Digital methods in a post-API environment

Jessamy Perriam (https://orcid.org/0000-0002-6362-8634)

The Open University, UK

Andreas Birkbak (https://orcid.org/0000-0001-9283-1677)

Aalborg University, Denmark

Andy Freeman (https://orcid.org/0000-0003-0851-202X)

Goldsmiths, University of London, UK

Correspondence details: Jessamy Perriam, Sociology Discipline, Faculty of Arts and Social Sciences, The Open University, Walton Hall, Milton Keynes, UK, MK7 6AA jess.perriam@open.ac.uk.

The Version of Record of this manuscript has been published and is available in the International Journal of Social Research Methodology, 25th October 2019, https://www.tandfonline.com/doi/full/10.1080/13645579.2019.1682840

Abstract

Qualitative and mixed methods digital social research often relies on gathering and storing social media data through the use of APIs (Application Programming Interfaces). In past years this has been relatively simple, with academic developers and researchers using APIs to access data and produce visualisations and analysis of social networks and issues. In recent years, API access has become increasingly restricted and regulated by corporations at the helm of social media networks. Facebook (the corporation) has restricted academic research access to Facebook

(the social media platform) along with Instagram (a Facebook-owned social media platform). Instead, they have allowed access to sources where monetisation can easily occur, in particular, marketers and advertisers. This leaves academic researchers of digital social life in a difficult situation where API related research has been curtailed. In this paper we describe some rationales and methodologies for using APIs in social research. We then introduce some of the major events in academic API use that have led to the prohibitive situation researchers now find themselves in. Finally, we discuss the methodological and ethical issues this produces for researchers and, suggest some possible steps forward for API related research.

Keywords

Digital methods, APIs, web scraping, ethics, Facebook, Netvizz, Twitter

Introduction

Digital methods, in the sense employed in this article, are a set of internet methods in that what they have in common is the ambition to make the most of the new data formats that arise with the wide uptake of the internet in social life. Digital methods researchers develop new tools inspired by internet technologies in order to be able to treat these new data formats in methodologically innovative ways (Rogers 2013). Digital methods are characterised by not keeping qualitative and quantitative methods separate. The hyperlink, for instance, affords a text to be available for qualitative interpretation while at the same time placed in a structure of link networks

that can be analysed with quantitative approaches. For these reasons, we believe digital methods constitute a productive challenge to other social science methods, including digital ones, and indeed a way for social research methods to develop in combination with digital social practices overall.

Digital methods researchers have highlighted the need to 'follow the medium' when collecting digital data in an age of rapidly shifting internet platforms, but in recent years it has become clear that these increasingly 'big business' platforms do not always wish to be followed. The so-called Application Programming Interfaces (APIs) that have provided social researchers with data access are shutting themselves off from academic use. After introducing what digital methods are and how APIs work, the main focus of the paper is to unpack what becomes of digital methods research in an age where API access can be retracted on short notice or where changes to API structures make them too unstable for longitudinal study. We use Facebook and Twitter as empirical examples of what has happened more specifically and relate that to ongoing research efforts. Finally, we suggest some pitfalls and potential ways forward for digital social research in an era of increasingly restrictive or even closing APIs.

How are APIs relevant to digital social research?

In the earlier years of qualitative digital research, the focus centred on studies of Internet cultures or digitised data. Richard Rogers (2009) posited that internet researchers could and should focus on and interrogate the internet itself, rather than the culture that formed around it or transposed itself onto it. He suggested that the

distinction between online and offline social life is often unhelpful. This dichotomy has similarly been described over the years as ‘virtual and real life’ or ‘online and offline’. Rogers argues that the data that populates the internet are ontologically distinct; that is, there are digitised data that have come from offline sources, and ‘natively digital’ data that originates in digital settings. Rogers advocates studying ‘natively digital’ data to observe and understand how social issues originate and circulate online. He coins the phrase ‘digital methods’ to describe research methods that take ‘natively digital’ data as objects of study.

As a result, Rogers goes on to suggest that social science research methods that have been produced and used before the common uptake of internet use may not be best suited to gathering and analysing ‘natively digital’ data. He suggests that ‘natively digital’ data require ‘natively digital’ research tools and methods. Much of Rogers’ work along with that of his colleagues at the Digital Methods Initiative (Marres and Weltevrede 2013, Marres and Rogers 2005) has focused on researching web objects, such as hyperlinks and user profiles, to conduct analysis of the social relations between different actors.

In particular, Marres and Rogers (2005) conduct and describe issue mapping as a digital method to observe who is linking to whom about a particular social issue. Rogers and Marres (2000) describe mapping actors in climate change debates based on which organisations link to others. Similar to a social network analysis approach, he describes how Greenpeace links to government organisations on their website, but this is not reciprocated. Conversely, a pharmaceutical company links to

Greenpeace on their website, but Greenpeace does not reciprocate. Issue mapping - while technically a pre-digital method[1] - allows social researchers to analyse networks of actors using natively digital data.

But what are the practicalities of natively digital methods? How could Rogers and Marres research and produce their findings on online climate change discussions? The answers lie in the tools created to facilitate the research. The climate change issue map was produced using a tool called Issue Crawler. Once provided with a website URL, the tool 'crawls' the website to look for links. From there, it crawls those linked websites, and so on, until the crawler reaches the limit specified by the researcher. The output can then be plotted on a network graph that shows the nodes and edges of those taking part in the debate. The hope is that the researcher can see where the issue centres, along with who is taking part, and who wishes to take part, in the discourse.

