Informed consent for linking survey and social media data - Differences between platforms and data types

Johannes Breuer[1], Tarek Al Baghal[2], Luke Sloan[3], Libby Bishop[4], Dimitra Kondyli[5], Apostolos Linardis[6]

Abstract

Linking social media data with survey data is a way to combine the unique strengths and address some of the respective limitations of these two data types. As such, linked data can be quite disclosive and potentially sensitive, it is important that researchers obtain informed consent from the individuals whose data are being linked. When formulating appropriate informed consent, there are several things that researchers need to take into account. Besides legal and ethical questions, key considerations are the differences between platforms and data types. Depending on what type of social media data is collected, how the data are collected, and from which platform(s), different points need to be addressed in the informed consent. In this paper, we present three case studies in which survey data were linked with data from 1) Twitter, 2) Facebook, and 3) LinkedIn and discuss how the specific features of the platforms and data collection methods were covered in the informed consent. We compare the key attributes of these platforms that are relevant for the formulation of informed consent and also discuss scenarios of social media data collection and linking in which obtaining informed consent is not necessary. By presenting the specific case studies as well as general considerations, this paper is meant to provide guidance on informed consent for linked survey and social media data for both researchers and archivists working with this type of data.

Keywords

informed consent, social media, surveys, data linkage

1. Introduction

Social media data have been a popular subject of study in the social sciences (as well as various other scientific disciplines) for quite some time as they have become a part of everyday life for many people and are used for a variety of activities that are of interest to social scientists, such as communication, information seeking, news consumption, and relationship management. Much of the research on the use and effects of social media data in the (quantitative) social sciences is based on survey data. When studying the use of media, however, several studies have shown that self-reports can be unreliable due to issues of social desirability or difficulties in recalling instances or patterns of usage (Araujo et al., 2017; Prior, 2009; Scharkow, 2016). A way to assess social media use more reliably is to use data obtained directly from the platforms. The types of data available depend on the platform, and they can be collected in different ways (see the following section). Notably, social media data can not only be used to study social media usage itself but also to investigate a variety of other topics, such as political communication or the formation and expression of opinions.

While social media data have several advantages compared to survey data, they also have certain limitations. Two important ones, especially for social-scientific research, are that they often lack in-depth explicit information about the individuals, e.g., regarding their socio-demographic attributes or attitudes, as well as relevant outcome variables, such as voting or purchasing behaviour or offline forms of civic engagement. To combine the unique advantages and deal with their respective limitations, data from surveys and social media can be linked (Stier et al., 2020). The linkage of surveys and social media data holds great potential and can be used to study a large variety of subjects (for a few examples, see the special issue 'Integrating Survey Data and Digital Trace Data' of the journal *Social Science Computer Review*).[7]

If researchers want to link surveys and social media data, there are several things they need to consider and address. One key issue is that of informed consent. As linked survey and social media data can be quite extensive, disclosive, and potentially also sensitive, obtaining informed consent is an important step in the process. While there can also be other legal bases for collecting and processing social media data for research, from an ethical perspective, obtaining informed consent is the preferable option for linking surveys and social media data (Menchen-Trevino, 2018). In this paper, we will discuss what researchers need to consider with regard to informed consent when they link surveys with social media data. Following some general considerations, we present experiences and solutions from three case studies in which survey data were linked with data from 1) Twitter, 2) Facebook, and 3) LinkedIn. We will compare these different cases and highlight similarities as well as differences between the platforms that are relevant for obtaining informed consent from participants. We also discuss cases in which obtaining informed consent is not required. More broadly, we discuss what to consider when ingesting such data into repositories and provide guidance on what researchers should pay attention to with regard to informed consent if they want to link surveys and social media data and subsequently archive them via a data repository. Accordingly, the considerations and suggestions in this paper are mostly targeted at researchers but are also relevant for staff at data archives who want to archive linked survey and social media data.

2. Linking surveys and social media data

There are two important factors that determine how survey data can be linked with social media data: 1) the type of social media data, and 2) the way(s) in which they were collected. Social media data can come from a wide range of platforms with very different purposes and attributes. In addition, the same platform can provide various types of data. Data from the platforms can include textual data (tweets, posts, comments, etc.), audio-visual material (images, video, etc.), network data (connections between users or content), or user profile information (name, location, occupation, etc.). Similar to the types of data, the ways in which they are collected or acquired can also vary (see Breuer et al., 2020). The most widely used approach is that researchers collect social media data themselves via Application Programming Interfaces (APIs) provided by the platforms or web scraping. However, they can also acquire data by entering into direct cooperation with the platforms or purchasing data from data resellers or market research companies. As an alternative to acquiring data via the platforms, researchers can also directly collaborate with users to collect social media data (see Halavais, 2019; we will discuss this option in more detail in a later section). Finally, it is also possible to reuse existing collections of social media data that have been created by other researchers and made available through data repositories or some other service. Importantly, the

type of data and how they are acquired affects how they can be linked. For example, the Terms of Service (ToS) of a platform API or contractual agreements with the platforms or data resellers may place restrictions on how the data can be used.

There are different ways in which social media data can be linked with survey data (see Stier et al., 2020). Depending on the type of social media data and how they are acquired, they can be linked with survey data on the individual level or an aggregate level, and they can be collected together for the same units of observation (*ex-ante linking*) or separately and linked subsequently (*ex-post linking* that uses existing survey and/or social media datasets). Within these types of data linking, different research designs are possible; for example, in the case of individual-level ex-ante linking, researchers can start with the survey and ask respondents to share or allow the collection of their social media data. Likewise, they can also first collect social media data and then invite users whose data they have collected to participate in a survey. In both cases, informed consent for collecting or using people's social media data and linking it with the survey data can be obtained as part of the survey.

The type of social media data that is collected, as well as the way in which it is supposed to be linked to survey data, determine what the informed consent needs to look like. In general, the informed consent for linking survey data with social media data needs to be in accordance with relevant local legal regulations, such as the General Data Protection Regulation (GDPR) in Europe, and should satisfy relevant ethical standards as defined by Institutional Review Boards (IRB), ethics committees, or the ethical guidelines of scholarly societies. GDPR requires a legal basis for processing personal data and, when linking data, informed consent is the standard.

Notably, when survey and social media data are linked, at least during data collection, identities are known, so the data are always personal (and sometimes also sensitive, e.g., when they include information about religious or political beliefs) and, thus, GDPR applies. Under GDPR, consent needs to be voluntary, informed, unambiguous, specific, and a clear affirmative action. 'Passive' consent, for example, the use of pre-ticked boxes, is not acceptable. Consent forms need to be in language suitable for the intended audience. Participants should be informed about: how any personal data collected about them will be used, stored, processed, transferred, who the data controller is (and their contact details), the legal grounds and purpose of the processing, any recipients of the personal data, the period of retention and their rights (including that they can complain to the Supervisory Authority; see, e.g., the UKDS GDPR guidelines[8]).

