Scraping social media data as platform research: A data hermeneutical perspective

Jünger, Jakob

Erstveröffentlichung / Primary Publication
Sammelwerksbeitrag / collection article

Empfohlene Zitierung / Suggested Citation:
Jünger, J. (2023). Scraping social media data as platform research: A data hermeneutical perspective. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 427-441). Berlin https://doi.org/10.48541/dcr.v12.25

Digital
Communication
Research.de

Recommended citation: Jünger, J. (2023). Scraping social media data as platform research: A data hermeneutical perspective. In C. Strippel, S. Paasch-Colberg, M. Emmer, & J. Trebbe (Eds.), *Challenges and perspectives of hate speech research* (pp. 427–441). Digital Communication Research. https://doi.org/10.48541/dcr.v12.25

Abstract: Working with social media data is a hermeneutic procedure systematically guided by doubts about the meaning of data at all stages of the research process, from data collection and preparation to data analysis and publication. A short walk through the automated data collection workflow, as it is implemented in the open-source software Facepager, highlights some of the epistemic peculiarities of the process. The paper encourages researchers to deal with technical details, errors, and restrictions in order to gain a deeper understanding of the organizing principles of the web. Technical limitations and hurdles should not solely be considered as problems to be solved, but also as indicators of social processes on online platforms. Scraping social media data touches on key aspects of platformization and, therefore, is not merely a data collection method, but also a means of examining the online world through a data hermeneutical lens.

DOI 10.48541/dcr.v12.25

Jakob Jünger

Scraping Social Media Data as Platform Research

A data hermeneutical perspective

1 Social science researchers and the platform ecosystem

While participation, inclusion, and empowerment were dominant topics in the early years of internet research (e.g., Scherer, 1998), the last decade has seen a focus on hostile, uncivilized, and deceptive behaviors (e.g., Ben-David & Matamoros-Fernández, 2016). To understand prosocial and antisocial behaviors, researchers have been working with data from social media platforms, including Facebook, Twitter, and YouTube, which provide application programming interfaces (APIs) that allow large-scale analyses of textual data (such as user comments), metrics (such as like and share counts), and network data (based on followers and hashtags). These data are not merely traces left behind by users; they are co-produced by users, platforms, and researchers (Driscoll & Walker, 2014; Vis, 2013).

In general, using social media data for research is not a neutral process—it promotes or hinders the development of platforms as researchers become part of the platform ecosystem. Reactivity and interactivity are embedded in scientific data collection and analysis processes (Marres, 2017, p. 190) both on a surface and a structural level. For example, on the surface level, every click on a YouTube or Tik-

Tok video by a researcher increases the view count. On the structural level, using APIs to amass large datasets increases the attention paid to the platforms studied. In fact, some researchers have stated that platform research has "facilitated these platforms' gradual societal acceptance" (Bruns, 2019, p. 1553). Furthermore, the findings from studies analyzing disinformation campaigns or hate speech can inform public debate and policy making as well as platform organizations and can eventually change the platform ecosystem.

Within this context, a firm grasp of social media data collection processes is crucial in order to understand how online platforms, users, and scientists shape communication datasets. The decisions made in this process have consequences for the interpretation of scientific findings in at least two ways:

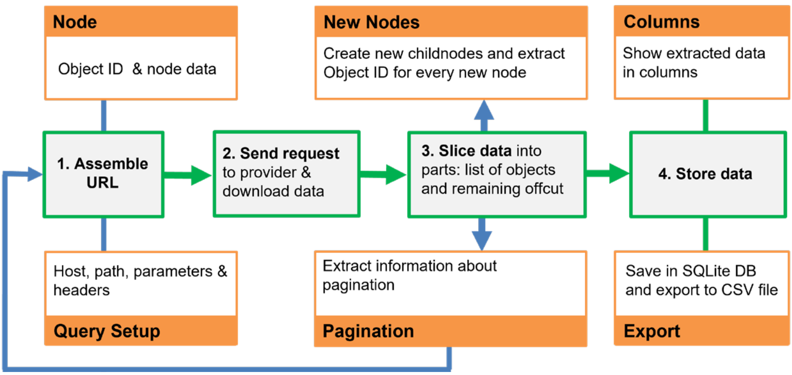
1. *Sampling:* No matter how much effort is exerted, samples of online content, to a certain degree, are always black boxes. For example, technical obstacles cause data dropouts without the exact causes always being known. In addition, populations are usually unknown – a list of all contents is not available – or cannot be defined because, for example, the boundaries of all possible communication situations are not sharply delineated. In this respect, it becomes necessary to assess what a sample actually represents.

