H2020-ICT-2018-2 /ICT-28-2018-CSA
SOMA: Social Observatory for Disinformation and Social Media Analysis

D2.2: Research data exchange solution

Project Reference No	SOMA [825469]
Deliverable	D2.2: Research Data exchange (and transparency) solution with platforms
Work package	WP2: Methods and Analysis for disinformation modeling
Type	Report
Dissemination Level	Public
Date	30/08/2019
Status	Final
Authors	Lynge Asbjørn Møller, DATALAB, Aarhus University Anja Bechmann, DATALAB, Aarhus University
Contributor(s)	See fact-checking interviews and meetings in appendix 7.2
Reviewers	Noemi Trino, LUISS Datalab, LUISS University Stefano Guarino, LUISS Datalab, LUISS University
Document description	This deliverable compiles the findings and recommended solutions and actions needed in order to construct a sustainable data exchange model for stakeholders, focusing on a differentiated perspective, one for journalists and the broader community, and one for university-based academic researchers.

Document Revision History

Version	Date	Modifications Introduced	
		Modification Reason	Modified by
v0.1	28/08/2019	Consolidation of first draft	DATALAB, Aarhus University
v0.2	29/08/2019	Review	LUISS Datalab, LUISS University
v0.3	30/08/2019	Proofread	DATALAB, Aarhus University
v1.0	30/08/2019	Final version	DATALAB, Aarhus University

Executive Summary

This report provides an evaluation of current solutions for data transparency and exchange with social media platforms, an account of the historic obstacles and developments within the subject and a prioritized list of future scenarios and solutions for data access with social media platforms.

The evaluation of current solutions and the historic accounts are based primarily on a systematic review of academic literature on the subject, expanded by an account on the most recent developments and solutions. Although marked by the research oriented agenda of the academic literature, the account as a whole takes the perspectives of both researchers and journalists and the broader community.

The results show that the methods for data exchange provided by the social media platforms are subject to increasingly strict restrictions of data access, making it difficult - if not impossible - to extract substantial social media data for thorough investigations.

APIs (Application Programming Interfaces) created for third-party business developers are the most used method to extract data from social media platforms for research purposes, but they are inherently flawed and limited, posing several challenges for researchers and journalists alike when it comes to data quality, research design and data reliability.

New targeted solutions for data access for researchers and journalists offered by the platforms in an effort to address issues caused by the API restrictions are for the most part insufficient and limited in scope and information richness of the data offered.

Instead of the current flawed model for data exchange, we propose alternative future scenarios for transparency and data exchange with social media platforms.

The first priority, albeit only feasible in the long term, is the establishment of a scalable data solution across and outside platforms, making it possible to conduct data research in a safe space adhering to ethical standards and without violating privacy laws. This scenario will be expanded on a future report to be handed in on April 30, 2020.

The second priority is the establishment of two dedicated API-based data access solutions, one specifically tailored to researchers and one for NGOs/journalists. Most current API offerings are for all developers, but tailored for commercial developers with terms of service that do not address research or journalistic purposes, nor require uploading obtained permissions. The two dedicated and differentiated APIs should have terms of service explicitly requiring the use of data in the public interest rather than commercial benefit, prohibiting problematic collaborations with companies like Cambridge Analytica.

As a short term action to achieve these future scenarios and move towards better conditions for data exchange with social media platforms, we recommend establishing protected data exchange sandboxes to test different solutions for data exchange that better balance the conflicting values of transparency and privacy.

Table of Contents

List of Figures

List of Tables

List of Terms and Abbreviations

Abbreviation	Definition
API	Application Programming Interface
GDPR	General Data Protection Regulation
IRB	Institutional Review Board
SSRC	Social Science Research Council

1 Introduction

Over the last decade, social media has evolved to become one of the most important drivers for acquiring and spreading information (Stieglitz et al., 2018, p. 156). This has led to an increasing accumulation of data about social media usage - data that has become a valuable commodity for the social media companies, as they can be used commercially to make predictive behavioral targeting (ibid.; Bechmann, 2013, p. 73).

The growth of social media usage and the availability of social media data has also opened up new academic research opportunities for analysing aspects and patterns in digital interaction and communication and for journalists to use data for investigative journalism (Stieglitz et al., 2018, p. 156). Social media has especially become important for research into computational social science that investigates questions using big data and quantitative methods for data mining, e.g. statistics and computational models such as machine learning (Batrinca & Treleaven, 2015, pp. 89-90).

However, the availability of social media data for academic research and journalism has changed significantly over the last years due to commercial pressures, as the social media companies have no interest in revealing what kind of data they have on users and exactly how they retrieved it (Batrinca & Treleaven, 2015, p. 89; Bechmann, 2013, p. 77). In addition, the most tools available are far from ideal, giving only superficial access to the raw data and requiring researchers to program analytics in a language such as Java (Batrinca & Treleaven, 2015, pp. 89-90).

Some social media data is accessible through Application Programming Interfaces (APIs), but most of the major social media companies are making it increasingly difficult for academics and journalists to obtain comprehensive access to their data, and only very few social data sources provide affordable data offerings to academia and researchers (Batrinca & Treleaven, 2015, p. 90).

For instance, social scientists around the world were up to recently able to use API data from Facebook to study online communities and investigate social media's impact on society. But as a reaction to several controversies - the Cambridge Analytica controversy especially - Facebook recently tightened the access restrictions to the APIs of its platforms (Bruns et al., 2018).

Furthermore, specific provisions implemented in Europe for data protection (such as the GDPR - General Data Protection Regulation) have been used by platforms as a shield to implement a generalized refuse to data access to researchers. The impact and implications of the GDPR and of the research exemptions built into the law on the activities of researchers engaging in social network analysis have become central in the academic discussion (see for instance Kotsios et al. 2019).

Thus, social media data access has become a major challenge for academic research, increasingly making social media platforms black boxes to researchers (Batrinca & Treleaven, 2015, p. 90; Bechmann, 2013, p. 77).

1.1 Purpose and Scope

The purpose of this report is to investigate past, present and potential future solutions for transparency and data exchange with platforms.

The report compiles the findings of a systematic review of existing academic papers with a focus on social media data access/exchange and on the basis of this review, discusses and recommends solutions and actions needed in order to construct a sustainable data exchange model for stakeholders in both university-based academic community and in the journalistic and broader community.

However, the main focus of the report will be on the academic community as they have a special status in the GDPR that provides a potential for a more extensive exchange model, benefiting knowledge in the greater society and breaking down knowledge barriers between the private platforms and society in general (Bechmann & Kim, 2020).

1.2 Structure of the report

The first part of the report is a systematic literature review investigating methods for data exchange with social media platforms. The literature review provides an account for the developments and historical obstacles within data exchange with platforms and an overview of present solutions available across highest impact social media platforms. Although the literature review is inherently marked by the research agenda of the papers included, the methods for data exchange with social media platforms are not limited to researchers and can be and especially have been applied by journalists and the broader community.

The second part of the report accounts for the latest developments and solutions for data exchange that are too recent to be included in the literature review. This section will take the perspective of both researchers and journalists and the broader community, and it will be based on recent literature and articles, own experiences with data access and meetings on the subject with representatives from platforms - primarily Facebook as this is the platform with the largest penetration in Europe (alexa.com) and the target of Cambridge Analytica (see Appendix 7.2 for list of these meetings).

The final section of the report consists of a prioritized scenario list that reflects on the potentials and challenges, and an evaluation of actions needed in order to carry out the use of such solutions.

2 Literature review

The following chapter contains a literature review on the developments within data access and present solutions for data exchange with social media platforms. A literature review is a systematic examination of the academic literature about a topic, critically analysing research findings, theories and practices (Efron & Ravid, 2019, p. 2). It will help us achieve a comprehensive, critical, and accurate understanding of the current state of academic knowledge on the subject across academic fields.

2.1 Methodology

With the aim of investigating solutions for social media platform data exchange with the highest impact, we define social media with three characteristics commonly emphasized when theorizing social media: Communication is de-institutionalized, the user is regarded as a producer, communication is interactive and networked (Bechmann & Lomborg, 2013, p. 767).

We identify the highest impact social media platforms with a starting point in the top 15 social networks based on the active number of global active users (Statista, 2019):

Table 1: Top 15 social networks globally

1. Facebook
2. YouTube
3. WhatsApp
4. Facebook Messenger
5. WeChat
6. Instagram
7. QQ
8. QZone
9. Douyin/Tik Tok
10. Sina Weibo
11. Reddit
12. Twitter
13. Douban
14. LinkedIn
15. Baidu Tieba

(Statista, 2019)

As we were not able to find statistics for Europe as a region, we have used the global accounts but excluded the platforms WeChat, QQ, QZone, Douyin/Tik Tok, Sina Weibo as they are primarily used in China (Statista, 2019). Hence, we identified the highest impact social media platforms in a European context as being Facebook, YouTube, WhatsApp, Instagram, Reddit and Twitter. It would have been interesting to also include LinkedIn in our accounts, but Linkedin is different from the other platforms as this is a platform specifically designed for professional purposes.

2.1.1 Literature review search

On March 12, 2019, academic publications related to data exchange with the above-mentioned social media platforms were identified and gathered across all academic fields using a very comprehensive search across over 400 international databases including Web of Science and Scopus (the full list of databases can be found in the references under Statsbiblioteket, n.d.). The publications were gathered via a systematic keyword search for academic articles, not including open science repositories and conference papers.

To achieve the objective of providing a comprehensive overview of solutions in and discussions on data exchange with social media platforms, the keywords *social media* (or the specific platforms Facebook, YouTube, WhatsApp, Instagram, Twitter, Reddit) paired with *data exchange* (or synonyms *data collection, data extraction, scraping, crawling* and *api*) were used to search the databases for academic articles (see Table 1).

The first round of search resulted in 51,905 academic articles. Additional keywords *data*, *collection* or *analytics* paired with *social media* (or one of the specific social platforms) in the title were used to filter the results to get 604 academic articles (see Table 1).

Table 2: Search terms for systematic literature search

Fields	Search terms		
Anywhere in text	"social media" OR "facebook" OR "youtube" OR "whatsapp" OR "instagram" OR "twitter" OR "reddit"	AND	"data exchange" OR "data collection" OR "data extraction" OR "scraping" OR "crawling" OR "api"
In title	"data" OR "collection" OR "analytics"	AND	"social media" OR "facebook" OR "youtube" OR "whatsapp" OR "instagram" OR "twitter" OR "reddit"

In the first manual review of the articles, 170 articles were completely excluded from the analysis. Some of these articles were duplicate versions, while others were deemed not relevant as they do not use or even mention methods for data exchange with social media platforms. The remaining 434 articles were included in the review.

2.1.2 Categorisation and in-depth analysis

All 434 articles were scanned to categorise the studies in regard to the data exchange method used and the social media platform studied. Some articles did not study specific social media but were rather methodological papers on data exchange or literature reviews of such and were thus categorized as *Other.* Other articles studied several different social media platforms or used different methods for data exchange and were placed in multiple categories.

Out of the 434 categorized articles, a total of 22 articles were determined to be relevant to include for further in-depth analysis of data exchange with social media platforms. Only methodical papers on methods for data exchange, and studies that compared methods for data exchange and heavily reflected upon the chosen method, were included for further review. Most studies simply mention the method used for data exchange and were thus not chosen for an in-depth review.

For the in-depth analysis, the chosen articles were reviewed with the purpose of accounting for the background of developments within data exchange, the academic view on the present solutions available, and data exchange solutions proposed for the future.

Results from the in-depth literature review and categorisation will be accounted for in the following sections. First, we will in general terms account for some of the different data resources offered by social media platforms and the collection methods that can be applied - with a larger focus on API solutions - and second, we will focus on each of the specific social media platforms.

2.2 Data resources and collection methods

Most social media platforms use different methods of access, many do not even provide convenient standardized ways for gaining access to data, and they are often subject to strict restrictions of data access and to further changes as the business models of the platforms change (Stieglitz et al., 2014, p. 91). In this section, we will account for some general terms that can be used to better understand the rugged jungle of different data resources and methods of access offered by social media platform.

Batrinca & Treleaven (2015) divide social media data resources into three categories; *freely available databases*, *data access via tools* and *data access via APIs*:

- Freely available databases cover repositories that can be freely downloaded. Batrinca & Treleaven (2015) mention Wikipedia's database dumps as an example (p. 94). Many of these databases are included in the Google Dataset Search tool (https://toolbox.google.com/datasetsearch).
- Data access via tools cover sources that provide controlled access to their social media data via dedicated tools, facilitating easy interrogation and stopping users from getting all the data from the repository. Batrinca & Treleaven (2015) further subdivide this category into two subcategories; *free sources*, and *commercial sources* (p. 94). Free sources are freely accessible repositories with tools protecting or limiting the access to the raw data, e.g. Google Trends and other tools provided by Google, or research tools extracting and visualizing data, e.g. Digital Methods Initiative (Amsterdam University) and QUT Digital Media Research Centre. Commercial sources are data resellers that charge for access to their social media data. For instance, Gnip provide commercial access to Twitter data through a partnership (see Section 2.3.4).
- Data access via APIs cover social media data repositories providing programmable HTTP-based access to the data via APIs, e.g. Twitter and Facebook. Batrinca & Treleaven (2015) point out that this arguably the most useful source of social media data for researchers (pp. 95-96) and given API's importance to researchers, the subject will be detailed in the next section (2.2.1).

