Civil Society Organizations' Data, Access, and Tooling Needs for Social Media Research

Samantha Bradshaw and Bridget Barrett

06/01/2022

Abstract

Civil society organizations working to combat online disinformation and influence operations need more transparency from social media platforms, access to more complete datasets, and user-friendly tools for media monitoring and analysis. We interviewed civil society representatives from 20 organizations working on these issues to better understand the kind of data access and tooling they need to conduct research, investigative journalism, and fact-checking and make policy recommendations. We identify and discuss four key domains: (1) Accessing and Improving APIs; (2) Making Transparent Transparency Reporting; (3) Breaking Down Legal and Policy Barriers to Data Access; and (4) Investing in Skills and Tooling for Social Media Research. Overall, access to data is highly uneven, both in terms of what data platforms provide access to and who has access to—or the right skillset to analyze—it.

Executive Summary

Civil society organizations working to combat online disinformation and influence operations need more transparency from social media platforms, access to more complete datasets, and user-friendly tools for media monitoring and analysis. We interviewed civil society representatives from 20 organizations working on these issues to better understand the kind of data access and tooling they need to conduct research, investigative journalism, and fact-checking and make policy recommendations.

In this report, we identified four key types of requests for data, access, and tooling necessary for civil society research. First, while APIs are appreciated and used by civil society organizations, in many cases, they cannot provide many of the variables that are necessary to their analysis (such as impressions, closed captions, and anonymized comments). Second, transparency reports are often used by civil society organizations to analyze questions of platform accountability, and many interviewees noted that there is potential to improve the kind of data and reporting in these reports, particularly in the context of geographic and language breakdowns. More data on content moderation staffing and practices would also help support the work of civil society organizations.

Third, our interviews also highlighted various legal and policy barriers that disproportionately impact poorer countries. Organizations in these countries need additional support in creating partnerships with universities and navigating legal agreements. Since platform relationships are foundational to data access, helping smaller organizations conducting legitimate research and advocacy to form these relationships will be crucial to their success. Finally, civil society organizations need easier access to data science and computational skill sets, as well as tools that help with the analysis of video content and direct messaging platforms. Many organizations do not have in-house capacity to carry out data science projects. Thus, creating more centralized repositories and training sessions can help improve civil society's capacity to conduct meaningful research.

In addition to these four types of research, we identified three cross-cutting themes that frequently intersected conversations about data access. First, civil society organizations found the lack of standardization across platform API access and data availability challenging. Because the digital ecosystem comprises many platforms, it is difficult to study a cross-platform phenomenon without interoperable tools or standards for API data availability. Second, civil society also noted the importance of cultivating personal relationships with platforms' employees to obtain data relevant to their domains of expertise. Without personal relationships, requesting access to data can feel like a lottery; civil society organizations doing legitimate work are not always approved for API access, or need to go through a third party to obtain access. Finally, interviews made clear that civil society research focuses on a broad range of issues at the nexus of platform governance, accountability, and transparency, with disinformation research being just one subset of a broader research agenda aimed at combating harm.. Taken together, these themes reflect the need for greater transparency standards and clearer processes for obtaining data, and that these standards and processes need to be designed for more use-cases than fact-checking and combating disinformation.

Despite greater calls for transparency, many organizations reflected on its risks and the need for privacy and clear processes for third-party research. They recognized that data can be misused by researchers and organizations; in some countries, organizations might also be pressured to share social media data with governments for surveillance, censorship, and coercion. Organizations suggested that platforms need to maintain the ability to vet, trust, and protect the organizations they give any data to. However, there is a much larger barrier for organizations who are not in North America or Western Europe and lack platform relationships or strong ties to academic research.

Overall, we see a need for connecting less resourced organizations with tools, data, and more resourced research institutions. Platforms should certainly invest more resources into developing relationships and providing skills training to civil society organizations in the Global South. But there is also a need to develop a research hub for civil society organizations doing diverse types of work around mis- and disinformation, Internet trust and safety, and platform governance and accountability. This type of hub could provide access to existing social media monitoring and analysis tools, provide training on how to use them, and foster connections between civil society organizations and research institutions.