2. *Operationalization:* The data structures that can be collected are prescribed by online platforms. Even though a multitude of data traces may be available, their meanings and creation contexts are more diverse than can be expressed, for example, by the number of likes. The data found are not necessarily the best, but only the best available indicators of theoretical constructs, such as communities or discourses.

Such uncertainties must be taken into account in the interpretation of research results. The more is known about the background conditions of the data-generating processes, the more stably the results can be interpreted. Working with social media data is a hermeneutical procedure systematically guided by doubts about the meaning of data at all stages of the research process, from data collection and preparation to data analysis and publication. Furthermore, the paper suggests a change of perspective, viewing technical limitations not solely as problems to be solved but also as indicators of social and organizational processes on online platforms.

In order to higlight some of the hermeneutical challenges, the following sections describe the automatic data collection workflow as we implemented it in Facepager (Jünger & Keyling, 2019). Facepager is a tool that can be used by non-programmers for automated data collection. By design, it is not a one-click-and-you-get-it-all solution; instead, it encourages researchers to deal with low-level API details, errors, and restrictions in order to gain a deeper understanding of the organizational and technical conditions of online platforms. The basic workflow consists of four steps: (1) assembling uniform resource locators (URLs), (2) downloading resources, (3) slicing and extracting data, and (4) storing and exporting data. The sketch of the workflow shown in Figure 1, and outlined in the following chapters, provides the background for delving into the epistemic dimension of social media data.

Figure 1: The Facepager process model



2 Step 1: Assembling URLs – indications about users and content

Whether they are implemented as classical webpages or originate from APIs, resources on the web are usually identified by URLs. When browsing the web, URLs are visible in the address bar. For example, the address of an Instagram page consists of the base path, “https://www.instagram.com,” followed by a path containing a handle, such as “smartdatasprint.” The dual function of URLs has been described within the context of semantic web applications (Sauermann et al., 2008). First, they are so-called endpoints for requesting documents or webpages containing information about users or posts. Second, they identify the described entities, such as the users, organizations, or posts.

Due to this dual function, scraping social media data always involves dealing with representations of entities instead of the entities themselves. Requested documents are representations of the platform's database content, which represents social entities. The situation is further complicated when an organization or a human is active under different accounts. Therefore, data accessed on the web provide indicators about behavior without a clear concept of what those data represent. As in Plato's allegory of the cave, we see the shadows of actors and must build hypotheses about their meanings based on the combination of the actors' moves and the platforms' infrastructure. We can only deal with the artifacts of data-generating processes leading to representations of something unknown.

Requesting the URL mentioned above will lead to a hypertext markup language (HTML) page that is rendered in the browser and shows information about a user profile. Different representations of the same data are usually attached to different URLs (Figure 2). For example, when adding the parameter "__a=1" to the Instagram URL a document containing JavaScript object notation (JSON) data instead of HTML data is delivered. These formats differ in important ways. HTML contains markup that is used to assemble the visual (or auditive) representation of a page; thus, the document contains the data that users see (or hear) on the user interface. JSON is a human- and machine-readable format containing data structured according to key-value fields. JSON is usually provided by API endpoints to enable the development of third-party apps in order to enhance platform functionality (Jünger, 2018).

While the structure of HTML pages changes frequently and must be explored by researchers to extract relevant data, API endpoints are documented on providers' pages and are relatively stable over time. The difference between the two access types has consequences for social media research because the documents (as well as the providers' databases) may contain different data points. Moreover, significantly different relations between the data points and different data contexts may become salient and eventually guide the process of knowledge production. As an example, conversation structures (e.g., threads containing replies to comments) are visible on the user interface. In contrast, reconstructing nested threads from API data gathered from platforms such as Facebook or VKontakte is partially impossible, although responses to hate

Figure 2: Three representations of the same Instagram page

Source: https://www.instagram.com/smartdatasprint/?__a=1

speech, for example, are important for analyzing toxic discourse dynamics, and conversations between users are essential for tracing community formation.