Issue Crawler has been in use for some time now and as Rogers points out, the Internet changes its attributes often and in circumstances outside of the researcher's control. It is the researcher's job to respond to these shifts in the medium. Rogers (2009) goes so far as to say we should 'follow the medium'. This approach worked well through the advent of Web 2.0 where early versions of social media platforms provided researchers with natively digital data to work with. Researchers and

[1] [1] Issue mapping is not a natively digital method. It has also been deployed as a method prior to the use of digital tools. Issue mapping has been particularly used in Science and Technology Studies projects that make use of Actor-Network Theory as a methodological framework (Callon 1986).

technologists worked together to produce tools such as ElFriendo (Rogers 2009), Netvizz (Rieder 2013) and T-CAT (Borra and Rieder 2014) to investigate networks and social relations occurring on MySpace, Facebook and Twitter, respectively. This was done by using preferences, userIDs and hashtags as the objects of research.

APIs are central to such efforts because they make it possible for the social media company to pass data from the platform to the researcher, usually in the form of a CSV or JSON file. This allows the researcher to conduct analysis or produce visualisations based on the data[2]. Calls to an API return data that represents a snapshot of the state of the system at any one moment in time. There are also cases where the 'liveness' of the API allows another level of investigation through repeated querying of the same endpoints to spot changes in the data over time that cannot be ascertained by a single query alone. This is necessary because although it is possible architecturally to maintain a full audit trail of every field, this has substantial costs attached. Losing the history of a datum is therefore often a commercial decision. Researchers can compensate by storing this data themselves and even analyse its use in real time. An example of this is the Facebook API URL endpoint (Facebook nd.a, Facebook nd.b) which allows the researcher to watch the engagement levels of a URL in real time.

[2] It should be recognized from the outset that accessing social media data for academic purposes is not without critique. In the Cambridge Analytica scandal, researchers gained access to the Facebook API (and therefore, Facebook data) under the then-legitimate pretence of academic research and passed on this data to actors who wanted to target specific users for political gain (Cadwalladr and Graham-Harrison 2018). While this may appear as an isolated case, we suggest that robust research ethics must always address the collection, use and disposal of social media data.

Until the last five years, researchers were able to analyse many social media platforms, usually on the proviso that they held an account and could gain developer access. Today, in order to access social media APIs of the likes of Facebook and Instagram, academic technologists (also referred to as developers) are required to submit applications outlining their purpose for using the API. This application process filters out those users who cannot easily provide a revenue stream for the platform or those who may critique the platform as part of their analysis. In the case of accessing the Instagram API from 2016 onwards, it is codified in the developer guidelines that the API must only be accessed to create tools and analytics for marketing or advertising purposes. Public, non-profitable research is excluded from Instagram's acceptable use cases. Companies may also place restrictions on the publishing of data used in a study thus making the study unreproducible by other researchers without again applying for access to the data set. If such access is not easily obtainable, the verification or falsification of results then becomes problematic. The same applies to studies that are produced by researchers within the companies themselves, using data not easily available to external researchers.

The reason why we see so many more academic studies of Twitter in comparison to Facebook-owned platforms is partly due to these access restrictions. This is not to say that academics face wholesale exclusion from researching Facebook-owned platforms. Instagram and Whatsapp have advertised funding opportunities for academic researchers interested in a mandated set of areas. The problem with these

funding opportunities arise around the independence and reproducibility of the research findings.

Returning to Rogers' call to 'follow the medium', what are we to do when the medium does not wish to be followed? There are a few actions available: rethink our reliance on APIs, subvert the rules or, route around the API. There are means of gathering social media data without the use of an API, but this is technically out of reach for many researchers. Additionally, gaining Ethics Review Board approval for this course of action would be very difficult at some institutions. What is ethical for a computer scientist in Denmark may not be considered ethical for a sociologist in Britain due to different institutional ecologies and histories. However, if there are no researchers attempting to subvert API restrictions, the academic community risks omitting Facebook and its stable of social media platforms from hands-on scrutiny, leaving more distanced academic critiques as the only option. In the next section, we illustrate what is at stake by looking at concrete examples of earlier and ongoing API-related research.

History of API research: The case of Netvizz

Shifts in API access with consequences for social research is by no means a new thing. Back in August 2009, Yahoo closed some of their search APIs after having sold off their search business to Microsoft[3]. APIs providing access to Yahoo functionalities such as term extraction had enabled at least 33 different 'mashup'

[3] In this section we draw on the important work of digital methods API pioneer Bernard Rieder as it has been recorded through his blog, thepoliticsofsystems.net.

tools, some of which served the purpose of academic research (Rieder 2009). Already at this relatively early point in time, digital methods scholar Bernard Rieder discussed the trade-offs of building tools that rely heavily on APIs controlled by private corporations:

> ”If service providers can close APIs at will, developers might hesitate when deciding whether to put in the necessary coding hours to built the latest mashup. But it is mashups that over the last years have really explored many of the directions left blank by “pure” applications. This creative force should be cherished and I wonder if there may be a need for something similar to creative commons for APIs – a legal construct that gives at least some basic rights to mashup developers…” (ibid.)

Despite already being aware of the precarious nature of API-based digital methods, Rieder went on to produce what would become one of the more well-known API-interface tools, Netvizz, in March 2010 (Rieder 2010). Netvizz was a Facebook app that allowed academics to extract Facebook data in a systematic way for research purposes (Rieder 2013). Briefly put, Netvizz let social science and humanities scholars with no coding experience download spreadsheets and graph files of not just their own network of Facebook friends, but also systematically collect data from Facebook groups and Facebook pages. Rieder remarks he ”was quite amazed how much data a third-party app could actually get from the platform” in the early days (Rieder 2010). Indeed, many researchers started to show interest in Netvizz, something which the currently 418 Google Scholar citations of Rieder’s paper about the tool attest to (Rieder 2013).

The Netvizz application is a useful case study when it comes to understanding the implications of API changes for qualitative and mixed methods digital social research due to its popularity among the research community and the fact that Facebook has changed its API multiple times since the beginning of Netvizz in 2010 - and Bernhard Rieder has been trying to play catch-up while documenting his experiences.