Depending on the social media data resource and the data access offered by the specific data resource, there are several different strategies you can use to collect the data. Butakov et al. (2018) identifies three different data collection strategies:

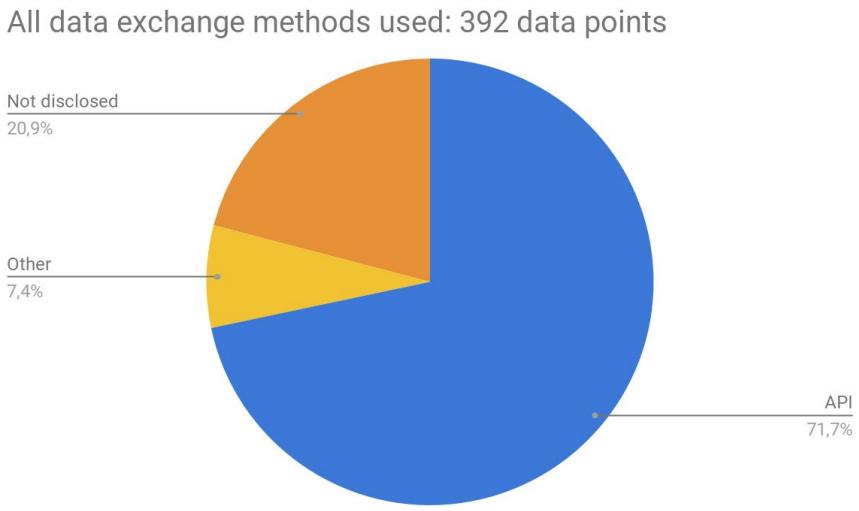
- *Search*: Collect historical data by performing a set of search queries to a social media platform that then retrieves the information and returns a collection of relevant data matching the specified query (Butakov et al., 2018, p. 392). This used to collect data of posts containing specific keywords or collect data of specific users.
- *Stream*: Continuously collect new data in accordance to a set of queries, for instance all new posts by a specific user or all new posts related to some specific topic (ibid.)
- *Traverse*: Collect data about transitions from entity to entity. This can for instance be used to collect data of users that share specific posts (ibid.).

2.2.1 Using APIs for data access

Many social media platforms make data on users and usage patterns available through Application Programming Interfaces (APIs) (Lomborg & Bechmann, 2014, p. 256). APIs are back-end interfaces that social media platforms provide to third party developers to create new add-on applications to the platform, thereby fostering the growth of an application ecosystem and enhancing the value of the platform through added functionality (Rieder et al., 2015, p. 2).

As a side effect, APIs provide access for researchers to collect data off a given social media platform often using JavaScript Object Notation (JSON) that makes the data searchable and sortable for subsequent data mining in for instance Python, R or dedicated network analysis tools such as Gephi. Researchers access the API through small software scripts to retrieve and store user data (Lomborg & Bechmann, 2014, p. 256).

The results of our literature search for research using social media data show that APIs are the most common method used to extract data from social media platforms (Appendix 7.1). An API was used 71.7 % of times data was extracted and only 7.4 % of times another method for data extraction was used:



But APIs are not designed to accommodate the needs of researchers, they are often subject to obscure restrictions of data access and to further changes of the platforms, and they pose a series of challenges for researchers that we will address in the following paragraphs (Rieder et al. 2015, p. 2 & Stieglitz et al. 2014, 91).

First of all, APIs have historically been made for business developers, not researchers. This means that researchers will have to pretend to be business developers in order to have access to the API, thus answering questions designed for the enhancement of app functionalities and not for instance uploading Institutional Review Board (IRB) and data agency permissions to carry out the research project.

Also, platform APIs require users to write in a programming language in order to access data (Batrinca & Treleaven 2015, p. 90). While this might be reasonable task for computer scientists, these skills are typically beyond most social science, health and humanities researchers that will need to collaborate with more technically skilled researchers in order to use APIs to collect social media data (ibid.; Lomborg & Bechmann, 2014, p. 258).

Another challenge is the fact that APIs evolve and change as new functionalities are added to the platforms or the underlying data models are changed (Halford et al., 2018, p. 3349). This rapid pace of change confounds the slower, more deliberate tempo of academic publishing and has major implications for research using data gathered through an API (ibid.; Driscoll & Walker, 2014, p. 1748).

For instance, academic papers may be referring to data collected years before the publication date when the API offerings were very different, making it difficult - if not impossible - to replicate a previous study if the data archived is not available for the larger community (ibid.).

When conducting longitudinal analysis or going back in time in data collection, the changing APIs also make it difficult to assure consistency in the data, and the researcher cannot see, if data patterns occur due to

changes in the API structure because they are not accounted for (Halford et al., 2018, pp. 3350-3351; Bechmann & Vahlstrup, 2015, p. 9).

Also, previously solid and well-tested data collection strategies and software scripts for accessing APIs may become obsolete due to the changing APIs (Lomborg & Bechmann, 2014, p. 260). The chance that any kind of research software can fully deal with a constantly changing set of issues is small and would require significant funding allocated for developers (Rieder et al., 2015, pp. 6-7).

Perhaps as a consequence of the changing APIs, many papers in our literature search describe the specific API used for data extraction as just "the API" of the specific social media platform and not by its version. As an example, 43.8% of the times an API was used to extract data from Twitter, it was disclosed as just "API" (Appendix 7.1.1). This lack of detail limits the generalizability of these studies as they are harder to reliably replicate or compare with any other studies (Driscoll & Walker, 2014, p. 1748).

Another challenge facing researchers working with APIs is the asymmetric relationship with the social media company, who shape the informational structure and control what is made available for analysis (Rieder et al., 2015, p. 19; Lomborg & Bechmann, 2014, p. 260). APIs are far from neutral tools, but rather subject to the company's shifting views of how sharing data with third-party developers can benefit their platform (Rieder et al., 2015, p. 5). This stands in stark contrast to traditional data collection methods where researchers produce their own data or work with described secondary sources of data (Halford et al., 2018, p. 3344).

Lomborg & Bechmann (2014) also point out that social media data collected through APIs has a built-in bias towards active users that contribute with content (p. 259). This provides APIs with a major blind spot concerning the ability to analyse user reading mode and the click-through patterns of the so-called "lurkers" who uses the social media platform frequently but rarely post to the stream themselves (Bechmann & Vahlstrup, 2015, p. 2). Thus, APIs as a single-standing method for data collection is not necessarily the most useful entry point for studying "typical users" (Lomborg & Bechmann, 2014, p. 259).

Finally, privacy issues are always present when data are collected through APIs - both legally and ethically.

Legally, social media data can be personal and sensitive data, even if it is technically public. When collecting the data through an API, the researcher does not know whether data that might seem mundane at the time of retrieval will become sensitive at a later point in time, and it must thus always be handled according to privacy laws (Lomborg & Bechmann, 2014, p. 261).

Also, it can also be questioned whether it is ethically right to collect, process, use, and report on social media data that may be public in principle, but might be perceived as highly personal by the users (Stieglitz et al., 2014, p. 91).The default public mode of many social media platforms may be understood to constrain user's expectations of privacy and to support analytic use but as Wheeler (2018) suggests, few users bother to read the terms, and users who give their consent to public broadcast through social media may have context specific intentions and expectations about the use of their content (p. 6).

Lomborg & Bechmann (2014) argue that researchers should seek informed consent before collecting data in order to respect human subjects' perceived privacy - at least in qualitative studies (p. 262). However, that is hardly an option for large-scale research with thousands of users involved, for which the legal and ethical challenges using APIs instead revolve around how data is anonymized - both to the researcher and when presenting results (ibid.).

These inherent flaws and limitations of social media APIs pose several general challenges for researchers when it comes to data quality, research design and data reliability. In the next sections, other challenges emerge as we focus on the specific methods for extracting data from each social media.

2.3 Twitter

Twitter is a microblogging site where users exchange short, 140-character messages called tweets, enabling rapid communication between its over 300 million monthly users, resulting in plenty of research attention (Morstatter & Liu, 2017, p. 1).

Over the years, Twitter has been covered by several different research disciplines and historically, most academic API-based research is carried out on Twitter (Lomborg & Bechmann, 2014, p. 257). Twitter's history of openness in terms of access to its database; coupled with the ease of collecting data through their public APIs[1], are some of the main reasons driving researchers' interest in Twitter data (Gayo-Avello, 2013, p. 650).

But API-based access is becoming more and more restricted, while concerns are raised about the data being problematic, flawed by demographic and population biases and unknown provenance, leaving fears that Twitter data may lead to poor research and over-confident conclusions (Lomborg & Bechmann, 2014, p. 257; Halford et al., 2018, p. 3342).

In our literature search, Twitter was the social media platform that was most often subject of studies via data analysis. Looking only at the studies that actually extract data, 63.3 % of the times data was extracted, it was from Twitter, and Twitter's APIs were used 78.2 % of the times data was extracted from Twitter, making the APIs by far the most used method for data extraction were the most used method to extract the data (Appendix 7.1.1). As mentioned in section 2.1.1, 43.8 % of the times an API was used to extract data from Twitter, it was disclosed just as "API" - perhaps as a consequence of the changing APIs.

2.3.1 The development of Twitter's APIs

Since Twitter first launched their API, the richness and structure of the data made available have changed considerably (Halford et al., 2018, pp. 3348 & 3350-3351).

Launched in 2006, Twitter was initially open about sharing data (ibid., p. 3348). Researchers could request privileged access to the API without restrictions, and as the practice spread, Twitter datasets and studies grew (Wheeler, 2018, p. 8). This resulted in the emergence of applications that access and process Twitter data, lowering the technical barriers and making it even easier for researchers to obtain datasets, thus fostering additional research (Congostoa, Basanta-Vala & Sanchez-Fernandeza, 2017, p. 29).

In 2011, Twitter altered its API structure, tightened its developer policies and shutdown several research tools, arguing that public archiving of datasets violated the API terms of service (Felt, 2016, p. 2; Wheeler, 2018, p. 8). The privileged research access to the APIs were phased out, and new limits of only one percent of daily Twitter traffic were imposed (Felt, 2016, p. 2).

As Twitter became a public corporation in 2013, the platform took further steps to secure its data flow (ibid., p. 4). In early 2013, Twitter changed its API in several important ways, shifting the security model so that all calls to the API required authentication, thus removing the ability to anonymously request query data, and implementing new API request limits across all API calls (Chudnov et al., 2014, p. 1). To researchers not experienced with software development, these changes made data much more difficult to access (ibid.).

In 2010, Twitter also began commercialising access to 'the Firehose' - the full, unrestricted stream of Twitter data (Driscoll & Walker, 2014, p. 1748). Access became limited to a few third-party data resellers, who in turn would sell the data at premium costs (Wheeler, 2018, pp. 8-9). After acquiring one of these

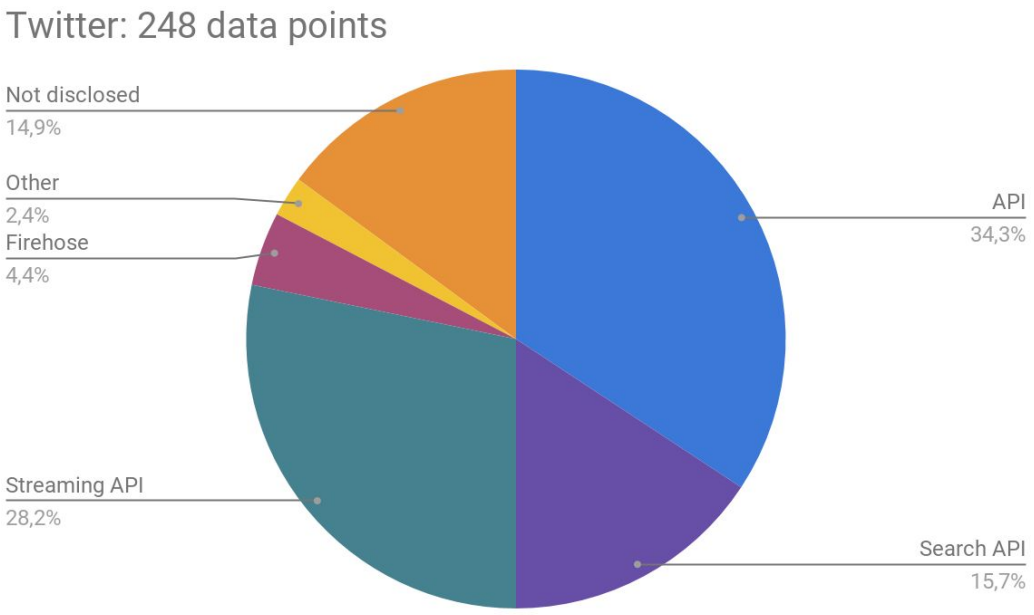
[1] In relation to Twitter understood as 'free' APIs contrary to their paid APIs

resellers - Gnip - in 2014, Twitter transitioned into only offering unrestricted data access through their own services, shutting off the full data streams at other resellers in 2015 (Halford et al., 2018, p. 3348).

As a result, full data access was suddenly reserved for companies and researchers with sufficient financial resources to pay premium data costs (Wheeler, 2018, pp. 8-9).

This tightening of API policies and implementation of data access restrictions had a negative impact on research, and there has been a decrease in the size and number of Twitter studies since 2011 (ibid., p. 8). Given the currently expensive access to the full Twitter data stream, research groups without substantial funding usually turn to tools that utilize Twitter's limited public APIs: the Streaming API and Search API (Felt, 2016, p. 4; Halford et al., 2018, p. 3348).

Looking only at the studies in the literature search results that extract data from Twitter, the Streaming API was used 28.2% of the times, the Search API was used 15.7% of the times, the Firehose was used 4.4% of the times, while 34.3% of times an undisclosed API was used (Appendix 7.1.1).



2.3.2 Search API

Twitter's Search API enables querying for tweets that include specific keywords or hashtags, user mentions, date created, etc. (Wheeler, 2018, p. 11). It is free to use and provides a maximum of 3200 tweets published in the past seven days, with a rate limit of 180 searches every 15 minutes (Lienemann et al., 2017, "Data Sources").

There are several important constraints to using the Search API for research, as data from the API is not complete and should not be considered the entirety of all public tweets matching the search criteria (Driscoll & Walker, 2014, p. 1749)

Firstly, the API is rate limited - that is, there is a limit to the number of times an individual user or application may execute a specific action within a given time frame. In the case of Twitter, the platform will block or permanently ban users violating the limits of 180 searches every 15 minutes (Wheeler, 2018, pp. 11-12).

Secondly, it is not possible to retrieve data from any arbitrary date or in real time - only from the last seven days before the current one (Gayo-Avello, 2013, p. 660). A more extensive archive is available at cost, e.g. Gnip (Wheeler, 2018, p. 11).

