Building OPUS, an Open Platform for social media User Studies

Henry Kautz **Sabit Ahmed**

Department of Computer Science
University of Virginia
Charlottesville, Virginia 22903 USA
henry.kautz@virginia.edu, bcw3zj@virginia.edu

Abstract

We are building an open platform for collecting, managing, and analyzing online behavior data from consented subjects. The platform supports informed subject consent, automated data donation, statistical and LLM-based analytics, and a wide variety of social media platforms.

Introduction

There has been much suggestive but little definitive research on how people's online behavior provides a signal of their mental and physical health. Studies to date have used just one or a few sources of data and studied a single condition; for example, Sadilek and Kautz (2013) used tweets to predict the spread of influenza. Furthermore, research on mental health and social media has been particularly hampered by reliance on coarse measures of online behavior such as total minutes of time spent by a subject on an app. As a result, healthcare applications based on users' online behavior have little empirical justification and debates about the degree and direction of causality between online behavior and health are unsettled (National Academies of Sciences, Engineering, and Medicine 2023). A major reason that research in this and other areas is at an impasse is that each research group needs to engineer their own data collection system; there is no open-source, secure, and robust framework for online behavior collection, management, and analysis. In response to this need, we are implementing OPUS, an **O**pen **P**latform for social media **U**ser **S**tudies, which will support research in psychology, healthcare, sociology, political science, and other fields. OPUS enables non-computing experts to collect, manage, and analyze subjects' online behaviorial data.

OPUS, illustrated in Figure 1, is a cloud-based infrastructure that asynchronously collects data from online accounts of subjects who have consented to participate in a research study. The system includes modules for many kinds of online behavior, including social media, discussion groups, search history, location history, and media consumption. As part of the consent process, subjects provide the system with access to their online accounts without revealing their logins or passwords to the system. A key feature of OPUS is its support of truly informed consent through a Subject Dashboard, where subjects can visualize their data in meaningful terms before releasing it to researchers. OPUS includes modules for data cleaning and integration and supports a variety of analytics platforms, including Python scipy.stats and sklearn, secure privately-hosted large-language-model based analysis, and export to external platforms such as REDCap and Qualtrics. A challenge at the core of research using online behavior is comparing the online behavior signal with real-world signals such as psychological surveys and physiological sensors. OPUS can bring in such information from MindTrails, a phone application developed at the University of Virginia (https://mindtrails.virginia.edu), and mindLamp (https://www.digitalpsych.org/mindlamp.html), a similar application developed at Harvard University.

Researcher Access to User Online Data Twitter provides an example of how researcher access to online data has changed over the years. In 2013, we collected hundreds of thousands of tweets from Twitter's public API, and showed that by analyzing the words and geotags of tweets and users' friends networks, the transmission of influenza could be predicted at both the individual and national level (Sadilek and Kautz 2013). Today, this work could not be carried out. First, Twitter removed geotags from public tweets and restricted the volume of tweets that researchers could download. When Twitter became X, its public firehose for research organizations was cut off and its researcher API removed. Similar restrictions on public API access were carried out by most social media platforms.

Countering the effect of loss of third-party access to users' online data was the European General Data Protection Regulation in 2018. Its "right to data portability" requires individuals to receive personal data that they have provided to a platform in a structured, commonly used, and machine-readable format. Today, every major platform allows users to download their own online behavior footprint, including users outside of Europe. This has led to an era where researchers can run studies by having consented subjects download and donate their data. The decrease in the number of subjects involved in any one study is partially compensated for by the increased granularity of data, the ability to link data from multiple online sources, and a reduction in

