CESSDA Work Plan 2020

WPT New Data Types

D4: Archiving Social Media Data

A guide for archivists and researchers

Document info

Dissemination Level	PU
Due Date of Deliverable	30/11/2020
Actual Submission Date	18/11/2020
Type	Report
Approval Status	Approved by Training Working Group Leader Irena Vipavc Brvar and Tools & Services Working Group Leader Mari Kleemola
Version	V1.2
Number of Pages	38 (incl. appendices; 34 without appendices)
DOI	10.5281/zenodo.5041072

Version history

Version	Date	Comment	Revised by
0.1	03/09/2020	First GESIS-internal draft created	Kerrin Borschewski
0.2	11/09/2020	GESIS-internal draft edited & extended	Johannes Breuer
0.3	19/10/2020	New structure for draft & shared with all partners	Johannes Breuer
0.4	26/10/2020	Structured text for first sections produced	Johannes Breuer, Kerrin Borschewski, Martin Vávra
0.5	30/10/2020	Sections added & extended	All authors
0.6	09/11/2020	Co-author comments addressed	All authors
0.7	11/11/2020	Document formatted and sent to reviewers	Johannes Breuer
0.8	17/11/2020	Review comments by Sara D. Thomson addressed	Johannes Breuer
0.9	18/11/2020	Review comments by Oliver Watteler addressed	Johannes Breuer
1.0	18/11/2020	Submitted to CMO	Johannes Breuer
1.1	26/03/2021	Addressed review comments by WGL & CESSDA MO	Johannes Breuer
1.2	22/06/2021	Minor edits based on feedback from WGL & CESSDA MO	Johannes Breuer

Author List

Organisation	Name	Contact information
GESIS - Leibniz-Institute for the Social Sciences	Johannes Breuer	johannes.breuer@gesis.org
GESIS - Leibniz-Institute for the Social Sciences	Kerrin Borschewski	kerrin.borschewski@gesis.org
The Czech Social Science Data Archive (CSDA)	Martin Vávra	martin.vavra@soc.cas.cz
GESIS - Leibniz-Institute for the Social Sciences	Libby Bishop	ElizabethLea.Bishop@gesis.org
University of Ljubljana, Social Science Data Archive (ADP)	Janez Štebe	Janez.Stebe@fdv.uni-lj.si
Slovak Archive of Social Data (SASD)	Katarina Strapcova	katarina.strapcova@savba.sk
TARKI Data Archive	Péter Hegedűs	hegedus@tarki.hu

Peer-review

Organisation	Name	Contact information
University of Edinburgh	Sara D. Thomson	sthoms13@exseed.ed.ac.uk
GESIS - Leibniz-Institute for the Social Sciences	Oliver Watteler	Oliver.Watteler@gesis.org

Contents

List of Tables

Executive Summary

Social media data are increasingly used in the social sciences. Archiving social media data is associated with a number of specific practical, legal (data protection, contracts with platforms, and intellectual property), and ethical challenges (privacy, informed consent) that researchers and archivists need to address. In addition, the documentation of social media data also has its own requirements. After introducing what social media data are and why archiving them is important, this guide discusses these different challenges and provides suggestions for addressing them. Based on these general considerations as well as examples of social media data that have been archived by CESSDA service providers (and how the identified challenges have been dealt with in those cases), the guide presents suggestions for metadata elements that need to be developed or extended for the proper documentation of social media data. Taking into account the various challenges in archiving social media data and how they may be addressed, the guide also presents practical recommendations for researchers working with social media data and the development of social media data collection tools with regard to the requirements for archiving them.

Abbreviations and Acronyms

API	Application Programming Interface
CESSDA	Consortium of European Social Science Data Archives
CV	Controlled Vocabulary
DDI	Data Documentation Initiative
DDI-C	DDI Codebook
DDI-CDI	DDI - Cross Domain Integration
DDI-L	DDI Lifecycle
DMEG	Data Management Expert Guide
DMP	Data Management Plan
EU	European Union
FAIR	FAIR principles – Findable, Accessible, Interoperable, Reusable
GDPR	General Data Protection Regulation
GESIS	GESIS - Leibniz-Institute for the Social Sciences
JSON	JavaScript Object Notation
SERISS	Synergies for Europe's Research Infrastructures in the Social Sciences
ToS	Terms of Service
UKDS	UK Data Service

1. Introduction

Using social media data for research has become increasingly popular in the social sciences as indicated, for example, by an increasing number of publications based on such data or the rapid growth of the field of computational social science that is characterized by the use of digital trace data of which social media data are a major part. The characteristics of social media data as well as the fact that they represent a new type of data for most social scientists mean that researchers are facing a set of challenges when it comes to working with these data. These challenges are present in every phase of the research data lifecycle: From the planning to the collection and analysis to the archiving and sharing phase. For the same reasons that social media data present challenges for researchers, archives that want to preserve these data and make them available to researchers also face challenges that they need to tackle in this process. These relate to practical, ethical, and legal issues as well as to questions of documentation.

The aim of this guide is to identify the challenges related to archiving[1] social media data, discuss examples of how these were addressed in cases in which social media data have been archived at institutions that are part of CESSDA, and present suggestions for solutions based on existing examples and experiences from the archives. The target audience for this guide are archivists and researchers who want to archive social media data. While this guide covers a variety of challenges in archiving social media data, the focus will specifically be on the issue of documentation, especially with regard to the identification and provision of appropriate metadata. Social media data differ from other data that are commonly offered by social science data archives in several important regards. Hence, it is necessary to find solutions for properly processing and documenting them to make them findable, accessible, interoperable, and reusable (FAIR). After introducing what social media data are, and why they should be archived in the first section[2], this guide discusses the practical, legal, ethical, and documentation challenges related to archiving social media data in section 2. The following section presents examples of archived social media data to illustrate how some of the identified challenges have been addressed by archives in practice. The fourth part focuses on the topic of documentation and presents suggestions for specific metadata elements for social media data.[3] This is followed by practical recommendations for researchers and data collection tools for social media (section 5). At the end of each of the substantive (sub)sections, there

[1] In general, the archiving of research data has three parts: preservation, documentation, and publication. As preservation is a more technical subject, the focus of this document will be on the aspects of documentation (especially also with regard to metadata) and publication (or sharing/making the data available). Nevertheless, when the term archiving is used in this document, unless noted otherwise, this includes all three dimensions: preserving, documenting, and publishing the data.

[2] For those who have little or no experience with metadata, Appendix A provides a short explanation of what metadata is and why the documentation of research is important (including some pointers to existing resources).

[3] Appendix B also provides a short general description of how social media documentation can be represented in DDI Lifecycle (DDI-L) and DDI - Cross Domain Integration (DDI-CDI).

is a summary of a few of the key points in the sense of actionable items or central take-home messages. The guide closes with a brief conclusion and outlook.

1.1 What are social media data and why are they interesting for the social sciences?

There are many definitions of social media. While reviewing and comparing them is beyond the scope of this guide, key defining attributes of social media are that they are online platforms that enable or facilitate interactions between users and allow users to produce and share one or more types of content. Accordingly, a broad and simple definition of social media is that they are web-based platforms that facilitate interactions between users and include or are built on user-generated content which can be of various types. What is also important to keep in mind for the definition of social media data is that the emergence of new platforms and the disappearance of others, as well as technological developments in general, have a substantial impact on what people understand as social media. What most people, including researchers, commonly refer to when they use the term social media are social networking sites (SNS), such as *Facebook* or *LinkedIn*, video platforms that are built on user-generated content, such as *YouTube* or *Twitch*, or platforms that are focused on public many-to-many communication, such as *Twitter* or *reddit* [4]. As these examples already show, social media platforms can be very different. This also means that the data they produce or provide are quite diverse. Hence, social media data can come in a variety of formats. Common types of social media data include textual data (posts, comments, etc.), audio-visual data (images, audio, videos), and network data (connections between users or content elements), but there are other types of social media data as well, such as user profile data (e.g., the number of contacts, certain indicators of activity, or profile information).

Social media data are "a valuable resource for researchers and an important cultural record of life in the 21st century" (Thomson, 2016, p. 1). In particular, social media data are interesting for research in the social sciences for a number of reasons. One is the growing popularity and relevance of these platforms. User numbers are continuing to grow, and for many of the users, social media platforms are increasingly important to stay informed and connected with others. Many topics that social scientists study are happening on or at least directly influenced by social media, such as political communication, social activism, social interactions in various groups, the search for information, or the formation and expression of opinions. The variety of social media platforms and the data they generate allow social scientists to study novel research questions or to find new ways of answering existing ones. Another benefit of social media data is that they can be collected unobtrusively. Compared to self-reported data from surveys, they are less prone to being influenced by social desirability. Also, social media data allow researchers to capture behaviour right when it happens, whereas

[4] Twitter has also often been called a microblogging platform but this term is not that commonly used in the literature anymore and it is at least debatable whether it can also be applied to reddit which rather resembles a message board or discussion forum.

surveys can only ask in retrospect. Hence, social media data are especially suitable for capturing immediate reactions, for example, to impactful societal events.

Furthermore, depending on how they are obtained, social media data can be much cheaper and also faster to acquire. As the variety of formats of social media data described above shows, social media data are generally more versatile than survey data in the sense that they can provide information that cannot be collected via surveys (or at least not easily and/or reliably). In addition, social media data can provide much larger samples than survey data, although there are various sampling biases (Sen et al., 2019).[5] Finally, compared to experimental studies, social media data are produced in a natural setting. This is the reason why these data are also often subsumed under the category of found data which differ from designed data (e.g., from surveys or experiments) as they were not produced for research in the data collection process. Importantly, in social science research, social media data do not have to be used as a substitute for survey data but can be used as an addition and be connected to survey data (Stier et al., 2020). Taken together, these attributes make social media data quite attractive for social science research which explains their increased use over the last few years.

