If present will place the user in a large-scale geography, but from 23rd May has been returned as 'null' (private field) due to EU privacy laws	N/A
User	user_lang	The user's choice of interface language	Twitter is available in 47 languages (at time of writing), may well not be the same as the language in which tweets are written, can change but most likely to be static	NEGLIGIBLE
User	utc_offset	The difference in hours and minutes between user time zone and UTC	If present will place the user in a large-scale geography, but from 23rd May has been returned as 'null' (private field) due to EU privacy laws	N/A
User	friends_count	The number of accounts this user is following	Changeable and dynamic, unlikely to be unique * unless extreme	NEGLIGIBLE*
User	screen_name	The screen name (aka handle) of a user	Screen name can change (dynamic) but is always unique, an individual identifier	HIGH
Geo	country_code	Two letter code of the country a tweet was issued from, or is about	May be derived from an exact point coordinate (lat/long), or from a place selected by a user such as a city. In the latter, this may be the country of the place from where the user is tweeting from, or a place that they are tweeting about. Either way, on its own this represents a high-level geography	NEGLIGIBLE
Geo	country	Name of the country a tweet was issued from or is about	May be derived from an exact point coordinate (lat/long), or form a place selected by a user such as a city. In the latter, this may be the country of the place from where the user is tweeting from, or a place that they are tweeting about. Either way, on its own this represents a high-level geography	NEGLIGIBLE
Geo	place_type	The nature of the location the tweet was issued in, or is about, such as a city or POI	Classification of place identified by user (either selected or derived from point coordinates) is generic and unlikely to be problematic	NEGLIGIBLE
Geo	full_name	Full name (string) of place e.g. 'San Francisco, CA'	Could lead to low level-spatial data if point coordinates, or user selection, results in identifying a city or town	VARIABLE
Geo	place_name	Short name (string) of place e.g. 'San	Could lead to low level-spatial data if point coordinates, or user	VARIABLE

		Francisco'	selection, results in identifying a city or town	
Geo	place_id	Unique ID (string) of place	Could lead to low level-spatial data if point coordinates, or user selection, results in identifying a city or town	VARIABLE
Geo	place_lat	Centre point of the location the tweet was issued in, or is about (latitude)	Gives a latitude value at the centroid of the location (e.g. centre of Manchester), may or may not be where the user was when tweet was posted, unlikely to be of use without corresponding longitude value	LOW
Geo	place_lon	Centre point of the location the tweet was issued in, or is about (longitude)	Gives a longitude value at the centroid of the location (e.g. centre of Manchester), may or may not be where the user was when tweet was posted, unlikely to be of use without corresponding latitude value	LOW
Geo	lat	Latitude of tweet location	Precise latitude of where user was when they tweeted, potentially could be at home or work, alternatively may be commuting. Either way has considerable potential to locate individuals in low level geographies, but this is significantly reduced without longitude value *risk is considerably higher with corresponding longitude	MEDIUM*
Geo	lon	Longitude of tweet location	Precise longitude of where user was when they tweeted, potentially could be at home or work, alternatively may be commuting. Either way has considerable potential to locate individuals in low level geographies, but this is significantly reduced without latitude value *risk is considerably higher with corresponding latitude	MEDIUM*