And thirdly, Search API results are not queried from the entirety of all tweets within those seven days, nor are they random samples of the overall Twitter activity (Driscoll & Walker, 2014, p. 1749) Rather, they are samples focused on relevance instead of completeness (Halford et al., 2018, p. 3348).

While we do not know how it is sampled, Driscoll and Walker's (2014) suggest that Search API results are skewed heavily towards central users and more clustered regions of the network (p. 1749). Also, Gayo-Avello (2013) suggest that because the Search API has been devised to power Twitter's search engine, it has a bias toward those users considered by the algorithm as more relevant (p. 660).

Whether these constraints will affect a research investigation depends on the nature of the data being queried (Halford et al., 2018, p. 3349).

2.3.3 Streaming API

Lienemann et al. (2017) suggests that Twitter's Streaming API is more useful to researchers than the Search API, as it is not rate limited, providing a stream of tweets as they are posted in real time ("Data Sources"). The results of the literature search suggest the same: When focusing on the papers that actually extract data for studies; the Streaming API was the most popular Twitter API, used 28.2% of the times data was extracted from Twitter (Appendix 7.1.1).

The volume of data extracted through the Streaming API is constrained by an undocumented upper limit known as the "streaming cap," which is believed and widely stated to be up to 1% of the entire Twitter stream at any point in time (Driscoll & Walker, 2014, p. 1750).

At present, there are two ways in which Streaming API provides real-time data: In a 1 % sample of all activity and or as filtered results (Halford et al., 2018, p. 3348).

You can retrieve a 1 % sample of all public tweets, in some papers also referred to as the Sample API (Morstatter & Liu, 2017, p. 3). Though we do not know how it is sampled, the company states it is random - a statement that has been validated in academic research (Halford et al., 2018, p. 3348; Morstatter & Liu, 2017, p. 3). This sample can be useful to get a look at 'what's happening on Twitter', but less so if the research aims at extracting data on a specific topic which will most likely not be part of the 1% sample (Halford et al., 2018, p. 3348).

For this purpose, the Streaming API allows users to harvest real time data via filtering by keywords, user IDs or location (Batrinca & Treleaven, 2015, p. 97). The API will provide a continuous stream of Tweets matching the search criteria, providing the researcher with only the most relevant data for their research task (Morstatter & Liu, 2017, pp. 2-3).

When 'total tweets' matching the query stays below the 1% limit, the Streaming API can return all of the tweets pertaining to that query (Wheeler, 2018, p. 12) - although Halford et al. (2018) suggests that there is no guarantee that the API will return all tweets for that search term even if these constitute less than 1% of all tweets (p. 3348). Once the volume surpasses the 1% limit, results are sampled, and how this sampling is done is not published by Twitter, providing severe issues of scientific documentation for the research community (Morstatter & Liu, 2017, p. 2)

Twitter states that the sampling of the filtered Streaming API is statistically representative, but several papers question this (Gayo-Avello, 2013, p. 660). Lomborg & Bechmann (2014) suggest that particularly

smaller samples collected from the Streaming API tend to misrepresent the volume of hashtagged content, implicating studies of hashtags and trends on Twitter (p. 260)

Both Morstatter & Liu (2017) and Driscoll & Walker (2014) investigate this by comparing the output from the Streaming API with the unsampled Firehose of all public tweets, only available at a premium cost.

Driscoll & Walker (2014) investigate bias when using the Streaming API to collect tweets during high-volume, short-term events and medium-volume, longer-term events.

For the former, they compared tweets from the Streaming API and Firehose collected during an extremely popular political debate, observing a significant data loss in the data from the Streaming API that indicate it is not an appropriate tool for studies that require comprehensive collections of tweets concerning high-volume events or topics (ibid., p. 1756).

For the latter, they compared tweets collected during the Occupy Wall Street protests, getting very comparable results, only experiencing data loss in the data from the Streaming API during a high-volume event in the time period (ibid., p. 1757). They conclude that the Streaming API excels at longitudinal data collection, but is a poor choice for massive, short-term events (ibid., p. 1762)

Comparing data from the Streaming API and the Firehose on the importance for hashtags and the distribution of the geotagged tweets, Morstatter & Liu (2017) also show evidence that the Streaming API can be biased, though the findings indicate that more coverage decreases bias (p. 5).

They propose a data-collection approach to collect data from the Streaming API which maximises coverage by splitting the keyword list among multiple queries. This approach will decrease the amount of bias simply because it generates more data and better coverage, as it becomes possible to gather tweets beyond the 1% limit, and because each individual query is designed to stay below the 1% limit, meaning each query should not be biased from the sampling (ibid., p. 8).

2.3.4 The Firehose

Providing the greatest access to data, Twitter's Firehose has access to 100% of all Twitter content. Whereas Firehose access was previously offered for a fee by several third party resellers, Twitter is now the sole provider of full, unrestricted data access after acquiring the reseller Gnip in 2014 (Lienemann et al., 2017, "Data Sources").

Through Gnip, Twitter offers access to the fully archived Search API and real-time data in the PowerTrack. Similar to the Streaming API, the PowerTrack provides a real-time stream of tweets matching a set of search criteria. However, with the valuable absence of the Streaming API's 1% data limit (Driscoll & Walker, 2014, p. 1751).

Hence, this is one way to overcome the limitations of the public APIs, despite it being very costly for research groups (Morstatter & Liu, 2017, p. 2). Another drawback is the significant computing resources - in terms of servers, network availability, and disk space - required to retain the Firehose data, thus favouring only the wealthy universities and labs (ibid.).

This makes full, unrestricted access to Twitter data unavailable to most researchers, also exemplified by the results of our literature search when looking at papers that used Twitter data extraction. Here, a Firehose was used only 4.4% of the times data was extracted from Twitter (Appendix 7.1.1).

Though having tremendous research potential, the full Firehose access is thus limited to clients capable of paying premium costs, privileging the research agendas of organizations and institutions that can afford it

(Wheeler, 2018, p. 10). That leaves researchers without significant resources to use the public APIs, allowing only the capture of up to 1% of the daily Twitter flow and limiting queries to real time or just one week back in time (Felt, 2016, p. 13).

The limited and flawed public APIs do not correspond well with the principles of social research, grounded in clearly understood populations, controlled and unbiased sampling and well-documented collection methods, none offered by Twitter (Halford et al., 2018, p. 3342). Instead, the APIs act as black boxes between the researcher and Twitter, preventing researchers from achieving total certainty about their results (Felt, 2016, pp. 2-3).

Various research tools have been developed to utilize the public API and work around their flawed nature, so researchers with limited funding can aggregate and analyze Twitter data. Felt (2018) compares three of them - Storify, Netlytic, and DMI-TCAT - and concludes that though requiring the most technical expertise, DMI-TCAT provides the most fertile data for analysis (p. 13). To date, DMI-TCAT is still up and running, and the source code can be downloaded via GitHub.

2.4 Facebook

Compared to the smaller Twitter platform (measured in number of users), there is far less attention paid to data-driven Facebook research despite it being the biggest social media site, with more than two billion users worldwide using it every month (Rieder et al., 2015, p. 3).

In our literature review isolating to papers extracting social media data, only 11% of the data-driven research examined Facebook, compared to 63.3% that examined Twitter (Appendix 7.1). For that reason, this section of the literature review will be less detailed than that on Twitter. For details on more recent developments in Facebook's APIs and current access points available, go to Section 3.2.1.

This significant imbalance in research interest in the platform may be explained by the restrictive architecture of Facebook (Rieder et al., 2015, p. 3). Though the public Twitter APIs have become more restrictive in recent years, Facebook's API is far more limiting (Felt, 2016, p. 5).

2.4.1 The development of Facebook's APIs

Over the years, there have been numerous and far-reaching changes to Facebook's API, and several access paths have been shut down. The company's first API was launched in 2006, called the REST API, and it was the main API for several years, receiving significant updates (Rieder et al., 2015, p. 6). The REST API also introduced the general setup of Facebook's API, where you have to create an app and register it to receive the necessary access credentials (ibid.).

In 2007, Facebook added a second access point, facilitating more complex interactions with the data pool. The Facebook Query Language (FQL) allowed for powerful filtering and concatenation, quite uncommon in the web-API space, and retrieving complex compound data such as large friendship networks became considerably easier and much faster (ibid.).

In October 2009, the Graph API was introduced by Facebook, and it remains as the primary way to get data out of the Facebook platform. The Graph API was a redesign of the REST API, now facilitating app development, moving the API closer to Facebook's architecture and interfaces (ibid.).

Like the REST API, the Graph API requires the user of the API to create an app in order to get access credentials to extract the data. Hence, researchers have to request permission through a Facebook app to collect nonpublic data from participants, meaning that a lot of status messages are harder to obtain than Tweets (Lomborg & Bechmann, 2014, p. 258; Batrinca & Treleaven, 2015, p. 97).

Using the Graph API thus involves developing a Facebook app and getting 'access tokens' from the users of the app, permitting the app to access the users' data (Batrinca & Treleaven, 2015, p. 97). During installation, the app will explicitly ask the user's permission to access certain data and depending on the access token and specific user's privacy settings, the app is then able to acquire data on the signed-in user (Rieder et al., 2015, p. 6).

From the implementation of the Graph API and until 2014, an access token from a Facebook user did not only allow the app to access the user's own data, it also allowed the app to access the user's friends list and data from user's friends. This meant that successful apps could collect enormous amounts of data, since these apps were often used by a large number of people and datasets retrieved through different apps could be easily merged (ibid.).

This general setup of the Graph API remained largely the same until 2014, where new versions were introduced and older access methods shut down (from 2015 onwards), moving towards a stronger protection of user privacy from third-party apps. The FQL was removed, the ability to combine datasets was now impossible, access to friendship relations and friend's data was removed and the News Feed access disappeared from all APIs. Also, apps now had to go through an obligatory review procedure if they asked for more than just the basic access permissions (ibid.).

Recently, Facebook has pushed towards even stronger privacy, and they have begun to curtail many data gathering possibilities. As these changes were not part of the literature search results - perhaps because they are too recent - the latest developments and newer solutions are instead accounted for in Section 3.2.1.

2.4.2 The Graph API

The data gathered through a Facebook app can be accessed by querying the Graph API, making researchers able to access historical data with fewer time limitations than Twitter's Search API (Batrinca & Treleaven, 2015, p. 97; Rieder et al., 2015, pp. 6-7).

When focusing on the studies extracting data from Facebook in the literature search results, the Graph API is used 23.3% of times data was extracted from Facebook, making it the most popular version of the API disclosed. For the most part, the type of API is however not disclosed which might be a consequence of the changing APIs, as accounted for in section 2.1.1 (Appendix 7.1.2).

Getting real time data from the Graph API is not straightforward, as the API does not offer real-time streaming access in the same way as Twitter's Streaming API does. Instead, you can retrieve the most recent posts repeatedly in short intervals to achieve near-real-time coverage (Stieglitz et al., 2018, p. 163).

However, the API is rate limited, limiting each user to 200 calls to the API per hour (Halford et al., 2018, pp. 3348-3349). Hence, there are significant risks of missing data and not detecting it, making it difficult to assess the reliability of the data set, when using the API to collect data (Lomborg & Bechmann, 2014, pp. 260-261).

As with the public Twitter APIs, there are significant considerations to be made, when using Facebook's API for research purposes, such as questions of completeness and representativeness in the data and the inability to assess reliability. We will return to an evaluation of new solutions for data access that has just recently been released and thus is not accounted for, used or evaluated in the existing literature in the review.

2.5 YouTube

YouTube is not only the most popular video sharing website, it has also become a platform where people express their opinions and participate in discussions and a tool for organizations to get out their message (Malik & Tian, 2017, p. 194; Shah, 2010, p. 226). Hence, YouTube videos and its related metadata, such as user comments, have become a data source that can be used in various fields of research (Malik & Tian, 2017, p. 195).

Google acquired YouTube in 2006, and they launched the YouTube Data API as one of Google's over 20 different APIs for developers (Bechmann, 2013, pp. 79-80). Ever since the acquisition, there have been many modifications in YouTube's interface, making data extraction harder. Shah (2010) mention this as a major problem as the developed research tools at the time are not adaptable to the constantly changing site and page structure of YouTube (p. 227).

The YouTube Data API is provided by YouTube for developers to integrate with in the same way as the Graph API as accounted for earlier. The API provides a search function, designed to simulate the search activity on the YouTube website and provide metadata results easily to be handled (Malik & Tian, 2017, p. 195).

It enables querying for videos that includes specific keywords or other attributes such as channel ID (collect videos uploaded by a specific channel), video category (collect the videos that only from a specific category), publish after/before and order the returned result by data/view count/rating, just like the normal search function on the website (ibid.).

Through the YouTube Data API, researchers can retrieve detailed information of each single video in the form of metadata. Some of these metadata are constant, such as video ID, length, upload channel, publishing time of video and comments, etc., while other metadata vary over time, such as view count, comment count, subscriber count, etc. Consequently, Malik & Tian (2017) separate metadata into two types: (a) Invariant Data and (b) Dynamic Data (p. 198).

Even though YouTube's Data API provides researchers with the opportunity to easily collect data, Malik & Tian (2017) conclude that few researchers have used large amounts of YouTube data to conduct analysis, as most either use very little data harvested manually or cherry pick the metadata attributes that are most easily harvested from the API (p. 195).

The results from the literature review when focusing on studies extracting data show that YouTube data was extracted only 3.3% of the times data was extracted for analysis (Appendix 7.1). Out of these, the API was used 61.5% of the time, while other methods include for instance using the search function from the website and collecting metadata manually (Appendix 7.1.3).

Malik & Tian (2017) suggest that the lack of research using large amounts of YouTube data is due to the fact that the Data API's search function is not designed to return large amounts of data (p. 196).

All metadata are tied to unique video identifiers called video IDs and to collect large volumes of metadata, you need a lot of video IDs as an API input. Using the search function in the YouTube Data API, you can easily generate a few hundred video IDs, but it is not scalable enough to provide access to large amounts of video IDs, and YouTube does not publish lists of the video IDs to use as input (ibid., pp. 196-197).

