HANDBOOK OF COMPUTATIONAL SOCIAL SCIENCE, VOLUME 2

Data Science, Statistical Modelling, and Machine Learning Methods

Edited by Uwe Engel, Anabel Quan-Haase, Sunny Xun Liu and Lars Lyberg

2

A BRIEF HISTORY OF APIS

Limitations and opportunities for online research

Jakob Jünger

Routledge
Taylor & Francis Group
LONDON AND NEW YORK

2

A BRIEF HISTORY OF APIS

Limitations and opportunities for online research

Jakob Jünger

1 Introduction

Online platforms such as Facebook, YouTube and Twitter offer a wide range of data for scientific research. Since many of the social media providers have set up application programming interfaces (APIs), extensive volumes of data can be collected automatically (Jünger, 2018; Keyling & Jünger, 2016). Social media data are attractive, *inter alia*, because they not only include already available communication, such as that from public media, but they also make organisational and interpersonal communication visible (Ledford, 2020). In addition, these data are process-generated (Baur, 2011, p. 1234), meaning that they are generated independently of scientific research and thus promise an authentic insight into human behaviour.[1] A wide range of studies in the social sciences exploit APIs for data collection and analysis. Thus, the establishment and development of APIs has significant implications for science.

This chapter starts by tracing the development of APIs, with a focus on the relationship of YouTube, Facebook and Twitter to science. Based on an extensive review of press releases, change logs and API references, three periods are distinguished. During the first period, of construction, the platforms established their APIs. An ecosystem of mashups, clients and organisations then evolved. In the following period, of conquest, the providers worked on securing their influence by strategic acquisitions and by placing restrictions on their APIs. For example, Twitter bought Tweetie and restricted the development of third-party clients by changing the terms of its services and API. In the third period, of concern, the political dimension of the APIs became apparent. For example, the events around the U.S. election in 2016 were reflected in changes in APIs and policies.

Comparing different services and historical epochs reveals both the variable and the constant principles of the APIs. Available endpoints, necessary skills and, not least, the regulation of the providers all play decisive roles in determining who can do what kind of research. This poses the threat of a divide between the data-haves and the data-have-nots (boyd & Crawford, 2012; Bruns, 2013). In the second part of the chapter, the different shapes of the APIs are evaluated from a social science perspective. Looking at APIs in relation to scientific demands reveals the factors that have to be considered when doing API-based research.

2 Tracing the development of APIs

An application programming interface works like a plug and a socket. For example, the USB specification defines the dimensions of the plugs and sockets, thus every plug should fit into a socket, whether it is a computer mouse, a storage stick or a fan for hot days. An API is the software counterpart of a hardware interface, and it defines how two software components will interoperate (Jacobson et al., 2012, p. 5). Many different functions of the computer system can be exposed by an API. On one hand, the *API provider* allows access to resources such as showing pictures on a monitor (output), the geolocated position of a device (input) or data on a hard drive (throughput). On the other hand, an *API consumer* uses these resources to create an application such as a computer game, an online map or a search machine. With regard to the collection and analysis of scientific data, web-based APIs are increasingly important. Here, the different parties are distributed over a network and communicate using the Hypertext Transfer Protocol (HTTP). These APIs implement the principles of representational state transfer (REST). Entities can be accessed by name (URLs), and data can be represented in different formats such as HTML, XML or JSON (Fielding, 2000). Typical actions include fetching and posting content, and these actions are called verbs or methods. Furthermore, cloud computing services, for example Amazon Web Services, provide data storage, machine learning capabilities and much more via web-based APIs. In a broad sense, every website server provides some sort of programming interface that is consumed by web browsers to display web pages. Furthermore, APIs can be implemented on network protocols other than HTTP, such as the WebSocket protocol, which allows faster bidirectional asynchronous transmission (Internet Engineering Task Force [IETF], 2011) or protocols for specific tasks such as MTProto, as used in the Telegram Messenger (Telegram, 2020).

Application programming interfaces are not only a type of software, but they constitute a contract between provider and consumer (Jacobson et al., 2012, p. 4). The contract assures the consumer that the interface will always work in the same way. In contrast to a standard web page, the core functionality of an API is relatively stable, and the data is structured. On this basis, it is reasonable to build software or hardware on an infrastructure provided by third parties. In consequence, an ecosystem may emerge around the central API provider. Social media providers introduced their APIs quite early, and Facebook, YouTube and Twitter, three of the most influential players, are the focus of this chapter. While not primarily developed for scientific research, their APIs, to some extent, allow access to the content and usage data on their platforms.

Thus, documentation of these interfaces is crucial for third-party developers. At the same time, the documentation and related policies give insights into the organising principles and are the basis for the historical reconstruction given in the following section. Media reporting, API references, platform and developer policies, press releases, weblogs and scientific literature are systematically analysed and accompanied by field research, that is, by testing the APIs. The internet archive was used to access API references from the earlier years (see Jünger, 2022, in press, for methodological details).

3 Three periods of API evolution

Software development is an ongoing process and not easily sliced into historical periods. With regard to APIs, the version numbers assigned by the operators can provide an initial orientation as they reflect changes in the technical infrastructure. Nevertheless, from a social science perspective, political and organisational changes are more important, and aiming at an overview

Table 2.1 Main issues in the evolution of APIs

Phase	Key issues
~2005: Period of Construction	Web 2.0
Main perspective: technology	Linking processes
	Authorisation services
	Data access
~2010: Period of Conquest	Acquisitions
Main perspective: economy	Competition
	Standardisation
	Access restrictions
~2015: Period of Concern	Intermediation
Main perspective: politics	Regulation
	Partnerships

of the different APIs and the consequences for research, the analysis needs to abstract from the many small steps of development. When looking at extensions and restrictions over the course of time, roughly three periods can be distinguished (see Table 2.1). The three providers were all founded around the year 2005, acquiring businesses and building their APIs to allow them to link into the diverse landscape of online services. Moving five years forward, to around 2010, more and more restrictions were introduced as the APIs matured. These restrictions controlled how third-party organisations could interact and profit from Facebook, YouTube and Google. While the providers opened their business in the first period, later they seemed to focus on conquering the ecosystem that had evolved around their APIs. In the third period, beginning around 2015, political issues increasingly arose, most prominently the role of the platforms and their APIs in the U.S. election. Thus, in addition to technological changes, the main issues changed as well. The three periods described as construction, conquest and concern, with their main issues, are briefly summarised in the following sections. What we will see is how organisations, technology and data became deeply intertwined with human communication behaviour and society. The interactivity of users, providers and platforms challenges the sciences (Marres, 2017, p. 33) and is reflected in the later development of the APIs. Therefore, the timeline in Table 2.1 will serve as the basis for distilling the principles that affect research.