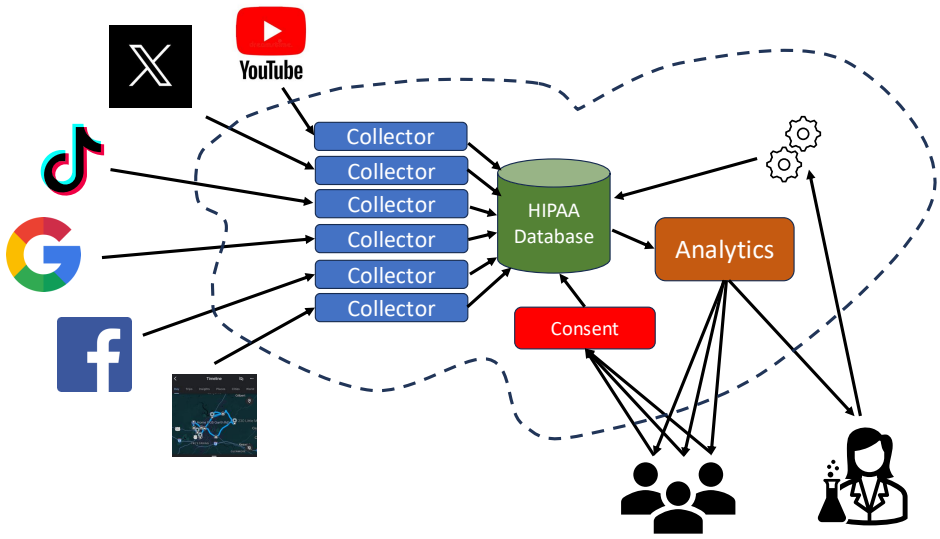

Figure 1: High-level design of OPUS. The researcher (far right) uses the Researcher Dashboard (gear icon) to configure a study, specifying the data sources, consent forms, and list of subjects. Subjects login to Subject Dashboards to consent to the study, authorize the relevant media sites, and review summaries of their data before final release to the researcher.

potential harm to subjects from having their data harvested without their control (Meyer et al. 2023). The challenge with data donation is the difficulty in including subjects who are not technologically sophisticated enough to run data donation software on their own computers. Most studies to date based on manual data donation have been limited to college students.

Architecture

The design of OPUS is inspired by an application that we built for collecting users' Google Takeout data for studies of anxiety and depression during the pandemic (Zaman et al. 2020). Our app first displayed a consent document that the subject signed electronically. It next invoked Google's OAuth API, which caused a Google page to appear where the subject signed in to their Google account. The app navigated to the Google Takeout page, where the subject selected the kinds of data designed (for the cited study, web search and YouTube history) and initiated the takeout process. Finally, the application downloaded the data from the subject's Google Drive.

social media	Facebook, Instagram X, Reddit, TikTok, Twitch, Discord
social connection	Tinder, Grindr TalkLife
entertainment	YouTube, Netflix, Spotify
location & places	Google Maps, Strava
search	Google history, Chrome history
physiological & EMA	MindTrails, MindLamp

Table 1: Online behavior data sources. EMA stands for Ecological Momentary Assessment, *e.g.*, push surveys.

OPUS generalizes this approach to include many social media platforms and GUIs for researchers and subjects. It contains the following modules:

Researcher Dashboard: A web interface for configuring a study, including specifying data sources and the consent protocol to employ, and for monitoring data collection. The initial version of the Dashboard appears in Figure 2.

Subject Dashboard: A web interface for subjects to perform initial consent with or without a researcher in attendance, according to study configuration; to review and visualize their own data; and to re-consent (if so configured) to release their data to the study.

Data collection: Collectors for social media, entertainment, web search, and other kinds of online data. See Table 1 for a list of data sources that OPUS will handle.

Data integration: A module for data cleaning; data type conversion; and data linking across multiple sources.

Data management: A secure HIPAA-approved database hosted by Google Cloud, plus options for the researcher to export data to other data store and analysis platforms.

We now provide selected details about the system.

Data Collection

OPUS supports several data collection methods:

API: In the past, many services offered an Oauth REST API for data download. Today, only a few social media services such as Reddit still provide an API. After the subject consents to the study on the Subject Dashboard, OPUS uses the services' APIs to access their data. On the first call to each API, the service's login and authorization page appears for the user to complete. After entering their login and password, OPUS runs autonomously and the Subject Dashboard may be closed. Note that OPUS never sees the logins or passwords itself.

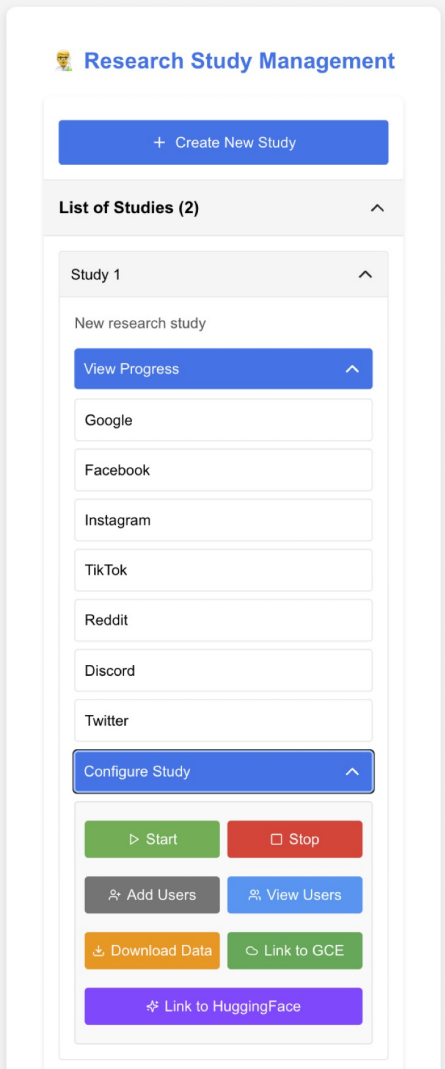

Figure 2: Researcher Dashboard, showing controls to configure, view progress, and export data from studies.