Key points

- The social media landscape is constantly changing, and the variety of platforms is mirrored by the variety of social media data types.
- Given the dynamics and diversity of social media, it is difficult to provide a generalizable and maintainable definition of social media.
- Social media data have several attributes that make them interesting for social scientists, such as their volume, variety, availability, and immediacy.
- At the same time, social media data also have certain limitations, such as specific sampling biases and disproportionate demographic representation (see, e.g., Sloan, 2017), which can be addressed by combining them with survey data.

1.2 What is the need for archiving social media data?

Most of the reasons why social media data should be archived are the same as for any other kind of research data: to increase the transparency and reproducibility of research. There are, however, some additional reasons why sharing and archiving social media data is important (see Weller & Kinder-Kurlanda, 2016). One important factor is the potential public and historical value of social media data (for a few examples, see the "Documenting the Now" initiative[6]). If researchers, for example, collect social media data on social movements (such as Black Lives Matter), impactful events (such as national elections), or global crises (such as

[5] Of course, social media data also have other (potential) limitations, such as the lack of information about individuals, e.g., regarding their sociodemographic characteristics or attitudes (see Stier et al., 2020).

[6] Home | DocNow Tweet Catalog, https://catalog.docnow.io/ (date of access: 06/11/2020).

the COVID-19 pandemic), sharing and archiving them is an important endeavour as it can enable future (then potentially historical) research as well as, for example, reuse for journalistic reporting. In such cases, sharing and archiving the data can be seen as being in the public interest. Another reason sharing and archiving is vitally important for social media data is the issue of inequalities in data access.

There are various ways in which researchers can access social media data (see Breuer et al, 2020). In general, researchers can…

- collect the data themselves,
- directly cooperate with platform providers,
- purchase the data from third parties (typically at corporate rates) or
- partner with platform users to collect donations of their own data exported from a platform (see Breuer et al., 2020).

As these data acquisition methods require different resources in terms of funds, contacts, or (computational) skills, they are not equally available to all researchers. These inequalities in data access can create a division among researchers into "the Big Data rich and the Big Data poor" (boyd & Crawford, 2012, p. 674). For example, it may well be that social media companies are more likely to establish direct collaborations only or preferably with (researchers from) more prestigious institutions. Among researchers who collect social media data themselves, the most common method is the use of Application Programming Interfaces (APIs) provided by the social media platforms. However, as these are provided and controlled by the platform, the type and volume of the data that can be collected through them can change substantially. A well-known example is the drastic reduction of the access to and functionalities of the *Facebook* Graph API in the wake of the Cambridge Analytics scandal which essentially means that most of the data from *Facebook* are not available to academic researchers anymore. In view of such developments and the general risk that APIs can be fundamentally altered or shut down altogether, some researchers have argued that the (computational) social sciences are facing an "APIcalypse" (Bruns, 2019) or may be entering a "Post-API age" (Freelon, 2018). This risk of losing widely used sources of social media data further highlights the relevance of sharing and archiving social media data for research.[7]

Generally, in an era marked by the spread of disinformation and increasing conflict between third-party commercial platforms and national and international regulators, archiving social media research data is vitally important. Archiving these data within the context of institutional archives is imperative to making progress towards holding social media

[7] The accessibility of social media data also has an impact on the prevalence of their use in research. A prime example is the popularity of Twitter for research in the social sciences. While it is by no means the most widely used platform, it has some attributes that make it especially attractive to researchers, such as its researcher-friendly API, the relatively public nature of tweets, or the brevity of individual posts which make them more manageable units of analysis. In addition, many researchers are also among the heavy users of the platform. This 'twofold popularity' of Twitter among researchers has made it somewhat of a 'model organism' for social media research.

platforms, governments, and corporations accountable. Research and archive institutions, such CESSDA service providers, are governed by stringent professional standards and possess decades of experience working with highly sensitive personal data. These institutions not only ensure the preservation of authentic records of social media data but also, more importantly, facilitate on-going, high quality research that will lead to evidence-based solutions to many of societal issues, such as those of disinformation and its effects or the impact of social movements.

Key points

- In addition to the usual reasons for archiving research, there are some specific additional ones for social media data, most importantly, their (potential) public and historical value and (potential) ability to reduce disparities in data access.
- Harvesting data from an API provided by a platform, one of the most commonly used methods of collecting social media data, is limited by restrictions imposed by social media companies that regulate access and make these data at risk of becoming unavailable in the event a platform decides to substantially change or completely close their services.
- The risk of API access routes becoming unavailable for researchers further increases the importance of social media data archiving.

2. Challenges to archiving social media data

When it comes to archiving social media, there are a number of factors that researchers and archivists need to keep in mind and address. These include practical, legal, and ethical considerations that determine how social media can be collected, shared, and archived as well as a lack of standardized approaches to documentation. In the following section, the key practical, legal, and ethical issues will be outlined, some general recommendations for how these can be addressed will be provided, and archivists and researchers will be pointed to relevant resources that provide some more detailed guidance on specific challenges. Further, the focus will be more on the topic of documentation and what archivists and researchers need to consider when archiving social media data. As there is a large variety of social media data, it is not possible to provide in-depth guidance for every type. Hence, the suggestions provided in the following sections are general recommendations that should be applicable to most kinds of social media data.[8] Notably, some of the resources that will be referred to in the following provide more specific suggestions for solutions for particular types of social media data (e.g., data from Twitter).

[8] Two useful resources for some discussion of and general guidance regarding the challenges related to archiving social media data are the SERISS "Report on legal and ethical framework and strategies related to access, use, re-use, dissemination and preservation of social media data" (Hagen et al., 2019) and the materials from the CESSDA webinar "Archiving Social Media Data - Challenges and Proposed Solutions" (available at https://zenodo.org/record/3875963#.XuvbWmgzZnI, date of access: 06/11/2020).

Another difficulty, besides the variety of platforms and data types, of striking a balance between broadly applicable recommendations and detailed case-specific guidance is the fact that the different social media data acquisition methods (see previous section and Breuer et al., 2020) have downstream effects also on how the data can be shared and archived. For example, if the data are collected via platform APIs, their Terms of Service (ToS) typically specify how the data can be shared. Similarly, there may be contractual agreements in the case of direct collaborations between researchers and platform providers that regulate how the data can be used. Importantly, one criterion for reusability requires that metadata includes information about the license under which the data can be reused. While relevant ToS passages or other contractual agreements can be described, they make it difficult to fulfil the criterion of using standard reuse licenses. Restrictions and agreed upon specific access conditions can, hence, reduce the reusability of the data. If data are acquired through cooperation with social media users (Halavais, 2019), the informed consent used in such studies affects whether or how the data can be shared (see section 2.3 on ethical challenges). Keeping the differences between social media data platforms, data types, and collection methods in mind, in the following sections, some of the key practical, legal, ethical, and documentation challenges in archiving social media data will be discussed and provide some general guidance on how these may be addressed by archivists and researchers.[9]

2.1 Practical challenges

As outlined in the previous sections, social media data can come in a variety of formats, and this is one of the practical challenges for archiving them. "Variety" is, in fact, one of the so-called "three Vs" of big data that apply to social media data as well. The format that social scientists and the archives catering to them are most familiar with is tabular data, in which, typically, one row is one unit of observation, and one column represents one variable. While social media data can be tabular, in many cases, they are not. Even data that can be easily represented in the traditional form of rectangular tabular data, such as data from user profiles, is often not in that format at the point of acquisition. For example, data collected through web scraping are usually unstructured. And while data acquired through APIs are normally structured, they are often not provided in the tabular data formats that social scientists are familiar with. One common data format that APIs deliver is JSON (JavaScript Object Notation). As JSON files essentially contain key-value pairs, they can be converted to tabular data and then exported to CSV (Comma-Separated Values) or other common file formats for tabular data. Nonetheless, this is at least one extra step that researchers or archivists need to take care of if the data should be archived in common formats for tabular data. Importantly, this only applies to numeric or textual data or data that can be represented by numbers and

[9] While this is beyond the scope of this guide, in order to better structure and facilitate the process of archiving social media data once this becomes a more common practice, there is a need to develop standard workflows and distributions of tasks and responsibilities between researchers and archivists for this area in the future.

characters (such as network data). When it comes to audio-visual data, the issue of formats (but also of volume, see below) becomes even more pronounced. However, as social scientists still only very rarely work with large-scale image or video data, and storing this data would require infrastructures that differ substantially from those currently in place at social science data archives, the focus in this section is on numeric and textual data that can also be represented in tables.

Besides variety, the other two Vs of social media data are volume and velocity. Volume refers to the size of the data. Similar to the types of social media data, their size can also vary substantially. While social media data can be small, e.g., if they are collected for qualitative research, they can exceed the size even of large cumulated survey data sets by several orders of magnitude. This refers to their dimensions in terms of rows and columns (cases and variables) as well as their file sizes. Very large survey data sets can be in the triple-digit megabyte range, whereas social media data can be in the triple-digit gigabyte or even larger (terabyte sizes). Again, this difference can become even bigger when audio-visual data (images, audio, video) are included/used.

The third V, velocity, describes the speed at which the data are generated and modified. First of all, the velocity of the data also affects the volume. For example, collecting even just a *sample* of posts, other content or user activities for one or two months normally generates much more data points than collecting survey data for the same amount of time (even with a large sample with a high response rate). In addition, the data may change quickly and repeatedly. Users may, e.g., edit or delete their posts, other content they produce or their profiles. Hence, for repeated or continuous collections questions regarding the frequency of updates and versioning of archived data arise. While cumulative data from survey programs is typically only extended once every few months or even years, depending on the number of users and how active they are, social media data can potentially be updated and extended within seconds. This shows that, unlike for survey data, truly continuous data collections are possible for social media data. As social science data archives typically deal with completed collections or at least data collection programs that are not updated and extended with such high frequency, making such data available requires new solutions for data storage and data access.