2012 and 2013 were big years for Netvizz in the sense that many of the core functionalities that researchers find interesting were added: Facebook page analysis and 'bipartite' graph files featuring both users and posts. The Netvizz app reached 60.000 unique users (Rieder 2015). In early 2015, the first major road block appeared: Facebook informed Rieder that the Netvizz app was going to be suspended among other things because it allowed for export of friend data. With the introduction of the Facebook API version 2.2, this data was indeed no longer available (ibid.). A cut-down version of Netvizz only able to export data from groups and pages was able to continue running until very recently.

In 2018, likely as a response to the Cambridge Analytica scandal, another major Facebook API update was introduced. With this, all existing apps had to be reviewed again, and it became clear that Facebook would no longer allow apps whose primary purpose was data download (Rieder 2018). One interpretation of this is that Facebook does not intend to support independent research, not even not-for-profit academic research supposed to serve the public good. At the time of revising this article (August 2019), Rieder tweeted that Netvizz is no longer

available for public use. The fact that it managed to function and be publicly available for so long is testament to Rieder's observation that Facebook's app reviews are highly automated and generic, clearly trying to tackle a very large volume of apps (and users) and perhaps not always receiving specific follow-up (ibid.).

Case stories from the TANTlab

The API changes Facebook make have very concrete consequences for digital social research. In April 2018, Facebook severely limited the API access to Facebook Groups and Events (Facebook 2018a). This is perhaps the most dramatic change for digital social research to date given that researchers have been relying on data from groups and events to understand everything from political mobilization to cultural consumption in a digital age. In TANTlab at Aalborg University in Copenhagen, Denmark, for instance, researchers used anonymized information about which Facebook events had overlapping attendance in order to map spheres of cultural life in the Danish capital (Abildgaard et al. 2017)[4]. To what extent are those who like to attend events at public libraries the same as those who go to classical music concerts? Such questions were of interest not just to researchers trying to understand cultural patterns, but also to public institutions such as the Royal Danish Theatre who are trying to understand their audience and identify new groups to attract (Munk et al. 2019). With the API changes shutting off access to who have RSVP'ed or left a

post or comment on a Facebook event page, this research has been made virtually impossible.

As a result of the new API restrictions, the first months of 2018 saw a TANTlab effort to collect a large dataset of public Facebook pages located in Denmark before the API changes hit. This was done in order to be able to critically evaluate the consequences of the API changes as they happened, and in order to have the most interesting data set possible before the restrictions were fully implemented (TANTlab 2019, Munk and Olesen, forthcoming). Still, such social media data sets quickly become outdated and researchers now no longer have the opportunity to update them.

Other projects have been interested in more qualitative analyses of the interactions taking place through posts and comments on Facebook pages and in Facebook groups. For example, Birkbak (2012) contrasted and compared how two Facebook groups became home to very different understandings of a 2010 snowstorm on the Danish island Bornholm, including its severity and the need for emergency help. Such work is greatly assisted by systematic access to all posts and comments via the Facebook API, but this is increasingly difficult. The Netvizz app removed the *Group* module in July 2018 in preparation of the review of the app by the Facebook corporation.

[4] One of the authors (Birkbak) is a member of the TANTlab and the accounts here are based on his experiences including those of other members of the lab elicited through a group

Access to posts and comments on Facebook *pages* is easier, and was still available until recently via Netvizz for those not able to script their own API call. But it is clear that Facebook intends the API access to serve businesses looking to maximize their business intelligence via Facebook (and thus their expenditure on Facebook ads). For example, the documentation explains that the intended 'common usage' is among other things to "provide aggregated, anonymized public content for competitive benchmarking, understanding what content resonates with people and identifying best practices" and to "provide tools to understand how a business's own brand, products, or services are being publicly talked about" (Facebook 2018b).

The focus of Facebook's API access is on businesses understanding themselves and their markets better, and there is no mention of public research (or other nonprofit activities for that matter). The API documentation has made it increasingly clear that Facebook's objective is to make it easier for (commercial) users to manage their own pages and groups and for commercial developers to program third party apps that use Facebook content in creative ways, while not to allowing for data extraction even when this is done for the purposes of academic research.

In 2017, TANTlab researchers Anders Kristian Munk and Asger Gehrt Olesen were asking how the controversy surrounding the HPV vaccine expressed itself on Facebook. The effort was part of a core interest in TANTlab in how public

interview taking place 17th December 2018.

controversies around science and technology can be studied with digital methods. Anders and Asger relied on their newly-won ability to write Python code that retrieved data through the Facebook API in order to collect data from Facebook groups and pages related to the HPV vaccine. However, by the end of 2017 they became aware that Facebook had announced so-called 'breaking changes' to their API by 5th February 2018. Breaking changes are different from other API changes in the sense that they break the functionality of existing apps if these apps do not adapt to the new API rules and requirements. In other cases, apps can simply choose to use an older version of the Facebook API in order to continue running, but this is not the case with breaking changes.

The biggest shift introduced by 5th February 2018 would be the deactivation of access to user data in the sense of being able to tie posts or comments on a Facebook page to an individual user and that user's activities elsewhere. For researchers in the TANTlab, lacking data on the level of individual users would severely impact the analytical interest of the Facebook data, so a project was launched to collect data in advance of the 5th February 2018, both in relation to the HPV controversy, but now also in relation to a much broader interest into what characterizes public Facebook activity in Denmark (Munk and Olesen, forthcoming).

As another TANTlab researcher, Anders Koed Madsen, expressed in our interview with him, what was so valuable for research purposes was the sort of "granularity" offered by the Facebook API that allowed researchers to explore patterns of

interactions on the level of users. As stated in the TANTlab data policy, this does not have to involve profiling individual users (TANTlab 2018). The interest is rather in being able to situate a given comment or statement in relation to activity elsewhere. For example, the degree of user overlap between different Facebook pages can say something about cultural consumption patterns without building a dataset where individuals can be identified.