Introduction

Digital platforms have become essential to almost every aspect of social, political, and economic life. But recent news has highlighted their potential for harm–to individual wellbeing,[1] users' safety and security,[2] and democratic institutions at large.[3] Due to the closed "black box" nature of platforms' algorithms, policy development processes, and enforcement methods, researchers have struggled to understand the scope and severity of information disorder and the myriad of ways platforms may cause harm. Amidst high-profile scandals and increasing government scrutiny, platforms have also erected barriers to data access–for example, by further limiting their APIs in terms of available data and membership requirements[4] or bringing lawsuits against researchers who scrape data from their services.[5]

Despite these difficulties, there have been several new initiatives launched to support data access. Some have fallen short, such as the Social Science One initiative, where the data Facebook shared with researchers erroneously left out half of all its US users.[6] Academics have identified clear gaps in platform access to data[7] and made recommendations for what data sharing between companies and platforms should look like[8]. Governments have also stepped up efforts to compel platforms to share data. Some ongoing policy proposals include the Platform Transparency and Accountability Act and the Social Media DATA Act in the US,[9] as well as Article 31 of the European Digital Services Act, both of which compel platforms to provide confidential access to vetted researchers.[10] Other organizations, such as the UN, have also laid out principles and guidelines for data sharing between platforms and third-party researchers.[11]

Civil society organizations are often overlooked in conversations about data access.. An important next step is to articulate what data access looks like for civil society organizations working on various issues pertaining to digital rights and democracy. In this report we seek to clarify what kind of data access, policies, and tools would best support their work.

Methodology

---

[1] Wells, Horwitz and Seetharaman. (2021). Facebook Knows Instagram is Toxic for Teen Girls, Company Documents Show. *The Wall Street Journal*.

[2] Scheck, Purnell and Horwitz. (2021). Facebook Employees Flag Drug Cartels and Human Traffickers. The Company's Response is Weak, Documents Show. *The Wall Street Journal*. See also: Schechner, Horwitz and Glazer. (2021). How Facebook Hobbled Mark Zuckerberg's Bid to Get America Vaccinated. *The Wall Street Journal*.

[3] Hagey and Horwitz (2021). Facebook Tried to Make Its Platform a Healthier Place. It Got Angrier Instead. *The Wall Street Journal*.

[4] Roose, K. (2021). Inside Facebook's Data Wars. *The New York Times*. See also: Shane, T. (2022). What Facebook gutting CrowdTangle means for Misinformation. *First Draft News*.

[5] Lapowsky, I. (2021). Platforms vs. PhDs: How tech giants court and crush the people who study them. *Protocol*.

[6] Timburg, C. (2021). Facebook made a big mistake in data it provided to researchers, undermining academic work. *The Washington Post*.

[7] Edelson, L. (2022.) Platform Transparency Legislation: The Whos, Whats, and Hows. *Lawfare*. Tromble, R. (2018).

[8] CDT. (2022). Improving Research Access to Digital Data: A Workshop Report.

[9] Persily, N. (2021). The Platform Transparency and Accountability Act: New Legislation Addresses Platform Data Secrecy. *Stanford Law School*. Lapowski, I. & Brody, B. (2021.) Lawmakers want to Force Big Tech to give Researchers More Data. *Protocol*.

[10] Engler, A. (2021). Platform data access is a lynchpin of the EU's Digital Services Act. *Brookings Institute*.

[11] UN Interagency Group (2021). Selection of Data from Online Platforms that Would Enable Better Understanding of Disinformation Online and Efforts to Counter it.

Between April and May 2022, we conducted interviews with 23 representatives from 20 civil society organizations working to combat disinformation and foreign influence operations. Interviews were semi-structured, with questions focusing on the organization's goals, the tools and data they used, and what they needed to work more effectively. Our goal was to understand what data or analytical tools civil society organizations would like to have that they cannot access, and what tools the broader research community could build to support more equitable global research.

Participants were recruited first from the authors' and Carnegie's existing network, followed by referrals. Participants were selected based on their regional expertise and geographical diversity (see Table 1 for an overview). Interviews were conducted in English and lasted between 30 minutes and an hour. Participants were offered a $100 USD honorarium for their time and participation. In our report, we do not directly attribute quotes or findings to interviewees. All interviewees were also given the opportunity to read and give feedback on this report prior to its publication.