Reverse engineering the URLs of HTML documents or reading API endpoint documentations is not merely informative from a technical point of view. The organizational principles of the platforms become visible, such as when usage scenarios for data processing are described in API references. In such scenarios, numerous references to marketing purposes and the data-centric business models of the providers appear. For example, Instagram provides two use cases for its API:

> The API is intended for *Instagram Businesses and Creators* who need insight into, and full control over, all of their social media interactions. If you are building an *app for consumers* or you only need to get an app user's basic profile information, photos, and videos, consider the Instagram Basic Display API instead. (Instagram, 2021, emphasis added)

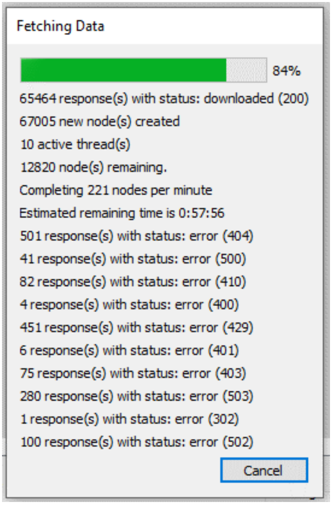
In contrast, academic research does not seem to be a relevant use case from the providers' perspective. In recent years, APIs have become gradually more restrictive (Jünger, 2021), with some scholars even talking about the "Post-API Age" (Freelon, 2018) or the "APIcalypse" (Bruns, 2019). Although Facebook has launched research partner programs, initiatives investigating disinformation and related issues, such as the Ad Observer (Edelson & McCoy, 2021) and the Instagram monitoring project of AlgorithmWatch (Kayser-Bril, 2021), reported they were shut down by Facebook. In consequence, as long as online platforms do not accept their ethical obligation to open up research that serves the public interest, researchers are forced to put even more effort into understanding the various pathways to online platforms' data.

3 Step 2: Downloading data and platform mediation

In a broad sense, the entire web can be considered an API (Fielding, 2000) since downloaded resources are processed by other software, such as a browser or dedicated research tools and scripts. Viewing an interface from the perspective of media theory (Marres & Gerlitz, 2016) highlights the fact that APIs are not solely technical infrastructures but also involve social processes, especially in terms of the processes of the provider organization and the rules governing usage. The interface is under the control of the API provider and is usually not specifically designed to serve scientific purposes. Thus, like the behavior of users, the analysis of content on social media platforms is mediated through these platforms.

This mediation has limitations, especially when rate limits slow down the collection process or when certain content is not available. Each platform has its own set of rules. In general, access is smooth if researchers behave like humans and download data slowly. However, efficiency is restricted with this method, and the research process gains little from automation. For example, Twitter restricts the number of requests for a list of followers to 15 per 15 minutes. Each batch of data contains up to 5,000 IDs; in the next step, profile information can be requested for up to 100 IDs at a rate of 900 calls per 15 minutes (Twitter, 2021a). Thus, crawling the followers of followers for network analyses can become a tedious task and should be carefully planned. Moreover, when crawling the web, a variety of status codes, such as redirects (302), rate limits (429), and server errors (500), are encountered (see Figure 3 for examples). The larger the dataset, the higher the chance of encountering errors. The status codes tell a story about how the web works and highlight the dynamic nature of changes on the web. What works today may already be obsolete tomorrow, and pages that were not working a moment ago may begin delivering data again a few seconds later. In other words, the dataset is a time-traveling slice of information with barely known parameters.

Figure 3: Typical web scraping status codes on German news pages (error rate = 2%; screenshot of Facepager)



Access restrictions apply to the content as well. For example, on Facebook, access to posts in groups and on pages has to be reviewed by the platform; and even then, the names of comment authors are not available through the basic API. Furthermore, the API only provides a sample of posts per page and "will return approximately 600 ranked, published posts per year" (Facebook, 2021). The sampling criteria are opaque, and posts with more likes and shares are presumably preferred (Ho, 2020). Researching highly active accounts, such as those of news media outlets or politicians, is thus potentially biased toward popular content. Moreover, even though deleted posts and moderation practices are crucial for the analysis of antisocial behaviors, these details are usually hidden from the interface. Careful reflection in terms of the platform architecture is indispensable when assessing the scope of research findings.

Although access restrictions can be study limitations, insights into the platforms can be gained when scraping social media data. For example, API results from Telegram (2021) include flags for restricted users, and placeholders for deleted content can be retrieved from Disqus (2021), a comment plugin occasionally embedded in news websites. Therefore, dealing with errors at a low level of data collection can offer fruitful insights into the platformization of human behavior.