Malik & Tian (2017) propose a framework for continuous collection of video IDs and related metadata, taking on the key challenge of collecting large amounts of YouTube metadata through the Data API.

As a first step, they use a small amount of video IDs as seeds to search the API and retrieve video IDs of videos that are related to the searched video, exponentially growing a list of video IDs in database (Malik & Tian 2017, p. 197). After gathering a huge number of video IDs, they use the Data API to retrieve detailed metadata of each video, systematically storing it into the database on a continuous basis to capture the complete evolution of metadata (ibid., p. 198).

2.6 Instagram

Despite having more monthly users than Twitter, the photo and video-sharing social media platform Instagram does not attract much research interest based on the results of our literature review, also confirmed by Domínguez et al. (2017, p. 325). Looking at the studies using data exchange from the literature search, only 2.3% of the data extractions for research was from Instagram and of these, the Instagram API was used 90% of the times (Appendix 7.1.4).

Instagram API allowed users to extract data published in the past or in real time. Compared to Twitter's APIs, the Instagram API had the advantage of allowing users to recover data from any moment in the past, providing more flexibility for researchers (Domínguez et al., 2017, p. 325). The Instagram API was however rate limited, allowing 500 calls per access token in a 1-hour time slot (Halford et al., 2018, pp. 3348-3349).

Facebook acquired Instagram in 2012, and Instagram has in the same way as Facebook tightened access through the API and is currently shutting down the Instagram API. The platform is transitioning to the new and much more restrictive Instagram Graph API, developed on the basis of the Facebook Graph API and inheriting all its structural solutions. These changes are too recent to be part of the literature search results and instead, the latest developments and current solutions are accounted for in Section 3.4.

2.7 Reddit

Reddit is a social media platform focused on topical issues and hosting different discussions consisting of user posts across hundreds of communities called *subreddits*. These discussions are aggregated on a home page to create "the front page of the web" that Reddit was founded to provide to its readers (Gaffney & Matias, 2018, p. 2).

As one of the largest forums on the web, Reddit has gained high visibility in the past years, also driving research attention to the site (ibid., p. 1). In our literature search however, only 1% of the data extractions were from Reddit, with all of these being from its API (Appendix 7.1.5).

Reddit provides an open API for anyone to freely mine data from the web site, without restrictions, and since 2015, Jason Baumgartner has provided researchers with a complete copy of the platform, frequently updated through the API (ibid.). The data is available at pushshift.io.

Subsequently, many researchers have adopted the dataset, and have used it to study a wide range of questions, but Gaffney & Matias (2018) discovered substantial gaps and limitations to Baumgartner's dataset, estimating that 0.043% of all comments and 0.65% of all submissions may be missing (pp. 1 & 4). The missing data represent risks to research validity and risk of bias in research using the data (ibid., p. 11).

After being made aware of these issues, Jason Baumgartner filled any gaps in the dataset and took steps to ensure the integrity of future data by double-checking for missing content (ibid.). Gaffney & Matias (2018) encourage researchers to use the dataset but check the integrity of the data before publishing results (p. 11).

2.8 WhatsApp

None of the 434 articles retrieved used data from WhatsApp (Appendix 7.1). This is most likely because the platform is end-to-end encrypted and does not offer any immediate solution with which researchers can extract data and understand behavioral and communicative patterns of users and other stakeholders (https://www.whatsapp.com/business/api). For recent developments, see section 3.6.

2.9 Review Conclusion

The results from the literature review show that APIs are still by far the most used method to extract data from social media platforms for research purposes, but as it has been accounted for in the review, these APIs are subject to increasingly strict restrictions of data access.

Batrinca & Treleaven (2015) single out the data restrictions and monetization of data access as the biggest concern for data-driven social media research (p. 115). Because of this, computational social science is becoming an exclusive domain for major companies, government agencies and privileged academic researchers with private data used to produce papers that cannot be critiqued or replicated (ibid.).

Hence, it is important that researchers have access to computational environments and social media data for experimentation, and Batrinca & Treleaven (2015) call for the establishment of such public computational environments and data facilities for quantitative social science research, where researchers can access data via a cloud-based facility (p. 115).

Such an environment in the form of custom APIs for researchers may also address concerns over the level of technical skills needed to access data via the APIs. As Bechmann & Vahlstup (2015) suggest, a generic multi-user system is needed, so social scientist and humanist researchers will not need assisting computer scientists to retrieve data from various social media APIs (p. 3).

Rieder et al. (2015) also call for a sustainable setting for social media research and the establishment of legal research rights for researchers to access and use social media data, equivalent of fair use principles or similar provisions (p. 19). Without better conditions for social media research, social media analysis can become impossible for researchers operating independently from commercial interests, making knowledge concerning the activities of billions a private entity (ibid.).

3 Recent developments and solutions

In the following section, we will account for recent developments and solutions within data exchange from the perspective of both researchers and journalists and the broader community, expanding on the literature review with developments too recent to be accounted for in the literature search results. Apart from literature and articles, this section is based on own experiences and meeting with the social media platforms on the subject (see Appendix 7.2).

3.1 Twitter

In a move to prevent spam from Twitter bots, Twitter has imposed new API rules, requiring developers to provide detailed information about how they use or intend to use the APIs from September 10th, 2018 (Roth & Johnson, 2018). Other than that, Twitter has not imposed any substantial API restrictions not accounted for in the literature review.

However, it is important to note that Twitter was one of the first platforms to systematically provide data grants calls for researchers. These grants were provided to a selected few researchers and included an extended access to data (Raffi, 2014).

3.1.1 Ads Transparency Center

Recently though, Twitter has launched a new tool for researchers and journalists to access data on ads on Twitter. Launched on June 28th of 2018, the Ads Transparency Center, enables users to freely search for any handle of a specific advertiser and view any ad campaigns that have run from that handle within the last seven days (Falck, 2018). It also consists of lists of certified advertisers for political campaigning, for the US, the EU, Australia and India and a list of certified issue advertisers for US issue advertising.

Initially, you were only able to access further details on ads with ‘political content’ from the US, but in March this expanded to all EU member states (in accordance with the EU Commission agreement in continuation of the HLEG Report on Disinformation - Buning et al., 2018), India, and Australia (Twitter Inc., 2019). Twitter distinguishes two kinds of political content: ‘political campaigning’ and ‘issue advertising’ (Twitter, n.d.). The further details available for ads that fall under these categories include billing information, ad spend, impression data per Tweet and demographic targeting data (Falck, 2018).

There are, however, several limitations to the Ads Transparency Center, as pointed out by the Office of the French Ambassador for Digital Affairs in their assessment of the center as a tool to counter disinformation (Office of the French Ambassador for Digital Affairs, ©2018).

For one, technical solutions chosen by Twitter for the Ads Transparency Center makes it cumbersome, technically difficult and potentially in violation with Twitter’s terms to do large-scale quantitative investigations.

For instance, the authorisation to use the tool expires regularly. Although you do not need login to access the Ads Transprency Center, access depends on an undocumented process of temporary authorization through the guest tokens in need of renewal after only a few dozen requests (ibid., “Twitter Ads Transparency Center Assessment”).

Also, all of the data is tied to the online interactive user interface, and cannot be downloaded. Hence, to conduct quantitative studies on Twitter ads, you will have to reverse engineer the user interface, requiring a lot of time and advanced programming skills, and also potentially violating Twitter’s Developer Agreement on reverse engineering (ibid.).

Secondly, the data integrity is poor. Though Twitter have pledged to showing promoted political content in the Ads Transparency Center indefinitely, all ads are still only available for seven days. Also, both Twitter and the advertiser can remove ads from the Ads Transparency Center, as deleted ads or ads that have been taken down by Twitter no longer will be available in the Ads Transparency Center (Office of the French Ambassador for Digital Affairs, ©2018, "Twitter Ads Transparency Center Assessment").

Also, only a subset of political ads are shown. Twitter offers a wide range of advertising formats, but only Promoted Tweets appear in the Ads Transparency Center. Political campaigning ads are only allowed to be promoted via Promoted Tweets and In-Stream Videos, but issue advocacy ads are not subject to such a restriction, meaning all In-Stream Video political ads and all issue advocacy ads except Promoted Tweets will not appear in the Ads Transparency Center (ibid.).

3.2 Facebook

As a reaction to the Cambridge Analytica data scandal, Facebook imposed new data access restrictions to their platforms in April, 2018, heavily restricting the access points accounted for in Section 2.3.

3.2.1 Recent developments

Cambridge Analytica used personal Facebook data collected via a Facebook app that also accessed app users' friends' personal data, collecting data on over 50 million Facebook users (Meredith, 2018). Although this type of data collection was already curtailed by Facebook in 2014, as accounted for in Section 2.3, the exposure of the scandal in 2018 pushed Facebook to impose further restrictions.

For instance, the review process for apps requesting access to user data was tightened, while the opportunity for apps to request access to 'personal information' was shut down (Schroepfer, 2018). Also, access to data from public pages was closed, while access to events and public groups was heavily restricted - all access points that were previously used by many researchers (Freelon, 2018, p. 665). These changes were met with heavy scepticism among academics, declaring important social media research at risk - not only on the topic of disinformation and democracy but other critical issues such as cyber bullying and hate speech were also impossible to investigate (Bruns et al., 2018).

In an effort to address issues caused by the new API restrictions - and simultaneously ward off negative press from the above-mentioned criticism - Facebook have sought to replace API access with more targeted solutions for data access for researchers and journalists (Bruns, 2019, p. 8).

3.2.2 Ad Library API

Facebook introduced the Ads Archive API in 2018 with public data on political ads run in the U.S. and expanded the API, now as the Ad Library API, in 2019 with data on all active ads on their platforms (Sullivan, 2019). As of August 2019, it is available in the EU, the US, the UK, Brazil, India, Ukraine, Canada and Israel.

The Ad Library is explored through a web interface and queried through an API. For each country, aggregated data on ads is available in the form of a web page with dynamic tables or a downloadable CSV file (Office of the French Ambassador for Digital Affairs,©2018, "Facebook Ads Library Assessment").

This report has data on all ads run in the country since the launch of the Ad Library in March, 2019, including Page ID, Page Name, amount spent on ads and number of ads in the library, however no information about the content of the ads is available (ibid.). In other words, it is difficult for researchers, journalists and other stakeholders to understand exactly what is included as issue advertisements. For instance, what is the algorithmic understanding of an issue ad included in the library.

Neither the API nor the reports provide any information on targeting criteria or any engagement data, such as clicks, likes or shares. Without this data, it is impossible to find out which audiences advertisers are paying to influence, and whether they have been successful at that (Mozilla, 2019b).

Also, using the API relies on a complex authentication mechanism designed to prevent full automation. Accessing and querying the API requires an access token that can be retrieved by creating a Facebook app and using a Facebook account to login in through the app (Office of the French Ambassador for Digital Affairs ©2018, "Facebook Ads Library Assessment").

Both the user creating the app and the user giving the access token and querying the API (they can be the same user) will have to go through the certification process needed to actually publish ads, and the access token is only valid for two hours, after which the user will have to renew it through the app. Also, prior to accessing snapshots of image or video ads, the user making the query must connect to Facebook in a web browser (ibid.). These technical solutions hinder any automatic download of data.

Since you cannot download data in bulk, and you cannot access all ad data at once and filter it down, it is impossible to get a complete picture of all the ads running on Facebook or determine if the API is comprehensive. This also makes it very difficult to evaluate ads in larger topics or regions, requiring months of data collection even though this is the way most researchers work with data as our literature review has pointed to (Mozilla, 2019b).

3.2.3 CrowdTangle API

CrowdTangle is an analytics platform that was created to give content creators the data needed to succeed, allowing them to track how content spread for a fee. In 2016, Facebook acquired CrowdTangle and in 2017, they made the CrowdTangle API free for news organizations (Hare, 2017). Although relying mostly on Facebook data, it is also available across Twitter, Instagram and Reddit (Office of the French Ambassador for Digital Affairs, ©2018, "Detection tools").

The CrowdTangle API tracks posts shared by public pages or verified public persons, measures their social performance and a Chrome extension can track how content is being spread and the accounts who shared the content (ibid.). For instance, the tool makes it possible to follow the development of disinformation and investigate how it originated.

According to our meetings, Facebook is now opening for free access to CrowdTangle API for research groups, but we have yet to see how this is going to be rolled out in practice. Here, a potential challenge would be if Facebook chooses to screen researchers not only as their status of legitimate and good faith academic researchers, but also screens (out) critical researchers or potential research interest that could harm Facebook's reputation and thereby business.

3.2.4 Social Science One

In the wake of the most recent API restrictions in April, 2018, Facebook launched the new initiative Social Science One - a partnership with seven US-based non-profit foundations and the non-profit Social Science Research Council created to provide selected researchers access to data to study the impact of social media on democracy and elections (King & Persily, 2018). In this sense there are similarities to the aforementioned data grants from Twitter.

The first request for project proposals ended a year later. More than 60 researchers were chosen in review process involving a peer review effort overseen by the Social Science Research Council (SSRC), an additional ethics and privacy review by a separate panel, and a final stage where the co-chairs at Social Science One selects the projects on the basis of the first two stages (Bruns, 2019, p. 10).

The accepted researchers were given access to the above mentioned CrowdTangle API and Ad Library API and promised access to an anonymized 'URL shares' dataset describing URLs that have been shared on Facebook (King & Persily April 28, 2019). A new request for proposals to access the CrowdTangle API and the Ad Library API respectively were released in May, 2019 (Social Science One, 2019).

Bruns (2019) criticises the initiative for being 'designed predominantly to benefit the corporation' with little room for academic independence, narrow terms of the work - elections and democracy - and an inherently intransparent review process with veto power for Social Science One chairs with a privileged relationship with Facebook (pp. 8-11).

In the context of this report, it should be stressed that Social Science One is primarily an American construct deriving from SSRC with a satellite review board of the different regions. This means that the standards and norms for review follow an American centric approach.