Principles for Maintaining Security

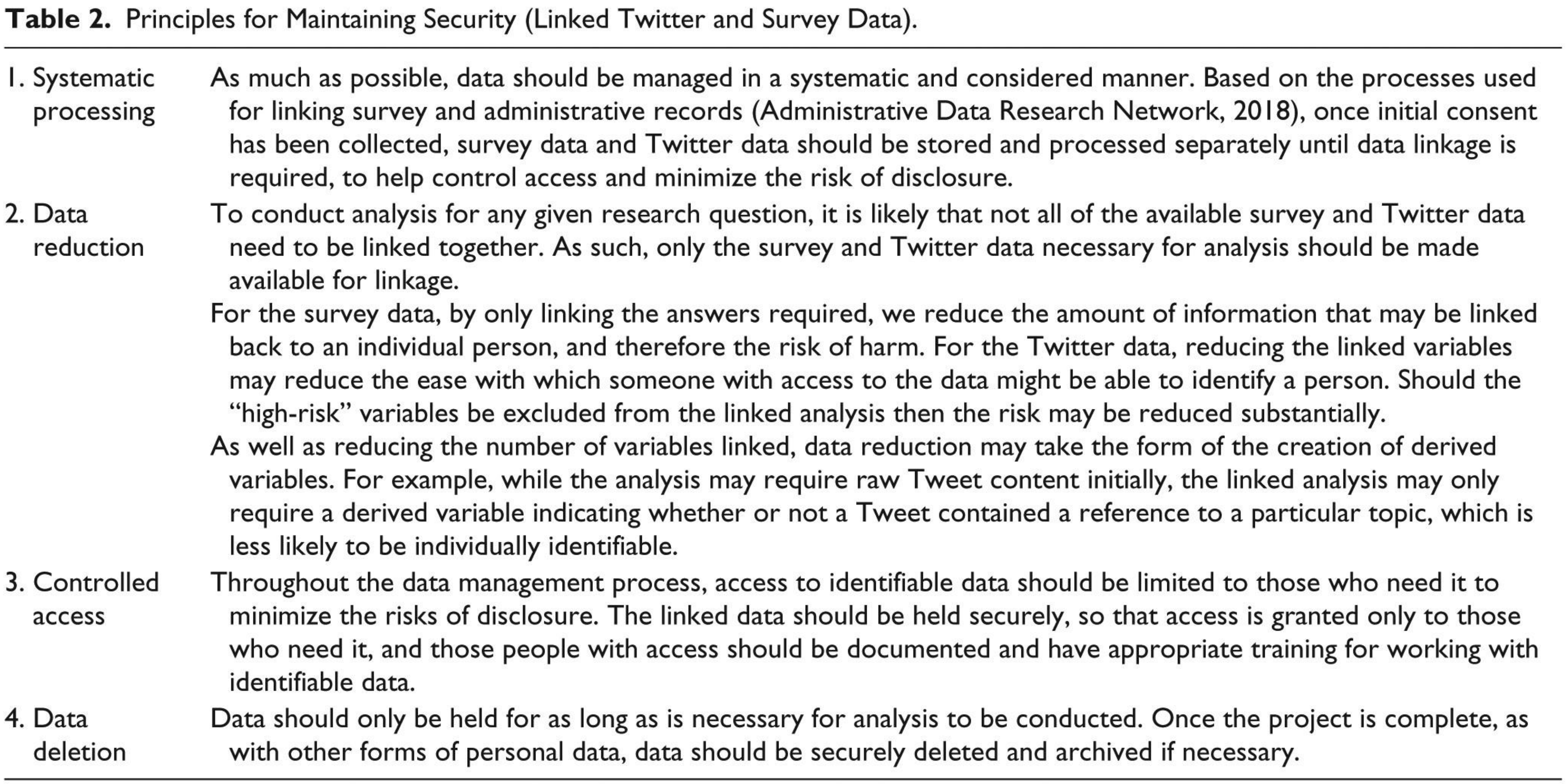
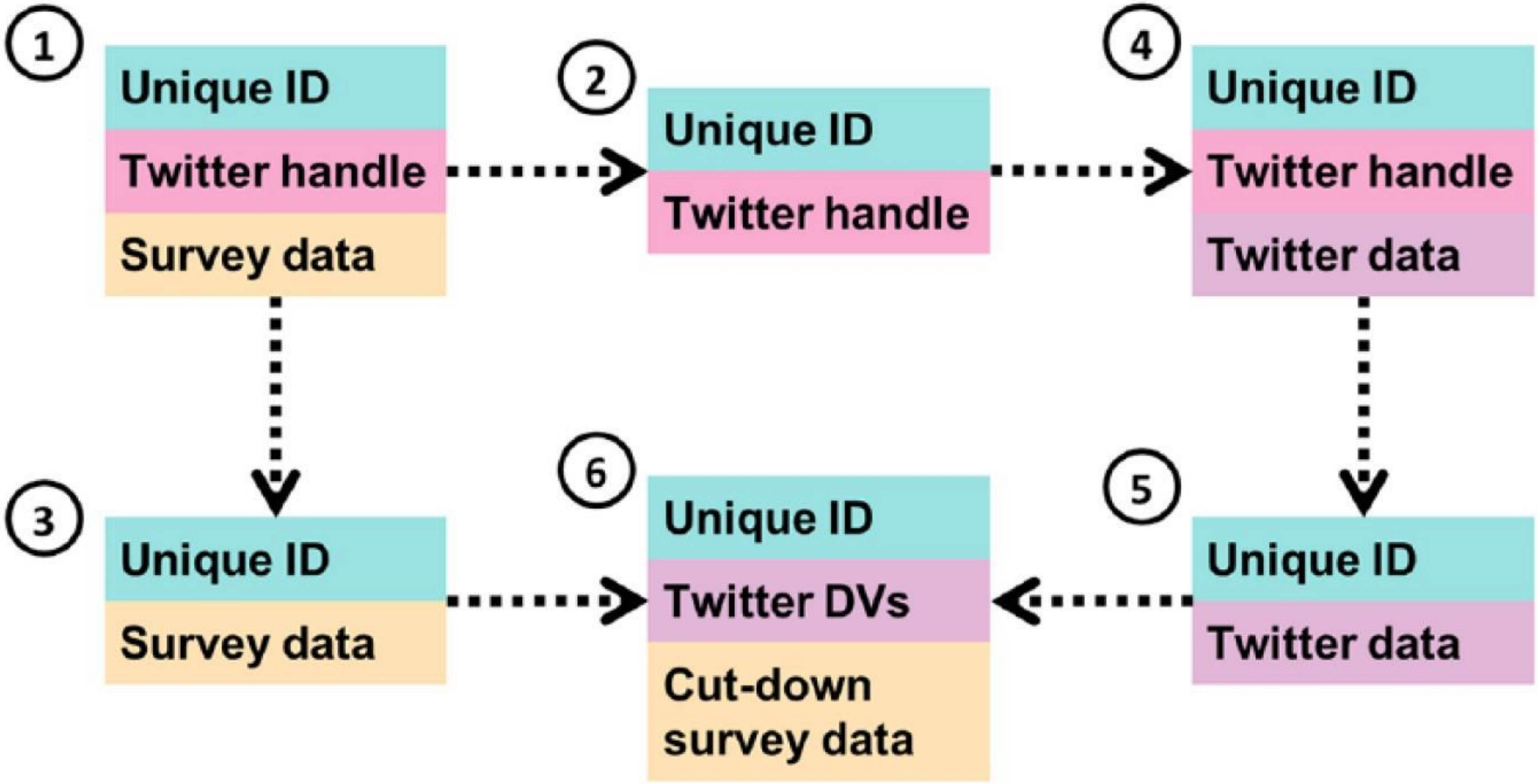