With regard to the volume and velocity as well as the sensitivity of social media data and constraints of using and sharing them by platform ToS or other contractual agreements/obligations, the traditional way of storing completed data sets and making them available for download (potentially with access restrictions in place) is not always feasible. Hence, there is a need for new models of data access. Van Atteveldt et al. (2020) provide some suggestions for such models. These include the publication of parts of a data collection as validation data sets, secure on-site or remote access, the publication of only metadata, and non-consumptive access or remote execution. For some of those, there already are solutions in place at archives for other types of data that can, potentially, be extended for social media data. An example of this are services for secure on-site access. These, however, do not

necessarily scale well and the expertise that the staff that runs those needs for social media data is different than for survey data. The idea of only publishing metadata highlights the importance of developing documentation standards for social media data (see sections 2.4 and 4 for more on this issue). Secure remote access is an area that many archives are currently actively working on. Of course, the different new models for data access can also be combined. For example, secure remote access can include options for remote execution which means that researchers develop their analyses with a small validation data set or simulated data with similar properties as the actual data, and then send their scripts to be executed on the archive servers, so that the researchers only see the results but not the actual data. Another reason for the development of new data access solutions for social media data is that researchers who work with those are used to different kinds of collecting and working with data. For example, they are usually familiar with the use of APIs, and it would, hence, also facilitate their use of archived social media data if archives also offered them to access the data they hold. Some of those data access options for social media data are currently being developed and implemented at the Social Media Archive (SOMAR) at the Inter-university Consortium for Political and Social Research (ICPSR) (Hemphill et al., 2018; 2019), and some CESSDA archives are also working on similar solutions. Of course, the provision of such new ways of data access requires time and resources (staff with the required expertise & the technical infrastructure).

Key points

- The "three Vs of big data" also apply to social media data: volume, variety, and velocity.
- The volume of social media data affects the size of data collections. These can be substantially larger than survey data, in terms of both file size as well as dimensions (number of cases and variables or rows and columns).
- The variety of social media data is a challenge for processing and documenting them.
- The velocity of social media data, or the speed at which they are generated, is much higher than for survey data. This greater velocity raises questions about collecting, updating, and versioning of data.
- These attributes of social media data as well as their potential sensitivity require new solutions for archiving and sharing, such as secure remote access

2.2 Legal challenges[10]

Several legal areas are relevant for archiving social media data: data protection for personal data, contractual agreements (e.g., with platform providers), and intellectual property rights, such as copyright and database rights (see RatSWD, 2020; Watteler, 2020).

[10] Disclaimer: None of the authors of this guide are legal practitioners, so the information presented in this section does not represent legal advice. Instead, the information provided here is meant to make archivists and researchers aware of some of the key legal questions that need to be taken into account for archiving social media data. To fully clarify specific legal questions, archivists and researchers should consult with lawyers.

2.2.1 Data Protection (GDPR)

In the EU, the General Data Protection Regulation (GDPR) and its national implementations are essential for dealing with personal data in archives, meaning that archive staff responsible for processing social media data should take into account the six privacy principles laid out in article 5 of the regulation, such as the lawfulness, fairness and transparency of data processing. Data protection relates to identified or identifiable natural persons, and there generally needs to be a legal basis for the processing of personal data. Out of the six lawful reasons the GDPR lists in Article 6 'informed consent' is usually the most important one used in the social sciences, for example, when conducting surveys. 'Informed consent' means the freely given, specific, informed and unambiguous indication of an individual's willingness to participate in a study. Studies that collect data for a large number of users, for example, by scraping the data or accessing them through platform APIs generally lack this kind of consent. And for such studies gathering informed consent after the fact may not be feasible. However, in such cases, the lack of explicit informed consent does not mean that the data cannot be shared and archived.

The GDPR also provides other options which allow for gathering and processing personal data without proper informed consent. One of these options is carrying out the data collection as a task in the public interest (Article 6 (1) f). Academic research can be based on this, but researchers need to consider that, in these cases, the rights of the individuals behind the data have to be weighed against the public interest. This is called the 'principle of proportionality' and essentially means that researchers cannot simply overrule or ignore fundamental rights. Researchers and archivists should also be aware of the fact that there are different implementations and legal interpretations of this exemption for research in individual EU countries.

Another relevant issue is that effective anonymization is not possible for social media data in many cases. What 'anonymization' means here is the reduction of information in the data that makes it close to impossible to reidentify the individual behind the data. For social media data "classical" anonymization algorithms invented for other types of data often do not work (Domingo-Ferrer, 2019).

2.2.2 Contracts with platforms - Terms of Service

What makes social media data special and different from, e.g., survey data with respect to legal questions is that, in addition to the interests of the scientific community (transparency, reproducibility, openness) and the individuals whose data are being used (privacy & data protection), archivists and researchers have to also take into account the interests of the commercial companies that own and operate social media platforms. Importantly, these interests may be at odds with those of the scientific community when it comes to data sharing and archiving (Breuer et al., 2020). The interests of the platform providers are reflected in the ToS of the platforms and their APIs, which typically also include sections on how their data

can be used (for an example of understanding the Twitter ToS in the work of archivists and researchers, see Littman, 2019). The fact that ToS vary significantly between social media platforms and are changing over time for individual platforms makes the situation more complicated.[11] Notably, although the question to what degree academic research is or can be bound by ToS of platforms or their APIs is an ongoing legal debate in many countries (and researchers and archivists have different views on this). In the cases in which social media data have been archived, researchers and archives have generally aimed for complying with ToS as much as possible. However, as this is still an emerging area of practice, recent and future court decisions[12] as well as possible changes in legislation may change this (Mancosu & Vegetti, 2020).

2.2.3 Intellectual Property Rights (copyright and database rights)

Social media users create content online and frequently also integrate materials of third parties in their messages, posts, tweets etc. Since, according to the Wikipedia definition, "copyright is a type of intellectual property that gives its owner the exclusive right to make copies of a creative work, usually for a limited time"[13] intellectual property rights might be relevant for social media data archiving. One possible way for archivists to deal with issues of copyright is to remove parts of the data (e.g., if they include copyrighted material). Unlike textual data, images or videos, there are also types of social media data for which copyright is not an issue. These include metadata like IDs or specific data types where it is hardly feasible to claim intellectual property rights, such as hashtags.

Importantly, databases can also be protected by copyright. Database rights (recognized as a special form of copyright law in the EU) recognize specific intellectual contributions needed for the design and creation of databases. In many cases, researchers use APIs for data collection, and these typically specify how platform databases can be accessed. This situation is different when web scraping is used to collect data as the legal implications of this are still being discussed and need to be clarified by further court decisions (RatSWD, 2020).

Overall, as there are quite a few legal questions with regard to archiving social media data, archives need to develop/extend legal expertise in this area (use of social media data) by educating their own staff and/or through making use of external legal consulting.

[11] E.g. Facebook changed ToS and restricted access to its API after the Cambridge Analytica scandal (Mancosu & Vegetti, 2020).
[12] See, e.g. recent US court decision in a case initiated by a group of researchers that violating ToS is not illegal in all cases: https://arstechnica.com/tech-policy/2020/03/court-violating-a-sites-terms-of-service-isnt-criminal-hacking/
[13] Wikipedia entry for 'Copyright', https://en.wikipedia.org/wiki/Copyright (date of access: 18/11/2020).

Key points

- GDPR does NOT prohibit archiving data; even personal data can be archived with particular technical and organizational means, such as access control, in place to protect the data.
- A comprehensive list of privacy and data protection legislation in countries around the world is provided by Kruse and Thestrup (2017).
- In order to be able to identify and deal with (potential) legal issues, archives need to assess the legal basis of the collected social media data and the rights to the data before they ingest them. In the process of archiving the data, depending on the type and nature of the data, archives may need to remove parts of the data (e.g., if they include copyrighted materials) and/or restrict access to the data (Watteler, 2020).

2.3 Ethical challenges

There are many ethical aspects to consider for the curation and sharing of social media data Two key concerns that will be the focus in this section are privacy and informed consent.[14] This focus follows some of the legal aspects discussed in the previous section.[15]

2.3.1 Privacy

For social media, the ambiguity related to privacy begins at the source: There is little agreement as to whether social media spaces are public, making data more readily available for research, or if are they private, making data available only with restrictions, or not at all (Markham & Buchanan, 2012; Webb et al., 2017). Given this lack of consensus, a precautionary principle is prudent: Even where a forum is technically open, if people expect it to be private and use it that way, then their privacy expectations need to be considered (though this may not rule out using the data with precautions).

[14] As the previous section should have shown, informed consent is also a legal issue as it is closely related to issues around personal data and data protection. Hence, it would also have been possible to cover questions related to informed consent under the legal challenges. However, it was decided to discuss informed consent primarily as an ethical challenge as ethical considerations regarding informed consent typically go beyond legal considerations. Put simply, the legal framework determines what you can and cannot do with the data, whereas ethical guidelines and recommendations provide guidance on what you should or should not do with the data.

[15] Of course, there are also other ethical questions that may need to be considered for the archiving of social media data, depending on the type of data and research questions. Generally, the core purpose of any ethical review is evaluating the researchers' actions will impact another person/group and whether that impact might be adverse, unwelcome, or harmful. What makes this challenging, especially for social media data, is that this has to be weighed against potential public interest in the research and the data. One critical issue that can be quite important for studies using social media data is the question whether the users who created the data represent a marginalised or otherwise vulnerable group in which case it's important, for some research approaches, to consider if it's appropriate to use the data or make conclusions about it without the input or collaboration of the communities in question. A helpful resource for such cases is the white paper by the *Documenting the Now* initiative on "Considerations for Archiving Social Media Content Generated by Contemporary Social Movements", https://www.docnow.io/docs/docnow-whitepaper-2018.pdf (date of access:16/11/2020).