3.1 Construction: linking technical processes (~2005)

The first APIs are said to have originated in the year 2000 from eBay and Salesforce (Lane, 2016). While eBay provided access to its marketplace (eBay, 2000), Salesforce sold what today is called cloud computing. Their API provided customer relationship management as a software service (Salesforce, 2000). Thus, from the beginning, APIs were related to commercial business operations. Some years later, at the end of September 2005, O'Reilly's article "What Is Web 2.0" elicited an echo that continues to this day. One of the key concepts associated with the term Web 2.0 is web services that allow for data-driven development and the mixing of applications (O'Reilly, 2005). An essential component of Web 2.0 mashups was Google Maps, which had been reverse-engineered by different users to integrate it into their own pages. Later, an official API was introduced by Google (Google, 2005).

Three key players entered the field at that time in relation to user-generated content, another component of Web 2.0. These were Facebook (2004), YouTube (2005) and Twitter (2006).

YouTube was the first of the three to publish an API. Only six months after their foundation, an XML-based API was made accessible. Two years later, after Google had acquired YouTube, the API was migrated to the JSON-based Google Data API in 2007. This version lasted for about five years. In the meantime, several functions were added to the API, for example an upload function (2008) and analytics (2012).

Two and a half years after its foundation, Facebook followed in 2006 with the Facebook API, Version 1.0. In the early years, they invested heavily in functionality and laid the cornerstones for their API. Access to the user feed (streams) and analytics allowed page providers to gain insights into their users' behaviour. An SQL-style language, the Facebook Query Language (2007), was established in addition to fixed endpoints. Facebook even published its source code, "initiating an industry-wide practice of controlled openness" (Bodle, 2011, p. 329).[2] In contrast to the other services, Facebook promoted integration in two ways. On the one hand, third parties could now integrate their applications into the Facebook website; on the other hand, Facebook services could be integrated into external applications. Furthermore, Facebook was gaining ground as an identity provider. Facebook Connect (2008) allowed users to log into third-party services with their Facebook account. While promoted as a comfort tool, at the same time Facebook acquired usage data about a broad variety of websites. These API-based technologies deeply embedded Facebook into the infrastructure of website providers.

Twitter also released an API only six months after the first tweet was posted. The API remained stable over a period of four years, and in fact, the API seemed to drive the development of Twitter from the beginning. Many third-party vendors built on this API, such as Summize, a search engine for product reviews. Summize would soon be acquired by Twitter and, as a result, a search API was added in 2008.

During this early phase, linking services through APIs was established as a basic principle of the web. While the hyperlink, as a core technology of the web, connects static resources, described by metaphors of space (Rogers, 2013, p. 46), APIs linked dynamic processes and flows of information. During this time, the companies became increasingly important for the internet economy (Doerrfeld et al., 2016), and the commercialisation of internet technology began to take place.[3]

3.2 Conquest: conquering the ecosystem (~2010)

After an ecosystem had emerged around the APIs, the providers reclaimed control. In 2011, Twitter had more than 750,000 registered third-party apps (Sarver, 2011). Instead of developing apps of their own, Twitter acquired some of the most popular clients. This included Tweetie, an unofficial app for using Twitter on iPhones (2010), and the social media dashboard application TweetDeck (2011). Moreover, the social media aggregation service Gnip was acquired (2014). At the same time, the opportunities for third-party clients were limited by the terms of the services and some changes in the architecture: "We've already begun to more thoroughly enforce our Developer Rules of the Road with partners, for example with branding, and in the coming weeks, we will be introducing stricter guidelines around how the Twitter API is used" (Sippey, 2012). The first version of the API was shut down in early 2013 after the possible impact had been evaluated with blackout tests (Twitter, 2013).

Three types of access restriction were implemented, step by step. First, with open authorisation, a new mechanism for logging into the API was introduced (#oauthcalypse). All requests now needed to be authorised, which broke with the former policy of more open access (Twitter, 2010). Second, third-party apps on mobile and entertainment devices, as well

as apps with more than 100,000 users, needed to go through app reviews before they could go into production mode (Twitter, 2012). This effectively prevented the development of new clients without the collaboration of Twitter. Third, rate limits were imposed on all endpoints. Only fixed amounts of data could be requested based on sliding time frames (Twitter, 2012). Access, therefore, became more complicated, causing some negative reactions from the ecosystem. Twitter was laying the foundation for their API, essentially as it continues to exist today.

At Facebook, the API was also being transformed into the form that it still has today. Some changes related to apps integrated into the Facebook website, in particular, Facebook Markup Language (FBML) was replaced with iframes. As with Twitter, the access rules were also worked out. From 2011 on, all requests had to go through open authorisation, and (presumably automated) app reviews were introduced. Nevertheless, one year later, the number of apps was over nine million (Facebook, 2012). With the introduction of the Graph API 2.0 in 2014, which was even before the first wave of the Cambridge Analytica scandal, Facebook introduced changes that limited access of third-party apps to friends who had the same app installed (Facebook, 2014, 2020). YouTube's situation was similar: the third version of the YouTube Data API, introduced in 2012, stipulated that all requests had to be authenticated with OAuth2. The older access methods, based on developer keys only, continued to exist in parallel until 2015 (YouTube, 2013). Eventually, older devices from third-party vendors such as Panasonic's smart TVs no longer worked (Golem, 2015). Furthermore, around this time, Google blocked a YouTube app developed by Microsoft. Microsoft had developed the app because there were no apps for Windows Phones, only for Android and iOS (Golem, 2013).