Table 2. Principles for Maintaining Security (Linked Twitter and Survey Data).

1. Systematic processing	As much as possible, data should be managed in a systematic and considered manner. Based on the processes used for linking survey and administrative records (Administrative Data Research Network, 2018), once initial consent has been collected, survey data and Twitter data should be stored and processed separately until data linkage is required, to help control access and minimize the risk of disclosure.
2. Data reduction	To conduct analysis for any given research question, it is likely that not all of the available survey and Twitter data need to be linked together. As such, only the survey and Twitter data necessary for analysis should be made available for linkage. For the survey data, by only linking the answers required, we reduce the amount of information that may be linked back to an individual person, and therefore the risk of harm. For the Twitter data, reducing the linked variables may reduce the ease with which someone with access to the data might be able to identify a person. Should the "high-risk" variables be excluded from the linked analysis then the risk may be reduced substantially. As well as reducing the number of variables linked, data reduction may take the form of the creation of derived variables. For example, while the analysis may require raw Tweet content initially, the linked analysis may only require a derived variable indicating whether or not a Tweet contained a reference to a particular topic, which is less likely to be individually identifiable.
3. Controlled access	Throughout the data management process, access to identifiable data should be limited to those who need it to minimize the risks of disclosure. The linked data should be held securely, so that access is granted only to those who need it, and those people with access should be documented and have appropriate training for working with identifiable data.
4. Data deletion	Data should only be held for as long as is necessary for analysis to be conducted. Once the project is complete, as with other forms of personal data, data should be securely deleted and archived if necessary.

Source: Sloan et al. (2020) *https://doi.org/10.1177/1556264619853447*

Flowchart: Splitting the Data



Source: Sloan et al. (2020) [*https://doi.org/10.1177/1556264619853447*](https://doi.org/10.1177/1556264619853447)

Derived Variables and Summary Measures

- If we reduce the granularity of the data, we can remove the risk of disclosure

- Consider…
  - Summaries of emotive states based on multiple tweets
  - Putting users into ordinal groups (e.g. deciles)
  - Introduce random error (replace values)
  - Aggregate to higher geographies

- All of these have disadvantages, not least a lack of transparency

- Could researchers request their own derivation approaches?

5: ANALYSIS

?

- That is for another day…
- If you want to learn about analyzing Twitter data you can, e.g., attend the GESIS workshop “Digital Trace Data in Social Science” (Dec 7-8, 2020)

6: ARCHIVING & SHARING

Why Archive Twitter (or any) Data?

- FAIR principles for stewardship of scientific research data
  - Findable
  - Accessible
  - Interoperable
  - Reuseable
- Funder and publisher requirements might matter too….

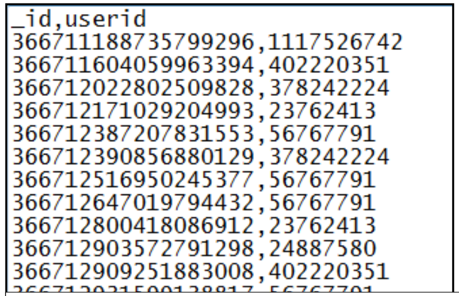
Boyd (2010) contends that this [digitally connected] era is characterized by a distinct set of affordances and dynamics. In particular, it affords persistence, replicability, scalability, and searchability of information. Papacharissi and Yuan (2011) add to this the affordance of shareability.