With traditional data, privacy can often be protected by de-identification, e.g., removing names, IDs, or unusual occupations. Sometimes similar tools can be applied to social media data, but often such data are more difficult to anonymise. As with consent (see the following subsection), sometimes the problem is simply large scale. Other obstacles are even greater, such as Twitter's requirement that content should be published unaltered and with attribution, making it relatively simple to search such content and possibly identify individuals (see Townsend & Wallace, 2016). Equally, it may be undesirable from a research perspective to modify data, as doing so may destroy its value for research.

In general, the debate about anonymisation of any data is shifting away from a binary framing (un/anonymised) to a recognition that because disclosure risk cannot be eliminated, it must be mitigated.[16] Publishing and sharing social media data can still benefit from anonymisation strategies, meaning the reduction of information by means such as aggregation. But often additional measures will also be needed, such as systematic approaches to risk assessment (e.g. disclosure risk reviews and Data Privacy Impact Assessments) and more reliance on data access controls, as the examples in the next section demonstrate.

2.3.2 Informed consent

Unlike survey data or data from experiments, social media data occur "naturally" in that they are not produced for research.[17] Depending on how the data were acquired, the users whose data are used for research and later shared and archived may not have given explicit informed consent to this use of their data. In some cases, platforms' ToS may permit, or imply, that third-party reuse is unrestricted. However, given the fact that users often do not read, or do not understand ToS, it cannot be (ethically) claimed that such acceptance constitutes informed consent, especially when data are reused for purposes different to the original. This relates to the issue of user awareness: Even if they read the ToS, users may not be aware that their data can be used for other purposes, such as research, or even the fact what value the data they produce can have for research.

While in some situations, consent is, indeed, problematic to acquire, in others, it can be implemented. A scenario in which this is relatively easy to achieve is when social media data are linked with survey data on an individual level (Stier et al., 2020). In such studies, obtaining informed consent for the collection or use of their social media data can be part of the survey. Sloan et al. (2020) provide a good template for obtaining informed consent for linking Twitter and survey data which can be adapted for other types of social media data as well.

In cases where it is not possible to gather informed consent from the individuals whose data have been collected (e.g., in the case of large-scale data collections via platform APIs), researchers should at least document and explain why this was not possible. There is a

[16] Notably, this is in line with the basis of the GDPR which also is a risk-based approach to data privacy.

[17] Accordingly, social media data are often also subsumed under the category of 'found', 'ready-made' or "process-produced data" (Grenz, 2020).

growing number of good examples with guidance for the ethical collection and archiving of social media data (e.g., Williams et al., 2017). What is important for archives is that the researchers who collected the data can document how they obtained informed consent or why it was not possible to obtain. Ideally, in either case, research with social media data should be approved by an ethics committee or institutional review board, and their approval should also be documented for the archiving process.

Key points

- Researchers should consider carefully if informed consent is feasible for the research.
- The "Short guide on legal and ethical issues for the researcher to consider when using social media for research" (Appendix A) in the Synergies for Europe's Research Infrastructures in the Social Sciences (SERISS) "Report on legal and ethical framework and strategies related to access, use, re-use, dissemination and preservation of social media data" (Hagen et al., 2019) can provide some helpful, practical guidance for researchers.
- The core principles for ethical research (respect, autonomy, beneficence, and justice) remain relevant, regardless of the data type.

2.4 Documentation challenges for social media data

When documenting social media data, some of the challenges that researchers and archivists need to address are the same as for other types of data. These include, e.g., the need for timely documentation as well as the resources (knowledge and time) required for proper documentation. However, there also are some documentation challenges that are specific to or at least more pronounced for social media data. One challenge refers to the collection of the data. When researchers collect data via APIs, e.g., they usually do not know the technical details of the data collection process. Depending on the type of request, some APIs only provide samples of the data, and the sampling algorithms are often black boxes for researchers. Accordingly, researchers using these API functions are not able to fully document the specifics of the data collection process. Related to this issue, it may not be possible for researchers to properly identify the target population of social media users if it is not fully transparent to them how API requests are processed and how data may be sampled by APIs (Kinder-Kurlanda et al., 2017; Thomson, 2016).

Another challenge of the documentation of social media data is that due to the novelty of such data, no specific standards and methods for their documentation exist. Therefore, researchers currently use different methods to document social media data, resulting in varying and inconsistent documentation. Furthermore, the standards commonly used in the social sciences lack elements that allow a detailed description of social media data (e.g., the different stages of the data collection and data processing). It is possible to describe such data superficially (on the study-level) to make data findable, but such documentation likely does not suffice for the reuse of the data. As Kinder-Kurlanda and colleagues (2017) note: "One solution to this issue is to extend the well-established standards for metadata and documentation of social

science survey data to also be applicable to social media data" (p. 5). The Data Documentation Initiative (DDI) has taken a step in that direction with their new specification (DDI - Cross Domain Integration (DDI-CDI)), which is currently under review. DDI-CDI is not specific to survey data but can be used for any data type (for more information see Appendix B).

Another example of special challenges when documenting social media data concerns the versioning of the data.

Data management procedures typically result in several versions of the data file. Version and edition management is one of the most important parts of provenance management (CESSDA Training Team, 2020), essential for reproducibility and trustworthiness[18]. It helps to:

1. Clearly distinguish between individual versions and editions and to keep track of their differences.
2. Prevent unauthorized modification of files and loss of information, thereby preserving data authenticity.

There are some difficulties connected to the application of versioning to social media data as these data sets are frequently/continuously updated which makes it necessary to design data workflow systems helping to track such frequent transformations of the data. The main challenge here is that the content of platforms is continuously changing and if archived data are expected to reflect those changes (e. g. deleting posts from the data set if they were deleted from the platform), the result is an increasing number of versions.

A suggestion by the Research Data Alliance[19] (for identifying the exact version of a subset as it was used during a specific execution of a work or even if the data source is continuously evolving) is that data are stored in a versioned and timestamped manner. Data sets should be identified assigning persistent identifiers (PIDs), thus, enabling timestamped queries that can be re-executed against the timestamped data store (Rauber et al., 2016).

Regarding versioning itself, the RDA working group says that "in large data scenarios, storing all revisions of each record might not be a valid approach. Therefore, in our framework, we define a record to be relevant in terms of reproducibility, if and only if it has been accessed and used in a data set. Thus, high-frequency updates that were not ever read might go - from a data citation perspective - unversioned" (Rauber et al., 2016).

Data Cite also published recommendations for citing rapidly changing data. Generally, there are four possible ways (Data Cite Metadata Working Group, 2017) for citing data sets, given they were stored and documented accordingly:

a) Cite a specific slice or subset (the set of updates to the data set made during a particular period of time or to a particular area of the data set).

[18] One example is the UKDS versioning policy (see "Version control and authenticity", https://www.ukdataservice.ac.uk/manage-data/format/versioning (date of access: 06/11/2020)).
[19] WG Data Citation - Making Dynamic Data citable,https://www.rd-alliance.org/wg-data-citation-making-dynamic-data-citable.html (date of access: 06/11/2020).

b) Cite a specific snapshot (a copy of the entire data set made at a specific time).
c) Cite the continuously updated data set but add an Access Date and Time to the citation.
d) Cite a query, timestamped for re-execution against a versioned database.

Notably, the "slice," "snapshot" and "query" options require unique identifiers. Option (c) necessarily means that following the citation does not result in access to the resource as cited. This limits the reproducibility of the work that uses this form of citation.

Key points

- Even if no specific standards for the documentation of social media data exist yet, researchers and archivists should generally make sure that the documentation of social media data is as detailed as possible.
- One example is that, while existing metadata standards do not include this, researchers using APIs to collect social media data should document the information from the ToS and the version number of the API they used (also see section 4).[20]

3. Examples of archived social media data from the CESSDA archives

Considering the challenges outlined in the previous sections, it is understandable that, so far, not that many social media data sets have been archived at CESSDA archives. A SERISS project survey among European social science data archives was carried out in June 2019 (see Hagen et al., 2019). Among 18 archives that answered only German GESIS (GESIS - Leibniz-Institute for the Social Sciences) and British UKDS (UK Data Service) had some significant collections of social media data at the time.[21] At the rest of the archives that responded, no social media data were deposited at the time of the survey. In the remainder of this section, we will look in more detail at a few of the examples of the social media data archived at GESIS and UKDS to discuss how some of the challenges presented in the previous sections were dealt with in these specific cases.

GESIS has several archived social media data sets in its catalogue. Notably, in some of these data collections, researchers from GESIS were directly involved. This facilitated the actual ingestion and documentation of the data and was especially helpful for early projects that

[20] Given that social media platform functionalities also tend to change, it may also be helpful to provide a copy of or link to the documentation of the most recent release notes about the platform architecture itself, as changes in functionality can have significant impact on interpreting the data in the future.
[21] In Switzerland at FORS (Swiss Centre of Expertise in the Social Sciences) there are few data sets on political communication and political behaviour in the catalogue that mention data from social media. At FSD (Finnish Social Science Data Archive) in Finland there is a qualitative data set containing messages from the Finnish Suunta E-Guidance Service.

served as pilots for the process of archiving social media data. One of those projects is the data set "German Bundestag Elections 2013: Twitter Usage by Electoral Candidates" (Kaczmirek & Mayr, 2015). As the ToS of the Twitter API do not allow the full tweet texts to be shared for this project, the archived data only include a) a list of the candidates, links to their Twitter and Facebook profiles and information some of their key attributes, and b) a list of tweet IDs that can be used to retrieve the tweets via the Twitter API (given that neither the Tweet nor the associated account have been deleted).[22] These data and how they were collected have been described in a detailed methods report (Kaczmirek et al., 2014). The same types of data were collected and archived again for the 2017 federal election in Germany (Stier et al., 2018a); the methodology is documented in Stier et al. (2018b). Here, in addition to the candidate information and the tweet IDs, the data also include "lists of organizations relevant during an election campaign, i.e. political parties and important gatekeepers, along with their respective Twitter and Facebook accounts".[23] Both of these data sets are accessible for academic research and teaching by registered users.