In parallel, another issue related to securing influence arose: standardisation. Already in 2007, Google had launched the provider-independent interface specification OpenSocial (Kraus, 2007). Over the years, several social networking organisations joined, among them LinkedIn, MySpace, the Google-owned Orkut, the German networking site StudiVZ and Xing. The project was later continued by the World Wide Web Consortium (W3C) (Jacobs, 2014), and in 2017, the W3C recommendation, Activity Streams 2.0, was published (W3C, 2017). Another initiative was launched by Mozilla in 2012. Their Social API was intended to provide a better integration of social web applications into browsers (Mozilla, 2019). Facebook joined and provided an extension for Firefox. One of the goals was to reduce the number of share buttons on websites.

Nevertheless, few of the providers built on the concepts of standardisation, and little is heard about these projects now. Furthermore, depending on who sets the standard, the dominant platforms can become even more dominant:

> Yet, this interoperability comes at a price as a handful of dominant SNSs utilise Open APIs and a growing number of social applications to solicit, collect, and open up user data for advertisers and data brokers that have much to gain from users' valuable data.
>
> (Bodle, 2011, p. 321)

For example, Facebook took a 30% slice off payments made through apps on its platform (Glaser, 2018). Thus, standardisation activities cannot hide the fact that suppliers are engaged in aggressive competition. For example, in 2012, Twitter blocked API access to images hosted by the Facebook-owned Instagram (Hernandez, 2012). In response, Instagram prevented the display of their images on the Twitter feed (Twitter, 2012). YouTube focused on a fight in the field of copyrights and took action, for example, against the MP3 conversion services (Zota, 2012). Common to all providers was the establishment of stronger access mechanisms and competition

to play a dominant role in the internet economy. Providing APIs pays off when the apps built on the APIs create users, attention and revenue.

3.3 Concern: the political dimension of technology (~2015)

The increasing economic influence of technology brought political consequences. Especially in recent years, it has become apparent that internet companies do not simply provide technology but are deeply embedded in societal contexts. Discussions regarding the rights and obligations relating to dealing with data are particularly controversial. On the one hand, actors outside the companies claim access for socio-political reasons. For example, Politwoops permanently monitors Twitter and archives the deleted tweets of politicians. Politwoops is run by the Open State Foundation in 30 countries. In 2015, Twitter stopped access to their API, stating, "Preserving deleted Tweets violates our developer agreement. Honoring the expectation of user privacy for all accounts is a priority for us, whether the user is anonymous or a member of Congress" (Twitter, 2015, cited from Trotter, 2015). The decision was subsequently heavily criticised by human rights organisations and internet activists (e.g. Accessnow, 2015). Eventually, Twitter reopened the access. In an opposing case, another nongovernmental organisation (NGO) enforced the blocking of data access for the third-party provider, GeoFeedia. According to the American Civil Liberties Union, GeoFeedia had offered U.S. authorities a product for monitoring protests that was based on geodata from Twitter. In consequence, data access was revoked not only by Twitter but also by Facebook and Instagram (Cagle, 2016).

These two cases demonstrate how the platforms have evolved from service providers to information intermediaries (Newman & Fletcher, 2018). Their role as intermediaries is intensively discussed in relation to the U.S. elections, the two anchor points being the data analytics firm Cambridge Analytica (CA) and the Russian company, Internet Research Agency (IRA). The first critical reports on Cambridge Analytica appeared in the *Guardian* on 1.12.2015. However, the wave did not really start rolling until over two years later when the *New York Times* and the *Guardian* took up the matter again (Cadwalladr & Graham-Harrison, 2018; Dachwitz et al., 2018; Rosenberg et al., 2018). Facebook was primarily affected by this. Aleksandr Kogan, an assistant professor of psychology in Cambridge, had collected data under the auspices of scientific purposes. Users had participated voluntarily in a personality quiz; however, the data was passed on to Cambridge Analytica. The psychometric services of Cambridge Analytica were then used by Steve Bannon in the Trump election campaign to target voters. Later, Facebook referred to the fact that, according to the terms of its service, such a transmission was not allowed. Nevertheless, against a backdrop of common practices and rules, an API was used, in the way APIs are used, to connect different businesses. Until recent years, the architecture of Facebook was explicitly designed for the integration of different services. For example, browser games such as Cow Clicker were directly embedded in the user interface but were served from third-party servers. For these games and apps to operate, some access to user data was crucial for authorisation (Bogost, 2018).

Furthermore, Google, Facebook and Twitter came under pressure because they did not stop the potential influence of the Russian Internet Research Agency in the 2016 U.S. election (Dawson & Innes, 2019). That company, also known as the "Trolls from Olgino", created polarising fake comments and booked political advertisements on the platforms. Although real people were acting here, the issue of bot communication gained popularity in the news. Thus, in the context of the IRA and CA, the role of APIs in data analysis became contested. In consequence, the organisation of the APIs was changed quite drastically (Reselman, 2018). While app

reviews were not new at that time, they became obligatory for all developers on the platforms. Policies were revised and, for example, the scope of data access was restricted around 2018. These changes were intensively discussed in the social sciences under the terms "post-API age" (Freelon, 2018) and "APIcalypse" (Bruns, 2019). Indeed, independent research is becoming harder as researchers also need to go through app review when accessing platform data (Bruns, 2018, 2019; Venturini & Rogers, 2019). While the dependencies increased, the companies also started explicit cooperations with science, for example with the Social Science One initiative (Puschman, 2019; Social Science One, 2020).