Jenny L Davis & Nathan Jurgenson (2014) Context collapse: DOI: 10.1080/1369118X.2014.888458

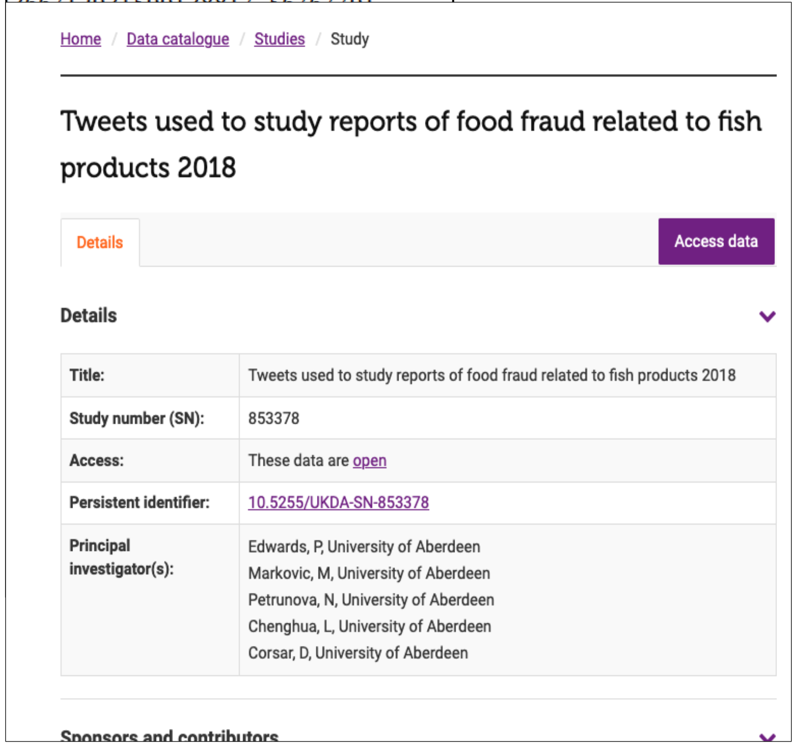
Archiving Twitter Data - Challenges

- Accepting Twitter’s Developer Policy is required for account and API
- “Express and informed consent required for (…) Republishing content accessed by means other than via the Twitter API or other Twitter tools
- You must maintain the integrity of all Twitter Content that you display publicly or to people who use your service.
- If you store Twitter Content offline, you must keep it up to date with the current state of that content on Twitter. Specifically, you must delete or modify any content you have if it is deleted or modified on Twitter.
- If you provide Twitter Content to third parties, including downloadable datasets or via an API, you may only distribute Tweet IDs, Direct Message IDs, and/or User IDs (except as described below).”
- Justin Littman - https://medium.com/on-archivy/twitters-developer-policies-for-researchers-archivists-and-librarians-63e9ba0433b2
- https://developer.twitter.com/en/developer-terms/policy

Archiving Twitter IDs – Current Practice



```
_id,userid
366711188735799296,1117526742
366711604059963394,402220351
366712022802509828,378242224
366712171029204993,23762413
366712387207831553,56767791
366712390856880129,378242224
366712516950245377,56767791
366712647019794432,56767791
366712800418086912,23762413
366712903572791298,24887580
366712909251883008,402220351
```