These consent requirements are similar, but not identical, to the ethical requirements of many ethical review bodies. An ethical review will typically also require addressing additional issues, such as the participation of children or vulnerable people. Finally, the formulation of the informed consent also needs to take into account the characteristics of the social media platform and the specific type(s) of data that should be linked with survey data. In the following section, we will focus on individual-level linking that starts with the survey. However, many of the considerations regarding the platform attributes and what implications these have for obtaining informed consent are also applicable to other kinds of social media data and linking approaches.

3. Differences among social media platforms and data types that are relevant for informed consent

There is a tendency to treat social media data (and platforms) as homogenous, and this extends into the literature on survey and social media data linkage. The assumption is that there are universal rules and protocols that can be applied to ensure informed consent for data linkage, but this is true only to a limited extent. The platforms have different purposes, the data are structured differently, the data are collected in different ways, and ascertaining a unique identifier for a respondent on a platform is simple in some cases and complicated in others. There are also complications concerning what is actually considered public, as some platforms allow anyone (logged in or not) to view data, others require a researcher to have an account and to log in, and some may even require there to be a link (e.g., following, friendship, connection) between the respondent and researcher before any data can be viewed. Beyond the technical questions of visibility and data access, users also have different expectations about how private specific types of information are on different platforms. Given the potential disconnect between users' views on the privacy and sensitivity of data and levels or ways of accessing platforms, it may be that, even though a respondent might consider their data to be 'public' and be happy to share it, the technological attributes of a platform can make that data difficult to access. To demonstrate this and discuss what it means for obtaining informed consent, we discuss three case studies of survey and social media platform data linkage covering three platforms: Twitter, LinkedIn, and Facebook.

3.1. Case study 1: Twitter

The first case study on Twitter data is based on two studies, one from the UK and one from Germany. The UK study detailed in Al Baghal et al. (2019) draws upon three representative surveys of the British adult population: the British Social Attitudes Survey 2015, the Understanding Society Innovation Panel 2017, and the NatCen Panel 2017.

The design of the German study was different from the UK study in several regards. The participants in this study came from a non-probability web-tracking panel in which participants have agreed to have their browsing behavior tracked. The panel is maintained by a professional market research company. For a project with a methodological interest in questions of data linking and a substantive interest in online news consumption, researchers purchased access to the web tracking data for one year. The participants of this panel were invited to different online surveys. In the first of these, those who reported having a personal Twitter account were asked for consent to link their Twitter data to their survey responses.

Public or Private?

If we consider social media platforms to sit on a continuum with 'public' at one end and 'private' at the other, then Twitter is quite firmly at the 'public' end. Notwithstanding debates about the 'imagined audience' of a tweet (Marwick and boyd 2010), Twitter is a broadcast medium through which tweets can be viewed by anyone. They are visible via search engines and can be viewed without having to log in to the site. Users can select to mark their tweets as *protected*, which means that their tweets are visible only to their followers (and users have to approve who these followers are), but these options are made clear to users - meaning that the public/private nature of a tweet is well defined on this platform.

Specifics of Informed Consent

Sloan et al. (2020) discuss their procedure for gaining informed consent for survey and Twitter data linkage. They identify five areas from Singleton and Wadsworth (2006) which need to be addressed: (a) why the data is being collected; (b) what will be done with it; (c) what is being collected; (d) secure data storage; and (e) maintaining anonymity. Accordingly, they developed the following consent statement:

> (a) As social media plays an increasing role in society, we would like to know who uses Twitter, and how people use it. (b) We are also interested in being able to add people's, and specifically your, (c) answers to this survey to publicly available information from your Twitter account such as your profile information, tweets in the past and in future, and information about how you use your account.
>
> (d) Your Twitter information will be treated as confidential and given the same protections as your interview data. (e) Your Twitter username, and any information that would allow you to be identified, will not be published without your explicit permission.
>
> Sloan et al. (2020, p. 65)

Any consent statement needs to address the specific types of data that a platform generates, using terminology that users will understand. In the extract above, the statement mentions tweets, profile information, and information about how the platform is used. These three broad areas simplify the complexity that underlies Twitter data. Notably, when extracting data from the Twitter API (see below), a single tweet can have over 150 attributes associated with it, covering everything from the content of the tweet itself to the number of followers the user has and various measures of geographical location. It is also not possible to explain the complexity of the analysis that this linked data will be subjected to. Sloan et al. (2020) acknowledge that there is a compromise here between complete information and the need to provide a practical and comprehensible explanation that enables participants to make an informed decision. Further information was provided in a series of help screens that participants could access if needed, covering:

> 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?
>
> What if I change my mind?

The language of the consent statement in the German study was based on the one developed by Sloan et al. (2020). The text was translated into German and slightly adapted to reflect the design and purpose of the study. Still, the wording is very similar to that used by Sloan et al. (2020). What

was different in this study compared to Sloan et al. (2020) was that the more detailed information about the data collection and handling was not provided via additional info screens but on a separate website that was linked in the consent statement in the online survey. The full text that was presented on that website is included as an appendix for this paper. Accordingly, the consent statement in the German study was the following (note: we translated the German text into English, trying to be as literal as possible with our translation):

> Since social media play an increasingly important role in society, we would like to know who uses Twitter and how people use Twitter. We are also interested in combining the answers from people, and also your responses from the survey with publicly available information from your Twitter account.
>
> Would you be willing to provide us with your Twitter username for this research project so that we can link your Twitter data with your responses from this survey for scientific purposes?
>
> Of course, your data will be treated confidentially and not used for commercial purposes. Your Twitter name will not be mentioned in any publication and all Twitter data will be protected by us with the same care as the data from the survey. You can find more information on how we process the data here [link to website with information].

Another feature of the consent statement for Twitter and survey data linkage is the need to specify that consent is being given to collect both *historic* and *future* data. As a microblogging platform, Twitter is not static, and the platform encourages frequent interaction with other users and continuous production of content - what Edwards et al. (2013) describe as *locomotive*. Because of the fast turnover of information, it is important that respondents are given a cue to consider their past behaviour and published content on the platform. Some users will have tweets going back years, and, unlike a biographical platform such as LinkedIn where users are encouraged to keep their profiles current, Twitter users are unlikely to monitor or regulate their past activity.

Unique Identifiers

Unique identifiers are essential for the data linkage process as they allow the researcher to identify an individual user on a given social media platform in an unambiguous manner. While this may seem obvious, there are two related issues : 1) Researchers must know what this identifier should be, and 2) when working with the unique identifiers, measures to protect participant privacy need to be taken. For Twitter data, the question of what the unique identifier should be is easy to answer. Twitter usernames are unique to each user, and the user can specify what this username should be. The username is often referred to as a Twitter handle, and they are the mechanism through which people tweet each other (a mention), and can be used as an alternative to a phone number or email address when logging into the site. It is reasonable to expect a survey respondent to know what their username is, although recall ability and accuracy will be determined by how heavily they use the platform and when they last logged in.

Al Baghal et al. (2020) detail the questions used in the same group of studies, as discussed by Sloan et al. (2020) above. The version used in the Understanding Society Innovation Panel 2017 is as follows:

> What is your Twitter username (e.g. @usociety)?
>
> *Soft check: Twitter username does not begin with '@' or contain spaces* 'Please check and amend. Twitter usernames should begin with an @ character and should not contain any spaces.'