4 Consequences for science

The establishment and development of APIs is not without consequences for science. Platform providers are data intermediaries who base their business models on the datafication of user behaviour (Dorfer, 2016; Cukier & Mayer-Schoenberger, 2013). Therefore, organisations like Facebook, Twitter and Google are embedded in an ecology of actors, expectations and discourse. Such "ecologies of communication" have been analysed from the perspectives of media logic (Altheide, 2013; van Dijck & Poell, 2013; Klinger & Svensson, 2014), mediatisation (Couldry & Hepp, 2013) and mediation (Livingstone, 2009). One of the basic assumptions of media ecology is that technology frames the behaviour of actors and thus formats the resulting communication artefacts (Altheide, 1994, p. 670). These artefacts, in turn, influence the following activities, for example when users comment on other users' comments, using the means provided by the platform. This results in a circular process in which users and platform operators work on the co-production of data (Vis, 2013). Scientists can be seen as a special kind of user, analysing data that is mediated by the platforms (Figure 2.1). Therefore, APIs are not neutral research instruments but rather what Marres and Gerlitz call interface-methods: "methods that we – as social and cultural researchers – can't exactly call our own, but which resonate sufficiently with our interests and familiar approaches to offer a productive site of empirical engagement with wider research contexts, practices, and apparatuses" (Marres & Gerlitz, 2016, p. 27). The following section builds on these concepts with a focus on the ecology of data. Analysing the mediating mechanisms reveals how user and platform behaviours are intertwined and also the conditions of scientific knowledge production.
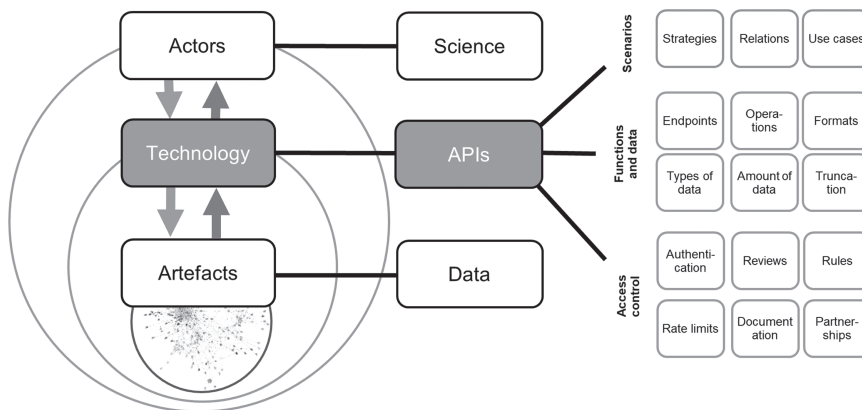

Figure 2.1   The ecology of data and the shapes of APIs

By comparing the development of APIs over the course of time and between platforms, 15 organising principles describing the APIs were inductively developed and then grouped into three categories (Table 2.2):

- *Scenarios*: The API providers publish use cases on their websites. These use cases, for example, encompass curating content or marketing and analysis. Furthermore, the providers follow different strategies. For example, when features appear in the API first, this points to an API-driven development. With regard to apps developed by third-party vendors, some of these provide core functionality, substituting for the user interfaces of the providers, while others enhance the platforms. All of these different aspects illustrate how the APIs are integrated into the data ecology.
- *Functions and data types*: The APIs differ in their endpoints and data formats. Some give access to historic data while most APIs are designed for (near) real-time access. The returned data are formatted, ordered and truncated in different ways, containing only minimal data, on the one hand, and fully hydrated objects on the other.
- *Access control mechanisms*: Authentication, authorisation and the rules of data access change over time. Moreover, documentation plays a key role in determining how easily social scientists find their way into the world of APIs. Reflecting on the different access control mechanisms highlights the limitations of social science research.

Each of these principles, and thus different API designs, presents both opportunities and limitations for online research, as discussed later.

4.1 Scenarios

The use cases for APIs promoted by the providers include both displaying content in third-party applications and generating content. Priority is given to marketing-related goals that promote interaction between organisations and customers. The APIs also enable integration of features such as "like" buttons or authentication functions ("Login with . . ."). Although the range of possible applications is broad, scientific analysis is not very prominent. However, Twitter, in particular, now lists research as a target group. Nevertheless, they point out that academic research is a means to "improve our service" (Twitter, 2020a). Scientific research usually needs only read-only access to APIs and is therefore covered by the scenarios, despite the many access restrictions (see section 4.3).

Based on the providers' own communication (blogs) and media coverage, it can be assumed that the providers have different strategies according to whether the development is API-driven or whether the API is rather a reaction to external requirements. In some cases, functions are available via the API earlier than via the web interface; for example, in 2009, the Twitter geotagging API was available before it was supported in the official client (Twitter, 2009). In contrast, before any API was made available for Instagram or Google+, users reverse-engineered the platform and developed unofficial APIs.[4]

With regard to the scientific use of APIs, another aspect of development is interesting: for cross-platform applications, the standardisation of APIs would be helpful. Two development approaches can be noted here. First, in some cases, the providers rely on semantic web technologies such as Microdata, which is embedded in web pages and thus allows structured access. The dominance of the platforms has probably contributed to the fact that media providers, for example, use the OpenGraphProtocol, which is preferred by Facebook and supported by other providers such as Twitter and many more (Facebook, 2020). On this basis, the preview images