These analytical moves rely very much on the ability to move back and forth between quantitative and qualitative ways of viewing the data (Venturini and Rogers 2019). In a new explorative project, Madsen seeks to identify physical locations that serve as meeting points between people who do not normally come in touch with each other. To give an indication of this, he looks at Facebook event data to determine whether users with diverging political loyalties are attending events at the same location (Madsen 2018). This enables him to use Facebook API data to inform city planning. But the changes made by Facebook prevents such analysis by blocking access to the user level. Even if posts and comments are still accessible, the content can no longer be systematically situated in relation to specific users. This makes it much harder to characterize a community of users, not to mention trying to detect bot activity.

As such, one thing that is at risk due to the current API changes is the ability to easily move between qualitative and quantitative moments in the analysis, leaving social research with the usual two camps of either doing hand-held qualitative

studies of social media through manual observation and interviews, or doing quantitative studies of those data that remain available.

Tactics for collecting Twitter data

The example of Twitter can help explain further what such developments mean for academic research. Twitter also offers some free API access, but it is only possible to search 7 days back in time, and even when doing so, you only access a sample of all tweets (Twitter 2019). The solution for digital social research has been to always collect Twitter data forward in time using tools like TCAT (Borra and Rieder 2014), but this has the obvious disadvantage of not being able to look back at events that were not predicted. While it may be easy enough to predict that planned events such as national elections will generate Twitter activity of interest to researchers, it is very difficult to catch emergent hashtag-oriented events such as #blacklivesmatter early enough for data to be collected in a timely manner to be complete enough for rigorous research.

Faced with such situations, researchers can decide to pay for Twitter data access through their Historical API. In TANTlab, for instance, researchers ended up paying $1000 for a historical data set covering 40 days of Twitter activity related to the controversial culling of Marius the Giraffe at Copenhagen Zoo. While this is not an overly large sum, such monetary restrictions no doubt limits the flexibility of public research, not to mention student projects. Even when funding is secured, purchasing data sets raises a host of technical questions that qualitative and mixed methods social science researchers are rarely equipped to handle on their

own. These questions include how to deal with data delivery through a number of .JSON files that cannot simply be opened in a spreadsheet and which are so large that a server needs to be set up to store them.

A second entry point for social researchers to interact with social media data in a world of increasingly strict API access is to set up official partnerships with the corporations themselves. So far such partnerships have been relatively rare and mostly oriented towards large-scale quantitative analyses driven by research questions already recognized by the corporations to be important (Cha et al. 2010). This may be due to the fact that Facebook already have an in-house research unit conducting (infamous) live experiments and running quantitative investigations, not least in order to disprove accusations made against social media, such as the echo chamber hypothesis (Bakshy et al. 2012).

However, Twitter (the corporation) recently approached the academic community with an invitation to come up with new 'health metrics' for public debate taking place on Twitter (2018). The reward would be data access and some funding:

> "Successful applicants will collaborate directly with Twitter's team, receive public data access and meaningful funding for their research. (…) Our expectation is that successful projects will produce peer-reviewed, publicly available, open-access research articles and open source software whenever possible." (ibid.)

Calls like this one may be the new normal of digital social research in an age of increasingly restricted API access. The upside is that corporations such as Twitter are starting to recognize that they have data sets of value to academic research

and acknowledge that there may be mutual benefits of collaborating with university researchers of different kinds (in supposition that developing health metrics for public debate requires not only quantitative skills but also theories of democratic politics and interpretive analysis). The downside is that academic freedom is greatly impeded if data access is limited to invitation-only events where the corporations have framed the questions in advance (see also Puschmann 2019).

Discussion: Methods for the post-API research landscape

In the face of API restrictions and the consequences described above, perhaps the best alternative is to set up a public research API, which would make data access available on equal terms to all researchers. In lieu of such an interface for researchers, one might ask what happens to public inquiry if some of the main media platforms of contemporary public debate are closed off from academic study. But robust public research APIs are not the only available strategy, and their appearance may be unlikely or require legislation. In this discussion section we therefore also touch upon interface methods and web scraping before discussing what a public research API might look like.

Interface methods

Increased API restrictions over recent years have meant that social scientists are considering how to frame their methods in light of this instability of the research object. In putting forward their standpoint of 'interface methods' Noortje Marres and Carolin Gerlitz (2016) suggest embracing this instability of digital data sources

and use it to reflect on the social science research methods that we attempt to shoehorn the data into. Morever, they suggest an inventive approach as detailed:

> "Instead of fixing the provenance and purposes of methods, we suggest that digital research requires us to embrace their multifarious character. Hence, instead of asking what the capacities of social digital methods are, and deciding with which agendas they are and are not in alignment, we advocate experimental inquiry into what makes their deployment productive for social inquiry…To adopt an 'interface methods' approach, however, means that we do *not* seek to decide which of these two states is more true – affinity or alienation – on general grounds. Rather we must determine what is the most productive relation *between* media and method." (Marres and Gerlitz 2016, emphasis original)

It is important to note that while Marres and Gerlitz still use API-gathered data in their examples given, they do not take a pure digital methods approach. A digital methods approach would demand that an entirely new methodology be adopted. Rather, with an 'interface methods' approach, they suggest examining both the specificities of the data produced by the platform and the rationales of existing social research methods to form a fruitful and grounded approach. What this means in an era of increased restrictions on API access is that we need to treat the restriction as a platform specificity and explore the development of specific meeting points between web platforms and social research methods, both of which will shift in nature along the way.

More recently Marres (2018) and colleagues have been furthering this work, developing a framework or protocol called 'situational analytics' which operates on the assertion that computational social science has a different notion of the 'situation'

being examined on digital platforms in comparison to those held by other social scientists. They raise the problem of gathering, categorising and analyzing digital discussions on issues by using a purely computational approach. Within situational analytics, a combination of approaches is proposed, for example scraping Twitter for tweets that mention the topic and include a link to a related YouTube video. From there, researchers form a corpus of videos to watch and manually analyze the audio visual content. Additionally, researchers scrape the description texts and categories that form the metadata attached to YouTube videos to produce a lexicon on this topic. This then allows researchers to visualise and analyze the relations between categories. Not only is this emerging 'situational analytics' approach an example of interface methods, it also highlights the inventiveness and manual analysis required to describe a social phenomenon beyond the textual data that can be scraped. To be sure, this approach can make use of APIs but it does not depend on them or valorise them.

