

Public Data Access Programs

A First Look



Mozilla



ALGORITHMIC
TRANSPARENCY
INSTITUTE



National Conference on Citizenship

Public Data Access Programs: A First Look

Assessing Researcher Data Access Programs
Under The Digital Services Act

Executive Summary

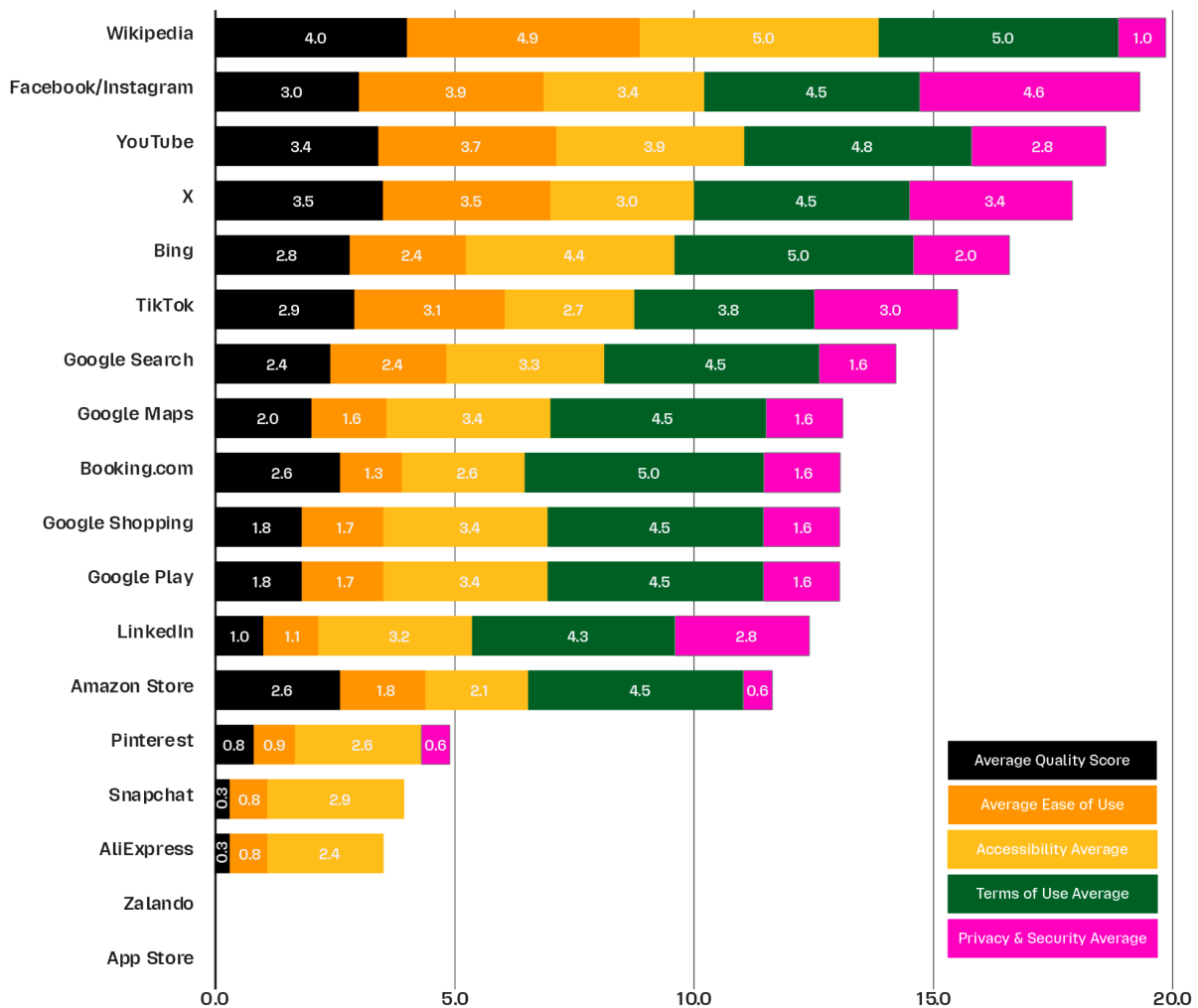
Overview

This report represents an in-depth effort to systematically evaluate the data access programs provided by the major online platforms who are regulated by the EU's Digital Services Act (DSA). We look specifically at access to publicly accessible data. Data access for researchers, journalists, and NGOs is critical to ensuring that threats on these platforms, for instance to civic discourse or public security, are identified, the companies are held accountable to the public, and ultimately that the rights of citizens are protected.

In this report we provide both a rubric for evaluating these programs based on a detailed set of criteria as well as the first set of scores using that criteria to produce a moment-in-time scorecard. We rate each platform based on an independent assessment of what the research community believes is needed to further research in the public interest. We examine what data these platforms make available to researchers, the usability of these systems, the accessibility of the programs to researchers, the terms and conditions applied to this access, and the security and privacy provisions they employ. Where possible we have tested and compared these systems to the public data we see on these sites, and otherwise we have reviewed public statements, forms, and documentation about each program. We have also reached out to each platform directly to ask clarifying questions and to seek additional information about access. Our grades reflect all of the platform offerings we could identify and examine as of May 2024 and represent a snapshot of the programs. Most of the programs detailed in this report are new, and several changed over the course of our research. We hope and anticipate that these programs will continue to expand and evolve through further use, scrutiny, and guidance by researchers and regulators.

Rankings

Each platform's data access program was assessed using 47 measurement questions grouped into five categories: Quality, Ease of Use, Accessibility, Terms of Use, and Privacy and Security. The aggregate results rank the platforms by adding the average scores for each category together into one composite score:



Based on these findings, we identify key recommendations for regulators both in the EU, and more broadly seeking to increase the impact of their rules, for researchers exploring how to make use of platform data, and for the platforms themselves to improve their offerings. We anticipate that this process can be repeated periodically to evaluate improvements over time for public interest research.