The organizations' geographical focus ranged widely: Five organizations worked in Europe, four globally, three in Africa, two in Latin America, and others in the United States, Taiwan, the Philippines, and Australia. Their methods for studying disinformation and influence operations also varied: Seven focused on fact-checking, while the others focused on tracking and analyzing narratives using quantitative, qualitative, and computational methods. Many organizations had aims beyond combating disinformation, including work on platform accountability, governance and ethics, but our interviews focused on the organization's work in the space of disinformation and foreign influence operations.

Table 1. Overview of Participants

Organization	Interviewee Location
Australian Strategic Policy Institute	Australia
CekFakta.com	Indonesia
Center for Democracy and Development West Africa	Nigeria
Center for Democracy & Technology	US
DebunkEU.org	Lithuania
Doublethink Lab	Taiwan
EU Disinfo Lab	Italy
Equis Research	US
FactSpace West Africa	Ghana
German Marshall Fund	US
International Fact Checking Network	US
Internet Lab	Brazil
Internet Sans Frontières	Senegal

Maldita	Spain
Odipo Dev/Mozilla Foundation	Kenya
Politix/other	Lithuania
Tattle	India
Verafiles	Philippines
VerificadoMX	Mexico
Who Targets Me	England

Findings

Our interviews with civil society allowed us to identify four key types of data access challenges: (1) Accessing and Improving APIs; (2) Making Transparent Transparency Reporting; (3) Breaking Down Legal and Policy Barriers to Data Access; and (4) Investing in Skills and Tooling for Social Media Research. In addition to these domains, civil society activists pointed to several high-level themes and ideas that came up frequently in our discussions.

The first of these is that platforms vary widely in how accessible they make their data. Some platforms do not have APIs or publish transparency reports. Even if platforms do have APIs,some are restricted and do not provide enough meaningful data for civil society to conduct research and advocacy. Even if platforms do publish transparency reports, some consist of aggregate statistics on a topic while others can include the content itself or a searchable interface. While many of the recommendations below are platform specific, civil society members frequently noted the need for more standardization across platforms, suggesting that every social platform should have a basic API containing information about account and content takedowns. Such standardization would also allow for much better and easier tooling across platforms.

The second high-level theme that came up repeatedly in discussions is that platform relationships are often foundational to data access. Civil society organizations discussed how their personal relationships with platforms helped build the trust required to enter data-sharing arrangements. In some cases, these relationships helped organizations access APIs by lending additional legitimacy to an application. In other cases, they allowed organizations access to additional data that would not be publicly available for other researchers or organizations. Without these relationships or streamlined processes in place, getting access can feel like taking part in a lottery. At the same time, the necessity of maintaining relationships with companies can also impede research, as certain requests for data might be self-censored to not harm working relationships.

Finally, while the interviewees were selected based on their work to combat influence operations, and specifically disinformation, many civil society organizations' work extended beyond this. Several of these organizations tackled disinformation as one component of combating the influence of authoritarian regimes, purposefully divisive or inflammatory narratives, and other types of harmful content, and their recommendations are informed by this broader scope. Disinformation was also part of a broader research agenda that focused on issues to do with platform accountability, governance and ethics, and the role of social media companies in the

digital public sphere. Thus, civil society organizations don't just want to fact-check individual pieces of content; they want to track narratives, identify key players in networks, and hold platforms and policy makers accountable. Limiting access, data, and tools to only what is necessary for fact-checking content may hinder broader efforts to prevent harm both on and off social media.

1. Accessing and Improving APIs

APIs (Application Programming Interfaces) are one of the primary ways platforms share information with third-party researchers. APIs allow individuals that have few programming skills to gather data about content and users on a platform. Thus, they are a critical avenue for data collection by civil society actors, who do not always have the computational skills to collect data by other means or the legal capacity to enter into complex data-sharing agreements with platform companies. However, they are not widely available; only some platforms provide APIs to researchers, and in recent years, platforms have restricted access to their APIs and limited the kinds of data available for third-party research. Civil society researchers noted their wish for every platform to have a functional research API and provided some specific feedback on CrowdTangle and Twitter.