One example of social media data from external projects and researchers is the data set "AUTNES Content Analysis of Party Facebook Pages 2013" which includes "Party Facebook postings during the six weeks of election campaign for the Austrian general election in 2013 (postings for all parties that passed the threshold for entering the parliament in 2013 ...)".[24] The codebook for this study also includes a brief section describing the data collection methodology. Importantly, all of these data sets include social media data from persons or institutions of public interest (mostly politicians and political parties) which makes these data different from social media data collected from regular users with regard to privacy and data protection. By contrast, the data set "Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness" contains "IDs of geotagged Twitter posts from within the United States" collected in 2014 and 2015.[25] The data are organized into many different files (including, e.g., tweet IDs and hashtags per day for U.S. states and counties) and are only available upon request. In addition to these data files, there is also a Python script for retrieving the tweets using the list of tweet IDs; a process called rehydration.[26] As these data are larger, more sensitive, and more complex than the other social media data sets mentioned here, the archiving was more complicated in this case. The full considerations as

[22] German Bundestag Elections 2013: Twitter Usage by Electoral Candidates, https://search.gesis.org/research_data/ZA5973 (date of access: 06/11/2020).
[23] Social Media Monitoring for the German federal election 2017, https://search.gesis.org/research_data/ZA6926 (date of access: 06/11/2020).
[24] AUTNES Content Analysis of Party Facebook Pages 2013, https://search.gesis.org/research_data/ZA6882 (date of access: 06/11/2020).
[25] Geotagged Twitter posts from the United States: A tweet collection to investigate representativeness, https://doi.org/10.7802/1166 (date of access: 06/11/2020).
[26] Python Script to rehydrate Tweets from Tweet IDs, https://doi.org/10.7802/1504 (date of access: 06/11/2020).

well as the steps taken to archive these data are described in full in a paper by Kinder-Kurlanda et al. (2017).

All of the previous data sets are examples of completed collections. As mentioned before, however, given the three Vs of social media data (volume, variety, and velocity), an alternative to completed collections that can be archived as files (and then downloaded) are continuous collections that are made accessible otherwise. The social media data collections for the German federal elections in 2013 and 2017 have been developed into such a continuous collection at GESIS. The *GESIS Social Media Monitoring*[27] that allows users to explore and access aggregated data from the social media data collected for the German General Election 2017, has been and is updated with newer data and will also provide data for the upcoming federal election in 2021. Another example is *TweetsKB* which provides a regularly updated "corpus of anonymized data for a large collection of annotated tweets"[28] (the data collection methodology is described in detail in Fafalios et al., 2018). As in the other examples, *TweetsKB* makes available only the tweet IDs and metadata and annotation data. These two examples illustrate alternative data access models. Instead of or in addition to archived files, they allow access to databases via their own APIs.

Similar to GESIS, the most common kind of social media data archived at UKDS is also data from Twitter. UKDS offers quite a few Twitter data sets that have been deposited by researchers. Again, due to the Twitter ToS the actual tweet texts cannot be archived. Instead, the tweet IDs plus some metadata (e.g., timestamps of the tweets) are stored.[29] In some cases also lists of hashtags or topics are included in the data sets. These data are freely available (or "open" as per UKDS data access policies[30]). Two examples of a UKDS data set containing tweet IDs is the "Brexit Twitter data 2017-2019" (Cram & Llewellyn, 2020) which is based on hashtags related to Brexit and "UK-EU referendum Twitter data" (Cram & Llewellyn, 2020) and consists of tweet IDs collected on the UK-EU referendum between September 2015 and August 2016.

The examples of social media data archived at GESIS and UKDS illustrate some of the challenges identified in the previous sections. However, these examples also show how some of these challenges can be addressed. For example, sharing only the tweet IDs is a common solution with regard to legal restrictions by the Twitter ToS. While this practice is in accordance with the Twitter ToS, it reduces the reusability of the data as well as the reproducibility of

[27] Social Media Monitor, http://mediamonitoring.gesis.org/ (date of access: 06/11/2020).
[28] TweetsKB - A Public and Large-Scale RDF Corpus of Annotated Tweets, https://data.gesis.org/tweetskb/ (date of access: 06/11/2020).
[29] This is a requirement for archiving data via the UKDS ReShare platform that is also laid out in the UKDS FAQ on depositing data: https://www.ukdataservice.ac.uk/help/faq/deposit.aspx#socialmedia (date of access 06/11/2020).
[30] UK Data Service Data access policy, https://www.ukdataservice.ac.uk/get-data/data-access-policy (date of access: 06/11/2020).

previous findings based on these data as tweets and user accounts may be changed or deleted between the original data collection and the so-called rehydration of the tweets via the tweet IDs. To improve data protection, one method that has been employed is the use of access restrictions (e.g., for the geotagged tweets archived at GESIS). Another one is the aggregation of data (as in the case of the GESIS Social Media Monitoring platform).

New modes of data access can also be a suitable way of addressing some of the practical, legal, ethical, and documentation challenges for social media data. *TweetsKB* and the *GESIS Social Media Monitoring* are two examples, but other solutions for sharing social media data are certainly possible. While this is not the focus of this document, we will briefly discuss them (again) in the conclusion and outlook section. Regarding the documentation of social media data, accompanying methodological reports or "data papers" (like the ones available for the GESIS social media data examples mentioned above) are a viable solution for providing relevant information and metadata. However, the fact that the catalogue entries do not capture all of these relevant pieces of information about the social media data and how they were collected shows that there is a need to develop best-practices and standards for the documentation of social media data in the social sciences.

Key points

- While the number of social media data sets archived by CESSDA service providers is still relatively small, some archives, especially GESIS and UKDS have some experience with archiving social media data and have several social media data sets in their holdings.
- The existing examples of archived social media data can be used as templates by researchers and archival staff.
- Given the volume, velocity, and variety of social media data, new modes of data access may be required for continuous collections.

4. Metadata for social media data

Some of the information (or metadata elements) needed for the documentation of social media data are the same as for survey data. Information on metadata elements that are generally needed to document data can be found in the CESSDA Metadata Model (CMM) (Borschewski et al., 2019). However, given the specific characteristics of social media data, documenting them to ensure transparency, reproducibility, and replicability requires many additional pieces of information. Likewise, some of the description elements used for survey data cannot be applied to social media data. Since there are so many different types of social media and hence different types of social media data, it is not possible to have a general solution that is applicable to every (present and future) case. Based on the discussion of the specific attributes of social media data as well as the existing examples of archived social media data, a list with an overview of the most basic elements that can be used to document social media data on the study-level is presented in the Table 1 (note: the table does not include statements about the repeatability of elements or whether they are mandatory).

Notably, the elements listed in Table 1 are pieces of information that are specific to social media data. These should be used in combination with "basic" metadata elements (for example, title, creator, collection dates, etc.). Given the differences between social media platforms as well as the fact that platforms change, the list presented in Table 1 is a non-exhaustive overview of possible information that should be documented.[31] Again, the focus here is on pieces of information that are specific to social media data and cannot be captured in full by existing metadata standards that have been developed for other types of data in the social sciences, such as survey data. Nevertheless, where possible, references are made to existing controlled vocabularies (CVs) from DDI as compatibility with those is essential to ensure interoperability. Importantly, in the "Remarks" column of Table 1, we briefly explain why and in what way an extension of existing metadata standards may be possible or desirable for social media data, even if CVs that could be applied exist.

Table 1: Suggested additional study-level documentation elements for social media data.

Description element/field	How to provide	Remarks
Data Type: Social Media Data	CV or free text	Useful to distinguish between studies/research about social media and studies and studies/research using social media data; if only keywords are used (e.g., social media or Twitter), both of these types of data will be found by search engines, so a field indicating that the data are from social media (not only about them) would be very helpful for searching and filtering data
Content Type	CV or free text	Information if the content of the collected data is in the format text and/or video(s) and/or image(s) and/or audio file(s)
Platform	Free text (e.g., Twitter, Facebook, etc.)	New platforms emerge frequently, therefore, a CV is not practical; researchers may also collect data from multiple platforms
Period the collected data were created	using standard date formats (e.g., DD-MM-YYYY + optional HH:MM)	the timeframe in which the data in the data set were created (meaning, e.g., when the posts were posted on the platform)
Collection method for social media data	CV	List of collection methods: E.g., API, web scraping, cooperation with platform, data purchased, collaboration with users (e.g.,

[31] While only study-level documentation was discussed here, information on the variable-level is also relevant for social media data. These can be provided in codebooks to, e.g., describe information on how data has been processed and what type of data original and derived variables contain. The codebook can also be used to provide methodological details about data (pre)processing which is especially important for textual data.

		data donation), other. The "AutomatedDataExtraction" code value and its categories from the DDI Alliance Controlled Vocabulary for Mode Of Collection[32] can serve as a good starting point here. Importantly, users should be able to select more than one option here.
If collection method = via API or web scraping: Tool(s) used to collect the data	Free text	name and version of tool(s) or software package(s) used to collect the data or for bespoke tools built by the researcher, source code and documentation or link to GitHub or similar code repository
If collection method = via API: Which version of which API was used?	Free text	API version number
If collection method = via API: **API Terms of Services**	Free text	Ideally a link to or copy of the ToS valid for the API when it was used should be provided
Type of units of observation	Free text	While much of the options for social media data could be captured, e.g., with DDI CV Analysis Unit[33] (especially the Media Unit code value and its categories), adding a free text option would allow capturing a wider range of social-media-specific cases, such as hashtags, mentions, or links.
Contractual agreements with third parties regarding (re-)use of the data	Free text	What are the contractual agreements regarding the (re)use of the data with third parties? These could be users, platform providers, data resellers, etc. Ideally, the corresponding documents (informed consent, contracts) should be made available or their relevant points cited or described (e.g., from contractual agreements with companies)
Sampling information	Free text	For survey data, the sampling information is typically captured in a CV[34], but since there are endless different possibilities for the sampling with social media data and the selection of content, free text is the only feasible option.