In the same lines, this could benefit American (Californian and Ivy League) universities and research groups affiliated with them, as collaborating with unknown researchers and universities pose a greater danger to Facebook that did not have strong ties to the research communities beforehand. This could have a chilling effect on the review as well, as the reviewers signing off on such collaboration also risk their relationship to Facebook then.

Last but not least the Social Science One datasets are made available using differential privacy, designed to make it impossible to disclose the identities of users (Dwork, 2008). Differential privacy inserts bits of noise in the dataset so that you cannot reverse engineer and disclose identities. The first version of the codebook for the URL dataset suggested access to URL shared by 20 people and above. However, audits suggested that this cluster size was not big enough to not disclose identities. Insisting on protecting on data level instead of safe space solutions means that also research done in small countries and markets are disfavoured in comparison to large countries and regions. The reason for this is that they identities are more easily disclosed with the same amount of data points (e.g. demographics, interests, region). Thus, an American dataset can release more data points than a Belgian or Danish equivalent.

Also, very recent developments suggest that the future of the initiative is at risk. Funders of the initiative and connected researchers are losing patience with Facebook, as the company has not yet provided the originally specified 'URL shares' dataset. On August 27, 2019, the funders sent a letter to SSRC recommending 'winding down the project' if Facebook cannot deliver the promised dataset by September 30, 2019 (Silverman, 2019).

Following the recommendations, SSRC has immediately paused all review processes in the project, stating that the funders may be willing to reinitiate their support to the program, if the complete dataset will be made available by September 30. In a statement, Facebook and Social Science One proclaim what they will continue working together to make data available to researchers (ibid.).

3.3 YouTube

To our knowledge, the access to YouTube data has not undergone any significant recent developments not accounted for in the literature review ("Revision History", n.d.).

In 2018, Google introduced a new transparency report for political ads called Political Advertising on Google. The report is explored through a web interface or a downloadable csv-file and provides information on political ads that have run in the regions of the EU, India and the US, including ads on YouTube (Smith, 2018).

However, there is no option to only investigate ads on YouTube. Google must adhere to the EU Commission Action Plan requiring an account of political ads, but due to the ambiguous nature of AdSense, Google may have chosen to interpret this across YouTube and Google Search. However, this is unclear from the report provided.

Also, contrary to Facebook's Ad Library, Google only provides a report and not an API, making it complicated - if not impossible - to explore a wide range of research questions, for instance with methods using natural language processing or vision algorithms.

Google has not been available for an interview in connection with the fact-finding for this report (see also Appendix 7.2).

3.4 Instagram

When announcing the most recent Facebook API restrictions in April, 2018, Facebook also advanced the previously planned shutdown of Instagram's public API - without due warning for third-party developers or researchers (Bruns, 2019, p. 2).

Parts of the API that were scheduled for deprecation on July 31st, 2018 - such as follower lists, relationships, and commenting on public content - were shut down in April instead, while public content reading was shut down in December and what is left will shut down in early 2020 (Constine, 2018).

Instagram simultaneously migrated third-party services to its Graph API, which is designed exclusively for Business Profiles, effectively shutting down data access for non-Business Accounts altogether (Gummadi, 2018; Bastos & Walker, 2018).

Although the access to Instagram data for researchers, journalists and the broader community is virtually non-existent, some data can be accessed through the CrowdTangle API. Relying mostly on Facebook data, the API also enables users to track Instagram posts on public profiles (Office of the French Ambassador for Digital Affairs, ©2018, "Detection tools").

However, in the same way as Youtube, Instagram (owned by Facebook) need to adhere to the committed agreement signed to inform good faith research on disinformation as outlined in the EU Commission Action Plan:

> ***E. Empowering the research community***
> *12. Support good faith independent efforts to track Disinformation and understand its impact*
> *- Information on collaborations with fact-checkers and researchers, including records shared*
>
> *13. Not to prohibit or discourage good faith research into Disinformation and political advertising on their platforms*
> *- Information on policies implementing this commitment*
>
> *14. Encourage research into Disinformation and political advertising*
> *- Information on policies implementing this commitment*
>
> *15. Convene an annual event to foster discussions within academia, the fact-checking community and members of the value chain*
> *- Report on the annual event*
>
> *(full Action Plan available at https://eeas.europa.eu/sites/eeas/files/action_plan_against_disinformation.pdf)*

To adhere to this commitment, Facebook decided not to make a seperate ad library for Instagram but instead made Instagram ads available through the Facebook Ad Library API ("How are ads...", n.d.). This

unification is not well-documented and thus entails additional challenges to the reliability and validity of any research activity.

3.5 Reddit

The Reddit API is public for anyone with no data restrictions, enabling users to collect full datasets from the site. As accounted for in section 2.6, a complete copy of Reddit is available and queryable at pushshift.io.

As mentioned previously, the CrowdTangle API also enables users to explore and track links shared on Reddit (Office of the French Ambassador for Digital Affairs, ©2018, “Detection tools”).

3.6 WhatsApp

Recently, WhatsApp like Twitter and Facebook has started collaborating with selected researchers in a research award program, especially in regard to the recent disinformation debate that has put pressure on the legitimacy of WhatsApp in countries such as Brazil (Chaturvedi, 2018). However, WhatsApp does not provide any data to the award recipients, rendering the program irrelevant in the discussion of data access (WhatsApp, n.d.).

End-to-end-encrypted communities such as WhatsApp pose a profound challenge to research by academics and journalists alike, as content is not made available and therefore cannot be object of scrutiny, when it comes to democratic challenges such as disinformation circulation.

Facebook (owner of WhatsApp) has limited the community to which a user/actor can broadcast content to, in an attempt to mitigate potential damage made by strategic actors. However, the end-to-end encryption remains a black-box for research and a more elaborate discussion is needed on how researchers (and journalists) can systematically evaluate whether communication adheres to democratic standards and values. This could work with a communication sphere that is non-encrypted (for academic research) when reaching a certain mass or containing systematically targeted ads with defined critical content.

4 Scenarios, solutions and actions needed

Based on the research literature review and the accounts for recent developments in data exchange models from the previous chapters, the following is a prioritized scenario list mapping different solution models including their potentials and challenges. The list will be made in the framework of long, mid term solutions and short term actions that will benefit the research and journalistic community in the pursuit of a heightened knowledge level on social media behavior and disinformation (and associated critical democratic topics) for the greater good of society.

4.1 Scalable data solution across and outside platforms (1st priority)

As accounted for previously, one of the problems in the current data exchange models is the changing level of data access, depending to some extent on cases of privacy breaches and tightened regulation, e.g. GDPR. As the report has shown, these changes together with the always incomplete data access have caused research to be less reliable and valid, and thus the policies designed and the decisions taken on top of such research to be equally flawed.

The ideal and first priority scenario in a long term model from a purely scientific point of view would be stable access to social media data in controlled and safe spaces that mitigate privacy risk for the data subjects involved. In such a scenario, the focus would be on securing access to legitimate researchers with strict obligations and tight rules for processing, rather than trying to make data non-identifiable. As attempts based on differential privacy have shown (Dwork, 2008), making data non-identifiable is a very hard task. This is especially the case when a major requirement is to guarantee sufficient and equal data exchange for all countries - not favouring some (larger) countries over other (smaller) countries.

Creating such a stable solution would require the data to be stored and accessed outside the social media companies in question, meaning the companies are not under liability, if data breaches happen or if public opinion shifts to disfavour research involving social media data.

Also, this solution will have to cover data from several different social media platforms, thus encouraging shared standards for data exchange, benefiting research using data from different social media platform (Stieglitz et al., 2014, p. 91).

Historically, precedence for such a solution has been set in health data registers such as DNA registers containing highly sensitive and personal data. Genomic data is hard to anonymize due to the unique features of a genome (Mittos, Malin & De Cristofaro, 2019) which to some extent is similar to social media data, e.g. pictures disclosing identity directly or indirectly. Also, DNA data subjects consent on behalf of relatives to have the profile stored, much like social media users consent on behalf of the ones they are connected to and communicate with. Consent procedures needs to be discussed in further details because the data that have historically been used the most by social media researchers is graph data, where connections between people are highly relevant for mapping circulation of information, such as disinformation, or accounting for the power of the actors in the network, e.g. identifying bridging hubs and influencers. Network data is therefore important in the mapping of disinformation flows and is a good example of data needed that is difficult to combine with privacy solutions on the data unit level.

However, significant actions are needed in order to implement such a solution. For instance, DNA registers are connected to a health sector that in most European countries is state funded. This is not the case with social media data collected by private companies. Hence, a new model for sector registers of collected data across private companies is needed (King & Persily, 2019) where access is administered, controlled and verified by official stakeholders in accordance to transparent requirements and without violating freedom

of science or reviewing the (critical) character of the research. Such requirements can be as high as needed in order to safeguard the privacy of the data subjects and the research ethics, but if all requirements are met, access cannot be denied on the grounds of, for instance, the critical character of the research (see also Franzke, Bechmann, Ess & Zimmer, 2020; Bechmann & Kim, 2020).

Safe space solutions, scenarios and discussions on the governing and control of these will be the topic of an upcoming SOMA report to be delivered in April, 2020.

4.2 Dedicated and differentiated APIs (2nd priority)

Currently, the APIs offered by social media platforms are available to *all* users but tailored for commercial business developers, with the terms of service written with commercial users in mind. The use of these APIs pose several challenges for researchers and journalists when it comes to data quality and data reliability. Additionally, the APIs have not previously allowed researchers to actually verify as such and show the permission (ethical and legal clearance) provided by IRBs and Data Agencies following GDPR (Bechmann & Kim, 2020).

In order to be able to differentiate between the type of data that businesses, NGOs/journalists and academic researchers have access to in a scalable and non-discriminatory way, two additional dedicated APIs would be one solution that is implementable as a midterm scenario.

We propose establishing different dedicated APIs - one for NGOs and journalists and one for academic university-based research due to the research exemptions built into the GDPR. This includes researchers' ability to limit or avoid restrictions on secondary processing and processing of sensitive data, to override the subjects' right to object to processing and erasure if relevant safeguards are implemented, and to collect some types of data without consent (Kotsios et al., 2019, pp. 2-3).

Such dedicated APIs should live up to the requirements below to truly support research and investigative journalism with an interest in both evidence-based (social) media literacy and heightening the level of societal knowledge. However, due to the research exemptions in the GDPR, the API for journalists and NGOs will not be able to return as much and as sensitive data as the dedicated academic research API, but should instead have a Graphical User Interface (GUI) usable for journalists and NGOs.

1. The APIs should have strict validation processes that verify identity, affiliation and permissions and ethical clearance, along with thorough guidelines for how to store, process and disseminate findings and inform researchers, journalists and NGOs about risk assessments and mitigation along with potential audits to check that these requirements are met (see also Franzke, Bechmann, Ess & Zimmer, 2020).
2. The APIs should offer detailed documentation of the data, such as API changes and graph structure, and documentation of potential problems with consistency and replicability to strengthen the method's reliability.
3. The APIs should have access to real time data with either no sampling or sufficient documentation on how this sampling is done to avoid severe issues of scientific documentation for the research community.
4. The APIs should have dedicated terms of service that explicitly require the use of data in the public interest rather than commercial benefit, prohibiting problematic collaborations with companies like Cambridge Analytica and preventing the platforms from shutting down research tools on the grounds of non-differentiated API terms of service violations.
5. The APIs should be open (with strict verification processes as suggested above) and free to use to avoid favoring only the wealthy and famous universities and publishers.
6. Data should be downloadable in bulks without expiration (when clearance have been provided) to empower thorough investigations of large topics or regions.

7. Not only aggregated data but actual content, engagement scores, graph data (or dividing actors in different groups depending on the position in the network) should be available to allow for proper investigations of communication and behavioral patterns.
8. Datasets should not be constructed by the platforms topically and designed to only answer some questions, thereby violating freedom of science. Instead APIs should be structured around data points and the number of data points could be made available depending on actor status, thereby trusting the verification process instead of designing privacy on the data point level (e.g. making data point limitations on top of GDPR).
9. The platforms should follow some general requirements in what and how data is made available in order to benefit research across platforms, as the research investigation and mapping of content circulation cuts across platforms.

4.3 Short term actions needed

Several actions are needed in the short term if we are to achieve the above mentioned scenarios. As minimum requirements in a short term perspective, we recommend moving the existing solutions towards:

1. Scalable solutions for extended data access for independent research and journalism.
2. Data shall be made available on the basis of the data type, not structured around specific questions to be asked.
3. Privacy protection shall be enforced on the level of access verification and processing requirements, not on the level of data units available per se for independent research and journalists.
4. EU Commission or trusted groups shall be encouraged to publish interpretations of or guidelines on social media access for different stakeholders (researchers and journalists) in the light of GDPR, and help facilitate discussions on legal solutions clearly defining and delimiting actors from business stakeholders.
5. Data access solutions shall be constructed on the basis of European regulation, and not with a starting point in American models of transparency, privacy and ethics.
6. EU Commission shall encourage and help facilitate collaborations and crowdsourced initiatives (also using data portability) to construct lists and shared repositories outside platforms in safe spaces (e.g. crowdsourced lists of debunked content from journalists and fact-checkers following the same data model).

A first step in designing the current solutions differently and with a higher degree of balance between transparency and privacy would be to test different solutions for data access in protected data exchange sandboxes to evaluate solutions on presumably conflicting values.

Such sandbox solutions could take many forms - from opening APIs in beta versions to constructing white room facilities with access to the firehose (for projects demanding extended access). In all cases, willingness from platforms, legal guidance from the EU Commission, and funding are needed both to test solutions for journalists/NGOs *and* for independent research. Here, current SOMA Centers of Excellence and similar infrastructures can in both cases be used as a starting point, following a European first approach.

5 Conclusion

The results of the research review show that the current model for data exchange with platforms is flawed, as social media platforms have been imposing increasingly strict restrictions on data access, complicating or even preventing the work of researchers and journalists investigating potential critical questions for democracy, namely the challenge of disinformation and the potential effects on voting behavior and the character of the public debate.