The use of the @ symbol on Twitter is the universal standard for addressing a user. Therefore, having an @ in the prompt for their username further clarifies what is required and should be understood by any Twitter user. The further check of ensuring there are no spaces is intended to avoid respondents confusing their username with their Twitter name (which is normally the actual name of a user).

Again, in the German study, the language was quite similar:

> Please enter your Twitter username (e.g. @gesis.org) into the free-text field.
>
> My Twitter username is: @_____

The instructions, as well as the additional soft check in the UK study, illustrate what can go wrong when linking survey and Twitter data via the username. People may misspell their usernames or even (intentionally or unintentionally) provide a handle that is not theirs. This happened in both of the studies that this case study is based on, meaning that the linkage of survey and Twitter data failed in these cases. To minimize data loss due to typos or the provision of a wrong username, one solution can be to have participants follow and/or send a direct message to a Twitter account created by researchers for the purpose of the study. While it is helpful to remind respondents of the expected format, at least in the German study, some respondents may have been confused by the @ symbol as they entered their email address instead of their username. This confusion is even more understandable when considering that an email address is what many users use to log into their Twitter account.

Another consideration is that usernames can change. Hence, if a substantial amount of time passes between obtaining informed consent and the username and collecting the data, the username may have changed. It is also possible that accounts are deleted in the meantime. One way to address the issue of changing user names is to obtain the user ID based on the username via the Twitter API. Unlike the username, the user ID is persistent.

To increase data privacy, Twitter usernames should only be used as unique identifiers when necessary. For the linking process, using a unique generic ID is preferable. In addition, the full survey and Twitter data should be kept separate. Sloan et al. (2020) present a workflow that ensures that there is no linked dataset that contains the full survey data and the full Twitter data. The German study went further and included an explicit reference to this on the website containing the extended information on the collection and use of the Twitter data (the full text can be found in Appendix A):

Only information that is no longer personally identifiable (e.g., how often you tweet, how often you address political issues on Twitter, etc.) is linked to the survey data.

Data Access

Except when the username has changed, a researcher can easily identify an individual user profile through searching on the Twitter website, using a search engine, or via the Twitter APIs. The latter method is widely used by researchers, and there are all manner of tools developed for researchers that draw on the APIs, allowing researchers to access historical data (the REST API) or current data (the Stream API). When collecting data via the Twitter API based on usernames, data for protected accounts cannot be collected via the API. A potential alternative method for collecting Twitter data that also allows accessing data for protected accounts is to have participants export their personal Twitter data archive (which is an option available via the Twitter account settings) and share it with the researchers. These data are not limited by the limitations of the API, which restricts the amount of historical data that can be accessed. However, this method of data donation (which we will discuss again for the next case study) means more effort for the participants and requires a safe solution for transferring the data to the researchers.

Rights to the Data

Another consideration that needs to be made is the question of who has the rights to the data. What is important to note here is that none of the authors of this paper are lawyers, so what we say here as well in the corresponding sections for the other two case studies are our personal views based on our experience as researchers and/or data archives personnel and should not be taken as legal advice. There are other sources that provide legal opinions on matters related to the use of social media data. One such example is the expert opinion included in the report on 'Big data in social, behavioural, and economic sciences' by the RatSWD [German Data Forum] (2020). While its focus is on web scraping, it also includes a short section on the 'Binding effect of the Twitter API terms of use'. Also, while the expert opinion was written for the German case, it includes several sections discussing EU law, including the GDPR.

In general, if social media data are collected via APIs, their Terms of Service (ToS) are an important thing to consider when assessing what can be done with the data. Notably, ToS can be somewhat open to interpretation, especially for the case of academic research. While this is also not based upon legal expertise, a blog post by Justin Littman provides a good breakdown of 'Twitter's Developer Policies for Researchers, Archivists, and Librarians'.[9] One aspect on which the Twitter API ToS and Developer Policies place restrictions is the sharing of the data. Notably, even when Twitter data are linked with survey data and informed consent is obtained via the survey, the Twitter data collected via the API are observed through a platform owned by a commercial company rather than directly provided by the individuals (as would be the case in a data donation scenario; see the next case study). This means that platform ToS and Developer Policies need to be considered by researchers and archivists when deciding how the data can be used and shared.

Data Sharing

The Twitter Developer Policies state that data accessed via the Twitter APIs cannot be shared in full with third parties. Most importantly, one of the requirements is that only the Tweet IDs can be shared (not the tweet text or the associated metadata). Hence, if researchers archive Twitter data, they typically only archive Tweet IDs (see Kinder-Kurlanda et al., 2017). In their FAQ[10], the UK Data Service also lists this as a requirement for depositing Twitter data. If other researchers want to reuse the data, they need to collect the tweets again based on the list of Tweet IDs; a process called rehydration. Of course, tweets and accounts can be deleted. Thus, while the use of Tweet IDs and rehydration respects the users' 'right to be forgotten', it reduces the reproducibility of research findings. An alternative to sharing Tweet IDs is to only share derived data. Of course, while this increases privacy protection, this option somewhat limits the reproducibility of findings based on such data as well as their potential reuse value. Only sharing derived data is the solution employed by the German study, which is described in the extended information on the collection and processing of the data:

> In accordance with the general terms and conditions of Twitter, we will not publish the data or pass it on to third parties. Only features derived from the data without any personal reference may be shared with other scientists under certain circumstances (e.g., which topics you are particularly interested in, how active you are on Twitter). We will never pass on information to third parties by which you can be directly personally identified.

Twitter also limits the number of tweet IDs (and user IDs) that can be shared but makes an exception for academic research:

> Academic researchers are permitted to distribute an unlimited number of Tweet IDs and/or User IDs if they are doing so on behalf of an academic institution and for the sole purpose of non-commercial research. For example, you are permitted to share an unlimited number of Tweet IDs for the purpose of enabling peer review or validation of your research.
>
> (Twitter, 2020)

The openness of Twitter in supporting academic studies is significant, and such allowances demonstrate an understanding of the needs of the research community by addressing issues concerning transparency and replication.

3.2. Case study 2: Facebook data

The second case study is based on the German project described in the previous case study. In the second online survey within that project, respondents who reported having a personal Facebook account were asked to install and use a browser plugin that collects public posts (as well as some metadata on them, such as the number of likes and other reactions they have received) from the users' personal Facebook feeds. Hence, what was collected was not content produced by a user but content by other sources (e.g., media outlets or other organizations) that the user is exposed to. The browser plugin was available for the desktop version of the Chrome and Firefox browsers and could be installed via the official plugin stores. A detailed description of the plugin and its use can be found in Haim and Nienierza (2019).

Public or Private?

Coming back to the hypothetical continuum between private and public for social media data, data from Facebook is more on the private end of this spectrum. While Twitter is generally meant and used for public communication, Facebook is more often used for personal communication. On the technical side, unless a user profile is public - which, unlike Twitter, is not the default case - their status update and profile information can only be seen by their Facebook friends. Although the data that the browser plugin collected - public posts from a user's news feed - can be considered less sensitive than posts made by the users themselves, they are private in the sense that only the users can access their personal Facebook news feed.