4 Step 3: Slicing and extracting data: A data hermeneutical perspective

After accessing and downloading resources, data wrangling begins. Web scraping involves extracting snippets of interest from many data fields and organizing them appropriately for analysis. In the case of HTML, data, such as the dates of posts, are often deeply nested in the hierarchical structure generated by the content management system of the platform. Boilerplate removal involves cutting away unnecessary content and omitting elements, such as webpage footers, headers, and metadata, to reduce the data to units of analysis, such as posts or comments. Alternatively, the elements of interest can be cut out from the data and transferred into a database file. Thus, data wrangling is a multistep process of slicing and extracting data.

The techniques used for data scraping "follow the medium" (Rogers, 2009) when selector languages are used to address the elements in the source code, because these languages also play a role in building webpages. One of the core technologies is cascading style sheets (CSS) selectors, which, on the production site, are used to specify the appearance of specific elements, such as the size, colour, and font of a comment text box. The same selectors can be used with R or Python packages or tools, such as Facepager, to grab content. In general, selectors define a path in the hierarchy of the HTML or JSON document to obtain data, such as the date field inside a comment element that is nested on the page. Sometimes different techniques and intermediate data conversion steps need to be combined. For example, collecting Twitter replies by scraping the interface is not straightforward. Progressing from an undocumented API endpoint to the date of a Twitter reply can be accomplished with Facepager by using a chain of modifiers, including transformation from JSON into HTML, and then parsing the timestamp into a formatted date object ("items_html|css:div.js-stream-tweet|xpath://div[@class='stream-item-header']//@data-time|timestamp").

In general, the hierarchical and technical structures of social media data pose a challenge since scientific data analysts are more used to working with tabular data. Shaping the data is the first step of the analysis, and it defines the units of analysis (cases) and their properties (variables). Even though data formats can be transformed into each other, the shape of the data may frame how researchers think about the world and what research questions are raised. Different data analysis frameworks require different data preparation steps. A multilevel regression problem implies the assembly of different levels of standardized data in the same dataset, a network analysis problem requires relational data, and a hermeneutical problem involves a rich textual representation of the same data. These perspectives come from different research frameworks with different epistemological foundations, such as interpretive and normative paradigms (Wilson, 1973). Going through the steps of shaping the data makes it clear that transforming the world under investigation into a research problem is not merely a measuring procedure but also a knowledge generation process. Considering data wrangling as reconstructive data hermeneutics can bring fruitful irritations with regard to scientific thinking and strengthen the link between one's own analysis and the analyzed artifacts.

Along with investigations into source code and data structures, insights into website architectures can be gained from the data wrangling process. Against the backdrop of static content in the early days of the web (O'Reilly, 2005), interesting issues arise, such as how interactivity and real-time responses are built with dynamic programming languages. Furthermore, the division of labor between diverse roles, like database engineers, frontend designers, and marketing officers, is inscribed into the source code. By following links to content delivery networks and metadata containing semantic web markup in the header, one can see how a page is embedded in a web of services. These metadata often follow Twitter or Facebook standards and are used, for example, to create previews of shared links. In this way, a simple webpage documents the infrastructure of the online ecology from the infrastructural roots to the data leaves of the platformization tree (van Dijck, 2020).

Amid all these issues, the interplay of creativity versus standardization stands out as a dominant theme, and it can be illustrated in the case of emojis. Emojis may become a nuisance when scraping data because their encoding goes beyond the range of standard codepoints used for representing alphanumeric signs. Starting as a small proprietary list of pictures on Japanese mobile devices in around 2000, big tech companies (e.g., Google and Apple) pushed for emojis to finally be included in the Unicode Standard in 2010 (Bergerhausen et al., 2011; Pardes, 2018). However, despite the standardization of code points, emojis are challenging in at least four ways. First, when transferring data between software or devices, care must be taken to choose the right encodings; otherwise, the output will contain cryptic letters or empty boxes. Some functions, for example, in R under Windows, still have limited Unicode support. Second, emojis and colored and animated variations are developed over time, and new emojis, such as the transgender pride flag, are constantly proposed (Unicode, 2021), mirroring societal developments. After new emojis are included in the standard, font designers, device manufacturers, and software developers lag behind and must decide whether, when, and how they will update their products. Third, the concrete representation of the emojis is left to the vendors, and there are diverse stylings across platforms. Fourth, even though the Unicode standard includes textual descriptions of the emojis, the interpretation is obviously open to users. For example, the "Folded Hands" emoji is known under the names "Thank You," "Please," and "Prayer" (Emojipedia, 2021), all of which bear quite different meanings. Overall, the emoji-related technical issues encountered during web scraping evoke a broad range of semiotical and social issues in the tension between standardization and innovation.