If we look across all platforms accounted for in the report, the current model consists of a mixture of freely available data (e.g. Reddit), data available through restricted APIs (e.g. Twitter's public APIs), data openly available through targeted tools (e.g. the ad libraries), extended data access grants and collaborations (e.g. the Twitter data grants and Social Science One datasets) and research collaborations without data access (e.g. WhatsApp).

Despite being by far the most used method for extracting data from the social media platforms, the APIs are inherently flawed and limited and pose several challenges for researchers and journalists alike when it comes to data quality and reliability. The newer and more targeted solutions for data access, such as the ad libraries provided by the platforms to address issues caused by the API restrictions, are for the most part insufficient and limited in scope and information richness.

Although these tools have to adhere to the same requirements across different platforms (Buning et al, 2018 and subsequent EU Commission Action Plan), the terms are implemented very differently, due to the co- and self-regulatory approach and the very generic requirements, complicating the use of data from different platforms. Even solutions involving APIs that are more scalable and dynamic (such as Facebook's Ad Library API) still lack an understanding of the work processes in research.

Research data grants instead highly control *what* researchers are studying (only some data sets designed to answer specific questions are made available), *who* is studying it (only the 'famous' and 'great' will get access), limiting the field and volume of research significantly and thereby hampering the ability to make adequate peer review processes in a critical mass.

Keeping peer review and reproducibility as golden scientific standards is complicated by the differential privacy models applied and the claimed proprietary status of the platform data. Also, researchers cannot know whether the data provided are full datasets or if potentially harming data have been deleted by the platforms. Furthermore, data access grants and collaborations make research vulnerable to changing public opinions, conflicts, disagreements and political landscapes inside and outside the organizations, causing delays in the data access and subsequent research or even risking shut down as the case with the recent developments within the Social Science One initiative (Silverman, 2019).

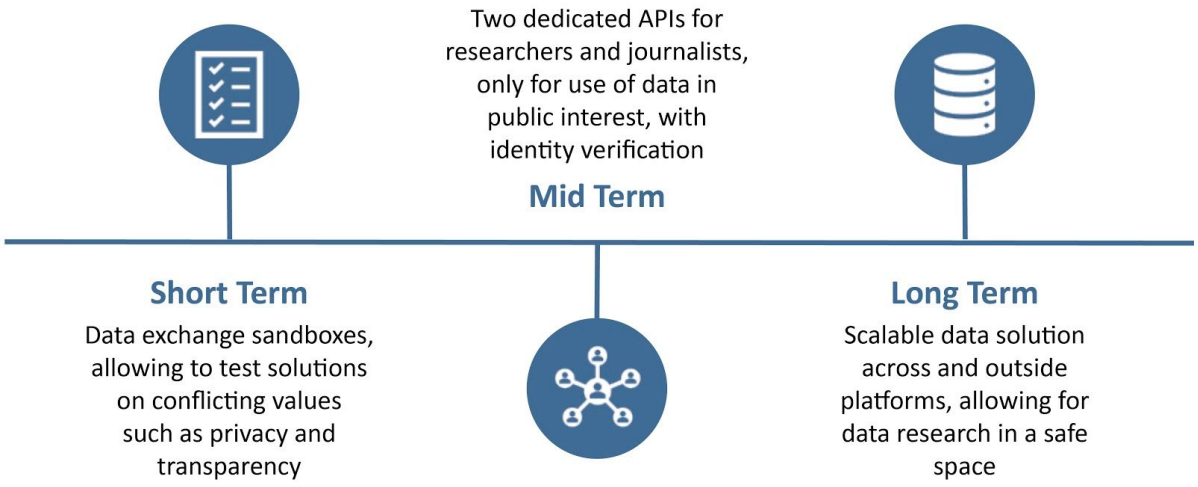
These restrictions on data access have for the most part been made in an attempt to mitigate against privacy breaches, especially in the light of the recent Cambridge Analytica scandal but also in the wake of GDPR taking effect.

However, if no data access is provided, researchers and journalists cannot detect and monitor whether harmful content circulation is taking place. Protecting privacy is not only about end-to-end encryption and protecting data but also about providing non-filtered access to controlled spaces for researchers and journalists to detect if privacy breaches are possible or actual privacy violations take place. While leaks are an issue, so is the lack of oversight that emerges when social media data cannot be analyzed by truly critical, independent scholars.

Such access is also necessary to serve other fundamental democratic values such as protecting individuals against systematic manipulation and election interventions. As the HLEG on disinformation (Buning et al, 2018) suggests, transparency and evidence-based research are needed to provide decently informed media literacy and policy initiatives.

Therefore, we propose alternative future scenarios for data exchange with platforms that will better support researchers and journalists and balance different democratic concerns and ideals. We recommend gradually implementing all scenarios and short term actions, as visualised in Figure 1.

Figure 1: Timeline for execution of the above mentioned solutions



The first step should be to test different solutions for data access in protected data exchange sandboxes to evaluate solutions on presumably conflicting values, such as transparency and privacy, and to find solutions with a higher degree of balance between these values.

The second step should be the establishment of two dedicated API-based data access solutions specifically tailored for researchers and NGOs/journalists respectively to have access in a scalable and non-discriminatory way. Among other things, the APIs should be free to use with strict validation processes, provide access to an abundance of different data (depending on status), not limited to answer specific questions, provide detailed documentation and have dedicated terms of service explicitly requiring the use of data in the public's interest rather than for commercial benefit, prohibiting problematic collaborations with companies like Cambridge Analytica and preventing the platforms from shutting down research tools using the APIs unless misconduct is detected.

The long-term step, and the most important one, is the establishment of a scalable data solution across and outside platforms, making it possible to conduct data research in a safe space adhering to highest ethical standards and without violating privacy laws. This scenario will be expanded on in a future report to be handed in on April 30, 2020.

6 References

Bastos, M. & Walker S. T. (2018, April 11). Facebook's data lockdown is a disaster for academic researchers [Online article]. Retrieved on August 20, 2019, from: https://theconversation.com/facebooks-data-lockdown-is-a-disaster-for-academic-researchers-94533

Batrinca, B. & Treleaven, P.C. (2015). Social media analytics: a survey of techniques, tools and platforms. *AI & Soc,* 30(1), 89-116

Bechmann, A. (2013). Internet profiling: The economy of data intraoperability on Facebook and Google. *MedieKultur: Journal of Media and Communication Research*, 29(55), 72-91

Bechmann, A. & Kim, J. Y. (2020, in press). Big Data: A Focus on Social Media Research Dilemmas. In R. Iphofen (Ed.) *Handbook of Research Ethics and Scientific Integrity.* Berlin: Springer.

Bechmann, A., & Lomborg, S. (2013). Mapping actor roles in social media: Different perspectives on value creation in theories of user participation. *New Media & Society*, 15(5), 765–781

Bechmann, A. & Vahlstrup, P. B. (2015). Studying Facebook and Instagram data: The Digital Footprints software. *First Monday*, 20(12)

Bruns, A., Bechmann, A., Burgess, J., Chadwick, A., Clark, L. S., Dutton, W. H., … Zimmer, M. (2018). Facebook shuts the gate after the horse has bolted, and hurts real research in the process. *Internet Policy Review*. Retrieved on March 13, 2019, from: https://policyreview.info/articles/news/facebook-shuts-gate-after-horse-has-bolted-and-hurts-real-research-process/786

Bruns, A. (2019). After the 'APIcalypse': social media platforms and their fight against critical scholarly research. *Information, Communication & Society*, DOI: 10.1080/1369118X.2019.1637447

Buning, M. et al. (2018). A Multi-dimensional Approach to Disinformation: Report of the independent high level group on fake news and online disinformation. Brussels: EU Commission.

Butakov, N., Petrov, M., Mukhina, K., Nasonov, D. & Kovalchuk, S. (2018). Unified domain-specific language for collecting and processing data of social media. *Journal of Intelligent Information Systems,* 51(2), 389-414

Chaturvedi, A (2018, July 4). WhatsApp launches research awards for social science and misinformation [Online article]. Retrieved on August 27, 2019, from: https://economictimes.indiatimes.com/tech/internet/whatsapp-launches-research-awards-for-social-science-and-misinformation/articleshow/64854666.cms?from=mdr

Chudnov, D., Kerchner, D., Sharma, A. & Wrubel, L. (2014). Technical Challenges in Developing Software to Collect Twitter Data. *The Code4Lib Journal*, 44, 1-1

Congostoa, M., Basanta-Vala, P. & Sanchez-Fernandeza, L. (2017). T-Hoarder: A framework to process Twitter data stream. *Journal of Network and Computer Applications*, 83, 28-39

Constine, J. (2018). Facebook restricts APIs, axes old Instagram platform amidst scandals [Online article]. Retrieved on August 19, 2019, from http://social.techcrunch.com/2018/04/04/facebook-instagram-apishut-down/

Domínguez, D. R., Díaz Redondo, R. P., Vilas, A. F., & Khalifa, M. B. (2017). Sensing the city with Instagram: Clustering geolocated data for outlier detection. *Expert Systems with Applications*, 78, 319–333

Driscoll, K. & Walker, S. (2014) Working Within a Black Box: Transparency in the Collection and Production of Big Twitter Data. *International Journal of Communication*, 8, 1745-1764

Dwork C. (2008) Differential Privacy: A Survey of Results. In Agrawal M., Du D., Duan Z., Li A. (Eds.) *Theory and Applications of Models of Computation, 5th International Conference, TAMC 2008 Xi'an, China, April 25-29, 2008 Proceedings*. Berlin: Springer.

Efron, S. E. & Ravid, R. (2019) *Writing the Literature Review: A Practical Guide.* New York, NY: The Guilford Press.

Falck, B. (2018, June 28). Providing more transparency around advertising on Twitter [Online article]. Retrieved on August 13, 2019, from: https://blog.twitter.com/en_us/topics/company/2018/Providing-More-Transparency-Around-Advertising-on-Twitter.html

Felt, M. (2016). Social media and the social sciences: How researchers employ Big Data analytics. *Big Data & Society*, 3(1), 1-15

Franzke, A. S., Bechmann, A., Ess, C. M., & Zimmer, M. (2020). Internet Research: Ethical Guidelines 3.0. AoIR (The International Association of Internet Researchers).

Freelon, D. (2018). Computational Research in the Post-API Age. *Political Communication*, 35:4, 665-668

Gaffney, D., & Matias, J. N. (2018). Caveat emptor, computational social science: Large-scale missing data in a widely-published Reddit corpus. *PLOS ONE*, 13(7), e0200162

Gayo-Avello, D. (2013). A Meta-Analysis of State-of-the- Art Electoral Prediction From Twitter Data. *Social Science Computer Review*, 31(6), 649-679

Gummadi, R. (2018, January 30). Instagram Graph API Launches and Instagram API Platform Deprecation | Facebook for developers. Retrieved August 14, 2019, from: https://newsroom.fb.com/news/2018/04/restricting-data-access/

Halford, S., Weal, M., Tinati, R., Carr, L., & Pope, C. (2018). Understanding the production and circulation of social media data: Towards methodological principles and praxis. *New Media & Society*, 20(9), 3341–3358

Hare, K (2017, March 8). Since Facebook made Crowdtangle free, more than 150 local newsrooms have adopted it [Online article]. Retrieved on August 13, 2019, from: https://www.poynter.org/tech-tools/2017/since-facebook-made-crowdtangle-free-more-than-150-local-newsrooms-have-adopted-it/

How are ads about social issues, elections or politics identified on Instagram? (n.d.). Retrieved on August 27, 2019, from: https://help.instagram.com/118613625676963

King, G., & Persily, N. (2018). *A New Model for Industry-Academic Partnerships*. Retrieved on August 16, 2019, from: https://gking.harvard.edu/partnerships

King, G., & Persily, N. (2019, April 28). First Grants Announced for Independent Research on Social Media's Impact on Democracy Using Facebook Data [Online article]. Retrieved on August 16, 2019, from: https://socialscience.one/blog/first-grants-announced-independent-research-social-media's-impact-democracy

Kotsios, A., Magnani, M., Rossi, L., Shklovski, I., & Vega, D. (2019). An Analysis of the Consequences of the General Data Protection Regulation (GDPR) on Social Network Research. arXiv preprint arXiv:1903.03196.

Lienemann, B. A., Unger, J. B., Cruz, T. B., & Chu, K.-H. (2017). Methods for Coding Tobacco-Related Twitter Data: A Systematic Review. *Journal of Medical Internet Research*, 19(3), e91.