Key Findings

Platforms have employed four distinct approaches to providing researcher's public data access

We observed that platforms have varying approaches to sharing public data with researchers, that can be summarized as:

1. **Formal programs:** five platforms have created new and specific systems and tooling for researchers investigating systemic risks in the EU.
2. **Existing programs and APIs:** five platforms provide or have repurposed existing APIs and researcher programs that can be used in the context of DSA-related research.
3. **Permission to scrape:** several platforms explicitly or implicitly permit data scraping as a means of data access.
4. **Data requests:** three platforms invite researchers to request specific data sets on an individual basis but do not document what exists or provide a standard mechanism for access.
5. **No efforts to provide access to public data?:** two platforms do not appear to have developed any formal mechanism for researchers to access public data at this time.

Data scraping has emerged as an offering from several prominent platforms

Alphabet (Google Search, Play Store, Shopping, and YouTube) explicitly acknowledges data scraping as a means for enabling researchers access to public data. Booking.com, Amazon, and Pinterest *implicitly* allow data scraping for non-commercial purposes through their terms of service. By offering researchers permission to scrape data, these platforms model a valuable approach and normalize the concept which may increase the value of scraping as a tool for public interest research in the future.

Public data access programs are new, and often lack visibility and documentation

Platforms have only recently introduced new programs or adjusted existing ones. At this time, few researchers have gained access, and fewer still have systematically probed the data to understand the benefits and limitations of the programs. Critically, access to

information, data documentation, and technical support all appear to be extremely limited at this time and in some cases very difficult to find even when they do exist.

We lack shared definitions for “public data” and who should get access to it

There exists no clear or agreed upon definition of what qualifies as “public.” Each platform has made different decisions about what data to share, and in many cases those choices may differ from researcher expectations. In addition, platforms have varying or ambiguous criteria for who is eligible to access data. This lack of clarity and standardization makes it difficult for the platforms to provide the best offerings and for researchers to conduct research.