Specifics of Informed Consent

Given that people generally consider Facebook data to be private and sensitive and, because the installation and use of the browser plugin required more effort than the provision of the Twitter handle in the first survey, the consent statement for the Facebook data was a bit more detailed:

> For many people, Facebook is an important source of information. As you probably know, the display of news items on Facebook is highly personalized. Since Facebook provides virtually no information about this, it is unclear how this selection is made.
>
> As independent scientific researchers, we are interested in how the personalized display of messages on Facebook works. To this end, we cooperate with researchers who have developed a browser plugin (for Firefox and Chrome) that collects public posts in the news feed of individual users. We would like to link the data we already have from the survey and web tracking with data on the public posts in your Facebook news feed.
>
> Would you be willing to install this browser plugin?
>
> The plugin only records posts from your news feed that have actually been publicly shared on Facebook and can, therefore, be seen by any Facebook user. Private posts, such as status updates from friends or private messages, are not recorded. Login codes and passwords are also not recorded. In addition, you can view the data collected from your news feed at any time and delete it if necessary. You can find more detailed information on data protection for the browser plugin here [link to a website with information].

Similar to the consent statement in the online survey, the information presented on the linked website was also a bit more extensive (see Appendix B).

Unique Identifiers

As Facebook user names are not unique (in most cases, people use their real names for their Facebook profiles) and because the data were not collected via the Facebook API (see the following section on this issue), a different unique identifier was needed to link the Facebook data with the survey responses. For that reason, participants were asked to generate a six-digit code in the survey: first letter of mother's first name, first letter of father's first name, first letter of own first name, day from date of birth, last letter of own hair colour, last letter of own eye colour. To create the link,

participants had to enter the code again as part of the installation process for the browser plugin.

Data Access

As described at the beginning of this subsection, a browser plugin that the participants had to install was used to collect the Facebook data. The plugin only collects current data, so it is not possible to access historical data. Users can also deactivate the plugin and delete data that has been collected with the browser plugin.

The use of the browser plugin was necessary in this study for two reasons: 1) it is the only way to directly capture exposure to content on Facebook via the news feed, and 2) data access via the Facebook API has essentially become unavailable to academic researchers as a consequence of the Cambridge Analytica scandal. As platform providers can substantially alter or even completely close APIs at any time, some researchers have argued that research with social media data may be facing an 'APIcalypse' (Bruns, 2019) or entering a 'post-API age' (Freelon, 2018). Asking users to install and use a browser plugin to collect Facebook data is one way of partnering with users to address this issue (see Halavais, 2019). Another option is a data donation model in which users export parts of their personal Facebook data archives and share them with researchers (see Thorson et al., 2019 for an example). Mancosu and Vegetti (2020) have also suggested a web scraping routine for collecting public Facebook data.

Rights to the Data

While privacy is less of a concern for public Facebook posts that cannot be directly associated with the user in whose news feed they appeared, a legal issue that needs to be considered for these data is copyright. As many of the public posts in users' news feeds come from media outlets or companies, many of them are protected by copyright.

Data Sharing

The fact that many of the captured posts are likely protected by copyright means that the full raw data cannot be easily shared. To increase data privacy, the survey data should only be linked with data derived from the posts, such as counts of different types of posts. While the users from whose feeds the posts were collected cannot be directly identified from these data, the issue of copyright, as well as the fact that identification of users cannot be ruled out completely, means that the raw data cannot be shared freely. For those reasons, the part on data access and sharing in the extended information document read as follows:

> The anonymized (aggregated) linked data, which includes your survey responses and web tracking data as well as information on public posts from your Facebook news feed, is used for scientific purposes only. Commercial use of the data is excluded. Access for third parties to the complete linked data will only be possible in a special secure environment.

3.3. Case study 3: LinkedIn data

The study on consent to link survey responses and LinkedIn data is being conducted during the fourteenth wave of the Understanding Society Innovation Panel (IP). To the best of our knowledge,

this is the first survey asking for LinkedIn linkage consent. Understanding Society has a focus on measuring labour market activity, and LinkedIn focuses on employment and businesses, being used largely as a professional networking site. In terms of scale, a recent survey in the UK by regulator Ofcom (2019) found that 16% of UK internet users used LinkedIn; however, its employment focus means users are mostly a subset of the population who are or would like to be economically active. About half of the UK population (based on *Understanding Society* data) is employed, suggesting that LinkedIn coverage of its target population could be higher than Twitter is for the population it targets (~25% of Internet users in the UK use Twitter).

LinkedIn is what Edwards et al. (2013) call *punctiform* – '[it] capture[s] the structure of social relations at particular moments and [is] therefore 'punctiform' in providing a snapshot of these relations.' Interestingly, Edwards et al. (2013) originally classified all social media as locomotive, and defined social media data, by definition, as *not* being punctiform; but, when comparing the information turnover and purpose of LinkedIn with a microblogging platform, such as Twitter, it is, indeed, quite static by comparison. LinkedIn, as a biographical profile site, does fit the description of being a snapshot of a user's career status.

Public or Private?

On the private-public spectrum, LinkedIn is perhaps the most public of all social media sites. The main purpose of the site is to network professionally, including looking for new business and employment opportunities. Having a private profile would naturally limit that objective. Moreover, this public nature of the profile has been recognized on a legal basis. Recent US litigation determined that such scraping was indeed legal (Woollcott, 2019; also see Mancosu & Vegetti, 2020), partly based on the understanding that LinkedIn profile data is owned by the users and that user profiles are public for the purpose of being accessed by others.

Specifics of Informed Consent

Given that if the profile is made public, it can be accessed and data scraped directly, the initial need to obtain consent is for ethical considerations. As the LinkedIn project grew out of the UK project on Twitter, the specifics of informed consent are based almost entirely on that project. There was a focus on the same five areas addressed with Twitter and Facebook for informed consent, and the language was similarly based on that developed by Sloan et al. (2020). The main changes were on being more LinkedIn-specific, including a focus on employment and education content. Accordingly, the language for the LinkedIn consent is as follows:

> We would like to know who uses LinkedIn, and how people use it. We are also interested in being able to link the information people have provided for this study to publicly available information from their LinkedIn accounts, such as their employment or education history, their connections, or information about their employer.
>
> Information collected from your LinkedIn account will be treated as confidential and protected in the same way as your interview data. Any LinkedIn information that would allow you to be identified will not be published.

> Are you willing to tell me the name of your personal LinkedIn account and for your LinkedIn information to be linked with the information you have provided for this study?

The additional help text included with this question provides information regarding what is being asked and ensures greater informed consent. This includes information on what data will be collected and why, who will have data access, and data security procedures. Again, this is largely based on the wording developed for the Twitter study. Besides changing the focus to LinkedIn information, the main difference with what was provided when asking to link Twitter data is the inclusion of a statement about GDPR, which came into effect after the UK Twitter study. Full wording for these help links is included in Appendix C.