Lomborg, S., & Bechmann, A. (2014). Using APIs for Data Collection on Social Media. *The Information Society*, 30(4), 256–265

Malik, H., & Tian, Z. (2017). A Framework for Collecting YouTube Meta-Data. *Procedia Computer Science*, 113, 194–201

Meredith, S. (2018, April 10). Facebook-Cambridge Analytica: A timeline of the data hijacking scandal [Online article]. Retrieved on August 13, 2019, from: https://www.cnbc.com/2018/04/10/facebook-cambridge-analytica-a-timeline-of-the-data-hijacking-scandal.html

Mittos, A., Malin, B. & De Cristofaro, E. (2019). Systematizing Genome Privacy Research: A Privacy-Enhancing Technologies Perspective. *Proceedings on Privacy Enhancing Technologies*, 1, 87–107

Morstatter, F., & Liu, H. (2017). Discovering, assessing, and mitigating data bias in social media. *Online Social Networks and Media*, 1, 1–13

Mozilla. (2019a, March 27). Facebook and Google: This is What an Effective Ad Archive API Looks Like [Blog post]. Retrieved on August 13, 2019, from:https://blog.mozilla.org/blog/2019/03/27/facebook-and-google-this-is-what-an-effective-ad-archive-api-looks-like/

Mozilla. (2019b, April 29). Facebook's Ad Archive API is Inadequate [Blog post]. Retrieved on August 13, 2019, from: https://blog.mozilla.org/blog/2019/04/29/facebooks-ad-archive-api-is-inadequate/

Office of the French Ambassador for Digital Affairs. ©2018. Disencyclopedia (https://disinfo.quaidorsay.fr/encyclopedia). Licensed under CC BY-SA (https://creativecommons.org/licenses/by-sa/2.0/legalcode)

Oussalah, M., Bhat, F., Challis, K., & Schnier, T. (2013). A software architecture for Twitter collection, search and geolocation services. *Knowledge-Based Systems*, 37, 105–120

Raffi, K. (2014, February 5). Introducing Twitter Data Grants [Blog post]. Retrieved on August 27, 2019, from: https://blog.twitter.com/engineering/en_us/a/2014/introducing-twitter-data-grants.html

Revision History. (n.d.). Retrieved on August 27, 2019, from: https://developers.google.com/youtube/v3/revision_history

Rieder, B., Abdulla, R., Poell, T., Woltering, R., & Zack, L. (2015). Data critique and analytical opportunities for very large Facebook Pages: Lessons learned from exploring "We are all Khaled Said": *Big Data & Society*, July-December, 1-22

Roth, Y. & Johnson, R. (2018, July 24). New developer requirements to protect our platform [Blog post]. Retrieved on August 22, 2019, from: https://blog.twitter.com/developer/en_us/topics/tools/2018/new-developer-requirements-to-protect-our-platform.html

Schroepfer, M. (2018, April 4). An Update on Our Plans to Restrict Data Access on Facebook | Facebook Newsroom. Retrieved August 14, 2019, from: https://newsroom.fb.com/news/2018/04/restricting-data-access/

Shah, C. (2010). Supporting Research Data Collection from YouTube with TubeKit. *Journal of Information Technology & Politics*, 7(2–3), 226–240

Silverman, G. (2019, August 27). Exclusive: Funders Have Given Facebook A Deadline To Share Data With Researchers Or They're Pulling Out. [Online article]. Retrieved on August 30, 2019, from: https://www.buzzfeednews.com/article/craigsilverman/funders-are-ready-to-pull-out-of-facebooks-academic-data

Smith, M. (2018, August 15) Introducing a new transparency report for political ads [Blog post]. Retrieved on August 27, 2019, from: https://www.blog.google/technology/ads/introducing-new-transparency-report-political-ads/

Social Science One. (2019, May 3). Request for Proposals for Fast Access to CrowdTangle and Ad Library Data. Retrieved August 16, 2019, from: https://business.twitter.com/en/help/ads-policies/restricted-content-policies/political-content.html

Statista. (2019). Most famous social network sites worldwide, ranked by number of active users (in millions). In *Statista - The Statistics Portal*. Retrieved March 1, 2019, from: https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/

Statsbiblioteket. (n.d.). Databaser. Retrieved on March 12, 2019, from https://www.statsbiblioteket.dk/databaseliste/databaser_view?start_letter=ALL

Stieglitz, S., Dang-Xuan, L., Bruns, A., & Neuberger, C. (2014). Social Media Analytics. *Business & Information Systems Engineering*, 6(2), 89–96

Stieglitz, S., Mirbabaie, M., Ross, B., & Neuberger, C. (2018). Social media analytics – Challenges in topic discovery, data collection, and data preparation. *International Journal of Information Management*, 39, 156–168

Sullivan, M. (2018, March 28). Facebook expands ad transparency beyond politics: Here's what's new [Online article]. Retrieved on August 13, 2019, from: https://www.fastcompany.com/90326809/facebook-expands-archive-ad-library-for-political-ads-heres-whats-new

Twitter. (n.d.). Political Content. Retrieved August 16, 2019, from: https://business.twitter.com/en/help/ads-policies/restricted-content-policies/political-content.html

Twitter Inc. (2019, February 19). Expanding transparency around political ads on Twitter [Blog post]. Retrieved August 13, 2019, from: https://blog.twitter.com/en_us/topics/company/2019/transparency-political-ads.html

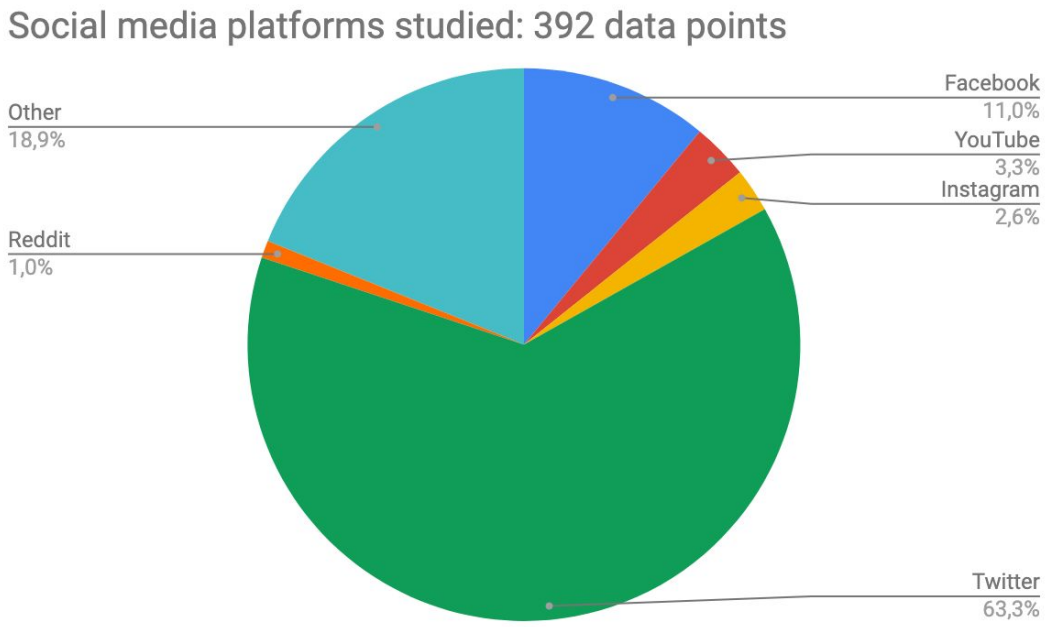
WhatsApp. (n.d.). WhatsApp Research Awards for Social Science and Misinformation [Blog post]. Retrieved August 22, 2019, from: https://www.whatsapp.com/research/awards/

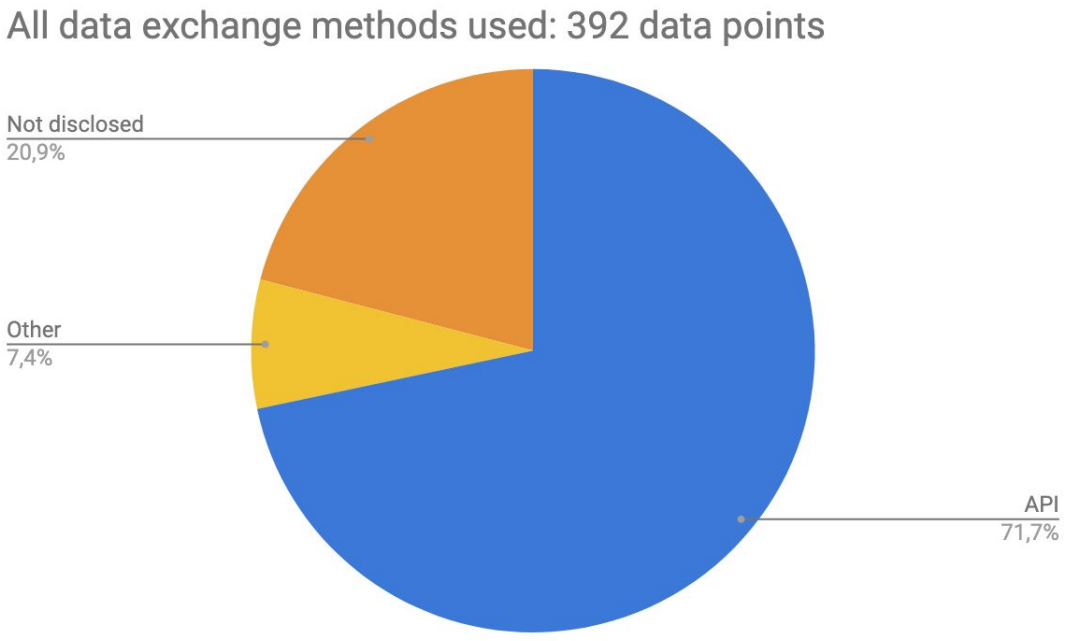
Wheeler, J. (2018). Mining the First 100 Days: Human and Data Ethics in Twitter research. *Journal of Librarianship and Scholarly Communication*, 6(2), eP2235

7 Appendix

7.1 Results from literature search

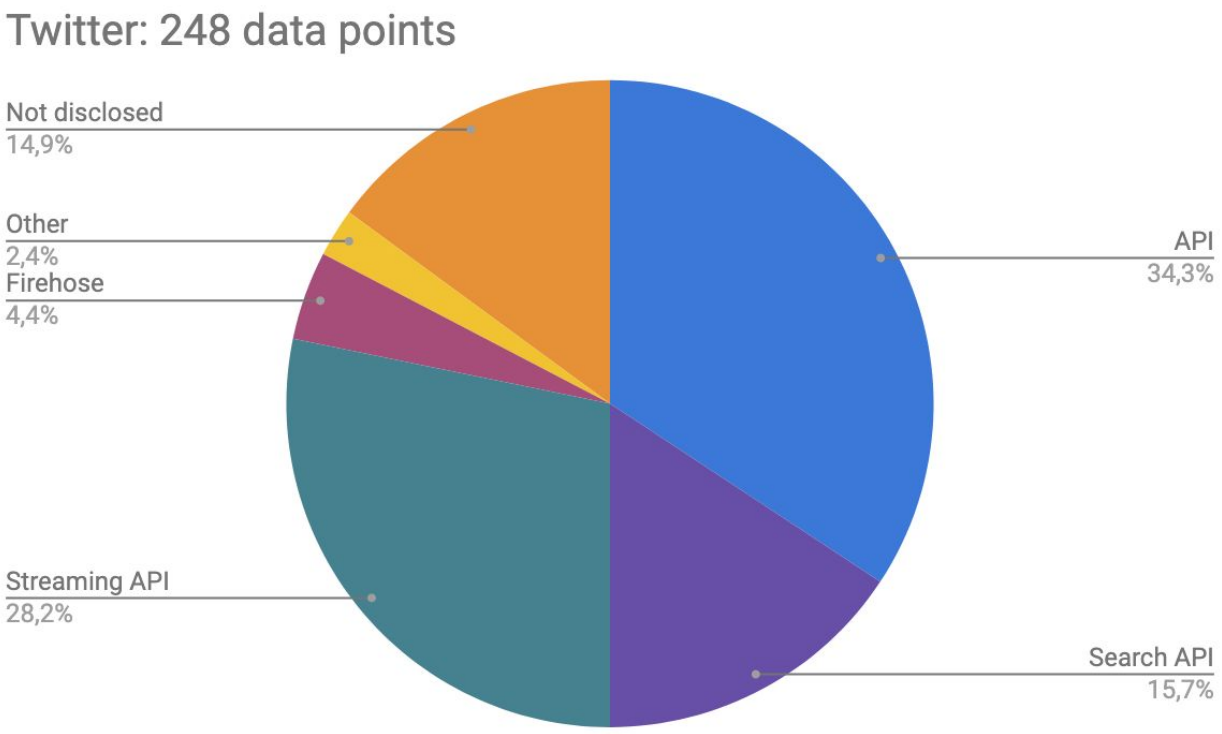
Results from categorising the 434 articles from the literature search in regards to the data exchange method used and the social media platform studied. 57 articles that did not study specific social media but were rather methodological papers on data exchange or literature reviews were still included in the literature review but are categorized as *Other* and are not part of the following charts. Some articles studied several different social media platforms or used different methods for data exchange and were placed in multiple categories. Hence, the number of data points in the following charts being 392, not 434.





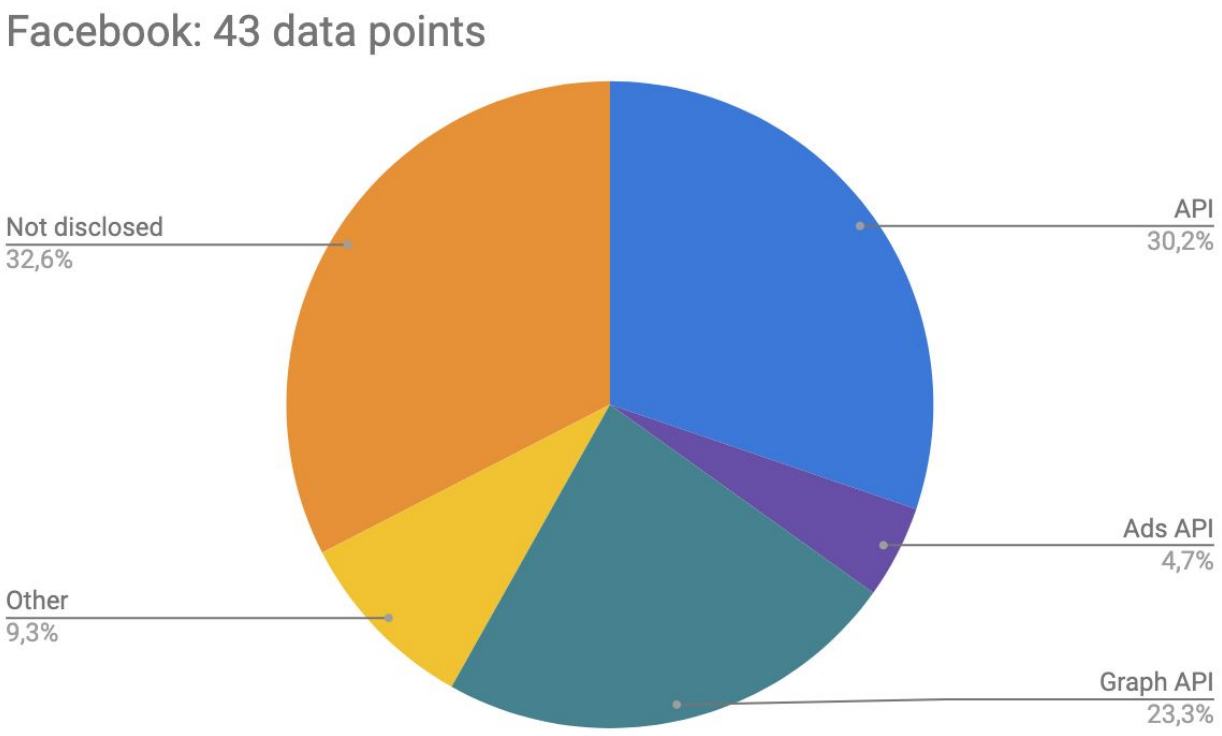
7.1.1 Twitter

Twitter	
API	85
Search API	39
Streaming API	70
Firehose	11
Other	6
Not disclosed	39
TOTAL	248