[32] See the entry for the DDI Alliance Controlled Vocabulary for Mode Of Collection, https://vocabularies.cessda.eu/vocabulary/ModeOfCollection?lang=en (date of access: 09/11/2020).
[33] See the entry for the DDI Alliance Controlled Vocabulary for Analysis Unit in the CESSDA Vocabularies Service, https://vocabularies.cessda.eu/vocabulary/AnalysisUnit?lang=en (date of access: 09/11/2020).
[34] See, e.g., the Sampling Procedure Entry for DDI in the CESSDA Vocabularies Service, https://vocabularies.cessda.eu/vocabulary/SamplingProcedure?lang=en (date of access: 09/11/2020).

As stated above, all of the metadata elements suggested in Table 1 relate to the study-level documentation. Documentation on the variable level is even more difficult for social media data as the data can come in a large variety of formats as different platforms and collection methods yield very different kinds of data.[35] For social media data, it is important to document the collection process of the data set in detail. In addition to the variables provided directly by the platforms and employed collection methods, research with social media data typically also involves the creation and use of derived variables. A common example is the detected sentiment of a text (or parts thereof). While discussing variable-level documentation for social media data is beyond the scope of this document, this example shows that it is important to also share pre-processing code for social media data sets (e.g., detailed information on data cleaning, weighting information, data validation, data analysis, tools and methods used for this, etc.) as well as the code used to collect the data if the researchers collected them themselves (via APIs or web scraping) to document relevant things like search terms or sampling strategies. Additional interesting pieces of information would, e.g., be information on data structure, relations between data, or quality measures. Moreover, in the case of linked survey and social media data, further documentation is required to, e.g., information on the level of linking (individual-level vs. aggregate) or the sampling procedure (see Stier et al., 2020).

Key points

- Given the specific characteristics of social media data, documenting them to ensure transparency, reproducibility, and replicability requires other pieces of information than survey data.
- For social media data, it is important to document the collection process in detail.
- Given the differences between social media platforms as well as the dynamic development of the social media landscape, it is difficult to define standard elements that should (and can) be provided for every type of social media data.
- It is also important to share and document the pre-processing code for social media data sets (e.g. detailed information on data cleaning and transformation as well tools and methods used for this) as well as the code used to collect the data, if the researchers collected them themselves (via APIs or web scraping), to document relevant aspects, such as search terms or applied filters.

[35] One example of "variable-level" documentation is the collection of tweets, considered as a text corpus, annotated with metadata. Exemplary description from Ljubešić et al (2017): "The corpus is structured into individual tweets, together with their metadata. The tweets in the corpus are tokenised, sentence segmented, word normalised, morphosyntactically tagged, lemmatised and annotated with named entities."

5. Practical recommendations for researchers and data collection tools

Since the target audience for this guide is not only archivists but also researchers, some practical recommendations for researchers that can help them in archiving their social media data is provided here. Related to that, suggestions for social media data collection tools regarding how they could facilitate the archiving of social media data are also presented. Many of these tools are also developed and maintained by researchers, so the other recommendations made here and elsewhere in the document should also be informative for the development of these tools. However, this section should also be relevant for archivists as it can be used to provide consulting for researchers who want to archive social media data (for some further guidance for archival staff, see Thomson, 2016).

As with other types of data, the general recommendation is that researchers should document as much as possible which, ideally, should also be as much as necessary to make the data reusable. For example, if researchers collect social media data themselves, they should document which tools they used (incl. version numbers) or, in case they used APIs, which versions of the API they used. To facilitate the assessment of the legal status of the data, researchers should also store a copy of the ToS of the platform and/or API that were valid when they collected the data. The existing archived social media data sets, their archive catalogue entries, codebooks, and accompanying methodological reports or publications can provide guidance and templates for the documentation of social media data by researchers.

The need for detailed documentation also highlights the importance of good data management for social media data. One helpful tool to achieve this is a data management plan (DMP). There are several resources available that aid researchers in creating a DMP. A very detailed one is the CESSDA Data Management Expert Guide (DMEG)[36] which also includes a short section on social media data.[37] Of course, the characteristics of social media data also have implications for the data management practices of the researchers who work with these data (Hemphill et al., 2020). While some of the established data management practices for survey data can also be applied to social media data, others have to be adapted or amended. As described in the previous sections, this, e.g., relates to questions of data storage, versioning, and documentation in general. Given these differences as well as the unique challenges related to archiving social media data, it is important that researchers plan ahead and allocate sufficient resources (by making use of an appropriate DMP).

In cases in which researchers collect social media data themselves (e.g., via APIs or web scraping), the tools they use to collect the data can also be used for documentation purposes.

[36] Data Management Expert Guide - CESSDA TRAINING, https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide (date of access: 06/11/2020).
[37] Resources for social media data - CESSDA TRAINING, https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/7.-Discover/Resources-for-social-media-data (date of access: 06/11/2020).

Many of the available tools for social media data collection already provide some relevant metadata, such as the version of the API that they use. While this may not be included in the data they generate, most tools provide some form of documentation that should contain such information. If the tools are not code-based[38], they should create logs that contain relevant pieces of information, such as search queries, applied filter criteria, etc. Again, many tools do this, but researchers may not always be aware of that. Another functionality that many data collection tools already offer that also facilitates archiving is the export of the data into common tabular data formats, such as CSV. More generally speaking, what would be helpful for researchers would be, if data collection tools would offer options to create some sort of "archiving package". Such a package would include the data in a format that archives commonly accept (e.g., CSV) as well as detailed metadata and additional relevant information (such as API or ToS versions, the original query, query timestamp, etc.).

The possibilities that data collection tools offer for data documentation show that, despite all of the challenges identified in the previous sections, the collection of social media data offers a lot of potential for automating data documentation processes at different stages of the data lifecycle. This is largely due to the use of computational methods for the collection, processing, and analysis of these data. Many tools and software packages offer detailed documentation of their functionalities as well options for generating metadata, both of which can be made use of to facilitate the documentation of social media data that is necessary for archiving them and making them findable, accessible, interoperable and reusable.

Key points

- Researchers should document their social media data and the associated collection process as much and as early on as possible.
- Established or commonly used DMPs may have to be adapted to accommodate the characteristics and requirements of social media data.
- The computational methods used for social media data collection can and should be leveraged to automate some of the data documentation, which can facilitate the archiving of social media data.

6. Conclusion & Outlook

Social media data are increasingly used in the social sciences, and their use holds great potential for answering novel research questions or finding new answers to existing ones. Despite the opportunities that social media data offer for research in the social sciences, working with them entails a set of unique challenges. While these affect all phases of the research data lifecycle, they have a number of specific implications for archiving the data.

[38] Notable, the long-term preservation of programming code is one additional challenge (that cannot be discussed in detail here).

Researchers and archivists who want to archive social media data need to address practical, legal, and ethical questions in the process, and also find or develop solutions for properly documenting social media data. The latter is especially relevant for making these data findable and reusable. Although there are not that many social media data sets archived at CESSDA service providers so far, the available examples can serve to provide some guidance on how some of the practical, legal, ethical, and documentation challenges can be addressed. Given the diversity of social media data as well as rapid technological developments that lead to changed or new types of social media data, it is very difficult to find universal solutions for archiving social media data. In practice, this means that, while documents like this one can provide some general orientation, in most cases, the archiving of social media data requires case-by-case decisions and consultation. For the development of guidelines and standards, finding the right balance between specifications and recommendations that are broadly applicable and ones that can be used as guidance for specific cases is a key task. The need for generalizability versus specificity depends on the type of guidance that should be provided. While consulting researchers on archiving their social media data certainly requires case-specific information and decisions, when it comes to documentation and metadata, there is a need for new standards or the extension of existing standards with social-media-data-specific metadata elements (as for example DDI-CDI) that can be used for all or at least most kinds of social media data. To address this, suggestions for the documentation of social media data and how these may be integrated into or combined with existing metadata standards for survey data have been provided in this guide. While these suggestions need to be further tested and evaluated in practice by archivists and researchers alike, we believe that they can serve as a basis for the development and implementation of documentation and metadata standards for social media data. As this is one of the most important issues for archives dealing with social media data, the focus of this guide is mostly on the topics of documentation and metadata. However, there are also other areas for which archives need to develop new solutions if they want to store and provide social media data. Apart from documentation and metadata, another crucial aspect is the provision of access to the data. As for other research data, archived social media data should be as FAIR (Findable, Accessible, Interoperable, and Reusable) as possible. Many of the suggestions and recommendations provided throughout this guide aim at making social media data FAIR, and address one or more of the four dimensions. Table 2 summarizes the key suggestions for social media data with regard to the FAIR criteria.[39]

[39] Notably, several of the suggestions and recommendations relate to more than one criterion. However, for the sake of clarity/readability, they are only mapped to one criterion (the one they are most closely or directly related to).

Table 2: Mapping the suggestions and recommendations for social media data to the FAIR criteria.