Unique Identifiers

Equally important when asking consent, however, is the need for additional data to be collected from the respondent to identify the correct LinkedIn profile to link to survey responses. Unlike usernames on Twitter, LinkedIn user IDs are largely not chosen by (and unknown to) individuals. When a user signs up, the site assigns a user ID based on the person's first and last name with an alphanumeric string appended (e.g., first-last-81341b34). These can be customized by users, but many users do not. After obtaining consent to link the data, survey questions can ask for this ID (as would be the case in Twitter linkage), but most respondents will not be able to provide an answer.

Rather, for most respondents, additional questions need to be asked to identify the correct LinkedIn profile from which to scrape data and link to survey responses. These can only viably be asked after consent has been obtained. It is possible to employ programming scripts (written, e.g., in Python or R) to search for profiles automatically using LinkedIn's search functionality. To limit search returns, and to correctly identify the respondent's profile, additional information about the LinkedIn profile needs to be collected. This information needs to include, at a minimum, the name the respondent has on their LinkedIn profile, but more information is needed to limit returns to the most likely matches to the respondent.

Another obvious identifier would be an employer listed on LinkedIn. However, the ability of these two fields to be limiting may be lacking, depending on the uniqueness of the name and employer combination. For example, 'Bill Gates Microsoft' returns only one profile. However, 'Tom Smith Tesco' returns 126 profiles. Additional questions about the profile should therefore be included but should be focused on what is likely included on profiles for most while avoiding overburdening respondents, especially given that all of the information requested is personal identifiers.

An initial set of possible questions are included in the fourteenth wave of the Understanding Society Innovation Panel (IP). In addition to profile ID (if known) and the respondent's name and most recent place of work listed on the profile, consenting respondents are asked for their profile job title, location, and most recent place of education listed.

Data Access

Collecting user data from LinkedIn to link to survey data is not as simple as it is from Twitter, as access to LinkedIn APIs is largely closed to research. However, collecting LinkedIn data also does not

require an additional plugin and user login, as in the case of Facebook. Rather, researchers can collect LinkedIn data directly from the website using established data scraping techniques (Haag, 2020) in programming languages such as Python or R. The set of identifiers provided by respondents is used in the LinkedIn search function, which returns a set of one or more profiles.

Given the lack of easily obtainable unique user identifiers and the need to scrape web pages, there may be multiple returns on search results using the set of identifiers collected after the initial consent question. To make these matches, we propose two methods. The first, deterministic linkage, requires exact matches on identifiers. These can include cases where the respondent knows their LinkedIn ID or where the identifiers provided yield only one return.

However, in some instances, an exact match may not be possible, for example, due to entry errors or where multiple returns exist on the set of identifiers used. Therefore, in these cases, we utilise probabilistic linkage methods that identify likely matches with a quantified level of uncertainty. Probabilistic linkage involves linking data based on statistical techniques that calculate from non-unique identifier sets the likelihood of links between records in each data source being correct, given the other links possible between records (Sayers et al. 2016; Doidge & Harron 2019). For each sample member, the most likely link (determined from linkage weights computed for each considered possible link) should be included in the linked dataset, although a similarity threshold below which 'best' links are considered incorrect is often applied to reduce linkage errors in the dataset. Hence, such methods are particularly useful for linking records when, due to entry errors and other sources of differences (for example, the University of Essex will not match with Essex University when deterministic linkage methods are used), identifiers for given subjects may be mismatched between data sources.

Rights to the Data

Given that the data is being scraped directly from websites and not through LinkedIn's API, considerations regarding the ToS of the API do not need to be factored in. Further, a legal precedent suggests that data on public profiles is open to all. However, that legal case was in the United States and may not hold if challenged in other contexts. Additionally, LinkedIn posts may contain copyrighted material that needs to be considered in data collection and curation.

Data Sharing

Again, unlike the Twitter project, since LinkedIn data is not collected via the API, the situation in regards to data sharing is less clear. Also, unlike the Twitter project, the work on LinkedIn has not focused on plans for archiving or comprehensive data sharing. The focus, rather, has been on the data collection and linkage of LinkedIn and survey data; the amount of work and programming required is non-trivial, in part due to lack of access to the LinkedIn API. Future expansions linking LinkedIn and survey data will place more efforts on ways to ensure efficient data sharing.

However, some data sharing is planned to generate processes and possible next stages for work. As noted above, we explain to respondents who will have access to what data from this process. We note that data from survey answers and LinkedIn information will be made available to researchers if they are able to present a strong scientific case to ensure that the information is used responsibly

and securely. We will also generate summary information from LinkedIn accounts, which would not allow identification and will have the same access controls as survey answers, which will be accessible by other researchers.

4. Platform attributes relevant for informed consent

As the case studies in the previous section have illustrated, social media platforms have specific attributes that are relevant for the formulation of informed consent. Some of these features are the same or similar across platforms, whereas others differ. Based on the case studies we have presented, the key similarities are: All three platforms offer different types of data that vary with respect to their (perceived) privacy and sensitivity, and all of them have complex data structures (whether extracted via API or scraping) that are too complicated to communicate to a lay audience, which means that informed consent will always be a compromise between a simplistic explanation of what the data is and how it will be used versus what the data *actually* are and how they will be *actually* used.

Despite these similarities, as discussed in the previous sections, there also are some clear differences between the platforms. Table 1 presents the differences between the platforms we have considered that need to be taken into account for creating informed consent statements and providing appropriate information to participants. In contrast to the description of the case studies in the previous sections, this table focuses on the platforms and their attributes rather than specific methods of data collection. Hence, while one of the comparison categories (unique identifiers) is the same, the others are different here.

Table 1. Differences between Twitter, Facebook, and LinkedIn that are relevant for the formulation of informed consent

	Twitter	*Facebook*	*LinkedIn*
Private/public	• Twitter is mostly used for public communication • If user accounts are not protected, much of the data is publicly visible	• Facebook data is generally considered more private • Unless users have public profiles (most do not), their activities and full profile information are only visible to logged-in users with whom they are connected	• LinkedIn is used mostly for public professional networking and job search • A US court decided public accounts are public-domain data, as the expectation is access by others
Dynamic nature of the content	• Twitter content is dynamic and changing • It is important to request access to historic and future data to get a fuller picture for individual users	• Facebook content is highly dynamic and changing • Whether researchers can access historical or future data depends on the data collection method	• LinkedIn data is less dynamic and volatile as users build a profile that is reasonably stable • It is not necessary to explicitly ask for historic data from LinkedIn users because the 'live' data is by definition historic
Unique identifiers	• User names are unique and can be used to link the data, but user names can change • User IDs are stable and can be accessed via the API with a list of usernames	• While there are user IDs, these are usually not known to users • Other identifiers need to be used to link the data	• A unique alphanumeric ID is assigned by the site, which can be customized • It is unlikely for users to know their LinkedIn ID, so there is a need to rely on other profile identifiers and to employ probabilistic linkage

While we have covered three platforms that differ in several important regards in our case studies, there are many other types of social media data that can be linked with survey data. Some of these types of data have properties with substantial implications for informed consent. To illustrate this, we will briefly discuss two such categories in the following section: aggregated social media data and social media data for figures of public interest.

5. Data from persons of public interest and aggregated data

The focus of the case studies presented in the previous section was on individual-level data for normal users of the platforms. However, beyond those presented in the case studies above and differing in several important regards, there are other types of users and forms of social media data that can be linked with survey data and also have implications for the issue of informed consent.