Table 2.2 Shapes of APIs and the main scientific demands

Category	Implementation	Main scientific demands
Scenarios		
Use cases	• Generate and view content • Marketing and analysis • Authentication and infrastructure	View content
Strategies	• API-driven development • Reactive implementation of APIs • Standardisation	Standardisation
Relations	• Substitutional clients • Complementary clients	Complementary clients
Functions and data types		
Endpoints	• Content: e.g. user-generated content • Analytics: e.g. metadata • Functionality: e.g. authentication methods	Content
Operations	• Post and get: HTTP methods • Query languages: SQL-like interface	Get and query
Data formats	• Structured snippets: e.g. JSON • Data(base) files: e.g. RSS feeds or data dumps	Structured
Types of data	• Live vs. historic data • Aggregated vs. non-aggregated data • Anonymised vs. personalised data	Non-aggregated complete historic data
Amount of data	• Fields: explicitly requested data • Hydrated objects: all available data included • Paginated data and streams: slice by slice	Garden hose & fire hose
Truncations	• Recency: only the last messages • Privacy: only unprotected messages • Ranking: only "relevant" messages	Differentiation between ranking information and creation time
Access control		
Authentication	• Open authorisation: standardised access • Key-based: dedicated authentication tokens • Open access: no authentication necessary	Key-based
Reviews	• Proposals • Screencasts	Institutional access
Rules	• Terms of services • Developer terms (+robots.txt)	Scientific rules
Rate limits	• Load-based • User-based • Request-based • Money-based	Pricey load-based access
Documentation	• Playgrounds & explorers • Software development kits • References	References
Partnerships	• Privileges • Bans	Privileges

are extracted when sharing links, and this basically also allows scientific web scrapers to access structured content.

The second approach concerns the standardisation of the APIs themselves (beyond the REST principle). Basic data structures and actions on social media platforms are very similar.

The development of the Open Social protocol, which was initially supported by Google, has now been promoted by the W3C under the name "Activity Streams" without, however, having a significant influence on the platforms as far as is traceable (Jacobs, 2014). Instead, own standards have been set. Here, it becomes clear that the providers prevent substitutive use of their APIs. Initially, clients developed by third parties, especially for Twitter, achieved a high degree of distribution; but now, third-party apps are limited to complementary usage scenarios. Not only do the usage guidelines prescribe this, but the limitations to access rates also effectively hinder growth in user numbers beyond the official clients.

Thus, if one considers use cases, presumed strategies and the relationships between operators and third parties, APIs appear to be a control instrument used for economic growth. Scientific criteria play only a marginal role at best, and therefore it is necessary to consider to what extent the data collection yields findings about the users investigated and to what extent it reflects the platform operators.

4.2 Functions and data types

Social media APIs not only provide data but they also expose functions such as the authentication of users or the provision of ad analytics to businesses. Nevertheless, for science, data access endpoints are the most important. While analytics could provide information not easily visible in the user interface, academic analyses mostly follow their own research questions and, thus, mostly require non-aggregated data. In contrast, analytic features are potential marketing mechanisms for the providers. While there are no "raw" data (Gitelman & Jackson, 2013, p. 2), every additional stage in the data-generating process challenges transparency, which is a core value for science. If scientists understand themselves as observers measuring the world, data analyses will usually only need read access, and therefore, only a limited set of operations is important. Write access changes the underlying world but can be helpful for recruiting participants through automatic invitation comments (e.g. Courtois & Mechant, 2014). This example highlights how APIs are not hidden backdoors, but rather, they are part of the social interaction infrastructure.

From this perspective, two principles must be weighed against each other. On the one hand, low-level access and low-level data allow flexibility, even for innovative research questions that require creative solutions for the operationalisation of theoretical concepts. On the other hand, using predefined query languages and endpoints not only enhances efficiency but also improves validity because the platform logic is transferred to the research procedures. However, since we do not know the effects of the platform logic in advance, usually the more comprehensive low-level data access is preferred. Ideally, we can use both to handle the amount of data, where a "garden hose" approach will use preselected cases and fields, and a "fire hose" approach will access all available data at once.

In relation to the formats, types and amount of data, the API designs are in stark contrast with academic research. Social media APIs are mostly designed to foster real-time interaction between users. For scientific research, on the other hand, it makes more sense to have access to historical data in defined periods of time to make it possible to compile a controlled sample from which to generalise to populations. Otherwise, inference statistics become unreliable. The truncation, sorting and aggregation of data based on largely unknown criteria is a particular challenge to research. For example, Facebook limits access to 600 posts per page, at maximum (Ho, 2020). Moreover, the unseen data is problematic, not only with regard to deleted data. For example, without knowing at what point in time a follower started following a page, the evolution over time can only be traced by ongoing live data collection. Moreover, scientists have to decide which perspective they will take: from the users' perspective, data should be ordered and

truncated in the same way that a user sees it. From the platform perspective, the data should be ordered by creation time and not truncated if custom analyses are to be implemented.

4.3 Access control

Access control mechanisms have become increasingly restrictive over time (see also Perriam et al., 2020), but at the same time, they have been standardised. In the field of social media APIs, there are hardly any open access options without authentication. Presumably, there are two explanations for this increasingly restricted access. First, the data is often personal data and must be protected against inappropriate access (Puschmann, 2019, p. 3). Especially in recent years, this has become particularly evident against the background of political discussion, and it has also been codified in corresponding regulatory documents such as the General Data Protection Regulation of the European Union. Second, these data are to be regarded as an economic asset of commercial enterprises, which thus not only protect their resources but also market them systematically.

The commercial property protection is particularly evident from the rate limits. A load-based limitation is quite reasonable when the operators bear the corresponding costs. Twitter offers higher contingents and more extensive access points for money (Twitter, 2020b). However, operators also often impose limits based on requests per time frame, which is not readily understandable given the strong economic position of the operators and the potential benefits they receive from third-party attention attribution. Limiting the number of requests to short time windows, such as 15 or 60 minutes, is inhibiting for scientific projects that rely less on continuous data access and more on one-off large-scale data collection.

As authentication mechanisms, on the one hand, personalised API keys are employed, and on the other, open authorisation has been established. OAuth allows very different access scenarios, particularly in a role separation between application developer and user (IETF, 2012). Thus, API-based third-party applications can make requests on behalf of users. In terms of scientific use for data collection, however, this procedure complicates the process, as corresponding methods must first be implemented. From a scientific point of view, authentication via API keys would be preferable if, for research projects, the developers of the API application (e.g. when used via Python or R scripts) were also the users of the application. However, open authorisation does allow for a division of labour, so that corresponding clients can be made available for non-programming projects, for example in applications such as NVivo, NodeXL or Facepager.