Semi-Automated Data Donation: Many services include in their personal data download pages an option to send the data to the user's Google Drive. In this method, after the subject consents to study participation using the Subject Dashboard, OPUS uses the Google Drive API to request permission to access the subject's Drive. This causes Google to pop up a page asking the user to login and confirm permission for OPUS to access their drive. When the subject does this, OPUS obtains an Oauth token. OPUS then sends the subject to the personal data download page for each service. Note that these pages may include Google Takeout, but Takeout actions are separate from Drive access. On each personal data download page, the subject specifies that the data should be downloaded to their Google Drive. At this point, the user's consent and authorization actions are complete and they may exit the Dashboard. Each service may take minutes to days to prepare the data and transfer it to the user's Drive. OPUS monitors the Drive and pulls the files into its database when they are ready. We call this collection method semi-automated because the subject needs to click through the proper options on each service's data download page.

Automated Data Donation: Even if detailed instructions are provided to the subject, there is a danger that the subject will make errors when clicking on the options on a service's data download page. OPUS provides browser automation to eliminate this problem; after the user authorizes the download page, automation takes over and completes the process. A technical challenge is that most phone web browsers, and in particular Safari and Chrome on iOS, do not support browser automation. OPUS works around this limitation using remote application streaming. The browser that is interacting with the various services is running on a server as part of a Kubernetes container along with an application streaming server. The browser is configured for automation by JavaScript injection with scripts for each service. The browser on the subject's device shows the server's GUI and passes clicks and typing back to the server.

The automation scripts need to be updated whenever a service changes its data download page. At first this may appear to a fatal obstacle to the automated data donation approach, but in fact it is easy to update scripts. ChatGPT can write a correct script. Using a simple prompt and an example of the page to be automated. When OPUS detects that a script has failed, it will eventually be able to update the script on the fly.

Facilitated Data Donation: For cases where none of the previous approaches are possible, OPUS supports manual data donation. The subjects are guided by text and videos step-by-step through downloading their data on their laptops and uploading it to OPUS. This approach was pioneered by Georgia Tech researcher Munmun De Choudhury and is currently the state of the art for data donation.

Improving Informed Consent

As noted earlier, the Subject Dashboard supports informed consent protocols as configured in the Researcher Dashboard. The traditional consent protocol for social media data has a subject consent a single time and requires the subject to quit the study in order to have any of their data deleted. It is difficult, however, for a subject to understand what their online data could be revealing about themselves in advance. For example, in our study on mental health during the pandemic (Zaman et al. 2020), subjects were told that the system would collect their history of watching YouTube videos and categorize the videos by topic. When analyzing the data, we discovered that predictions were best when the set of categories included "adult content". Subjects may have been reluctant to consent if they had known that this would be one of the categories. OPUS improves informed consent by supporting re-consent protocols where subjects see a visualization of their data and are able to delete some or all of it before it is incorporated in the study. Although this option may limit the scope of data collection for some subjects – and thus can be disabled by the researcher if so desired – we believe that in general it will lead to subjects being

more willing to trust and participate in the study. This belief is backed by our recent survey of potential study subjects (Wilson-Lemoine et al. 2025).

Data Analysis

OPUS supports on-the-fly data analytics in order to present the status and emerging trends of data collection on the Researcher Dashboard and to visualize subjects' contributed data on their Subject Dashboards. In addition to standard algorithms for statistical analysis, OPUS can make use of the enormous power of multi-modal large language and image models for clustering and interpreting social media data. A recent paper by Lyu et al. (2025) showed that OpenAI's GPT-4V outperforms specialized algorithms on almost every social media analysis task. Open source models such as LLaMA (Touvron et al. 2023) are running about a year behind OpenAI's best models. Therefore, by 2026, OPUS will be able to use an open-source model running on a private compute cluster in order to provide LLM-powered analytics without fear of HIPAA-protected data being exposed to a commercial service.

Use Cases

We are developing two studies using OPUS in conjunction with researchers in psychology and medicine at our university. These studies address issues of national significance, would be difficult to carry out without OPUS, and will stress-test and refine our design decisions.