Web scraping

Web scraping uses custom scripts to download large amounts of data via the browser interface, something which most companies do not allow and try to make difficult. In some cases researchers have used scraping to overcome limits to the amount of data that can be returned via APIs (Hernandez-Suarez et al. 2018). Deen Freelon, author of the research tool 'fb_scrape_public', refers to scraping as a central technique in the 'Post-API' world, despite its technical and legal/ethical difficulties (Freelon, 2018):

> "Researchers of social and other online media content should start by doing two things as they brace themselves for the uncertainty ahead. First, they should learn how to scrape the web; and second, they should understand the potential consequences of violating platforms' TOS by doing so." (Freelon 2018)

As web scraping is a technique used regularly by marketers and journalists, there are plenty of tools already available including commercial 'point and click' desktop tools, browser extensions and cloud based tools designed to be used by researchers with limited technical skills (Bradshaw 2017). Additionally, custom scrapers can be built to retrieve the HTML source code of web pages (Russell and Klassen 2019). Although web sites can be designed to make scraping difficult, in principle anything that can be seen in a web browser can be scraped given sufficient coding skills.

Scraping can have the problem that related data which may be included in API calls is not directly accessible in the interface without making subsequent scrapes of other pages. An example would be the Twitter API, which includes related objects such as user profiles with each tweet. If the same data were scraped it would require extra scraping to get the user profile data. As well as the additional overhead introduced in doing this, there may be some delay between between the initial and subsequent scrape, in which case the related data may have changed (e.g. account deleted, follower count changed, etc.). An additional concern is that since the actions of the scraping software can be detected by the platform, the retrieved data may be altered deliberately by the platform.[5] This can add overhead to the project as data

[5] The technique of misleading a web scraper (or any bot seen as malicious by the site owner for that matter) is called a honey pot or honey trap, as described in detail by Gržinić, Mršić, & Šaban, 2015.

needs to be validated through human cross-checking.

As well as the increased human and technical overhead that can be introduced by using scraping techniques, there are the legal implications in this process, as in most cases, it will violate the Terms of Service (ToS). Although this may seem a straightforward ethical constraint that invalidates the method entirely, even a cursory investigation of the history of web scraping reveals the area to be heavily contested and in many cases untested in legal process. In the few cases that have come to court in the US for example, judges have taken a lenient view where the scraping has been performed against user generated data that is publically viewable. In the case of HiQ versus Linkedin, where HiQ violated Linkedin terms of service by scraping user details from the site, judges ruled in favour of HiQ by supporting HiQ's counter claim that Linkedin's blocking of their scrapers amounted to anti-competitive practices (HiQ Labs, Inc. v. LinkedIn Corporation).

Freelon suggests that after exhausting all ToS compliant methods, researchers "should carefully consider the potential benefits and harms of using methods that violate a system's ToS". His primary concern is with the ability to violate user privacy, but he also points out that a ToS is designed to protect the companies' commercial interests. Although the former is something that is probably within the grasp of most researchers and ethics committees, it is the latter that will prove more problematic for future researchers. A possible starting point for such considerations would be the HiQ judgement which hinged on the data being publicly viewable. By logging into a

web site to perform the scrape, the legal position may become more complicated for academic researchers.

For other public researchers such as activists, journalists and artists the situation may be more fluid. Whotargets.me is a browser extension that scrapes political adverts from user's social news feeds and sends them to a central location for analysis. A report is supplied back to the user about how they have been microtargeted by political groups and how their own targeting compares to other geographic and demographic groups. Once again, this seemingly violates the ToS, but the service claims 20,000 users in over 80 countries, with their work heavily cited in the press.

Journalists and activists cite public interest as justification for violation of ToS, and as their domain is usually centered around political and social issues, they are less likely to end up in court as the adverse publicity could damage brand reputation. Similarly, some practitioners have used arts practices with some degree of liberty, and despite their projects being affected by API changes (Seppukoo.com, Fbresistance.com), some have perservered (SuicideMachine.org).

As Marres and Weltevrede (2013) point out there is a methodological aspect to scraping that it “makes it possible to render traffic between the object and process of social research analytically productive”. The process of designing and operating a scraper requires a detailed reading of the interface code and data infrastructures that support the social media network, which potentially keeps the researcher closer to

the system and can be an important part of understanding its technicity and its culture (see also Venturini and Rogers 2019).

Public research APIs

While we have tried to suggest that web scraping and interface methods provide interesting and worthwhile alternatives, the idea of public research APIs should certainly also be pursued. Social media researcher Axel Bruns published an open letter to Facebook and Twitter in April 2018, proposing four initial guidelines for such scholarly API access:

> "So how should API access be managed to ensure that (...) independent, critical research in the public interest can be conducted while protecting ordinary users' privacy? We see four key points here: 1) Straightforward scholarly data access policies; 2) Custom APIs for research purposes; 3) Accept the use of research data repositories; 4) Open and transparent engagement with the research community." (Bruns 2018)

Despite the letter being signed by hundreds of academics and researchers neither company has made any form of response at the time of writing. Bruns makes the case that recent abuses of user data via API's are not a reason to shut down public APIs (see also Bruns 2019). To the contrary, such incidents are even more reason to allow independent researchers access to social media data. Often their research aims align with those of the companies themselves, for example to study hate speech. An indication of why social media companies feel they can handle this problem by themselves can be seen in the technological optimism of Mark Zuckerberg, CEO of Facebook, when asked about identifying hate speech on his platform:

> "I am optimistic that over a five-to-10-year period, we will have AI tools that can get into some of the linguistic nuances of different types of content to be more accurate in flagging content for our systems, but today we're not just there on that" (Pearson 2019)

For the technology companies, the solution to a problem is often more technology, and currently artifical intelligence and in particular machine learning is the technological fix *par exellence*. Machine learning can avoid the expense involved in training and deploying human researchers across vast networks to curate their content. However, given the concerns over algorithmic inequality (Noble 2018, Eubanks 2018) and the failure of AI to achieve the objectives that researchers believed were within grasp only a short time ago (Marcus 2018), it does not seem to be a simple question of allowing the machines to 'catch up'. As Axel Bruns (2018) puts it in his open letter: "Now more than ever, strong independent research on these platforms is urgently needed: rigorous, ethical research access to platform APIs actually *protects* users and enhances evidence-based social media literacy."