Rapid response and exploratory research are not readily enabled

A significant portion of the research programs offered at this time require detailed applications and vetting or specific and time-bound research questions. The response time for access applications is variable, and rarely immediate. As a result, research into real-time emerging threats is limited and challenging at this time. This is significant as a primary use case of public data access is for rapid response and monitoring research, as for example with Meta’s CrowdTangle or Twitter/X’s API prior to May 2023.

Technical and policy limitations on data access may hinder research quality

Many of the programs impose rate limits on the volume of data that may be accessed, the speed at which data may be collected, and the exact data that may be stored locally by researchers. These limitations are likely to limit research that aims to define the scale of specific risks on a platform, or that processes and analyzes media posted to a platform.

Topline Recommendations

We have identified 25 recommendations for regulators, platforms, and researchers based on the findings of this research which are detailed later in this report. Our highest priority recommendations for regulators and platforms are outlined below:

Regulators should provide clear frameworks for sharing public data

Regulators globally should provide clear and harmonized frameworks for public data sharing between tech companies and public interest researchers. Such frameworks should clarify what constitutes public data for public interest research purposes. They should also ensure safe harbor for independent data collection for public interest research.

Currently, the European Union has the most developed legal framework under the Digital Services Act. While some questions related to public data sharing will likely be addressed in the forthcoming Delegated Act on Article 40, the European Commission should provide

further guidance in relation to data sharing under Article 40.12. This guidance should, inter alia, clarify what constitutes “publicly accessible data” for these research purposes and set a baseline for the specifications and documentation of data sharing, so that researchers, auditors, and regulators can effectively evaluate these programs. Several major platforms already permit data scraping as a form of data access for research, presumably because this is both easy to implement and demonstrates a clear connection with what is “public”. Critically, guidance should also ensure that researchers physically located outside of the EU will have access to data needed to conduct research related to systemic risks in the European Union.

Platforms should facilitate real or near-real time monitoring

Exploratory and rapid-response research requires timely and flexible access to public data in response to real-world events. This work is often done by non-academically affiliated researchers including journalists. Platforms should facilitate the use of public data for real or near-real time monitoring and ensure that all relevant researchers conduct this research. Platforms should allow organization-based access to research programs, in addition to project-based access. Interactive dashboards are an ideal mechanism for empowering non-technical researchers, particularly from journalism and civil society, to investigate public data in real-time.

Researchers should actively request data and document their experiences.

Many of the platforms we evaluated reported to us that they have received very few requests for data. While the “vetted” researcher process under the Digital Services Act is still being built, the public data sharing programs are already in place. Academic researchers, civil society researchers, and journalists need to explore and make data requests in order to expand a shared understanding of what kinds of research can be accomplished with platform data, and to develop a clearer sense of what kinds of data are on the menu for interrogation. Many of these programs are not widely publicized - in some cases we only discovered the existence of a program after multiple efforts to contact a platform. Through various efforts to catalog the programs including this one, more information is now available to find and apply for access. The experiences with these public data programs will be helpful for better envisioning non-public, more heavily “vetted” data sharing programs.

In conclusion

This report has tried to take a first look at the public data sharing programs of some of the world's largest tech platforms. This report focuses on the new or revised data sharing offerings released in the context of the EU Digital Services Act Article 40.12, requiring designated services to share publicly accessible data, but we also consider data sharing more broadly, and so this report is not intended to be an assessment of compliance with DSA 40.12 programs.

We find that reasonable progress has been made to increase transparency and access to public data by some of the most prominent platforms, but that significant work remains to effectively enable research across all the platforms that impact society.