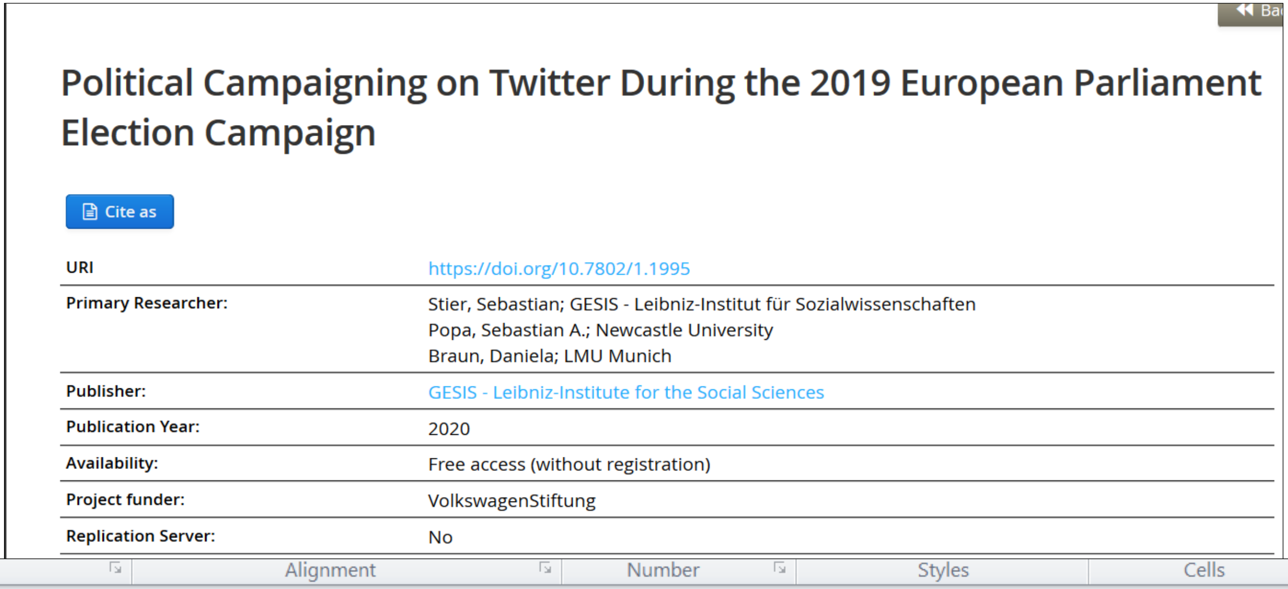
Home / Data catalogue / Studies / Study

Tweets used to study reports of food fraud related to fish products 2018

Details

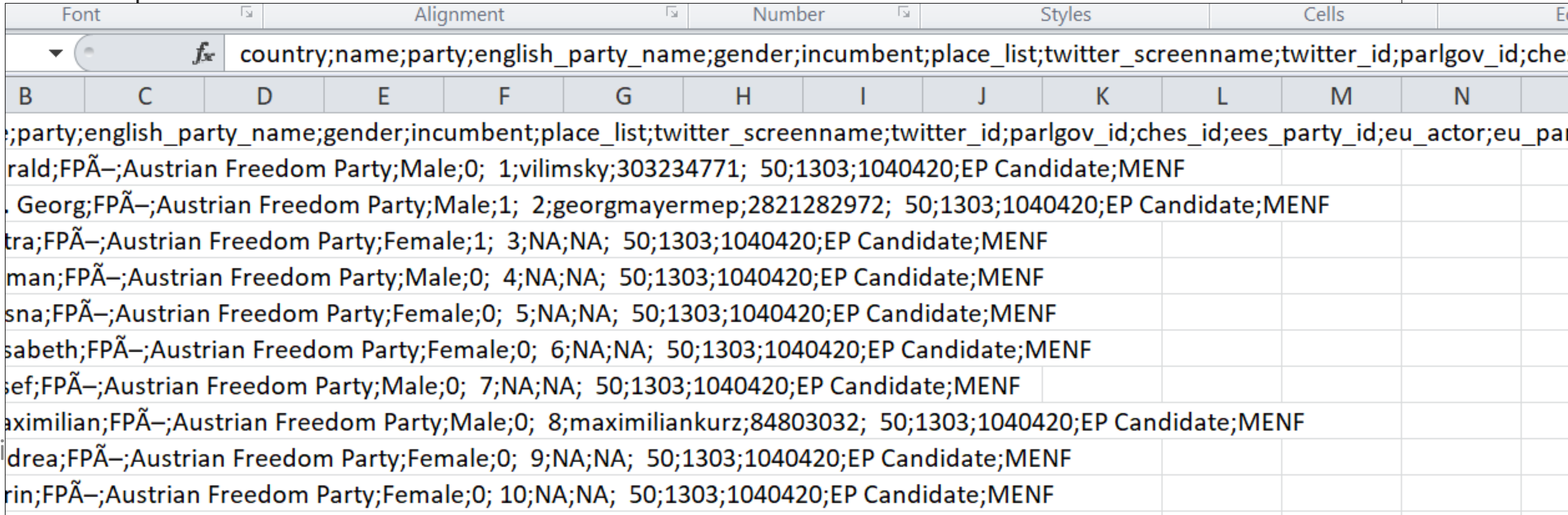
Title:	Tweets used to study reports of food fraud related to fish products 2018
Study number (SN):	853378
Access:	These data are open
Persistent identifier:	10.5255/UKDA-SN-853378
Principal investigator(s):	Edwards, P, University of Aberdeen Markovic, M, University of Aberdeen Petrunova, N, University of Aberdeen Chenghua, L, University of Aberdeen Corsar, D, University of Aberdeen

Sponsors and contributors



Political Campaigning on Twitter During the 2019 European Parliament Election Campaign