The first type that we want to discuss here are social media data from figures of public interest or institutions. Such data are often collected in the context of elections. For example, social media data collections for politicians and other relevant public actors (parties, public authorities, etc.) for the German federal elections in 2013 (Kaczmirek and Mayr, 2015) and 2017 (Stier et al., 2018) have been published via the GESIS data archive.[11] As the politicians are figures of public interest, at least when they use their professional social media accounts, it is not necessary to obtain their informed consent. While the data can be considered personal, what is important to also keep in mind in this context is that informed consent is only one of the possible legal bases for processing such data according to GDPR. Another one is a task carried out in the public interest, which is certainly something researchers can claim when studying the social media activities of politicians or other public actors in the context of elections. Also, if the data are generated by institutions, such as public authorities, they are also typically not personal data. These criteria are also important for questions regarding the publication of social media data. For example, the decision flow chart for the publication of Twitter communications by Williams, Burnap, and Sloan (2017) suggests that tweets by organisations and public figures can generally be published.

The second type of data is aggregated social media data from public figures that is published through other means than completed data collections available for download via a repository. The collection of social media data around federal elections in Germany has since been converted into an ongoing project with the GESIS Social Media Monitoring.[12] Instead of providing completed collections for specific elections, this platform offers aggregated data for user-defined periods of time, topics, or types of actors. Importantly, aggregated social media data can also be linked with individual-level survey data. In that case, there would be no one-to-one matching but a one-to-many-linking. Examples could be to link survey data to data on the volume or sentiment of tweets about a specific topic for a certain region and period of time. Of course, if aggregated data is used, it is not possible to gather informed consent for the linking from the individuals whose data was used to create the aggregate values.

A service that is similar to the GESIS Social Media Monitoring in several regards is The Social Web Observatory.[13] The Social Web Observatory is an initiative aiming to help researchers, mainly from the social sciences and digital humanities, to investigate information diffusion in the social web. The project aims to monitor various sources of information, such as websites and the most popular social

media platforms (Facebook, Instagram, Twitter). Users can gather data about different entities, such as politicians or other public actors, by using a wide variety of sources, such as keywords, hashtags, monitoring of websites. The material retrieved through a keyword search can be analyzed based on parameters that allow the extraction of indicators, such as the emergence of trends, emotions, attitudes about a phenomenon, event, or product (Tsekouras et al., 2020). Similar to the GESIS Social Media Monitoring, the data can also be aggregated over different time periods. As part of an informal collaboration between the Clarin: el[14] and SoDaNet[15] infrastructures, members of the EKKE / SoDaNet research team have set up entities to follow the campaign of political parties and candidates for both municipal and national elections in Greece between May and July 2019 by providing information about their official Facebook or/and Twitter accounts, Wikipedia pages, and relevant keywords. Again, similar to the GESIS Social Media Monitoring, users cannot extract raw data from the Social Web Observatory. Instead, processed or aggregated data, such as the number of articles, comments, or tweets or information about the domains containing the articles and comments are provided. Cases in which only aggregated data are used and shared are the second type of social media data collection that does not require informed consent from individuals.

Besides the Social Web Observatory and the GESIS Social Media Monitoring, which are geared towards social scientists, there also are other continuous social media collections. One example of those is *TweetsKB*[16] (Fafalios et al., 2018), which is a "corpus of anonymized data for a large collection of annotated tweets" that includes "metadata information about the tweets as well as extracted entities, sentiments, hashtags and user mentions" (description on the TweetsKB website). All of the services presented here are data sources that can serve as alternatives to data collections via web scraping, APIs, or data donation, as presented in the case studies. While researchers have no direct control over the actual data collection, these services can provide comprehensive data that can also be linked with survey data with the added benefit that the linking, in this case, does not require researchers to obtain informed consent from the individuals whose data are included in these collections.

6. Conclusion

The three case studies discussed in this paper provide examples of how informed consent for social media and survey data linkage can be obtained. However, there are clear differences in what information needs to be given to participants, depending on the platforms in use. Social media platforms are not homogenous in the way that they are used by individuals, the purposes they serve, or the manner in which they are structured and interacted with, both by content creators and the wider public. Accordingly, it is no surprise that it is difficult to provide concrete guidance on informed consent that can be applied to all platforms and types of data. This is further exacerbated by the fact that platforms can change or disappear, and new ones emerge.

However, despite the fact that providing general solutions for informed consent for linking surveys and social media data is not possible, the cases and aspects we have discussed should serve as guiding points for researchers and archivists working with such data. It is worth noting that the informed consent process detailed for the Twitter case study has been adopted and modified for later projects - indicating that there is value in adapting the work of others.

Based on what we presented in the paper, some of the general recommendations for informed consent for linking surveys and social media data are to take into account and address what types of social media data are collected and by what means, how private and sensitive they are, how exactly they will be linked to the survey data, how they are stored and can be accessed, and whether current, future, or historic data are required and collected.

Acknowledgment

The work of Johannes Breuer, Libby Bishop, Dimitra Kondyli, and Apostolos Linardis on this paper was funded by the Consortium of European Social Science Data Archives (CESSDA) as part of the WP2020 project 'New Data Types'. The work of Luke Sloan and Tarek Al Baghal on this paper is associated with the funded ESRC project 'Understanding [Online/Offline] Society: Linking Surveys with Twitter Data' (ES/S015175/1). Johannes Breuer wants to thank Pascal Siegers and Sebastian Stier for their assistance in writing the informed consent (including the extended data privacy information) for the German study which case study 2 and parts of case study 1 in this paper are based on.

References

Araujo, T. *et al.* (2017) 'How much time do you spend online? Understanding and improving the accuracy of self-reported measures of internet use', *Communication Methods and Measures*, 11(3), pp. 173–190. doi: https://doi.org/10.1080/19312458.2017.1317337.

Breuer, J., Bishop, L. and Kinder-Kurlanda, K. (2020) 'The practical and ethical challenges in acquiring and sharing digital trace data: negotiating public-private partnerships', *New Media & Society*, 22(11), pp. 2058–2080. doi: https://doi.org/10.1177/1461444820924622.

Bruns, A. (2019) 'After the "APIcalypse": social media platforms and their fight against critical scholarly research', *Information, Communication & Society*, 22(11), pp. 1544–1566. doi: https://doi.org/10.1080/1369118x.2019.1637447.

Doidge, J. C. and Harron, K. (2018) 'Demystifying probabilistic linkage', *International Journal of Population Data Science*, 3(1). doi: https://doi.org/10.23889/ijpds.v3i1.410.

Edwards, A. *et al.* (2013) 'Digital social research, social media and the sociological imagination: surrogacy, augmentation and re-orientation', *International Journal of Social Research Methodology*, 16(3), pp. 245–260. doi: https://doi.org/10.1080/13645579.2013.774185.

Fafalios, P. *et al.* (2018) 'TweetsKB: A Public and Large-Scale RDF Corpus of Annotated Tweets', in Gangemi, A. et al. (eds) *The Semantic Web*. Cham: Springer International Publishing, pp. 177–190.