Table of Contents

Executive Summary	2
Overview	2
Rankings	3
Key Findings	4
Topline Recommendations	5
In conclusion	7
Table of Contents	8
Background	9
Article 40 of the Digital Services Act	9
Approach	10
Criteria for Researcher Data Access Evaluation	10
Detailed Assessment Measures	11
Platform Evaluation Strategy	14
Observations	16
Quality	16
Accessibility	20
Security and Privacy	22
Recommendations for public data sharing programs	26
For Regulators	26
For Platforms	26
For Researchers	27
Challenges and Limitations	28
Conclusion	29
Platform Appendix	31
AliExpress	32
Amazon Store	33
Apple's App Store	34
Bing	35
Booking.com	36
Meta: Facebook & Instagram	37
Google Maps	38
Google Play	39
Google Search	40
Google Shopping	41
LinkedIn	42
Pinterest	43
Snapchat	44
TikTok	45
Wikipedia	46
X/Twitter	47
YouTube	48
Zalando	49

Background

Article 40 of the Digital Services Act

Under the Digital Services Act (DSA), the European Union initially designated 19 entities as Very Large Online Platforms (VLOPs) or Very Large Online Search Engines (VLOSEs):

- AliExpress
- Amazon Store
- Apple's App Store
- Bing
- Booking.com
- Facebook
- Google Maps
- Google Play
- Google Search
- Google Shopping
- Instagram
- LinkedIn
- Pinterest
- Snapchat
- TikTok
- Wikipedia
- X (formerly Twitter)
- YouTube
- Zalando

[Further services](#) have been designated since, but evaluating those additional platforms was beyond the scope of this research. These VLOP/SEs are subject to certain obligations under [Article 40 of the DSA](#), including that they make public data accessible to researchers studying "systemic risks to the European Union" and are "independent from commercial interests." Section 12 of Article 40 states:

"Providers of very large online platforms or of very large online search engines shall give access without undue delay to data, including, where technically possible, to real-time data, provided that the data is publicly accessible in their online interface by researchers, including those affiliated to not for profit bodies, organisations and associations..."

This research is conducted in the context of Article 40.12 being in force, but looks more broadly at what access to public data is available. This assessment is intended to be understood as the interpretation of the research community needs, and **cannot and should not** be interpreted to assess compliance under the Digital Services Act.

Approach

To develop a standardized methodology for evaluating public data access programs, we first identified the primary criteria that are relevant to researchers in collaboration with a network of stakeholders, like civil society research organizations and independent platform accountability experts. We have defined five distinct categories of measures as the criteria employed throughout the evaluation.

Criteria for Researcher Data Access Evaluation

Quality	An evaluation of the quality and comprehensiveness of the data provided, based on a range of criteria derived from real-world data usage for research (see below) as well as a detailed audit of the public data visible on each of the platforms. The grades for quality will help policymakers and the public to understand the breadth of the data they are sharing.
Ease of Use	An evaluation of the functionality of the system that provides data, based on real-world researcher practices and concerns. This set of grades will examine the practical usability of the system and data compared to expectations and best practices. These grades will demonstrate whether or not the tooling and limitations provided by the platforms are sufficient to actually conduct the critical research protected by the DSA.
Accessibility	An evaluation of the processes for gaining access to public data by researchers and other stakeholders. These grades will make clear whether or not the platforms are facilitating broad, diverse, and meaningful access to the community that DSA Article 40.12 is, in our view, designed to serve.
Terms of Use	An evaluation of the terms under which researchers can use the data. This set of grades will look both at how well the terms communicate what researchers can and cannot do, as well as compare those provisions to the practical concerns researchers have when conducting and publishing research.
Privacy and Security	An evaluation of the provisions established by the platforms and the provisions imposed on researchers in order to maintain the security of the data and protect the privacy of individuals who create or are referenced in the data available for research.

Detailed Assessment Measures

For each category, we developed a set of assessment measures in collaboration with a network of stakeholders including civil society researchers, academics, and independent platform accountability experts, who have studied these online platforms and made use of the various methods for accessing and analyzing public platform data.

For each measure, we developed a description, guiding questions, and a specific assessment question that was answered for each platform. Each assessment question is designed to be answered affirmatively for a pro-researcher outcome.