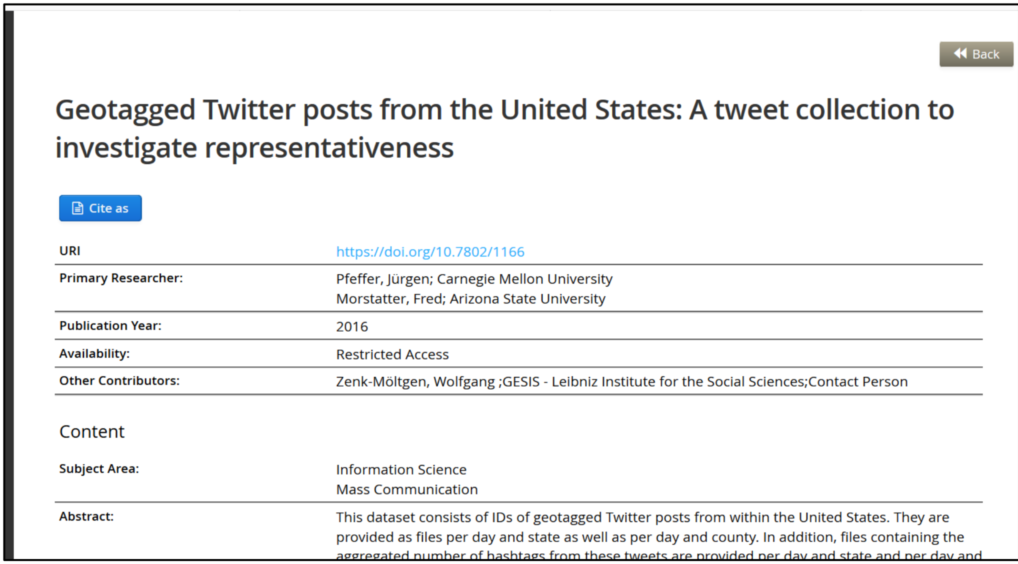
URI	https://doi.org/10.7802/1.1995
Primary Researcher:	Stier, Sebastian; GESIS - Leibniz-Institut für Sozialwissenschaften Popa, Sebastian A.; Newcastle University Braun, Daniela; LMU Munich
Publisher:	GESIS - Leibniz-Institute for the Social Sciences
Publication Year:	2020
Availability:	Free access (without registration)
Project funder:	VolkswagenStiftung
Replication Server:	No



```
country;name;party;english_party_name;gender;incumbent;place_list;twitter_screenname;twitter_id;parlgov_id;che
;party;english_party_name;gender;incumbent;place_list;twitter_screenname;twitter_id;parlgov_id;ches_id;ees_party_id;eu_actor;eu_pa
rald;FPÃ–;Austrian Freedom Party;Male;0;  1;vilimsky;303234771;  50;1303;1040420;EP Candidate;MENF
. Georg;FPÃ–;Austrian Freedom Party;Male;1;  2;georgmayermep;2821282972;  50;1303;1040420;EP Candidate;MENF
tra;FPÃ–;Austrian Freedom Party;Female;1;  3;NA;NA;  50;1303;1040420;EP Candidate;MENF
man;FPÃ–;Austrian Freedom Party;Male;0;  4;NA;NA;  50;1303;1040420;EP Candidate;MENF
sna;FPÃ–;Austrian Freedom Party;Female;0;  5;NA;NA;  50;1303;1040420;EP Candidate;MENF
sabeth;FPÃ–;Austrian Freedom Party;Female;0;  6;NA;NA;  50;1303;1040420;EP Candidate;MENF
sef;FPÃ–;Austrian Freedom Party;Male;0;  7;NA;NA;  50;1303;1040420;EP Candidate;MENF
aximilian;FPÃ–;Austrian Freedom Party;Male;0;  8;maximiliankurz;84803032;  50;1303;1040420;EP Candidate;MENF
drea;FPÃ–;Austrian Freedom Party;Female;0;  9;NA;NA;  50;1303;1040420;EP Candidate;MENF
rin;FPÃ–;Austrian Freedom Party;Female;0; 10;NA;NA;  50;1303;1040420;EP Candidate;MENF
```

Archiving Twitter with Access Controls

- “Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness”
- No tweet content, only IDs - to comply with Twitter Terms of Service
- Data accessible (by request) but not public because of no consent and reidentification risk
- Archived in SowiDataNet-*datorium*
  - Findable – Pfeffer, J. and Morstatter, F. (2016)
  - Preserved – DOI - (http://dx.doi.org/10.7802/1166)
  - Reproducible Python scripts, tools ,documentation
- [As open as possible, closed when necessary]()



Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness

URI	https://doi.org/10.7802/1166
Primary Researcher:	Pfeffer, Jürgen; Carnegie Mellon University Morstatter, Fred; Arizona State University
Publication Year:	2016
Availability:	Restricted Access
Other Contributors:	Zenk-Möltgen, Wolfgang ;GESIS - Leibniz Institute for the Social Sciences;Contact Person

Content

Subject Area:	Information Science Mass Communication
Abstract:	This dataset consists of IDs of geotagged Twitter posts from within the United States. They are provided as files per day and state as well as per day and county. In addition, files containing the aggregated number of hashtags from these tweets are provided per day and state and per day and