The suggestion here is that acceptable use policy for this API should be based on public interest considerations rather than commercial benefit. Given that the letter also calls for "open and transparent engagement with the research community" and that he has yet to receive any response from Facebook or Twitter, we can see that this approach is problematic as well. One of the issues is that social media companies have no commercial imperative to engage with an unprofitable and even risky activity. However, public research APIs may be the only option left for researchers who are not prepared to risk violating ToS, and given the large number

of signatories to Bruns' letter, pressure is building that could force social media companies to respond.

Such pressure could grow by extending Bruns' call beyond academic researchers into a more general group of public researchers that includes journalists and activists. This could encourage a wider debate framed in terms of transparency, as has been seen with fairness, accountability and transparency in machine learning, which has progressed from activism to an ACM conference topic in the space of a few years (https://fatconference.org/ and previously https://www.fatml.org/). The pressure should be directed towards general minimum acceptable standards of transparency and accountability for all social media platforms.

To illustrate the benefit of such transparency and how it can translate into accountability, consider the URL Share Count offered by Facebook, which offers to return the number of times a URL has been shared inside their system. This unauthenticated endpoint is very simple to use and can be queried repeatedly and was initially used by web sites to show a count of the number of times the page had been shared. However, this feature enabled the technical capability for the tool used by Buzzfeed to identify that fake news was being shared more than real news on Facebook in the run up to the 2016 US elections (Silverman 2016). This is a discovery that has since prompted much public debate and academic analysis of the phenomena (see also Bounegru et al. 2018).

Thus, a potential starting point for development of consensus on transparent research versions of social media APIs is meta research into social media studies that have used APIs to collate the type of API endpoints used and types of data retrieved. Future studies utilising digital methods could help outline the basic API access requriements that respect user privacy whilst enabling research to continue. In other words, we propose an empirical approach to designing robust public research APIs by scanning the field for exemplary cases where data retrieved through APIs has been used to public benefit. Such work could also support and inform other academic efforts at making existing social media data archives sharable (Weller and Kinder Kurlanda, 2016). In the same spirit, it could be part of the ToS of such an API that data is available in standardised formats for use by other researchers.

The experiences of Bernard Rieder and of TANTlab outlined above indicate that a robust public research API for social media should first and foremost prioritize stability, since many researcher hours have been spent trying to catch up with the latest adjustments to social media platforms. In securing stability, inspiration could be drawn from the open source community, which has been able to maintain among other things the Linux operating system software that is so stable that it is often preferred over commercial products. An open source approach to a public research API may include the development of basic standards for data formats and data access, allowing new features to be implemented without interrupting access to existing data types, and further allowing easier data exchanges and interoptability between data from different platforms. Such work could draw not only on open

source practices, but also on the development of internet software and the web itself, including the common internet protocols and programming languages recognized across developer communities.

Public research APIs are an area where researchers of all varieties may be compelled to collaborate, examine and share their ideas, and we would encourage a wider and more formal engagement to ensure not only that academic researchers are included in areas that clearly and urgently need more research, but also to make a positive case for direct engagement with social media corporations for the benefit of the companies, researchers and the public. In some cases, these efforts may be led by legislation in countries with identified social and political issues that have related social media effects. In a similar way that ethical issues around online gaming has resulted in legislation that has brought changes in gaming platforms and moves towards transparency (forbes.com, 2019), it may be the case that the desire to track URL sharing, special interest group formation or the spread of disinformation may be the driving force in getting social networks to provide transparency APIs.

Conclusion

Although the Cambridge Analytica revelations are seen as one cause of recent restrictions in API access to Facebook, we can only speculate or rely on the published accounts of whistleblowers (Frenkel et al. 2018) to understand the discussions held in boardrooms and the corporate thinking that guides decisions to close or restrict APIs. However, in this paper, we have tried to make clear that API restrictions have a detrimental impact on academic social research. Digital methods

form part of the suite of qualitative and mixed methods available to researchers who wish to study social phenomena in digital settings. Part of what has made the digital methods approach innovative is the ability to use coding skills to gather data from new platforms and websites.. However, the ability to conduct digital methods-led research hinges on access to APIs, which is decreasing due to varying commercial, ethical and political factors.

We have traced important parts of the history of API-based digital methods, including recent examples of how increasingly restrictive social media APIs have made academic research more difficult. This is happening in a time of privacy violations by commercial actors that call for more, not less, public research on social media data flows. We suggest it is imperative for academic and allied non-academic researchers, activists and journalists to gather together to call for greater access to social media data in order to research and share knowledge about the important issues pressing publics around the world. We support Axel Bruns' idea for a public research API developed through open collaboration between scholars and social media companies. We believe the specification of such an API should be based on the empirical grounding of key cases where public access to social media data has proven valuable, such as, but not limited to, projects that trace how misinformation travels. Furthermore, we propose that due to the open source software movements and the common standards on the web, developers are well equipped to build a public research API with a robust, open and non-commercial basic infrastructure that can operate across different social media platforms and data repositories.

References

Abildgaard, Mette Simonsen, Andreas Birkbak, Torben Elgaard Jensen, Anders Koed Madsen, og Anders Kristian Munk. 2017. “Five Recent Play Dates”. *EASST Review* 36 (2).