Quality	
Coverage Public	Does the platform provide access to the standard public data that is visible to users?
Coverage Inferred	Does the platform provide access to data that is inferred about or appended to users and content?
Coverage Aggregate	Does the platform provide access to aggregate data about users and content?
Media	Does the platform provide access to media content for analysis?
Recency	Does the platform provide real-time data on activity?
Historical	Does the platform provide full historical access to its data?
Granularity	Does the platform provide granular access to individual data attributes?
Completeness	Does the platform provide access to removed content?
Consistency	Is the data provided by the platform consistent over time?
Chronology	Does the data provide time-series chronological information?
Ease of Use	

API Access	Does the platform provide an API to enable researcher access to data?
Dashboard Access	Does the platform provide an interactive dashboard to enable researcher access to data?
Download Access	Does the platform provide data via downloadable archives?
Speed Rate Limiting	Does the platform enable unlimited frequency of requests for querying the data access system?
Volume Rate Limiting	Does the platform provide unlimited access to the volume of data a researcher can collect and analyze?
Independent Storage	Does the platform allow the researcher to extract and store the data independently?
Access Restriction	Does the platform allow the researcher to access the research API / system from anywhere?
Combinability	Does the platform permit joining data with externally sourced data or tools?
Search	Does the platform provide keyword search and filtering capabilities?
Boolean Logic	Does the platform provide the ability to apply boolean logic to search and filter data?
Documentation	Does the platform provide detailed documentation on the use of the data platform?
Language Accessibility	Does the platform provide documentation in EU local languages other than English?
Feedback	Does the platform provide a mechanism for submitting feedback and feature requests for the system?
Support	Does the platform provide a support service to researchers?
Accessibility	

Eligibility	Are all types of researchers eligible for access to the researcher data offering?
Exclusivity	Are researchers able to apply without unreasonable demonstration of approvals and credentials?
Geo-limiting	Are researchers able to apply from anywhere in the world?
Vetting	Are the researcher vetting requirements fair and appropriate? ¹
Data Access Affordability	Is the access to the data provided free of charge?
Security Requirements	Are the security requirements for data access reasonable? ²
Infrastructure Affordability	Is it affordable to maintain the infrastructure required to work with the data?
Application Ease	Is the access request process expected to be completed in a reasonable amount of time?
Rapid-response Feasibility	Can rapid-response research be conducted?
Collaboration Friendly	Does data access permit data-sharing within and across institutions?
Simplicity	Is the application process simple, user-friendly, and easy?
Appeals	Are researchers able to understand and appeal decisions around access?
Flexibility	Is it possible to revise or extend research use-cases on existing access approvals?

¹ Vetting requirements were judged based on the questions: Are the requirements for approval onerous or limiting in ways that make it difficult to get access? Are IRB approvals required in order to get access?

² Security Requirements were judged based on the questions: Does the platform require levels of security which are not easily maintained by individual researchers? Does the platform require the use of specific tools or processes? Does the platform require processes or procedures which negatively impact the research process?

Longevity	Is the researcher access available for a significant and reasonable period of time?
Terms of Use	
Fairness	Are the terms of use reasonable given the research landscape?
Clarity	Are the terms of use easy to review and comprehend?
Autonomy	Do the terms ensure researcher autonomy and independence?
Technical Limitations	Do the terms ensure there are no significant technical limitations that could impact research?
Privacy and Security	
Personal Data Protection	Does the data access ensure adequate protection of personal data?
Aggregation Privacy	Are data aggregation limits sufficient to protect against re-identification?
User Privacy Expectations	Does the data access uphold our inferred ³ user expectations about public and private forms of data?
Data Security	Does the data access system provide adequate security protocols? ⁴
Oversight	Is there sufficient oversight of researcher activity to ensure compliance with security and privacy protocols?