- CrowdTangle was widely applauded for its capabilities throughout our interviews. However, civil society organizations would benefit from the ability to track individual users on CrowdTangle, such as the ability to pull data on what accounts liked and shared specific content and who is connected to whom, as well as the ability to construct networks of accounts. Participants appreciated that privacy concerns may make this difficult.

- Platform APIs rarely provide impression data. One of the most common requests was for more detail on the impact that posts had, specifically impression metrics. How often was a piece of content seen, and how often was it engaged with? What was the frequency? Such requests were made for all platform APIs.

- Researchers asked for greater transparency into the user analytics of pages and popular accounts,including declared demographics (if available), average activity on the page, and data on comments or replies (such as how many there are on average, as well as the ability to pull the text of the comments).

- Currently, platforms give administrators of groups and chats wide leeway in making a group "private," particularly on Facebook. Some of these groups seem to be private in name only, as they may be extremely large and seem to approve all requests to join. Researchers asked for API access on all platforms to groups, pages, and chats that pass a certain user threshold.

- APIs typically give researchers the ability to pull data based on specific attributes, and researchers would like to have more control in designing the samples of the data they pull. Specifically, API users requested the ability to sample randomly, as well as the ability to sample representatively based on different attributes. For example, if studying a three-month period, they might pull a weighted sample based on how much content was

posted each month. API users expressed a desire to have the ability to create such weights for more–or all–attributes.

- Researchers repeatedly requested an API for analyzing TikTok content. At the very least, researchers wished that TikTok would make accounts more searchable and provide advanced search tools, particularly for high-impact accounts.

- Civil society organizations working on platforms often need to see when content previously appeared, but this is extremely difficult for image and video-based content. Creating reverse image searches within APIs and advanced search interfaces would help ameliorate this issue.

- While current APIs are well-suited to text-based content, civil society organizations need transcripts of videos or closed captions to be included in any API for a platform that hosts video content.

2. Making Transparent Transparency Reporting
In recent years, platforms have also begun sharing data through transparency reporting. Transparency reports give users insight into content moderation practices, government requests, and coordinated inauthentic behavior (CIB), among other topics. These reports typically include aggregate statistics from the platform on a certain topic, though sometimes they include content, account names, or searchable interfaces. Civil society made several recommendations to improve the data provided through transparency reporting:

- Platforms do not provide access to data that has been removed or taken down, including posts and user accounts that are removed from platforms for violating policies. Civil society organizations interested in auditing platform enforcement would benefit from access to datasets of removed content and accounts.

- Platforms do not provide country- or language-level breakdowns in their transparency reporting. Civil society organizations interested in the amount of hate speech or COVID-19 misinformation being removed in their country, for example, would benefit from data broken down at the country and language level.

- Platforms do not provide enough data about content moderation staff. Civil society organizations interested in tech justice and labor rights would benefit from data about content moderation staff, such as how many women and people of color are employed by platforms, as well as data on the languages represented among staff in each country.

- Platforms do not provide equal access to data about CIB. Civil society organizations interested in tracking and monitoring information operations would benefit from access to datasets of accounts and content taken down by platforms for CIB. This is particularly relevant for many civil society organizations from smaller and non-Western countries, who lack access to CIB data. They would also benefit from a baseline measure for both

the number of active users in a country and the day-to-day use of specific hashtags to better analyze the impact of CIB on digital conversations.

- Facebook has recently included lists of trending topics in their transparency reporting. Interviewees found these reports limited, but a helpful and relevant first step, and would like to see other platforms release similar reports. Civil society organizations also thought it would be helpful to include "top 10" and "top 20" lists broken out by country and language to better understand country-specific dynamics of digital communication.

- Participants requested all transparency reports–related to CIB, COVID-19 misinformation, hate speech, and other removals–be structured in a similar format.Participants reported dedicating significant hours to creating and cleaning spreadsheets and data to put it into functional, meaningful formats that they can then wrangle to analyze trends.

- Some platforms have created political ad libraries–searchable databases of ads that run on their platform. However, ad libraries are not available in every country where ads are being deployed. Further, there are several limits to the kinds of data civil society actors can collect. Who Targets Me has developed a detailed schema for data that all platforms should make available about political ads. This schema covers the various recommendations that other interviewees also made about political advertising libraries.