One of the changes in recent years has been the expansion of proactive reviews. Without an evaluation, which can also lead to complete rejection, only a small amount of (own) data can be accessed in a sandbox mode. From the scientific point of view, in principle, reviews are to be welcomed as a quality management mechanism; however, such reviews follow the criteria of the providers and not scientific criteria (Bruns, 2019, p. 10). While legal regulations such as the German copyright law acknowledge the special position of scientific analyses (e.g. §60d UrhG), this is hardly found among the platforms' rules. In addition, media companies or other commercial enterprises have the resources to enter into privileged partnerships, but individual scientific projects rarely do. The application process is a challenge, as is the process of going through app reviews, for example, when screencasts have to be created or sample analyses submitted, as is the case with YouTube. Against this background, the establishment of specific scientific partner programmes triggers a twofold response (Bruns, 2019; Puschman, 2019). On the one hand, the appreciation of scientific work is certainly welcome, but on the other hand, higher-resource research institutions are again being favoured. A more general access for academic institutions would be valuable.

Low-threshold access would also be helpful for teaching purposes. Moreover, when learning how to use APIs, the documentation is important. All social media providers offer extensive support in the form of software development packages (SDKs) or web interfaces for testing the APIs (API explorers). In the social science context, however, the reference of the endpoints is especially important, first, because appropriate programming knowledge for the integration of SDKs cannot always be assumed, and second, because the transparency of the collection and analysis is a crucial scientific requirement. This is especially significant from an external perspective, as the providers are closed systems with numerous unknown internal mechanisms.

Overall, it turns out that the access control of APIs is not exactly science-friendly. Specific access requirements should be considered in advance when designing data collection projects, although they may already have changed between the time a project is planned and implemented.

5 Conclusion

Programming interfaces have become an integral part of certain areas of social science research. Among other things, these interfaces allow structured and comparatively convenient access to communication data on online platforms. Nevertheless, certain peculiarities have to be considered, in particular because the APIs are under the full control of the providers and are not always compatible with scientific criteria. There is the possibility that scientific research may be based on the availability of data rather than on substantive criteria. One reason for the considerable prominence of Twitter studies in comparison to studies on other platforms such as Instagram, YouTube or WhatsApp could be, for example, Twitter's ease of availability (see Jünger, 2022, in press).

From a comparative perspective – both between different platforms and over time – it is possible not only to identify the variations and characteristics of the different APIs but also to benchmark these against scientific criteria. The construction period starting around the year 2005 is characterised by the providers building and extending their interfaces. In this way, they became anchored in the ecosystem of the internet, which became even more apparent in the years after 2010. The APIs were consolidated and access restrictions were established that fostered complementary use while hindering substitutive use. Finally, a few years later, and especially in the context of the first reports on Cambridge Analytica in 2015, the political dimension of interfaces became increasingly evident. Overall, a transformation has taken place from a technical, via an economic, to a political perspective on APIs.

A consideration of usage scenarios, functionality and access mechanisms reveals that the scientific use of APIs must always be reflected epistemologically. It is only in recent years that science has been increasingly perceived as a partner. However, the scientific requirements for data access continue to contrast with current implementation. Data are often not as comprehensively accessible as would be necessary to assess the quality of samples and the derived findings. Nevertheless, working with APIs enables otherwise virtually impossible insights to be gained into the processes on online platforms. It is important to recognise that the results reflect not only user behaviour, but also the platforms' mechanisms and the scientists' decisions. APIs are both a research tool and a research object. In this respect, it would be advisable in future to conduct more comparative or cross-platform studies.

Acknowledgement

Notes

1 For a critical discussion of this assumption see Jensen (2014), Webster (2011) and Vis (2013).
2 See https://github.com/facebookarchive/platform for an old snapshot of the Facebook platform's PHP code.
3 For a critical discussion of commercialisation see Fuchs (2014).
4 See https://github.com/mislav/instagram for an example of an unofficial Instagram API or https://github.com/jmstriegel/php.googleplusapi for an unofficial Google+ API.

References

Accessnow. (2015). *Hey @twitter! Bring back #politwoops*. https://act.accessnow.org/ea-action/action?ea.client.id=1921&ea.campaign.id=42617

Altheide, D. L. (1994). An ecology of communication: Toward a mapping of the effective environment. *The Sociological Quarterly*, *35*(4), 665–683. https://doi.org/10.1111/j.1533-8525.1994.tb00422.x

Altheide, D. L. (2013). Media logic, social control, and fear. *Communication Theory*, *23*(3), 223–238. https://doi.org/10.1111/comt.12017

Baur, N. (2011). Mixing process-generated data in market sociology. *Quality & Quantity*, *45*(6), 1233–1251. https://doi.org/10.1007/s11135-009-9288-x

Bodle, R. (2011). Regimes of sharing. *Information, Communication & Society*, *14*(3), 320–337. https://doi.org/10.1080/1369118X.2010.542825

Bogost, I. (2018, March 22). My cow game extracted your Facebook data. *The Atlantic*. www.theatlantic.com/technology/archive/2018/03/my-cow-game-extracted-your-facebook-data/556214/

boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, *15*(5), 662–679. https://doi.org/10.1080/1369118X.2012.678878

Bruns, A. (2013). Faster than the speed of print: Reconciling 'big data' social media analysis and academic scholarship. *First Monday*, *18*(10). https://doi.org/10.5210/fm.v18i10.4879

Bruns, A. (2018, April 25). Facebook shuts the gate after the horse has bolted, and hurts real research in the process. *Internet Policy Review*. 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*, *22*(11), 1544–1566. https://doi.org/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*. www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election

Cagle, M. (2016). *Facebook, Instagram, and Twitter provided data access for a surveillance product marketed to target activists of color*. www.aclunc.org/blog/facebook-instagram-and-twitter-provided-data-access-surveillance-product-marketed-target

Couldry, N., & Hepp, A. (2013). Conceptualizing mediatization: Contexts, traditions, arguments. *Communication Theory*, *23*(3), 191–202. https://doi.org/10.1111/comt.12019

Courtois, C., & Mechant, P. (2014). An evaluation of the potential of Web 2.0 APIs for social research. In G. Patriarche, H. Bilandzic, J. L. Jensen, & J. Jurišić (Eds.), *Audience research methodologies: Between innovation and consolidation* (pp. 212–224). Routledge.