Platform Evaluation Strategy

To answer the questions, we reviewed documentation available on the VLOPs' and VLOSEs' websites (e.g., terms of use, researcher program descriptions, researcher program applications, API documentation, etc.) and documented our findings in an Airtable database. We referenced existing research into data access programs, including the [Platform Transparency Tools Tracker](#), launched by [Anna Lenhart](#) and Annika

³ We inferred user expectations about privacy subjectively by contemplating for a given platform surface whether or not we would assume such data would be seen by people other than those we intended. For example, we assume a public TikTok video can be seen by the world, but our comments on a Facebook friend's personal photo would not.

⁴ Data security was assessed by reviewing the policies described and applied, and where applicable, the technical requirements associated with the data access.

Springsteel and maintained by the Institute for Data, Democracy & Politics at George Washington University.

We also reached out to each VLOP/SE describing our research and requesting additional information.

Scoring

We developed a strategy for scoring each platform on each measure using a scale from 0-5, based on the relative extent to which they met our criteria. A score of zero indicates the criterion was not met; a score of five indicates the criterion was fully met. A score of one through four indicates the criterion was partially met. In many cases, platforms score 0 for a measure because they do not offer a program or service, or because it is impossible to measure what they offer at this time. We internally mark scores that are not measurable in order to evaluate averages with and without these measures, however the final averages and ranking include every score including those which were “not applicable.” Finally, in several cases it was necessary to infer based on context scores where we know enough about the programs and the data, and felt this approach more accurately reflected relative value than using an absolute determination based on public information. Thus on the whole, we try to give a fair representation of the relative value on these measures between platforms even though the scores were determined subjectively.

Observations

Our assessment of the VLOP/SEs yielded key insights into the different platforms' approaches to making publicly available data accessible to the research community.

Quality



Time series data is missing from virtually all providers

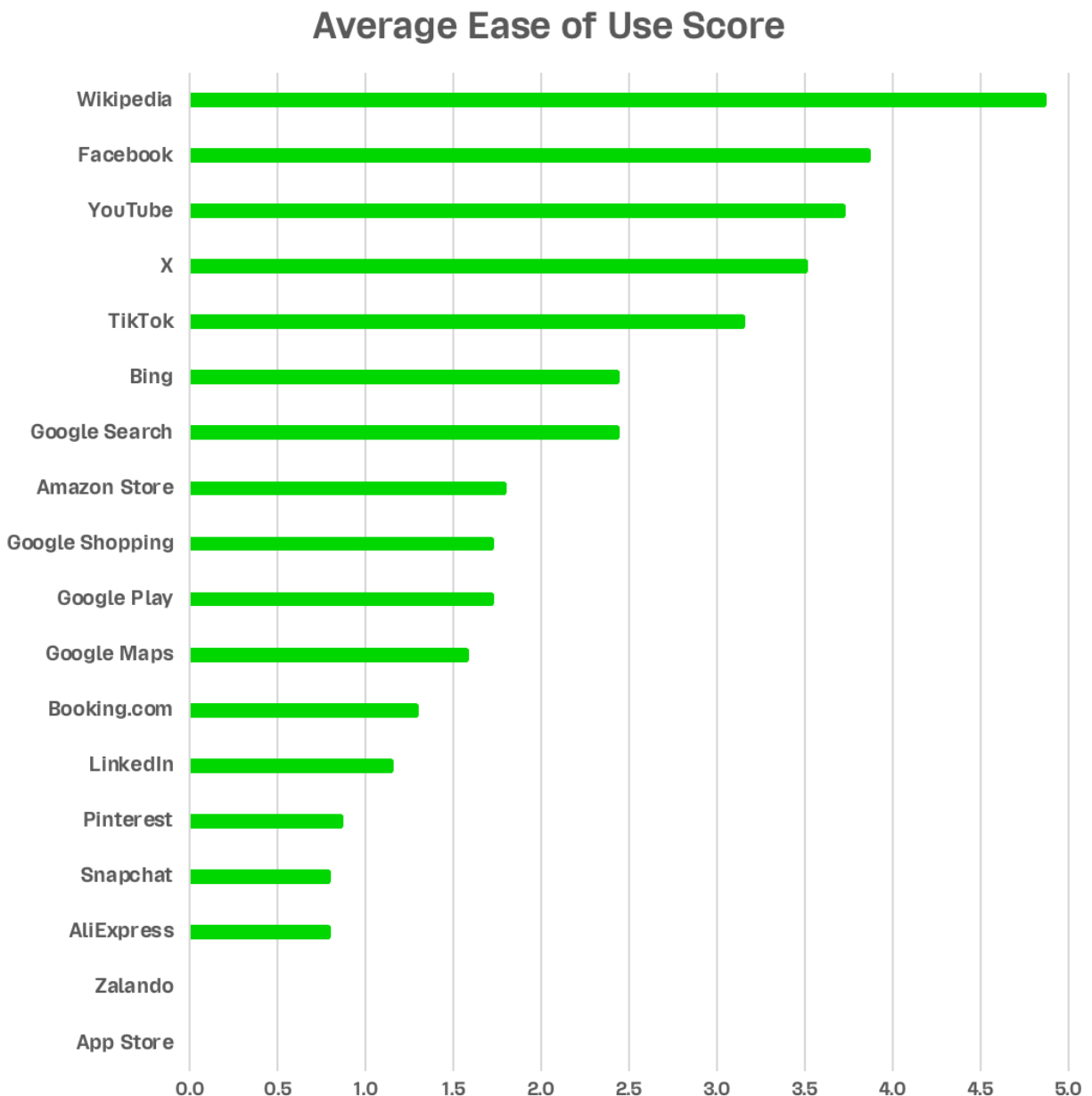
Researchers need access to information about what has happened over time to understand trends and spikes in interest in subjects, and to characterize the growth in reach and influence of problematic actors. Previously, CrowdTangle provided this data about Facebook and Instagram, but the Meta Content Library has lost this feature, and none of the other platforms except Wikipedia maintain such change-over-time data. Bing's search API appears to provide search volume over time similar to Google Trends data.