- Platforms conduct a significant amount of research into their products and the effect they have on trust, safety, and democracy. However, this research is rarely shared with the public or civil society organizations who would benefit from transparency into this work. Making internal research available for replication, oversight, and accountability would benefit the work of civil society organizations who work on issues of tech justice and digital human rights.

3. Legal and Policy Barriers to Data Access

Platforms can sometimes inadvertently erect legal barriers to data access. Additionally, several researchers and organizations have faced lawsuits for collecting platform data in ways that breach company terms of service agreements, such as web scraping. Civil society organizations noted the following legal and policy barriers to data access and made recommendations to address these challenges.

- While university researchers are usually approved to use APIs for academic research, civil society organizations do not always qualify. Creating processes for civil society organizations to be vetted for access to data can create a more equitable research environment. Currently, many civil society organizations seek to partner with universities to acquire access. This arrangement, however, only exacerbates inequalities between wealthier countries with many universities that have existing relationships with platforms and poorer countries that lack such institutions and relationships.

- Civil society organizations do not always have the expertise or resources to negotiate complex and time-consuming data-sharing agreements with platform companies. This is

particularly difficult for civil society organizations in countries where English is not the primary language, since they must often work with English-based contracts. Thus, the barriers to enter in data-sharing arrangements are often higher for civil society activists working in the developing world.

- Once organizations have established data-sharing arrangements with platforms, there are many limitations on how the data can be used. For example, platform data collected by civil society cannot be shared in legitimate legal cases to prosecute harassment or hate speech. Data also cannot be made public or used to create other tools that researchers might be able to use to help their work. While legal restrictions are important for preventing the misuse of platform data, there needs to be a deeper discussion about when making data public is appropriate.

4. Investing in Skills and Tooling for Social Media Research

Many interviewees raised issues related to the limitations created by platforms' policies, APIs, and other transparency initiatives. Beyond this, however, there were other issues due to a lack of resources, tools, and necessary skill sets. There was significant inequity in these areas, though this inequity was only partly related to geographic context. Civil society organizations that were created by journalists often lacked employees with the requisite data science skills. However, many such fact-checking organizations already have the support of the International Fact Checking Network, which provides a library of tools as well as mentorship opportunities, data access, and training. Multiple organizations raised the importance of this hub for collecting and sharing fact-checks on the COVID-19 pandemic and the war in Ukraine. Any new initiative aimed at combating disinformation could seek to partner with and support the IFCN. Participants additionally recommended the following:

- Computational and data science skills are in high demand; employees with these skill sets are thus expensive to employ. Organizations who can only employ a single computer scientist are unable to continue their work on planned projects if that employee leaves. Civil society would benefit from more training and skills to build capacity in this space.

- Rather than hire in house, organizations can outsource their data science and computational needs. Such relationships are also expensive, but less dependent on a single employee. Such outsourcing often depends on existing relationships and knowledge of what companies exist and which of their products might be applicable to specific use-cases. Civil society would benefit from a hub that would provide data science and computational capacity to organizations who cannot afford private outsourcing.

- Encrypted apps were repeatedly discussed by participants as difficult to conduct research and fact-checks on. Currently, organizations working on these platforms primarily use either manual collection of posts or referrals from friends, family, or other networks (such as engaged readers or accounts made for receiving content to fact check). Civil society would benefit from the creation of tools for collecting encrypted app content, potentially

from crowdsourcing or by making large groups or highly forwarded URLs public or transparent to researchers.

- Video content was also repeatedly discussed as difficult to analyze at scale. Currently, most organizations working on video content must watch an entire video to determine its relevance to their research. Civil society would benefit from tools that provide closed captions or otherwise automate the video analysis process.

Summary of findings

In this report, we outlined three high-level themes and four key types of requests for data, access, and tooling necessary for civil society research. Civil society organizations found the lack of standardization across platform API access and data availability challenging. They also noted the importance of having personal relationships with employees of platforms to obtain data relevant to their domains of expertise. Additionally, interviews made it clear that civil society research does not only focus on disinformation, but on broader issues around technology and its impact on society Taken together, these themes reflect the need for greater transparency standards and clearer processes for obtaining data, and that these standards and processes need to be designed for more use-cases than combating disinformation and fact-checking.