FAIR dimension	Suggestions & recommendations
Findability	• Define and use specific study-level metadata elements (e.g., data type) • Use those in addition to standard metadata elements (e.g., title, creator, collection dates)
Accessibility	• Provide sufficient documentation of the data, their collection, and pre-processing • There is a need to develop new modes of access for social media data as they are a) larger (and possibly also more complex) and b) more difficult to anonymize and, hence, more sensitive: e.g., secure remote access or access via APIs for data access (offered by archives)
Interoperability	• Use interoperable formats to archive social media data (e.g., JSON or CSV) • Where possible, use existing controlled vocabularies (CVs), e.g., from DDI for documentation
Reusability	• Ensure that legal requirements are met and that privacy concerns are taken into account when processing the data for archiving and sharing • If possible & applicable, data collection and processing scripts/code should also be shared • Variable-level documentation should describe the data processing (e.g., how variables based on textual data were created)

While developing sustainable solutions, standards, and best practices for archiving social media certainly takes time and requires the allocation of sufficient resources, the growing importance of social media data in the social sciences makes this a worthwhile undertaking. The suggestions and recommendations in this guide are meant to facilitate this process by giving archivists and researchers some general guidance on what they need to consider and take care of for archiving social media data and outlining the next steps that should be taken to establish the archiving of social media data as common practice.

7. References

Bargmeyer, B. E., & Gillman, D. W. (2000). *Metadata standards and metadata registries: An overview*. https://www.bls.gov/osmr/research-papers/2000/pdf/st000010.pdf

Bishop, L., & Gray, D. (2017). Ethical challenges of publishing and sharing social media research data. In K. Woodfield (Eds.), *Advances in Research Ethics and Integrity* (Vol. 2, pp. 159–187). Emerald Publishing Limited. https://doi.org/10.1108/S2398-601820180000002007

Block, W. C., Andersen, C. B., Bontempo, D. E., Gregory, A., Howald, S., Kieweg, D., & Radler, B. T. (2011). *Documenting a wider variety of data using the data documentation initiative 3. 1: Best Practices, Examples, and Recommendations for Extending the Standard.* (DDI Working Paper Series - Longitudinal Best Practices). https://doi.org/10.3886/DDILONGITUDINAL01

Borschewski, K., & Zenk-Möltgen, W. (2017, Juli 20). *Facilitating metadata capture and reuse in the social sciences with the example of social media data*. ESRA 2017: 7th Conference of the European Survey Research Association. https://www.europeansurveyresearch.org/conference/programme2017?sess=186#318

Borschewski, K., Förster, A., Friedrich, T., Zenk-Möltgen, W., Miranda, P., Moura Ferreira, P., Banovic, J., Bradić-Martinović, A., Malic, L., Ala-Lahti, H., Jääskeläinen, T., Moilanen, K., Hagen, S., Jakobsen, M., Storviken, S., Try Laundal, A. M., Utaaker Segadal, K., Balkan, L., Barbalet, S., … Bolton, S. (2019). *CMM CESSDA metadata model (Version 1.0)*. https://doi.org/10.5281/zenodo.3543756

boyd, danah, & Crawford, K. (2012). Critical questions for big data: Provocations for a cultural, technological, and scholarly phenomenon. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Breuer, J., Bishop, L., & 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), 2058–2080. 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), 1544–1566. https://doi.org/10.1080/1369118x.2019.1637447

CESSDA Training Team. (2020). *CESSDA Data Management Expert Guide*. CESSDA ERIC. https://zenodo.org/record/3820473

Corti, L., & Bishop, L. (2020). Ethical issues in data sharing and archiving. In R. Iphofen (Eds.), *Handbook of Research Ethics and Scientific Integrity* (pp. 403–426). Springer. https://doi.org/10.1007/978-3-030-16759-2_17

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2019). *Managing and sharing research data: a guide to good practice*. SAGE Publications Limited.

Cram, L., & Llewellyn, C. (2017). *UK-EU referendum Twitter data*. [data collection]. UK Data Service. http://doi.org/10.5255/UKDA-SN-852513

Cram, L., & Llewellyn, C. (2020). *Brexit Twitter data 2017-2019.* [data collection]. UK Data Service. http://doi.org/10.5255/UKDA-SN-854098

Data Cite Metadata Working Group. (2017). *DataCite Metadata Schema Documentation for the publication and citation of research data.* Version 4.1. Data Citee.V. http://doi.org/10.5438/0014.

DDI Alliance. (2020). *MRT Report: DDI-CDI - Cross Domain Integration (DDI-CDI) Features and Status.*

Domingo-Ferrer, J. (2019). Personal big data, GDPR and anonymization. In A. Cuzzocrea, S. Greco, H. Larsen, D. Saccà, T. Andreasen & H. Christiansen (Eds.), *Flexible Query Answering Systems* (Vol. 11529). Springer, Cham. https://doi.org/10.1007/978-3-030-27629-4_2

Fafalios, P., Iosifidis, V., Ntoutsi, E., & Dietze, S. (2018). TweetsKB: A public and large-Scale RDF corpus of annotated Tweets. In A. Gangemi, R. Navigli, M.-E. Vidal, P. Hitzler, R. Troncy, L. Hollink, A. Tordai & M. Alam (Eds.), *The Semantic Web* (pp. 177–190). Springer, Cham. https://doi.org/10.1007/978-3-319-93417-4_12

Grenz, T. (2020). Processualizing data: Variants of process-produced data. *Canadian Review of Sociology/Revue Canadienne de Sociologie*, *57*(2), 247–264. https://doi.org/10.1111/cars.12280

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

Hagen, S., Bishop, L., Koščík, M., Vavra, M., Štebe, J., Ryan, L., Payne, E., Løvlie, A., Rød, L.-M., Straume, Ø., & Høgteveit Myhren, M. (2019). *Report on legal and ethical framework and strategies related to access, use, re-use, dissemination and preservation of social media data* (Deliverable 6.3). https://seriss.eu/wp-content/uploads/2019/11/D6.3-Report-on-legal-and-ethical-framework-and-strategies... FINAL.pdf

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

Harzenetter, K. (2018). Metadata and documentation at the variable level. In S. Netscher & C. Eder (Eds.), *Data Processing and Documentation: Generating High Quality Research Data in Quantitative Social Science Research* (pp. 45-51). https://doi.org/10.21241/SSOAR.59492

Hemphill, L. (2019). Updates on ICPSR's Social Media Archive (SOMAR). https://doi.org/10.5281/ZENODO.3612677

Hemphill, L., Hedstrom, M. L., & Leonard, S. H. (2020). Saving social media data: Understanding data management practices among social media researchers and their implications for archives. *Journal of the Association for Information Science and Technology*, Advance online publication. https://doi.org/10.1002/asi.24368

Hemphill, L., Leonard, S. H., & Hedstrom, M. (2018). Developing a Social Media Archive at ICPSR. *Proceedings of Web Archiving and Digital Libraries* (WADL'18). http://hdl.handle.net/2027.42/143185

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

Kaczmirek, L., Mayr, P., Vatrapu, R., Bleier, A., Blumenberg, M., Gummer, T., Hussain, A., Kinder-Kurlanda, K., Manshaei, K., Thamm, M., Weller, K., Wenz, A., & Wolf, C. (2014). *Social media monitoring of the campaigns for the 2013 German Bundestag Elections on Facebook and Twitter.* (GESIS-Working Papers No. 31). https://www.gesis.org/en/services/sharing-knowledge/publications/archive/gesis-working-papers

Kinder-Kurlanda, K., Weller, K., Zenk-Möltgen, W., Pfeffer, J., & Morstatter, F. (2017). Archiving information from geotagged tweets to promote reproducibility and comparability in social media research. *Big Data & Society*, *4*(2), 1–14. https://doi.org/10.1177/2053951717736336

Kruse, F., & Thestrup, J. B. (Eds.). (2017). *Research data management—A European perspective*. De Gruyter. https://doi.org/10.1515/9783110365634

Littman, J. (2019, January 10). *Twitter's developer policies for researchers, archivists, and librarians*. On Archivy. https://medium.com/on-archivy/twitters-developer-policies-for-researchers-archivists-and-librarians-63e9ba0433b2

Ljubešić, N., Erjavec, T., & Fišer, D. (2017). *Twitter corpus janes-tweet 1.0*. http://hdl.handle.net/11356/1142

Mancosu, M., & 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. https://doi.org/10.1177/2056305120940703

Mannheimer, S., & Hull, E. A. (2018). Sharing selves: Developing an ethical framework for curating social media data. *International Journal of Digital Curation*, *12*(2), 196–209. https://doi.org/10.2218/ijdc.v12i2.518

Markham, A., & Buchanan, E. (2012). *Ethical decision-making and internet research: Recommendations from the AoIR ethics working committee* (Version 2.0). https://aoir.org/reports/ethics2.pdf

Netscher, S., & Eder, C. (2018). *Data processing and documentation: Generating high quality research data in quantitative social science research.* (GESIS Papers No. 22). https://doi.org/10.21241/SSOAR.59492

Pomerantz, J. (2015). *Metadata*. The MIT Press.

RatSWD [German Data Forum] (2020): Big data in social, behavioural, and economic sciences: Data access and research data management. *RatSWD Output, 4*(6), German Data Forum (RatSWD). https://doi.org/10.17620/02671.52

Rauber, A., Asmi, A., Van Uytvanck, D., & Proell, S. (2016). Identification of reproducible subsets for data citation, sharing and re-use. *Bulletin of IEEE Technical Committee on Digital Libraries, 12*(1), 6–15.