Taken together, the various challenges in data processing encourage a shift in perspective. The first reaction to technical problems might be an urge to fix the problem at hand. If one sits back for a moment, one can see through the code and the data into the social and organizational world of online platforms. From a data hermeneutic perspective, technical hurdles, because they are traces of social processes, become a subject of social science research.

5 Step 4: Storing and exporting data: Addressing replicability and platform rules

Data storage decisions have long-term consequences. The first decision to be made is whether to archive downloaded JSON or HTML data or extracted tabular data. Saving downloaded data can lead to large repositories, especially if media files have been collected. However, refinements and secondary analyses are possible if it becomes clear only later which data fields need to be analyzed. Facepager stores downloaded JSON data in an SQLite database. SQLite is an open-source database management system, and the files can easily be accessed with R or Python packages. Downloaded HTML data can be saved as files. The difference between the data formats for storage and analysis further demonstrates that data are always representations and lack a unique reference. In this sense, there is no such thing as raw data (Gitelman, 2013).

Since APIs and websites are constantly changing, corresponding documentation for downloaded and processed data needs to be prepared. Just a few months later, the structure and meanings that were obvious during the collection stage are often no longer apparent. A simplified compilation of the extracted data has the advantage that common data formats, such as CSV files, can be used; furthermore, documentation complexity is reduced. The reduction and documentation steps are fruitful for reducing errors and understanding the data. For example, in this step, it becomes apparent that Twitter IDs are very large and cannot be handled as numbers by Excel. Without being sensitive to such details, confusing paradoxes can sneak into the analysis. Automated data collection should, therefore, not be rashly outsourced to service providers. Even though this first decision about storage formats and documentation involves some effort, in the context of scientific analysis, it is important for the reproducibility and comprehensibility of the subsequent findings.

Another decision related to data formats concerns what is stored and for how long. Social media data often originate from users and demand thoughtful handling to balance legal regulations, platform terms, and ethical principles with the scientific research mandate. Data collection triggers complex considerations about the interplay between the involved actors and the processes of knowledge production in the context of social systems. Carefully reading platforms' terms of service, ethical guidelines, and copyright and data protection regulations can be inspiring, as more questions are raised than answers are given. For example, what can and should be done about deleted content is not obvious. On this point, the Twitter developer terms include the following regulation:

> If *Twitter* Content is deleted, gains protected status, or is otherwise suspended, withheld, modified, or removed from the Twitter Applications (including removal of location information), *you* will make all reasonable efforts to delete or modify such Twitter Content (as applicable) as soon as possible, and in any case within 24 hours after a written request to do so by Twitter or by a Twitter *user* with regard to their Twitter Content, unless prohibited by applicable *law or regulation* and with the express written permission of Twitter. (Twitter, 2021b, emphasis added).

Once collected, data are arranged into academic datasets. The removal of cases, as demanded by the Twitter terms, potentially obstructs reproducibility and destroys findings. Especially in research fields dealing with antisocial behaviors, censorship, and platform regulation, it is expected that content will constantly appear and disappear—the (dis)appearance itself is part of the research interest. If the research outcomes are not merely filed away, they will eventually change the world under investigation, for example, by fueling political debates. When contrasted with ongoing discourses about user privacy, the replicability of research, and political regulation, the quoted Twitter terms illustrate how the four mentioned actors—platforms, users, legal regulators, and "you"—struggle with their roles in the platform economy. Who can legitimately make what claims and who bears what responsibility when handling social media data is subject to permanent negotiation.

6 Conclusion

Careful reflection on the interplay between users, platforms, and researchers is essential to making sound sampling decisions based on online traces and to finding interpretable operationalizations of theoretical concepts. A short walk through the automated data collection workflow offers a vague idea of the epistemic puzzles and peculiarities to be explored. At first glance, assembling URLs appears to be nothing more than a technical process. However, if one takes a closer look, questions arise as to what these addresses actually locate and the kinds of realities that different data formats represent. Although download errors and access restrictions can be perceived as annoyances, they also invite researchers to reflect on the social and organizational conditions of the web. Meanwhile, data wrangling—reconciling data structures with academic thinking—makes the tension between creativity and standardization visible. Finally, deciding on storage options is accompanied by considerations of replicability and the rules of data ownership. Thus, scraping social media data touches key aspects of platformization and, therefore, is not merely a method of collecting data but also a means of studying the online world through a data hermeneutical lens.