Use-Case: Social Media Impacts On Adolescent Development

Perhaps the biggest research question around social media is its impact on adolescent development and mental health. Most people believe that social media is harmful to youth, and four states have already passed legislation restricting social media accounts for young people. There have been hundreds of studies and meta-studies on the relationship between social media use and mental health, and in 2023, the National Academies of Sciences, Engineering, and Medicine held a study on the subject (National Academies of Sciences, Engineering, and Medicine 2023). Surprisingly, the Academies concluded that, "Available research that links social media to health shows small effects and weak associations, which may be influenced by a combination of good and bad experiences. Contrary to the current cultural narrative that social media is universally harmful to adolescents, the reality is more complicated." A similar report by the Lancet Commission on Self Harm (Moran et al. 2024) stated, "research on the effects of social media has so far produced mixed results. In fact, for some young people, it might have benefits, facilitating connections for isolated people, providing online support networks, and delivering therapies." The main weaknesses of the studies to date are that they are too coarse grained, based on gross quantities such as screen time rather than the *content* that is viewed, and only consider one or a few social media platforms in each study. OPUS's data collection capabilities will substantially advance research in this area by facilitating collection of comprehensive data from social media interactions, search histories, and media consumption. In our use case focus groups we will engage youth from the start of the project using the principles of Youth Participatory Action Research (Cammarota and Fine 2010).

Use Case: Veterans' Mental Health

Another study examines the mental health of disabled veterans, with a specific focus on the unique needs of those in rural or underserved areas and issues related to veteran suicide (McIntire et al. 2021). Veterans in these contexts often face significant barriers to accessing traditional mental health services due to geographic isolation, limited infrastructure, and stigma surrounding mental health. OPUS's comprehensive data collection capabilities can facilitate research that connects online behavioral patterns with mental health outcomes, such as changes in mood, social engagement, and help-seeking behaviors. By analyzing data streams from social media interactions, search histories, and media consumption, researchers can identify early warning signs of mental distress or social withdrawal that are often precursors to suicide risk.

Next Steps

A first version of OPUS with limited functionality will be released in summer 2025. We are recruiting a community of other social media researchers to help evaluate each release and suggest improvements. We are planning in-person and remote workshops to bring together researchers in computer science, psychology, healthcare, and social science who could make us of OPUS and contribute their own ideas and code. By 2027, we plan to have our full vision for OPUS implemented and in use at multiple sites.

References

Cammarota, J.; and Fine, M. 2010. Youth Participatory Action Research: A Pedagogy for Transformational Resistance. In Cammarota, J.; and Fine, M., eds., *Revolutionizing Education*, 9–20. Routledge.

Lyu, H.; Huang, J.; Zhang, D.; Yu, Y.; Mou, X.; Pan, J.; Yang, Z.; Wei, Z.; and Luo, J. 2025. GPT-4V(ision) as A Social Media Analysis Engine. *ACM Trans. Intell. Syst. Technol.*, 16(3).

McIntire, K. L.; Crawford, K. M.; Perrin, P. B.; Sestak, J. L.; Aman, K.; Walter, L. A.; Page, D. B.; Wen, H.; Randolph, B. O.; Brunner, R. C.; Novack, T. L.; and Niemeier, J. P. 2021. Factors increasing risk of suicide after traumatic brain injury: A state-of-the-science review of military and civilian studies. *Brain Injury*, 35: 151–163.

Meyer, M. N.; Basl, J.; Choffnes, D.; Wilson, C.; and Lazer, D. M. J. 2023. Enhancing the ethics of user-sourced online data collection and sharing. *Nature Computational Science*, 3: 660–664.

Moran, P. A.; Chandler, A.; Dudgeon, P.; Kirtley, O. J.; Knipe, D.; Pirkis, J.; and et al. 2024. The Lancet Commission on Self-Harm. *The Lancet*, 404(10461): 1445–1492.

National Academies of Sciences, Engineering, and Medicine. 2023. *Social Media and Adolescent Health*. Washington, DC: The National Academies Press.

Sadilek, A.; and Kautz, H. 2013. Towards Understanding Global Spread of Disease from Everyday Interpersonal Interactions. In *Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI)*, 2783–2789.

Touvron, H.; Martin, L.; Stone, K.; et al. 2023. LLaMA 2: Open Foundation and Fine-Tuned Chat Models. https://ai.meta.com/llama/.

Wilson-Lemoine, E.; Kautz, H.; Petz, K. D.; Ahmed, S.; Barnes, L. E.; and Teachman, B. A. 2025. Examining openness to donating personal data from digital devices for mental health research among college students. In *Annual Conference of the Society for Digital Mental Health*.

Zaman, A.; Zhang, B.; Hoque, E.; Silenzio, V.; and Kautz, H. 2020. The Relationship between Deteriorating Mental Health Conditions and Longitudinal Behavioral Changes in Google and YouTube Usages among College Students in the United States during COVID-19: Observational Study. *JMIR Mental Health*, 7(11): e24012.