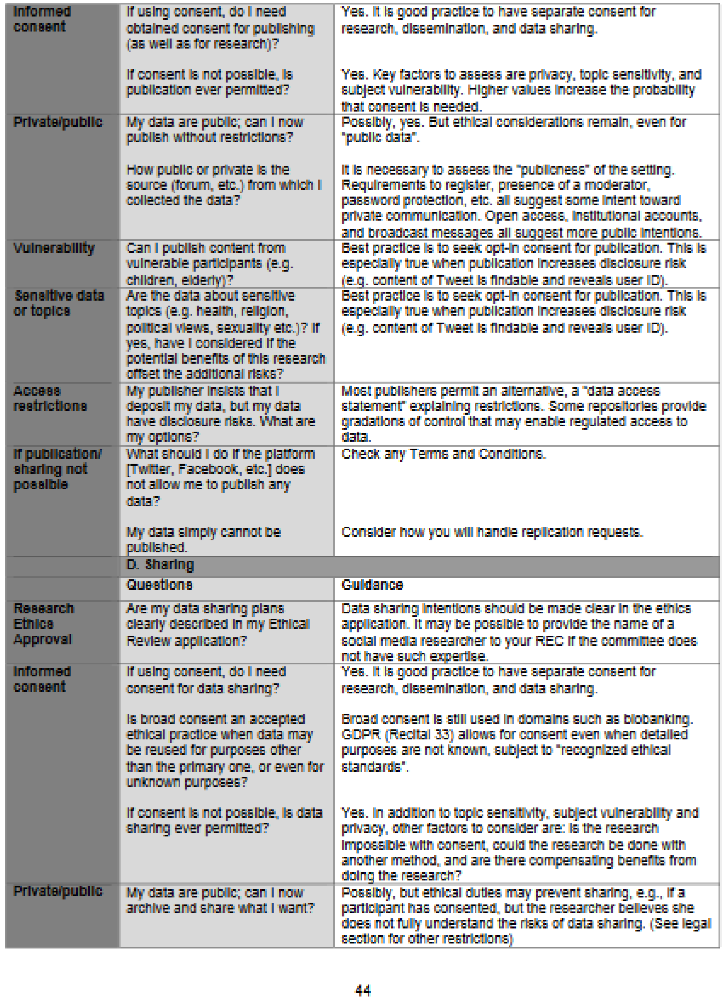
A “solution”, but not satisfactory

- Consent rates low with surveys, and even lower without (and hard work)
- Archiving only Tweet IDs does not meet standard of replication, partly due to deletions - 30-80% persistence rate over four years
  - (Zubiaga, A., “A longitudinal assessment of the persistence of Twitter datasets”, 2018)
- Who counts as third party? Anyone not you? Your team? Your institution? Your research network? Your archiving consortium?
- Treats all tweets the same – public/private, institution/individual
- Collaboration with platforms – better quality, but greater “digital divide” (and usually focus is on research access, not sharing)
  - see Bruns (2019) & Puschmann (2019) Information, Communication, & Society papers
- And finally, solution is a moving target because Terms can (and do) change

Special Considerations – Off -Twitter Matching

- **“We limit the circumstances under which you may match a person on Twitter to information obtained or stored off-Twitter.** Off-Twitter matching involves associating Twitter Content, including a Twitter @handle or user ID, with a person, household, device, browser, or other off-Twitter identifier. You may only do this if you have express opt-in consent from the person before making the association, or as described below.
- In situations in which you don’t have a person’s express, opt-in consent to link their Twitter identity to an off-Twitter identifier, we require that any connection you draw be based only on information that someone would reasonably expect to be used for that purpose. In addition, absent a person’s express opt-in consent you may only attempt to match your records about someone to a Twitter identity based on:
  - **Information provided directly to you by the person.** Note that records about individuals with whom you have no prior relationship, including data about individuals obtained from third parties, do not meet this standard; and/or
  - **Public data.**“