Cukier, K., & Mayer-Schoenberger, V. (2013). The rise of big data: How it's changing the way we think about the world. *Foreign Affairs*, *92*(3), 28–40.

Dachwitz, I., Rudl, T., & Rebiger, S. (2018, March 21). Was wir über den Skandal um Facebook und Cambridge Analytica wissen [What we know about the Facebook and Cambridge Analytica scandal]. *Netzpolitik.org*. https://netzpolitik.org/2018/cambridge-analytica-was-wir-ueber-das-groesste-datenleck-in-der-geschichte-von-facebook-wissen/

Dawson, A., & Innes, M. (2019). How Russia's internet research agency built its disinformation campaign. *The Political Quarterly*, *90*(2), 245–256. https://doi.org/10.1111/1467-923X.12690

Doerrfeld, B., Sandoval, K., Wood, C., Lauret, A., & Anthony, A. (2016). The API economy: Disruption and the business of APIs. *Nordic APIs*. https://nordicapis.com/api-ebooks/the-api-economy/

Dorfer, L. (2016). Datenzentrische Geschäftsmodelle als neuer Geschäftsmodelltypus in der Electronic-Business-Forschung: Konzeptionelle Bezugspunkte, Klassifikation und Geschäftsmodellarchitektur [Data-centric business models as a new business model type in electronic business research: Conceptual

reference points, classification and business model architecture]. *Schmalenbachs Zeitschrift Für Betriebswirtschaftliche Forschung*, *68*(3), 307–369. https://doi.org/10.1007/s41471-016-0014-9

eBay. (2000). *Ebay launches new initiative to provide expanded e-commerce solutions*. https://web.archive.org/web/20050208210046/https://pages.ebay.com/aboutebay/thecompany/2000/november.html

Facebook. (2012). *Amendment No. 4 to Form S-1*. www.sec.gov/Archives/edgar/data/1326801/000119312512175673/d287954ds1a.htm

Facebook. (2014). *Facebook platform changelog*. https://web.archive.org/web/20140714133607/https://developers.facebook.com/docs/apps/changelog/

Facebook. (2020). *Facebook platform migrations*. https://developers.facebook.com/docs/graph-api/changelog/archive/migrations

Fielding, R. T. (2000). *Architectural styles and the design of network-based software architectures* (Dissertation). University of California, Irvine.

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

Fuchs, C. (2014). *Social media: A critical introduction*. Sage.

Gitelman, L., & Jackson, V. (2013). Introduction. In L. Gitelman (Ed.), *"Raw data" is an oxymoron* (pp. 1–14). The MIT Press.

Glaser, A. (2018, March 20). Another whistleblower says Facebook knew for years that its data-sharing policies were a huge problem. *Slate*. https://slate.com/technology/2018/03/another-whistleblower-says-facebook-knew-for-years-its-data-sharing-policies-were-a-huge-problem.html

Golem. (2013). Google blockiert Youtube-App für windows phone [Google blocks Youtube app for Windows Phone]. *Golem*. https://www.golem.de/news/microsoft-google-blockiert-youtube-app-fuer-windows-phone-1301–96636.html

Golem. (2015). Youtube-App auf manchen Geräten nicht mehr nutzbar [Youtube app no longer functional on some devices]. *Golem*. https://www.golem.de/news/api-update-youtube-app-auf-manchen-geraeten-nicht-mehr-nutzbar-1504-113628.html

Google. (2005). *The world is your JavaScript-enabled oyster*. https://googleblog.blogspot.com/2005/06/world-is-your-javascript-enabled_29.html

Hernandez, B. A. (2012, July 26). Instagram: Twitter to blame for broken 'find friends' feature. *Mashable*. https://mashable.com/2012/07/26/instagram-twitter-to-blame-for-broken-find-friends-feature/

Ho, J. C.-T. (2020). How biased is the sample? Reverse engineering the ranking algorithm of Facebook's graph application programming interface. *Big Data & Society*, *7*(1). https://doi.org/10.1177/2053951720905874

Internet Engineering Task Force. (2011). *The WebSocket protocol. RFC 6455*. https://tools.ietf.org/html/rfc6455

Internet Engineering Task Force. (2012). *The OAuth 2.0 Authorization framework. RFC 6749*. https://tools.ietf.org/html/rfc6749

Jacobs, I. (2014, December 16). Opensocial foundation moves standards work to W3C social web activity. *W3C Blog*. www.w3.org/blog/2014/12/opensocial-foundation-moves-standards-work-to-w3c-social-web-activity/

Jacobson, D., Brail, G., & Woods, D. (2012). *Apis: A strategy guide: Creating channels with application programming interfaces*. O'Reilly.

Jensen, K. B. (2014). Audiences, audiences everywhere: Measured, interpreted, and imagined. In G. Patriarche, H. Bilandzic, J. L. Jensen, & J. Jurišić (Eds.), *Audience research methodologies: Between innovation and consolidation* (pp. 227–239). Routledge.

Jünger, J. (2018). Mapping the field of automated data collection on the web. Data types, collection approaches and their research logic. In M. Welker, C. Stützer, & M. Egger (Eds.), *Computational social science in the age of big data: Concepts, methodologies, tools, and applications* (pp. 104–130). Halem.

Jünger, J. (2022, in press). Die Macht der APIs. Online-Plattformen als Kontextfaktoren wissenschaftlicher Forschung [The power of the APIs. Online platforms as context factors of scientific research]. In A. Kostiučenko & M. Kuhnhenn (Eds.), *Die Macht des Kontextes*. de Gruyter.