Consistency in the data over time is difficult to verify

Researchers need to trust the data they will use when attempting to publish, and almost none of the providers make clear guarantees about the consistency of the data. Meta has released a new tool called the Content Library, which has documentation that specifically acknowledges that the data it returns will not be consistent from day to day or truly representative of what is on the platform for a given search⁵.

⁵ <https://developers.facebook.com/docs/content-library-and-api/content-library/#additional-information>

Ease of Use



There are no significant access restrictions except for Meta

Access restrictions are minimal across the board, except for Meta's offering. Most providers make it easy to connect to the data regardless of where you are and do not require special software or strategies for using the data. Meta's Content Library is a significant exception because it requires using both a VPN to connect and applies significant limitations on how to collect and analyze data from their API, such as only showing the first 1000 results in the dashboard, and automatically deleting data from the researchers' research tool every 30 days.

APIs are the primary mechanism of data access

APIs are widespread with the highest ranking providers, and are the ideal mechanism for making data accessible to researchers in a systematic manner, but require some technical skill to leverage.

Dashboards are an unmet need

Dashboards for public data do not exist for any services except Meta. This kind of accessible exploratory tool is widely used by non-technical and non-academic researchers at newsrooms and NGOs. While such tools are expensive to build and maintain, they would significantly expand usability, especially in support of real-time, rapid-response, and targeted research. A partial exception exists for Wikipedia, because the platform itself is effectively an exploratory dashboard in which all the data and history is explorable on the web.