Bakshy, E., Rosenn, I., Marlow, C., & Adamic, L. (2012). The role of social networks in information diffusion. In *Proceedings of the 21st international conference on World Wide Web - WWW ’12* (p. 519). Lyon, France: ACM Press. https://doi.org/10.1145/2187836.2187907

Birkbak, A. (2012). Crystallizations in the Blizzard: Contrasting Informal Emergency Collaboration in Facebook Groups. In *Proceedings of the 7th Nordic Conference on Human-Computer Interaction: Making Sense Through Design* (pp. 428–437). New York, NY, USA: ACM. https://doi.org/10.1145/2399016.2399082

Borra, E., & Rieder, B. (2014). Programmed method: developing a toolset for capturing and analyzing tweets. *Aslib Journal of Information Management*, *66*(3), 262–278. https://doi.org/10.1108/AJIM-09-2013-0094

Bounegru, L., Gray, J., Venturini, T., & Mauri, M. (2018). A Field Guide to'Fake News' and Other Information Disorders: A Collection of Recipes for Those Who Love to Cook with Digital Methods, Public Data Lab, Amsterdam.

Bradshaw, P. (2012). *Scraping for Journalists (2nd edition)*. Leanpub. Retrieved from https://leanpub.com/scrapingforjournalists

Bruns, A. (2018, April 25). Facebook Shuts the Gate after the Horse Has Bolted, and Hurts Real Research in the Process. Retrieved January 30, 2019, from https://medium.com/@Snurb/facebook-research-data-18662cf2cacb

Bruns, A. (2019) After the 'APIcalypse': social media platforms and their fight against critical scholarly research, Information, Communication & Society, 22:11, 1544-1566, DOI: 10.1080/1369118X.2019.1637447

Cadwalladr, C., & Graham-Harrison, E. (2018, March 17). Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach. *The Guardian*. Retrieved from https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Callon, M. (1986). Some elements of a sociology of translation: Domestication of the scallops and the fishermen of St. Brieuc Bay. In J. Law (Ed.), Power, action, and belief: A new sociology of knowledge (pp. 196–223).

Cha, M. (n.d.). Measuring User Influence in Twitter: The Million Follower Fallacy, 8.

Constine, J. (n.d.). Facebook pays teens to install VPN that spies on them. Retrieved January 30, 2019, from http://social.techcrunch.com/2019/01/29/facebook-project-atlas/

Eubanks, V. (2018). *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor*. New York, NY: St. Martin's Press.

Facebook. 2018a. "Breaking Changes - Graph API - Documentation". Facebook for Developers. 2018. https://developers.facebook.com/docs/graph-api/changelog/breaking-changes/.

Facebook. 2018b. "Features - App Development - Documentation". Facebook for Developers. 2018. https://developers.facebook.com/docs/apps/review/feature.

Facebook. (n.d.-a). Overview - Graph API - Documentation. Retrieved January 30, 2019, from https://developers.facebook.com/docs/graph-api/overview

Facebook. (n.d.-b). URL - Graph API - Documentation. Retrieved January 30, 2019, from https://developers.facebook.com/docs/graph-api/reference/v3.2/url

Feinberg, A. (2019, January 17). Jack Dorsey Has No Clue What He Wants. Retrieved January 30, 2019, from https://www.huffingtonpost.com/entry/jack-dorsey-twitter-interview_us_5c3e2601e4b01c93e00e2a00

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

Frenkel, S., Confessore, N., Kang, C., Rosenberg, M., & Nicas, J. (2019, January 29). Delay, Deny and Deflect: How Facebook's Leaders Fought Through Crisis. *The New York Times*. Retrieved from https://www.nytimes.com/2018/11/14/technology/facebook-data-russia-election-racism.html

Gržinić, T., Mršić, L., & Šaban, J. (2015). Lino—An Intelligent System for Detecting Malicious Web-Robots. In N. T. Nguyen, B. Trawiński, & R. Kosala (Eds.), *Intelligent Information and Database Systems* (Vol. 9012, pp. 559–568). https://doi.org/10.1007/978-3-319-15705-4_54

Hernandez-Suarez, A., Sanchez-Perez, G., Toscano-Medina, K., Martinez-Hernandez, V., Sanchez, V., & Perez-Meana, H. (2018). A Web Scraping Methodology for Bypassing Twitter API Restrictions. *ArXiv:1803.09875 [Cs]*. Retrieved from http://arxiv.org/abs/1803.09875

HiQ Labs, Inc. v. LinkedIn Corporation, N.D. Cal. 2017

Leingruber,T., & Stumpel, M. (n.d.). FB Resistance Artists. Retrieved February 1, 2019, from http://fbresistance.com/

Lovink, G. (Ed.). (2013). *Unlike us Reader: social media monopolies and their alternative*. Amsterdam: Inst. of Network Cultures.

Madsen, A. (2018). *Doing Data Together*. Project Description. Retrieved from http://vbn.aau.dk/en/projects/doing-data-together(fd8e3c38-a47b-44ee-ab37-73ee382a83bb).html

Marcus, G. (2018). Deep Learning: A Critical Appraisal. *ArXiv:1801.00631 [Cs, Stat]*. Retrieved from http://arxiv.org/abs/1801.00631

Marres, N. (2018, November). *Situational analytics: Why social research must become inventive in a digital age*. Oxford Internet Institute. Retrieved from https://www.oii.ox.ac.uk/videos/situational-analytics-why-social-research-must-become-inventive-in-a-digital-age/

Marres, N., & Gerlitz, C. (2016). Interface Methods: Renegotiating Relations between Digital Social Research, STS and Sociology. *The Sociological Review*, *64*(1), 21–46. https://doi.org/10.1111/1467-954X.12314

Marres, N., & Rogers, R. (2005). Recipe for Tracing the Fate of Issues and their Publics on the Web. http://research.gold.ac.uk/6548/1/Marres_05_Rogers_recipe_copy.pdf

Marres, N., & Weltevrede, E. (2013). SCRAPING THE SOCIAL?: Issues in live social research. *Journal of Cultural Economy*, *6*(3), 313–335. https://doi.org/10.1080/17530350.2013.772070

Merrill, J. B., & Tobin, A. (2019, January 28). Facebook Moves to Block Ad Transparency Tools —… [text/html]. Retrieved January 30, 2019, from https://www.propublica.org/article/facebook-blocks-ad-transparency-tools

Munk, A. K., Madsen, A. K., & Jacomy, M. (2019). Thinking Through The Databody - Sprints as Experimental Situations. In Åsa Mäkitalo, Todd E. Nicewonger, Mark Elam (eds.) *Designs for Experimentation and Inquiry: Approaching Learning and Knowing in Digital Transformation*. London: Routledge.