Sen, I., Flöck, F., Weller, K., Weiß, B., & Wagner, C. (2019). A total error framework for digital traces of humans. *arXiv*. https://arxiv.org/abs/1907.08228

Sloan, L. (2017). Who Tweets in the United Kingdom? Profiling the Twitter Population Using the British Social Attitudes Survey 2015. *Social Media + Society*, *3*(1), Advance online publication. https://doi.org/10.1177/2056305117698981

Sloan, L., Jessop, C., Al Baghal, T., & Williams, M. (2020). Linking survey and Twitter data: Informed consent, disclosure, security, and archiving. *Journal of Empirical Research on Human Research Ethics*, *15*(1–2), 63–76. https://doi.org/10.1177/1556264619853447

Stier, S., Bleier, A., Bonart, M., Mörsheim, F., Bohlouli, M., Nizhegorodov, M., Posch, L., Maier, J., Rothmund, T., & Staab, S. (2018a). *Social media monitoring for the German federal election 2017* (ZA6926 Data file Version 1.0.0). GESIS Data Archive. https://doi.org/10.4232/1.12992

Stier, S., Bleier, A., Bonart, M., Mörsheim, F., Bohlouli, M., Nizhegorodov, M., Posch, L., Maier, J., Rothmund, T., Staab, S. (2018b). *Systematically monitoring social media: The case of the German federal election 2017.* (GESIS Papers No. 4). https://nbn-resolving.org/urn:nbn:de:0168-ssoar-56149-4

Stier, S., Breuer, J., Siegers, P., & Thorson, K. (2020). Integrating survey data and digital trace data: Key issues in developing an emerging field. *Social Science Computer Review*, *38*(5), 503–516. https://doi.org/10.1177/0894439319843669

Thomson, S. D. (2016). *Preserving social media*. Digital Preservation Coalition. https://doi.org/10.7207/twr16-01

Townsend, L., & Wallace, C. (2016). *Social media research: A guide to ethics*. University of Aberdeen. https://www.gla.ac.uk/media/Media_487729_smxx.pdf

Van Atteveldt, W., Althaus, S., & Wessler, H. (2020). The trouble with sharing your privates: Pursuing ethical open science and collaborative research across national jurisdictions using sensitive data. *Political Communication*, Advance online publication. https://doi.org/10.1080/10584609.2020.1744780

Watteler, O. (2020). *Archiving social media data - challenges and proposed solutions - legal issues [Webinar].* CESSDA webinar. https://doi.org/10.5281/zenodo.3875962

Webb, H., Jirotka, M., Stahl, B. C., Housley, W., Edwards, A., Williams, M., Procter, R., Rana, O., & Burnap, P. (2017). The ethical challenges of publishing Twitter data for research dissemination. *Proceedings of the 2017 ACM on Web Science Conference (WebSci '17)*, 339–348. https://doi.org/10.1145/3091478.3091489

Weller, K., & Kinder-Kurlanda, K. E. (2016). A manifesto for data sharing in social media research. *Proceedings of the 8th ACM Conference on Web Science (WebSci '16)*, 166–172. https://doi.org/10.1145/2908131.2908172

Williams, M. L., Burnap, P., & 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), 1149–1168. https://doi.org/10.1177/0038038517708140

Appendix A - What is metadata & why is documentation of research data important?

Data Documentation "summarizes the description of research data, their content and structure, the context of data collection, data processing as well as the research data themselves" (Netscher & Eder, 2018, p. 5).

The documentation of data contains relevant context information on data and provides answers to important questions concerning the data, as for example: When was the data collected? How was it collected? What is it about? Etc. Documentation makes the process of data generation transparent and makes it possible to reuse and reproduce data, and fosters data sharing, which is one of the criteria of good scientific practice and is nowadays required by many academic journals and funders.

Data is documented with metadata. The most well-known and simplified definition of metadata is "data about data". An alternative and broader definition is provided by Pomerantz, who describes metadata as "a statement about a potentially informative object" (Pomerantz, 2015, p. 26). Both definitions, however, are quite general and Pomerantz's definition seems to ignore the fact that metadata can itself be seen as data as Bargmayer and Gillman (2000) point out: "We don't know when data is metadata or just data. Metadata is data that is used to describe other data, so the usage turns it into metadata" (Bargmeyer & Gillman, 2000, p. 1).

The quality of data description has a decisive influence on data reusability. Good metadata helps to make data FAIR (Findable, Accessible, Interoperable, and Reusable) – it enhances data findability, interpretability, interoperability, administration, exchange, reuse, and replicability. The more documentation about the context of the data is provided, the better it can be used. As a minimum, there should be sufficient metadata to make the data findable but also understandable and reusable by other researchers (CESSDA Training Team, 2020, p. 19).

A relevant distinction in this context is that between structured and unstructured metadata. Unstructured metadata is textual metadata without any standardized semantics or syntax. What this means in practice is that the research process is described in as much detail as necessary, without using any pre-existing standards, e.g., in codebooks or the interviewer instructions. This might lead to a richer and more detailed description of the survey than using standardized metadata schemas since there are no restrictions, predefined forms or limited vocabulary. However, unstructured metadata may be imprecise, ambiguous and it takes a lot more effort and resources to process the (possibly very detailed) information. Having no given schema for metadata, may also be very problematic when it comes to the search for data. Considering the rapidly increasing amount of archived and published data, a standardized data documentation becomes more and more important. Structured metadata implies that you have a standardized form that is filled with the associated information by researchers and

archives. Structured metadata is also more machine-readable. Standardization of metadata in schemas facilitates the search for keywords and increases efficiency. It will also make data more comparable and enhance its reusability. "Metadata containing information about workflows and process documentation is most effective for publishable research when it is standardized, and also deposited and made available to other researchers" (Thomson, 2016, p. 14). It is always good regarding usability if common metadata standards are applied when creating metadata, since structured and standardized metadata enables the exchange and reuse of metadata, as well as the development of related software tools.

There are many different types of structured metadata. Metadata that complies with a community standard can be expected to increase the reusability. Various metadata standards relevant for the social sciences and humanities can be accessed through the Digital Curation Centre website.[40] Probably the most widely used metadata standard in social sciences is the one published by the Data Documentation Initiative (DDI).[41] The DDI metadata standard has different versions, all of which are discipline-specific for the social sciences as well as machine-readable. These attributes and the general standardization of DDI means that data documented with them are better able to meet the FAIR criteria. On the other hand, however, these standards are less flexible, which can make it more difficult to apply them to new data types, such as social media data.[42]

Information can be distinguished into study-level documentation (information on the project which initiated data collection, on investigators, on topics contained in data set and methodological information on data collection, and the data management process) and data-level documentation – information both on the whole data file and on segments (variables) in the file. This archiving guide focuses on study-level documentation of social media data, but it should be kept in mind that variable-level documentation is generally important for data intelligibility too.

Similar to study-level documentation, documentation of a data set on the variable level can be provided in different forms, for example, as a separate supplementary published text file (codebook) or as an online portal based on a data set and its documentation (XML-file) (Harzenetter, 2018, p. 46). Corti et al. (2019) recommend that variable-level documentation should be embedded within a data set when possible and additional variable level

[40] Social Science & Humanities | DCC, https://www.dcc.ac.uk/resources/subject-areas/social-science-humanities (date of access: 06/11/2020).
[41] Welcome to the Data Documentation Initiative | Data Documentation Initiative, https://ddialliance.org/ (date of access: 06/11/2020).
[42] While this means that storing social media data in general repositories, such as Figshare, can be easier as they use more general and less discipline-specific metadata, the data stored there are generally less findable and reusable than data sets stored in curated repositories with discipline-specific metadata standards. To illustrate this issue, we present a brief general description of how social media documentation can be represented in DDI Lifecycle (DDI-L) and DDI - Cross Data Integration (DDI-CDI) in Appendix B.

documentation should be recorded in a structured metadata format, such as XML, to be machine-readable, if possible.

Similar recommendations can be found in the CESSDA Data Management Expert Guide (DMEG): "Metadata is often embedded into the data file (e.g., in the form of variable names and variable and value labels, different kinds of notes and content of supplementary variables). So, the structure of your data also contributes to the clarity of your data documentation" (CESSDA Training Team, 2020).

Appendix B - Representation of social media data documentation in DDI-L and DDI-CDI

As mentioned above, some of the information that needs to be documented for social media metadata is equivalent to the information documented for survey data. This information can, therefore, be easily documented in DDI Lifecycle (DDI-L) and partly also in DDI Codebook (DDI-C). Such information would for example include elements in DDI-L version 3.2 on the collection event as for example collection date, collector organization, data source, mode of collection etc. DDI-L also offers the possibility to capture information on data cleaning, weighting, and data appraisal under the "processingevent" module and information on the generation instructions in the "processinginstructionscheme". Furthermore, the instrument can be generally described in DDI-L version 3.2; however, this is more suitable for survey data. But still, overview information as the name, type, description etc. can still be documented (Block et al., 2011; Borschewski & Zenk-Möltgen, 2017). Therefore, the basic information concerning a social media data study is covered with DDI-L. But for social media data, several data-specific pieces of information need to be documented which differ from survey data. Furthermore, the specific needs and differences of social media data compared to survey data are not yet reflected in the DDI CVs. For more information, see also Block et al. (2011). New data types, such as social media data can be more easily documented in the DDI - Cross Data Integration (DDI-CDI) specification, which was opened for public review in April 2020.[43] It is aligned with DDI-C and DDI-L and geared towards the description of new forms of data (including social media data). "The intention is that DDI-CDI be a tool which can supplement systems using earlier versions of DDI, enabling them to better handle new types of data."[44] With DDI-CDI, it is possible to document a variety of research data in different formats coming from different sources, independent of their scientific and policy domain (DDI Alliance, 2020).[45]

[43] Public Review: DDI - Cross Domain Integration (DDI-CDI), https://ddialliance.org/announcement/public-review-ddi-cross-domain-integration-ddi-cdi (date of access: 06/11/2020).
[44] DDI – Cross Domain Integration: Introduction, p. 5, https://ddi-alliance.bitbucket.io/DDI-CDI/DDI-CDI_Public_Review_1.zip (date of access: 06/11/2020).
[45] DDI – Cross Domain Integration: Introduction, https://ddi-alliance.bitbucket.io/DDI-CDI/DDI-CDI_Public_Review_1.zip (date of access: 06/11/2020).