While APIs are appreciated and used by many civil society organizations, organizations are often unable to access many variables that are necessary to their work (such as impressions, closed captions, or anonymized comments). Transparency reports are often used by civil society organizations to analyze issues of platform accountability, and many interviewees noted the potential for improving the kind of data and reporting in these reports, particularly in the context of geographic and language breakdowns. More data on content moderation staffing and practices would also help support the work of civil society.

Our interviews also pointed to various legal and policy barriers that disproportionately impacted poorer countries. Organizations in these countries need additional support in creating partnerships with universities and navigating legal agreements. Since platform relationships are foundational to data access, helping smaller organizations form these relationships to conduct legitimate research and advocacy will be crucial to their success. Finally, civil society organizations need easier access to data science and computational skill sets, as well as tools that help with the analysis of video content and direct messaging platforms. Many organizations do not have in-house capacity to carry out data science projects. Thus, creating more centralized repositories and training sessions can help improve civil society organizations' capacity to conduct meaningful research.

Conclusion

Civil society organizations are critical to understanding and combating harmful content online. Fact-checking organizations debunk claims for their readers while simultaneously creating archives of the types of false information spreading. Investigative researchers track hateful and divisive content and reveal foreign influence operations. Journalists hold platforms and government accountable for online harms from disinformation campaigns to other issues of trust and safety online. But despite the important role these organizations play, there are many

challenges around data access, tooling and skills for social media research and investigative journalism.

We spoke to 20 civil society organizations who all voiced the need for more and better data access to accomplish their goals. Many of these organizations pointed to the importance of standardized data access across all social media platforms, including a comprehensive API. Here, organizations frequently discussed the need for ways to collect and analyze data in encrypted messaging apps, as well as video content on YouTube and TikTok. Such calls for more (and better) data have been repeatedly made by academics and civil society organizations –in this way, this study is yet another confirmation of this need. In a step in the right direction, both YouTube and TikTok have recently announced researcher APIs.[12] Unfortunately, Meta may be sundowning Crowdtangle.[13]

Civil society organizations find transparency reporting beneficial to their work. Multiple organizations spoke of the usefulness of reports on coordinated inauthentic behavior and top trending topics. They noted that these types of high-level reports would help improve platform transparency and accountability, especially if they provided country- and language-specific data. Organizations also suggested that access to content that has been removed for violating policies, as well as better information on removal policies and processes such as content moderation staff representation, would be useful to their work.

Despite greater calls for transparency, many organizations reflected on its risks. They recognized that data can be misused by researchers and organizations, and that in some countries, organizations might be pressured to share social media data with governments. Organizations suggested that platforms need to maintain the ability to vet, trust, and protect the organizations they give any data to, which is why relationships are so fundamental to civil society social media research. However, organizations who are not in North America and Western Europe face significant barriers to developing these relationships. Platforms should invest more resources into developing relationships and providing skills training to civil society organizations in the Global South.

Given the risks of transparency to users and the difficulty for platforms to try to vet civil society organizations, we see a need for connecting less resourced organizations with tools, data, and more resourced research institutions. The International Fact-Checking Network plays such a role for fact-checkers, but international civil society organizations doing more diverse types of work combating harms online don't have a clear equivalent. Such a hub for combating online harms could provide access to existing social media monitoring and analysis tools, provide training on how to use them, and foster connections between civil society organizations and research institutions.

---

[12] Pappas, V. (2022.) Strengthening Our Commitment To Transparency. *TikTok Blog*.
https://newsroom.tiktok.com/en-us/strengthening-our-commitment-to-transparency
https://research.youtube/
[13] Lapowski, L. (2022.) Why Big Tech's Transparency Tool is Dying. *Protocol*.
https://www.protocol.com/bulletins/crowdtangle-brandon-silverman-testimony

This study confirmed the high-level need for more data from platforms in the form of APIs and transparency reporting and clarified the types of data and access that civil society organizations prioritize. We found that many organizations need tools to tackle direct messaging and video content and either don't know where to find them or can't afford what is available. There is a need to connect civil society organizations and research institutions to share data access and resources, particularly between wealthy and poor countries.