Can I just tick a box please?.....No, but…

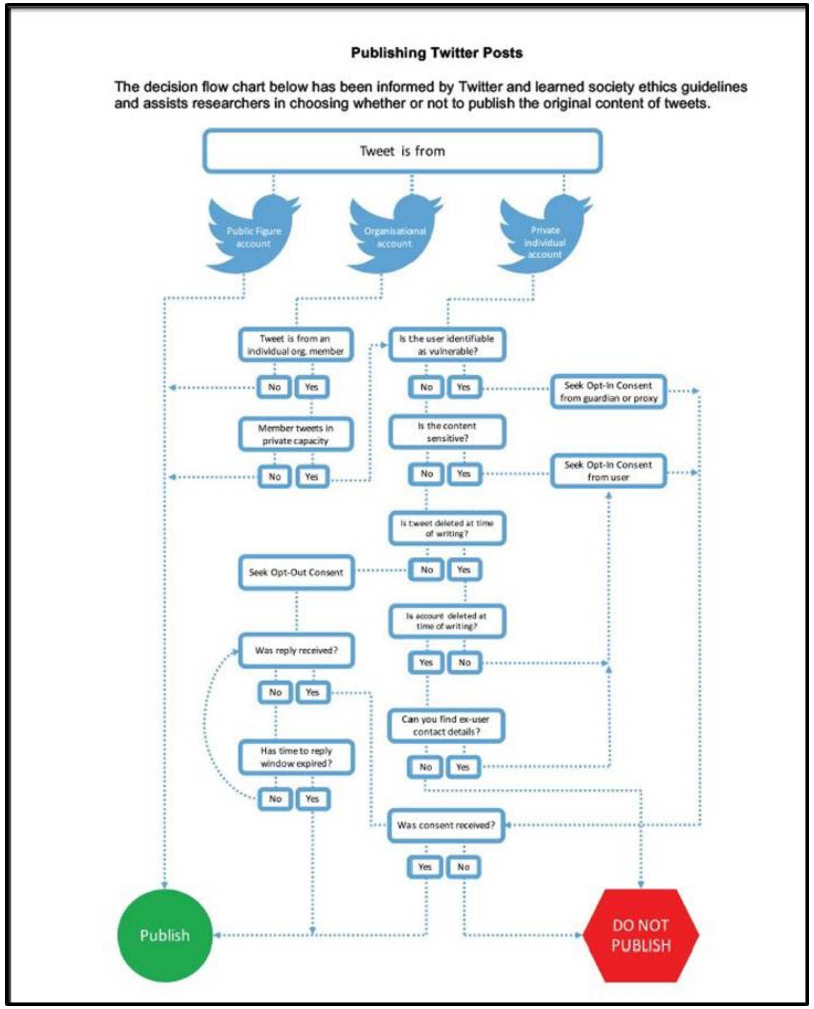


Informed consent	If using consent, do I need obtained consent for publishing (as well as for research)? If consent is not possible, is publication ever permitted?	Yes. It is good practice to have separate consent for research, dissemination, and data sharing. Yes. Key factors to assess are privacy, topic sensitivity, and subject vulnerability. Higher values increase the probability that consent is needed.
Private/public	My data are public; can I now publish without restrictions? How public or private is the source (forum, etc.) from which I collected the data?	Possibly, yes. But ethical considerations remain, even for "public data". It is necessary to assess the "publicness" of the setting. Requirements to register, presence of a moderator, password protection, etc. all suggest some intent toward private communication. Open access, institutional accounts, and broadcast messages all suggest more public intentions.
Vulnerability	Can I publish content from vulnerable participants (e.g. children, elderly)?	Best practice is to seek opt-in consent for publication. This is especially true when publication increases disclosure risk (e.g. content of Tweet is findable and reveals user ID).
Sensitive data or topics	Are the data about sensitive topics (e.g. health, religion, political views, sexuality etc.)? If yes, have I considered if the potential benefits of this research offset the additional risks?	Best practice is to seek opt-in consent for publication. This is especially true when publication increases disclosure risk (e.g. content of Tweet is findable and reveals user ID).
Access restrictions	My publisher insists that I deposit my data, but my data have disclosure risks. What are my options?	Most publishers permit an alternative, a "data access statement" explaining restrictions. Some repositories provide gradations of control that may enable regulated access to data.
If publication/ sharing not possible	What should I do if the platform [Twitter, Facebook, etc.] does not allow me to publish any data? My data simply cannot be published.	Check any Terms and Conditions. Consider how you will handle replication requests.
	D. Sharing	
	Questions	Guidance
Research Ethics Approval	Are my data sharing plans clearly described in my Ethical Review application?	Data sharing intentions should be made clear in the ethics application. It may be possible to provide the name of a social media researcher to your REC if the committee does not have such expertise.
Informed consent	If using consent, do I need consent for data sharing? Is broad consent an accepted ethical practice when data may be reused for purposes other than the primary one, or even for unknown purposes? If consent is not possible, is data sharing ever permitted?	Yes. It is good practice to have separate consent for research, dissemination, and data sharing. Broad consent is still used in domains such as biobanking. GDPR (Recital 33) allows for consent even when detailed purposes are not known, subject to "recognized ethical standards". Yes. In addition to topic sensitivity, subject vulnerability and privacy, other factors to consider are: is the research impossible with consent, could the research be done with another method, and are there compensating benefits from doing the research?
Private/public	My data are public; can I now archive and share what I want?	Possibly, but ethical duties may prevent sharing, e.g., if a participant has consented, but the researcher believes she does not fully understand the risks of data sharing. (See legal section for other restrictions)

Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation
Matthew L Williams, Pete Burnap, Luke Sloan
Sociology, First Published May 26, 2017
https://doi.org/10.1177/0038038517708140

Appendix A: Short guide on legal and ethical issues for the researcher to consider when using social media for research

SERISS WP6-D3 Report (not guidelines)
https://seriss.eu/wp-content/uploads/2019/11/D6.3-Report-on-legal-and-ethical-framework-and-strategies…__FINAL.pdf



Publishing Twitter Posts

The decision flow chart below has been informed by Twitter and learned society ethics guidelines and assists researchers in choosing whether or not to publish the original content of tweets.

Tweet is from
Public Figure account
Organisational account
Private individual account
Tweet is from an individual org. member
No
Yes
Member tweets in private capacity
No
Yes
Is the user identifiable as vulnerable?
No
Yes
Seek Opt-In Consent from guardian or proxy
Is the content sensitive?
No
Yes
Seek Opt-In Consent from user
Is tweet deleted at time of writing?
No
Yes
Seek Opt-Out Consent
Is account deleted at time of writing?
Yes
No
Was reply received?
No
Yes
Can you find ex-user contact details?
No
Yes
Has time to reply window expired?
No
Yes
Was consent received?
Yes
No
Publish
DO NOT PUBLISH

cessda

We very much appreciate the free and open source tools for Twitter data collection that we introduced in this workshop. If you use them (or any other free academic software like packages for R or Python), please cite them!

This workshop was supported by CESSDA and is part of its 2020 Work Plan *New Data Types*.

cessda.eu

@CESSDA_Data