Munk, A. K., Abildgaard, M. S., Ren, C., & Olesen, A. G. (2019). Kulturlivet på facebook #1: Begivenheder og brugere. (s. 1-11).

Munk, A. K. & Olesen, A. G. (forthcoming). Performing Danish Facebook. STS Encounters. Forthcoming Special Issue on Engaging the 'Data Moment'.

Nelson, A., 2019. Statement from Social Science Research Council President Alondra Nelson on the Social Media and Democracy Research Grants Program [online]. *Social Science Research Council*. Available from: https://www.ssrc.org/programs/view/social-data-initiative/sdi-statement-august-2019/ [Accessed 30 Aug 2019].

Newmeyer & Dillion LLP - Scott L. Satkin. (2018). Data Scraping: Theft or Fair Game? | Lexology. Retrieved January 30, 2019, from https://www.lexology.com/library/detail.aspx?g=5e951f2d-55c7-42a3-a539-fbe88165ea5a

Noble, S. U. (2018). *Algorithms of oppression: how search engines reinforce racism*. New York: New York University Press.

Pearson, J. (n.d.). Mark Zuckerberg Says Facebook Will Have AI to Detect Hate Speech In '5-10 years' - Motherboard. Retrieved January 30, 2019, from https://motherboard.vice.com/en_us/article/7xd779/mark-zuckerberg-says-facebook-will-have-ai-to-detect-hate-speech-in-5-10-years-congress-hearing

Puschmann, C. (2019) An end to the wild west of social media research: a response to Axel Bruns, Information, Communication & Society, 22:11, 1582-1589, DOI: 10.1080/1369118X.2019.1646300

Rieder, Bernhard. 2009. “Some Yahoo APIs Close, Mashups Too”. *The Politics of Systems* (blog). 13 August 2009. http://thepoliticsofsystems.net/2009/08/some-yahoo-apis-close-mashups-too/.

Rieder, Bernhard. 2010. “netvizz – facebook to gephi”. *The Politics of Systems* (blog). 22 March 2010. http://thepoliticsofsystems.net/2010/03/22/netvizz-facebook-to-gephi/.

Rieder, B. (2013). Studying Facebook via data extraction: the Netvizz application. In *Proceedings of the 5th Annual ACM Web Science Conference on - WebSci ’13* (pp. 346–355). Paris, France: ACM Press. https://doi.org/10.1145/2464464.2464475

Rieder, Bernhard. 2015. “The End of Netvizz (?)”. *The Politics of Systems* (blog). 23 January 2015. http://thepoliticsofsystems.net/2015/01/the-end-of-netvizz/.

Rieder, Bernhard. 2018. “Facebook’s App Review and How Independent Research Just Got a Lot Harder”. *The Politics of Systems* (blog). 11 August 2018. http://thepoliticsofsystems.net/2018/08/facebooks-app-review-and-how-independent-research-just-got-a-lot-harder/.

Rogers, R., & Marres, N. (2000). Landscaping climate change: A mapping technique for understanding science and technology debates on the World Wide Web. Public Understanding of Science, 9(2), 141–163. https://doi.org/10.1088/0963-6625/9/2/304

Rogers, R. (2009). *The End of the Virtual: Digital Methods*. Amsterdam: Vossiuspers UvA.

Rogers, R. (2013). *Digital methods*. Cambridge, Massachusetts: The MIT Press.

Russell, M. A., & Klassen, M. (2019). *Mining the Social Web, 3e* (3rd ed. edition). Beijing Boston Farnham Sebastopol Tokyo: O′Reilly.

Silverman, C. (2018). This Analysis Shows How Viral Fake Election News Stories Outperformed Real News On Facebook. Retrieved January 30, 2019, from https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook

Seppukoo » About. (n.d.). Retrieved February 1, 2019, from http://www.seppukoo.com/about

Forbes.com, Tassi, Paul. 2019. "EA Surrenders In Belgian FIFA Ultimate Team Loot Box Fight, Raising Potential Red Flags" https://www.forbes.com/sites/insertcoin/2019/01/29/ea-surrenders-in-belgian-fifa-ultimate-team-loot-box-fight-raising-potential-red-flags/#78c53f083675

TANTlab (2018). Data Policy. https://www.tantlab.aau.dk/Data+Policy/

TANTlab (2019). The Atlas Of Danish Facebook Culture. https://www.tantlab.aau.dk/Projects/Atlas+of+Danish+Facebook+Culture/

Twitter. 2018. "Twitter Health Metrics Proposal Submission". https://blog.twitter.com/en_us/topics/company/2018/twitter-health-metrics-proposal-submission.html.

Twitter. 2019. "Pricing". https://developer.twitter.com/en/pricing.html.

Venturini, T. & Richard Rogers (2019) "API-Based Research" or How can Digital Sociology and Journalism Studies Learn from the Facebook and Cambridge

Analytica Data Breach, Digital Journalism, 7:4, 532-540, DOI: 10.1080/21670811.2019.1591927

Web 2.0 Suicide Machine - Meet your Real Neighbours again! - Sign out forever! (n.d.). Retrieved January 30, 2019, from http://suicidemachine.org/

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