Jakob Jünger is Junior Professor for Digital Media and Computational Methods at the Institute of Communiction at the University of Münster, Germany. https://orcid.org/0000-0003-1860-6695

References

Ben-David, A., & Matamoros-Fernández, A. (2016). Hate speech and covert discrimination on social media: Monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of Communication*, *10*, 1167–1193.

Bergerhausen, J., Poarangan, S., & Anderson, D. (2011). *Decodeunicode – die Schriftzeichen der Welt* [Decodeunicode – characters of the world]. Schmidt.

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

Disqus. (2021). *API documentation*. https://disqus.com/api/docs/forums/listPosts/

Driscoll, K., & Walker, S. (2014). Big data, big questions. Working within a black box: Transparency in the collection and production of big Twitter data. *International Journal of Communication*, *8*, 1745–1764. https://ijoc.org/index.php/ijoc/article/view/2171

Edelson, L., & McCoy, D. (2021, August 10). We research misinformation on Facebook. It just disabled our accounts. *New York Times*. https://www.nytimes.com/2021/08/10/opinion/facebook-misinformation.html

Emojipedia. (2021). *Folded hands*. https://emojipedia.org/folded-hands/

Facebook. (2021). *Page feed*. https://developers.facebook.com/docs/graph-api/reference/v10.0/page/feed

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

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

Gitelman, L. (Ed.). (2013). *Infrastructures series. "Raw data" is an oxymoron*. MIT Press.

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), 1–15. https://doi.org/10.1177/2053951720905874

Instagram. (2021). *Instagram graph API*. https://developers.facebook.com/docs/instagram-api/

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. (2021). A brief history of APIs. How social media providers shape the opportunities and limitations of online research. In U. Engel, A. Quan-Haase, S. X. Liu, & L. Lyberg (Eds.), *Handbook of computational social science* (pp. 17–32). Routledge. https://doi.org/10.4324/9781003025245-3

Jünger, J., & Keyling, T. (2019). *Facepager* (Version 4.0.4) [Computer software]. https://github.com/strohne/Facepager/

Kayser-Bril, N. (2021). *AlgorithmWatch forced to shut down Instagram monitoring project after threats from Facebook*. https://algorithmwatch.org/en/instagram-research-shut-down-by-facebook/

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

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

Pardes, A. (2018, January 2). The WIRED guide to emoji. *Wired*. https://www.wired.com/story/guide-emoji

Rogers, R. (2009). *The end of the virtual: Digital methods*. Amsterdam University Press.

Sauermann, L., Cyganiak, R., Ayers, D., & Völkel, M. (2008). Cool URIs for the semantic web: W3C interest group note 03 December 2008. *W3C*. https://www.w3.org/TR/cooluris/

Scherer, H. (1998). Partizipation für alle? Die Veränderung des Politikprozesses durch das Internet [Participation for all? Changes in political processes through the Internet]. In P. Rössler (Ed.), *Online-Kommunikation: Beiträge zu Nutzung und Wirkung* [Online communication. Contributions about its use and effects] (pp. 171–188). Westdeutscher Verlag.

Telegram. (2021). *API > TL-schema > restrictionReason*. https://core.telegram.org/constructor/restrictionReason

Twitter. (2021a). *Rate limits: Standard v1.1*. https://developer.twitter.com/en/docs/twitter-api/v1/rate-limits

Twitter. (2021b). *Developer terms. Developer agreement and policy*. https://developer.twitter.com/en/developer-terms/agreement-and-policy

Unicode. (2021). *Submitting emoji proposals*. https://unicode.org/emoji/proposals.html

van Dijck, J. (2020). Seeing the forest for the trees: Visualizing platformization and its governance. *New Media & Society, 23*(9), 2801–2819. https://doi.org/10.1177/1461444820940293

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

Wilson, T. P. (1973). Theorien der Interaktion und Modelle soziologischer Erklärung [Theory of interaction and models of sociological explanation]. In Arbeitsgruppe Bielefelder Soziologen (Ed.), *Alltagswissen, Interaktion und gesellschaftliche Wirklichkeit, Band 1, Symbolischer Interaktionismus und Ethnomethodologie* [Everyday knowledge, interaction, and social reality, volume 1, symbolic interactionism and ethnomethodology] (pp. 54–79). Rowohlt.