Jünger, J. (2022, in press). Verhaltens–, Forschungs– oder Datenschnittstellen? Drei Perspektiven auf die sozialwissenschaftliche Bedeutung von Application Programming Interfaces (APIs) [Behavioural, research or data interfaces? Three perspectives on the significance of Application Programming Interfaces (APIs) for social science]. In E. Koenen, T. Birkner, C. Pentzold, C. Katzenbach, & C. Schwarzenegger (Eds.), *Digital communication research. Digitale Kommunikation und Kommunikationsgeschichte*. DGPuK.

Keyling, T., & Jünger, J. (2016). Observing online content. In G. Vowe & P. Henn (Eds.), *Political communication in the online world: Theoretical approaches and research designs* (pp. 183–200). Routledge.

Klinger, U., & Svensson, J. (2014). The emergence of network media logic in political communication: A theoretical approach. *New Media & Society*, *17*(8), 1241–1257. https://doi.org/10.1177/1461444814522952

Kraus, J. (2007, November 2). OpenSocial makes the web better. *Google Official Blog*. https://googleblog.blogspot.com/2007/11/opensocial-makes-web-better.html

Lane, K. (2016). *API evangelist history of APIs*. https://history.apievangelist.com/

Ledford, H. (2020). How Facebook, Twitter and other data troves are revolutionizing social science. *Nature*, *582*(7812), 328–330. https://doi.org/10.1038/d41586-020-01747-1

Livingstone, S. (2009). On the mediation of everything: ICA presidential address 2008. *Journal of Communication*, *59*(1), 1–18. https://doi.org/10.1111/j.1460-2466.2008.01401.x

Marres, N. (2017). *Digital sociology: The reinvention of social research*. Polity.

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

Mozilla. (2019). *Social API*. https://developer.mozilla.org/en-US/docs/Archive/Social_API

Newman, N., & Fletcher, R. (2018). Platform reliance, information intermediaries, and news diversity. A look at the evidence. In M. Moore & D. Tambini (Eds.), *Digital dominance: The power of Google, Amazon, Facebook, and Apple* (pp. 133–152). Oxford University Press.

O'Reilly, T. (2005). *What is web 2.0: Design patterns and business models for the next generation of software*. www.oreilly.com/pub/a/web2/archive/what-is-web-20.html

Perriam, J., Birkbak, A., & Freeman, A. (2020). Digital methods in a post-API environment. *International Journal of Social Research Methodology*, *23*(3), 277–290. https://doi.org/10.1080/13645579.2019.1682840

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. https://doi.org/10.1080/1369118X.2019.1646300

Reselman, B. (2018, July 5). It's the end of the API economy as we know it. *Programmableweb*. https://www.programmableweb.com/news/its-end-api-economy-we-know-it/analysis/2018/07/05

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

Rosenberg, M., Confessore, N., & Cadwalladr, C. (2018, March 17). How Trump consultants exploited the Facebook data of millions. *The New York Times*. www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html

Salesforce. (2000). *Products & services*. https://web.archive.org/web/20000303134956/www.salesforce.com/info/products.html

Sarver, R. (2011, March 11). Twitter development talk: Consistency and ecosystem opportunities. https://groups.google.com/forum/#!topic/twitter-development-talk/yCzVnHqHIWo

Sippey, M. (2012, June 29). Delivering a consistent Twitter experience. *Twitter Developers Blog*. https://web.archive.org/web/20120630064926/https://dev.twitter.com/blog/delivering-consistent-twitter-experience

Social Science One. (2020). *Our Facebook partnership*. https://socialscience.one/our-facebook-partnership

Telegram. (2020). *MTProto mobile protocol*. https://core.telegram.org/mtproto

Trotter, J. K. (2015, June 3). Twitter just killed politwoops. *TKTK*. http://tktk.gawker.com/twitter-just-killed-politwoops-1708842376

Twitter. (2009, November 19). Think globally, tweet locally. *Twitter Blog*. https://blog.twitter.com/official/en_us/a/2009/think-globally-tweet-locally.html

Twitter. (2010, August 30). Twitter applications and OAuth. *Twitter Blog*. https://blog.twitter.com/official/en_us/a/2010/twitter-applications-and-oauth.html

Twitter. (2012, December 5). Instagram photo-rendering issue. *Twitter*. https://web.archive.org/web/20121208053613/http://status.twitter.com/post/37258637900/instagram-photo-rendering-issue

Twitter. (2013, March 29). API v1 retirement: Final dates. *Twitter Developers Blog*. https://web.archive.org/web/20130502185948/https://dev.twitter.com/blog/api-v1-retirement-final-dates

Twitter. (2020a). *Twitter data for academic research*. https://developer.twitter.com/en/use-cases/academic-researchers

Twitter. (2020b). *Pricing. API access that scales with you and your solution*. https://developer.twitter.com/en/pricing

Van Dijck, J., & Poell, T. (2013). Understanding social media logic. *Media and Communication*, *1*(1), 2–14. https://doi.org/10.12924/mac2013.01010002

Venturini, T., & Rogers, R. (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. https://doi.org/10.1080/21670811.2019.1591927

Vis, F. (2013). A critical reflection on big data: Considering APIs, researchers and tools as data makers. *First Monday*, *18*(10). https://doi.org/10.5210/fm.v18i10.4878

Webster, J. G. (2011). The duality of media: A structurational theory of public attention. *Communication Theory*, *21*(1), 43–66. https://doi.org/10.1111/j.1468–2885.2010.01375.x

World Wide Web Consortium. (2017). *Activity streams 2.0: W3C recommendation 23 May 2017*. www.w3.org/TR/activitystreams-core/

YouTube. (2013). *Implementing OAuth 2.0 authentication*. https://web.archive.org/web/20150316081145/https://developers.google.com/youtube/v3/guides/authentication

Zota, V. (2012, June 20). YouTube geht gegen MP3-Konvertierungsdienst vor [YouTube takes action against MP3 conversion service]. *Heise*. www.heise.de/newsticker/meldung/YouTube-geht-gegen-MP3-Konvertierungsdienst-vor-1621474.html