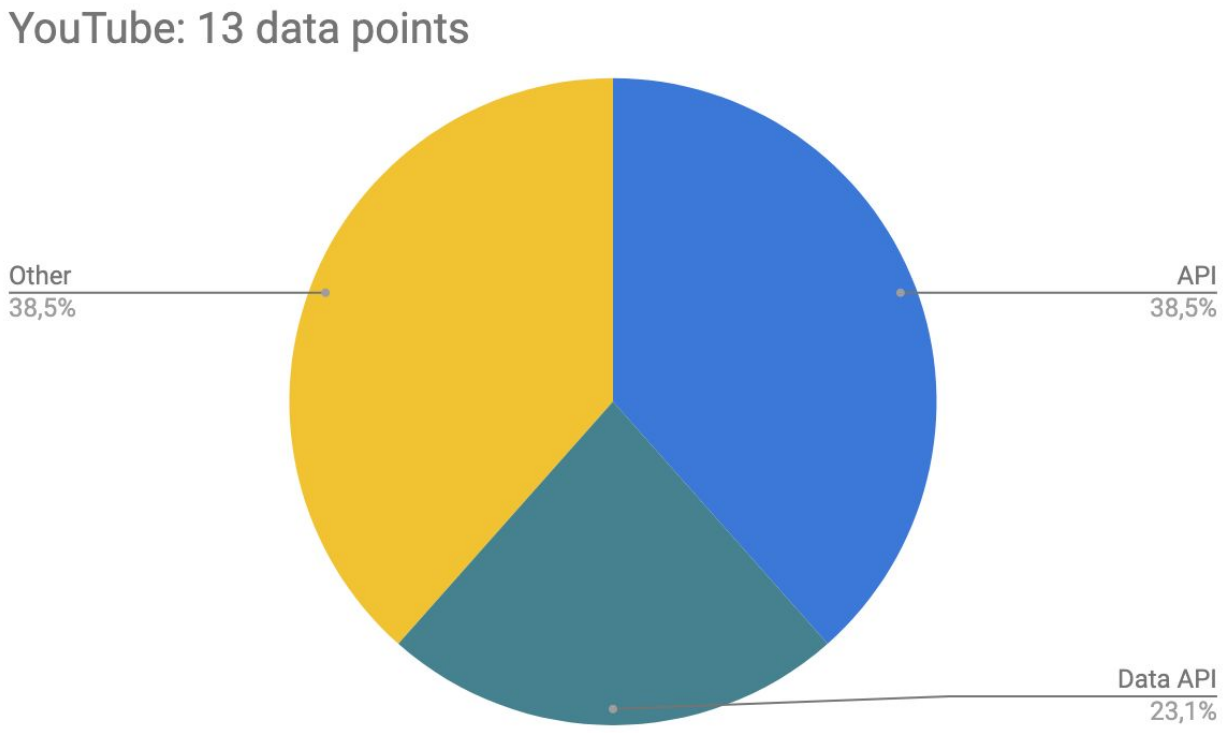
7.1.2 Facebook

Facebook	
API	13
Ads API	2
Graph API	10
Other	4
Not disclosed	14
TOTAL	43



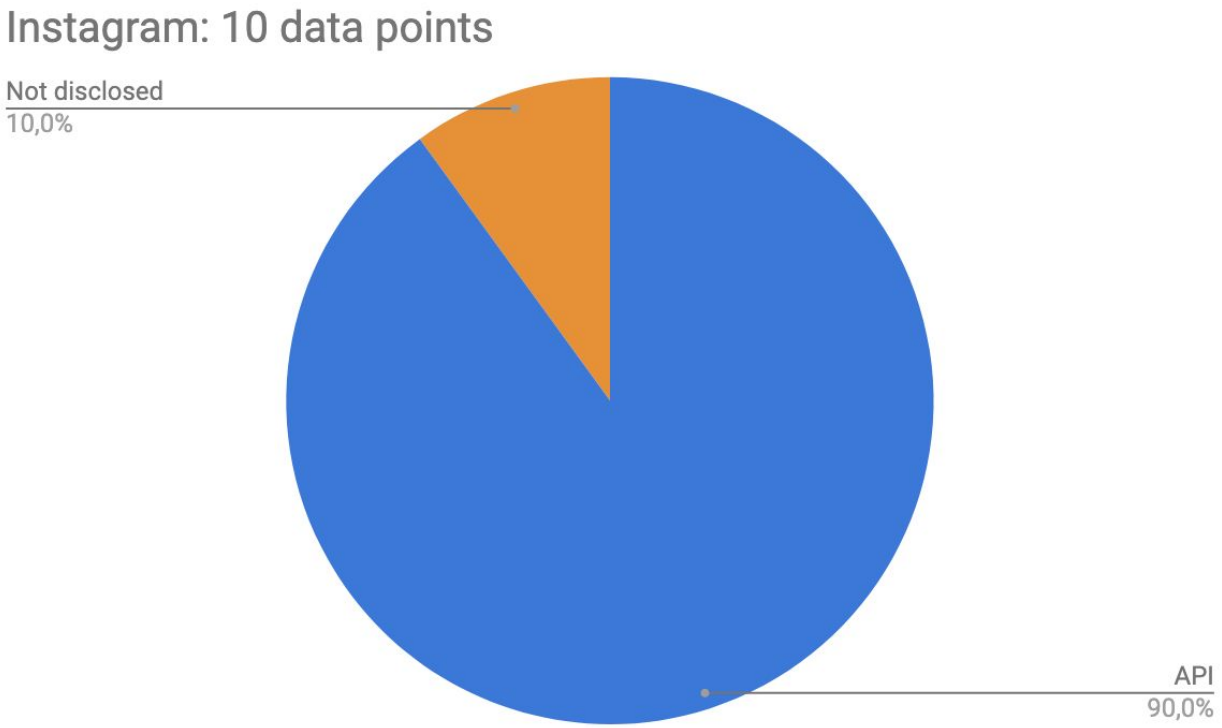
7.1.3 YouTube

YouTube	
API	5
Data API	3
Other	5
TOTAL	13



7.1.4 Instagram

Instagram	
API	9
Not disclosed	1
TOTAL	10

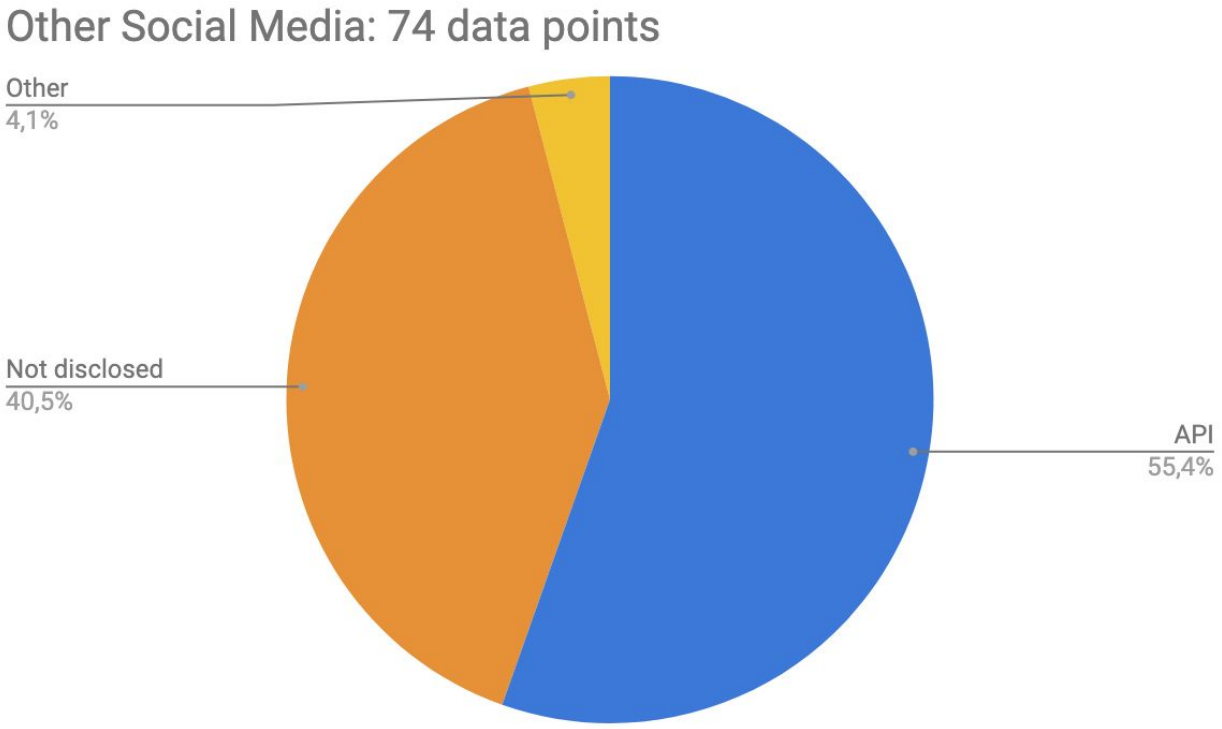


7.1.5 Reddit

Reddit	
API	4
TOTAL	4

7.1.6 Other social media platforms

Other social media platforms	
API	41
Not disclosed	30
Other	3
TOTAL	74



7.2 Meetings with social media representatives

Interview between Anja Bechmann & Facebook on fact-checking and research challenges
Organizer: Anja Bechmann, Aarhus University and Royal Flemish Academy of Belgium for Science and the Arts
Time: June 11, 2019
Participants: Representatives from Facebook's Election Research Commission and EU Affairs at Facebook
Description: Interview on the topics of disinformation, fake news and information disorder and Facebook's work in combating these issues with a focus on research challenges in regards to data access.

NOTE: Invitations were also sent to Google representatives but they did not respond.

Interview between Anja Bechmann & Facebook's Brussels office on fact-checking and research challenges
Organizer: Anja Bechmann, Aarhus University and Royal Flemish Academy of Belgium for Science and the Arts
Time: June 17, 2019
Participants: Representatives from EU Affairs at Facebook
Description: Interview on the topics of disinformation, fake news and information disorder and Facebook's work in combating these issues with a focus on research challenges in regards to data access.

Meeting on access to online platforms' data
Organizer: The European Commission
Time: June 18, 2019
Participants: Meeting between the European Commission representatives, SOMA representatives, and representatives from Social Science One
Description: The purpose is to discuss access to online platforms' data and share experiences with this from a European perspective and possible solutions going forward.

Facebook moderation workshop in Berlin - Oversight board discussions
Organizer: Facebook
Time: June 24-25, 2019
Participants: Representatives from Facebook's moderation team, Facebook's Global Affairs and Governance team, Facebook's Strategic Initiatives team, invited participants from NGOs, research, policy makers and other key stakeholders (app. 40 participants excl. Facebook representatives)
Description: Discuss and review the construction of an independent oversight board to handle disputes in moderation decisions made by the Facebook moderation team based on community guidelines.

Unconference on fighting disinformation: Research hackathon with Facebook and Social Science One
Organizer: The Office of the French Ambassador for Digital Affairs
Time: June 28, 2019
Participants: Members of Facebook's Election Research Commission team, Facebook's data science team, and selected european researchers granted access through Social Science One or with an extended interest in data exchange/applying
Description: The main goal for studying will be to identify practical hypotheses that can be investigated using available data and tools, and prototype ways to test them. Members of Facebook's Election Research Commission team will join the hackathon to introduce the most recent data and tooling available from Facebook, via the company's collaboration with Social Science One and the Social Science Research Council. This will allow participants to generate more likely to be answerable research hypotheses and includes, by order of availability:

1. The Ads Library, which is available to all ID-verified researchers and developers.
2. The CrowdTangle API, which will be available to all participants during the hackathon and to registered researchers after it.

3. The data in the Research Tool, which will be available to all participants during the event only, is synthesized and sampled version of the Social Science One URL shared data set. This dataset will allow researchers to understand the shape of the data available and its potential, as well as understand how to conduct research in a differentially private system. In order to test the hypotheses, the full data will be made available to researchers who are accepted into the Social Science One program through the Social Science Research Council.

Discussion on challenges and solutions related to research into digital political ads in the EU
Organizer: The Mozilla Foundation
Time: July 2, 2019
Participants: Classified
Description: In the run-up to the European Parliament elections Facebook, Google and Twitter implemented their commitments in the EU Code of Practice on Disinformation to increase political ad transparency, including labelling of political ads and creating publicly accessible political ads libraries that are searchable through appropriate interfaces (APIs). However, these measures have been subject to criticism by both the European Commission and independent researchers as being not fit for purpose. In March, the Mozilla Foundation and a cohort of independent researchers published five guidelines that these APIs must meet in order to truly support election influence monitoring and independent research. In response to this initiative, VP Ansip of the European Commission has invited the Foundation to organise a meeting with signatories of the open letter to better understand what categories of data and levels of data aggregation would enable relevant research, consistent with data protection rules, and to clarify what is realistically achievable. Hence, our discussions will focus on the following questions:

1. Which design flaws do the current AdApi's and ad archives have?
2. What data do researchers actually need to in order to better understand disinformation campaigns and monitor the implementation of the Code of Practice?
3. To what extent do these categories of data pose any data protection challenges?
4. What are some follow-up measures that could help to achieve greater transparency around political advertising in the EU and beyond? (incl. potential regulatory action, new models/institutions for responsible data sharing)?