Freelon, D. (2018) 'Computational Research in the Post-API Age', *Political Communication*, 35(4), pp. 665–668. doi: https://doi.org/10.1080/10584609.2018.1477506.

German Data Forum (RatSWD) (2020) 'Big data in social, behavioural, and economic sciences: Data access and research data management', *RatSWD Output Paper Series*. doi: https://doi.org/10.17620/02671.52.

Haag, F. (2020). 'LinkedIn Scraping with Python', *Medium*, 28 February. Available at: https://medium.com/federicohaag/linkedin-scraping-with-python-d8d14519602d (Accessed: 28th May 2020)

Haim, M. and Nienierza, A. (2019) 'Computational observation: Challenges and opportunities of automated observation within algorithmically curated media environments using a browser plugin', *Computational Communication Research*, 1(1), pp. 79–102. doi: https://doi.org/10.5117/CCR2019.1.004.HAIM.

Halavais, A. (2019) 'Overcoming terms of service: a proposal for ethical distributed research', *Information, Communication & Society*, 22(11), pp. 1567–1581. doi: https://doi.org/10.1080/1369118X.2019.1627386.

Kaczmirek, L. and Mayr, P. (2015). 'German Bundestag Elections 2013: Twitter Usage by Electoral Candidates'. *GESIS Data Archive, Cologne, ZA5973 Data file Version 1.0.0*. doi: *https://doi.org/10.4232/1.12319*.

Kinder-Kurlanda, K. *et al.* (2017) 'Archiving information from geotagged tweets to promote reproducibility and comparability in social media research', *Big Data & Society*, 4(2), p. 205395171773633. doi: https://doi.org/10.1177/2053951717736336.

Mancosu, M. and Vegetti, F. (2020) 'What You Can Scrape and What Is Right to Scrape: A Proposal for a Tool to Collect Public Facebook Data', *Social Media + Society*, 6(3), Advance online publication. doi: https://doi.org/10.1177/2056305120940703.

Marwick, A. E. and boyd, danah (2011) 'I tweet honestly, I tweet passionately: Twitter users, context collapse, and the imagined audience', *New Media & Society*, 13(1), pp. 114–133. doi: https://doi.org/10.1177/1461444810365313.

Menchen-Trevino, E. (2018). 'Digital trace data and social research: A proactive research ethics' in Foucault Welles, B. and González-Bailón, S. (eds.) *The Oxford Handbook of Networked Communication*. Oxford: Oxford University Press, pp. 519–538.

Prior, M. (2009) 'The immensely inflated news audience: Assessing bias in self-reported news exposure', *Public Opinion Quarterly*, 73(1), pp. 130–143. doi: https://doi.org/10.1093/poq/nfp002.

Sayers, A. *et al.* (2016) 'Probabilistic record linkage', *International Journal of Epidemiology*, 45(3), pp. 954–964. doi: https://doi.org/10.1093/ije/dyv322.

Scharkow, M. (2016) 'The accuracy of self-reported internet use—a validation study using client log data', *Communication Methods and Measures*, 10(1), pp. 13–27. doi: https://doi.org/10.1080/19312458.2015.1118446.

Sloan, L. *et al.* (2020) 'Linking Survey and Twitter Data: Informed Consent, Disclosure, Security, and Archiving', *Journal of Empirical Research on Human Research Ethics*, 15(1–2), pp. 63–76. doi: https://doi.org/10.1177/1556264619853447.

Stier, S *et al.* (2018). 'Social Media Monitoring for the German federal election 2017', *GESIS Data Archive, Cologne, ZA6926 Data file Version 1.0.0*. doi: https://doi.org/10.4232/1.12992.

Stier, S. *et al.* (2020) 'Integrating Survey Data and Digital Trace Data: Key Issues in Developing an Emerging Field', *Social Science Computer Review*, 38(5), pp. 503–516. doi: https://doi.org/10.1177/0894439319843669.

Thorson, K. *et al.* (2019) 'Algorithmic inference, political interest, and exposure to news and politics on Facebook', *Information, Communication & Society*, Advance online publication. doi: https://doi.org/10.1080/1369118x.2019.1642934.

Tsekouras, L. *et al.* (2020) 'Social web observatory: A platform and method for gathering knowledge on entities from different textual sources', in *Proceedings of the 12th language resources and evaluation conference*. Marseille, France: European Language Resources Association, pp. 2000–2008. Available at: https://www.aclweb.org/anthology/2020.lrec-1.246 (Accessed: 12th November 2020).

Twitter (2020) Developer Agreement and Policy. Available at: https://developer.twitter.com/en/developer-terms/agreement-and-policy (Accessed: 12th November 2020).

Williams, M. L., Burnap, P. and Sloan, L. (2017) 'Towards an Ethical Framework for Publishing Twitter Data in Social Research: Taking into Account Users' Views, Online Context and Algorithmic Estimation', *Sociology*, 51(6), pp. 1149–1168. doi: https://doi.org/10.1177/0038038517708140.

Woollacott, E. (2019) 'LinkedIn Data Scraping Ruled Legal.' *Forbes*, 10 September 2019. Available at: https://www.forbes.com/sites/emmawoollacott/2019/09/10/linkedin-data-scraping-ruled-legal/#30bdd8311b54 (Accessed: 26th October2020)

Endnotes

[1] Johannes Breuer is a senior researcher at GESIS – Leibniz Institute for the Social Sciences in Germany and can be reached via email: johannes.breuer@gesis.org

[2] Tarek Al Baghal is Senior Research Fellow and Associate Director of Understanding Society, Questionnaire Design, Essex University UK and can be contacted at talbag@essex.ac.uk

[3] Luke Sloan is Deputy Director of the Social Data Science Lab and Professor at the School of Social Sciences, Cardiff University UK. He can reached via email at SloanLS@cardiff.ac.uk

[4] Libby Bishop the Coordinator for International Data Infrastructures in the Data Archive at GESIS-Leibniz Institute for Social Sciences in Germany and can be reached at elizabethlea.bishop@gesis.org

[5] Dimitra Kondyli is a senior researcher at National Centre for Social Research (EKKE) – Institute of Social Research in Greece and can be reached via email: dkondyli@ekke.gr

[6] Apostolos Linardis is a senior researcher at National Centre for Social Research (EKKE) – Institute of Social Research in Greece and can be reached via email: alinardis@ekke.gr

[7] https://journals.sagepub.com/toc/ssce/38/5

[8] https://www.ukdataservice.ac.uk/manage-data/legal-ethical/gdpr-in-research.aspx

[9] https://medium.com/on-archivy/twitters-developer-policies-for-researchers-archivists-and-librarians-63e9ba0433b2

[10] https://www.ukdataservice.ac.uk/help/faq/deposit.aspx#socialmedia

[11] https://www.gesis.org/en/services/finding-and-accessing-data

[12] http://mediamonitoring.gesis.org/

[13] https://socialwebobservatory.iit.demokritos.gr/#/about

[14] https://www.clarin.gr/en

[15] https://www.sodanet.gr/

[16] https://data.gesis.org/tweetskb/

Appendix A - Website text with extended information on Twitter data

[Project/study name] Data Protection Information: Twitter data

Your Twitter data are collected by

[Name + address of institution]

Below you will find all information about our data collection that is relevant to you. You can contact us at the above address or via the email address [project email address] if you need more information about our research project.

What information is collected about my Twitter account?

We will only collect information about your Twitter account that is publicly available. This includes information about your account (such as your profile description, who you follow and who is following you), the content of your tweets (including text, pictures, videos, and links), and background information about your tweets (e.g. when you tweeted, what kind of device you used for it or - provided you have enabled this feature - the location from where you posted). We will collect information about your past tweets and will regularly update this information with current tweets for the duration of our study.

What is this information used for?

We use the data exclusively for scientific research. Linking your Twitter data with the survey data allows us to better understand your activities on the Internet and your opinions. With additional data from social media we can...

- better understand who uses Twitter and for what purposes.
- investigate whether Twitter contains scientifically relevant information and how good the quality of this information is.
- identify topics that people are concerned about but which are not part of our surveys.
- gather information in addition to that from the survey to capture attitudes and opinions of the population.
- test assumptions about the relationship between the use of social media and political attitudes and behavior.

What do you do to protect my personal information?

All information is stored and used in accordance with the General Data Protection Regulation (GDPR). Since the information from Twitter is publicly available, it is impossible to completely anonymize the collected data. Only information that is no longer personally identifiable (e.g. how often you twitter, how often you address political issues, etc.) is linked to the survey data. In accordance with the general terms and conditions of Twitter, we will not publish the data or pass it on to third parties. Only features derived from the data without any personal reference may be shared with other scientists under certain circumstances (e.g. which topics you are particularly

interested in, how active you are on Twitter). We will never pass on information to third parties by which you can be directly personally identified.

Who will have access to the data?

The anonymized linked data, which includes both your survey responses and your Twitter information, will be used for scientific social research purposes only. Commercial use of the data is excluded. Access to the complete linked data will only be possible in a special secure environment.

Your rights

You can withdraw your consent to the collection of your Twitter data at any time. To do so, just send an email to [email address for the project] or a written letter to

[name + address of the institute]

Please note that your Twitter username must be mentioned in the email or letter, otherwise we cannot correctly assign your data for deletion.

With regard to your personal data, you can make use of the following rights at any time:

- Right of access to information

- Right of rectification

- Right to deletion ("right to be forgotten")

- Right to limit processing

- Right to data transferability

You also have a right of appeal to a data protection supervisory authority.

Contact person

With all general questions and requests concerning data protection at [name of institution] you can contact:

[name + address of data protection officer]

Appendix B - Website text with extended information on Facebook data

[Project/study name] Data Protection Information: Facebook data

Your Facebook data are collected by

[Name + address of institution

Note: The browser plugin used in the study was created and maintained by an external collaborator whose contact details were provided here]

Your data will be transmitted for analysis to

[Name + address of institution running the study/project]

Below you will find all information about our data collection that is relevant to you. You can contact us at the above address or via the email address [project email address] if you need more information about our research project.

What information is collected about my Facebook account?

Only posts from your Facebook news feed that have been publicly shared are collected. Private posts, such as status updates from friends, are not collected. The following data is collected:

- the author of the public post in your news feed,
- date and time when the post was created,
- if applicable, the person or page who publicly shared that post on Facebook,
- contained text, contained image or video file, contained links,
- number of reactions (e.g. likes) and number of comments to the post, and
- position of the post within the news feed.

Personal login information, such as email address, login codes and passwords, are also not collected. Although only public posts from your news feed are collected, we cannot exclude the possibility that the data collected may still contain personal information (for example, if one of your Facebook friends posts publicly and tags you or others in these public posts). We anonymize such information or delete it before the data are analyzed.

What is this information used for?

We use the data exclusively for scientific research. Combining the data on public posts in your Facebook news feed with survey and web tracking data enables us to better understand your activities on the Internet and your opinions. With additional data from Facebook we can...

- better understand who gets exposed to which news on Facebook.
- investigate whether the Facebook news feed contains scientifically relevant information and how good the quality of this information is.
- identify issues that people may be concerned about but which are not part of our surveys
- test assumptions about the relationship between the use of social media and political attitudes and behavior.

What do you do to protect my personal information?

All information is stored and used in accordance with the EU General Data Protection Regulation (EU-GDPR). The collected data are encrypted and transmitted to research servers, all of which are located in Germany. In addition, you have the possibility at any time to view all data collected about you via the page [website for the browser plugin] after entering your personal identification (which you generate yourself in the questionnaire and the browser plugin). Through that website, it is also possible for you to delete your Facebook data. If you do not want the public posts from your Facebook news feed to be collected, you can also deactivate the plugin. By simply clicking on the respective symbol (in the upper right corner of your browser) you can deactivate and activate the plugin. Only information that is no longer personally identifiable is linked to the survey and web tracking data (e.g. how often you have seen news from a particular provider in your Facebook news feed).

Who will have access to the data?

The anonymised (aggregated) linked data, which includes your answers from the survey and web tracking data as well as information on public posts from your Facebook news feed, will only be used for scientific research. Commercial use of the data is excluded. Access for third parties to the complete linked data will only be possible in a special secure environment.

Your rights

You can withdraw your consent to the collection of your Twitter data at any time. To do so, just send an email to [email address for the project] or a written letter to

[name + address of the institute]

Please note that your Twitter username must be mentioned in the email or letter, otherwise we cannot correctly assign your data for deletion.

With regard to your personal data, you can make use of the following rights at any time:

- Right of access to information

- Right of rectification

- Right to deletion (“right to be forgotten”)

- Right to limit processing

- Right to data transferability

You also have a right of appeal to a data protection supervisory authority.

Contact person

With all general questions and requests concerning data protection at [name of institution] you can contact:

[name + address of data protection officer]

Appendix C - LinkedIn additional help links and text

What information will you collect from my LinkedIn account?

We will only collect information from your LinkedIn account that you have made publicly available. This may include information from your profile (for example your work or education history and your connections), the profiles of your connections (such as information about your employer), and posts you have made (including text, images, videos and web links). We will update this information.

This information will be collected and stored for as long as they are useful for research purposes. You can withdraw your consent at any time. If you do so, we will not collect any more of your LinkedIn data and will make no further links. However, previously collected data which has had your identifiers removed will be kept.

What will the information be used for?

The information will be used for social research purposes only. Adding your LinkedIn information and your survey answers will allow researchers from universities, charities and government to better understand your experiences, such as with work and education.

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

* Understand who uses LinkedIn and how they use it

* See what LinkedIn information can tell us about people and their work

* Collect information about things we don't ask in our survey

* Understand what happens between waves of the survey

Who will be able to access the information?

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

Summary information from your LinkedIn account which would not allow you to be identified will have the same access controls as your survey answers. At no point will any information that would allow you to be identified be made available to the public without your express permission

What will you do to keep my information safe?

All information we collect will be held in accordance with current data protection legislation (GDPR).

To keep your information safe, researchers will only be able to access the matched survey answers and detailed LinkedIn 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. Summary information from your LinkedIn account which you cannot be identified from will have the same level of protection as your other survey answers.