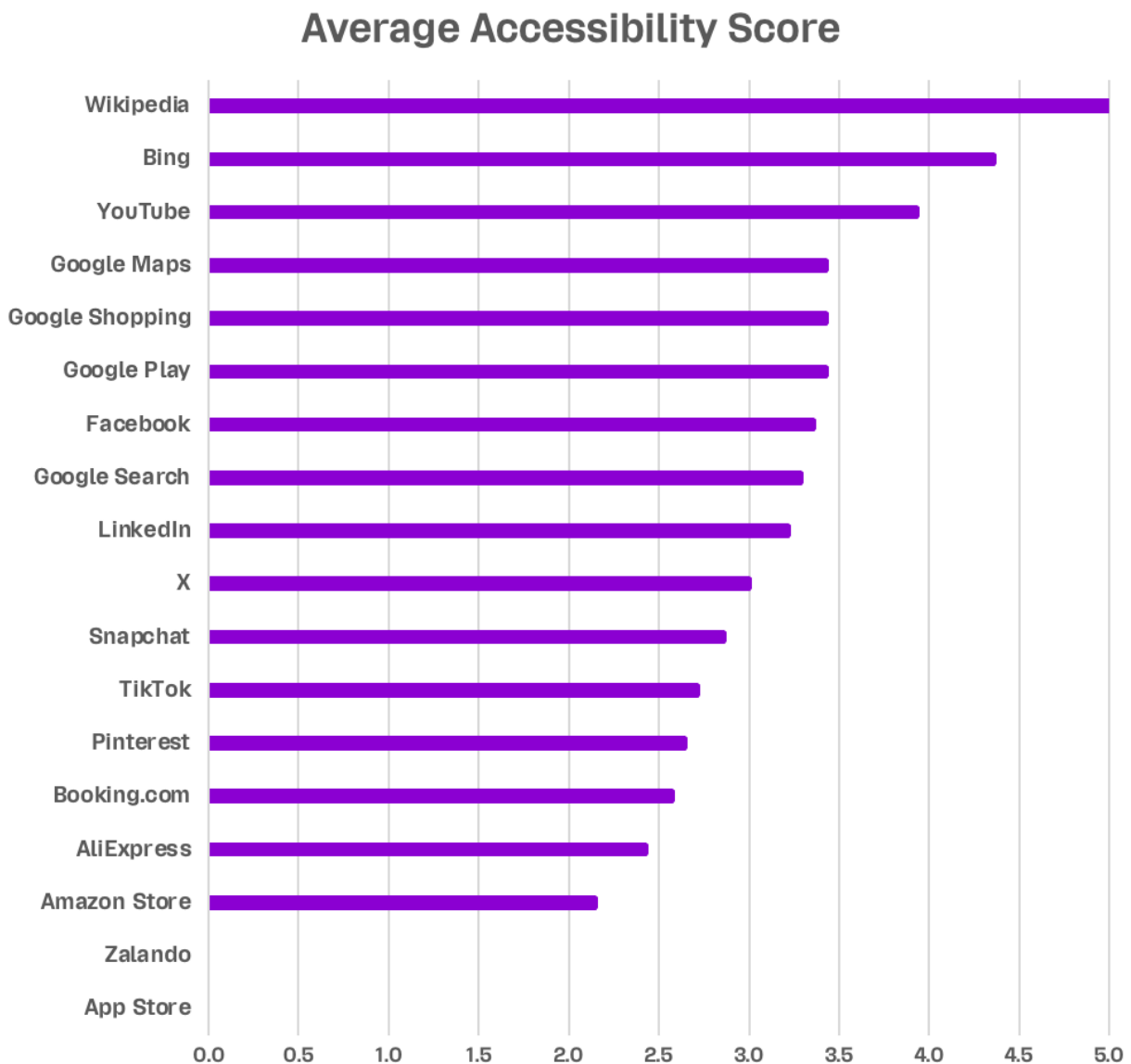
Customer support for researchers is still in its infancy

Documentation, feedback systems, and technical support are extremely limited overall. The major providers do offer some of these components, though effectively only in English, and not in the other major languages spoken in the EU.

Rate-limits may significantly impede research

The majority of programs have significant rate limitations for the volume of data and speed to access it. While the restrictions may be intended to prevent abuse of these systems, the measures are impractical and significantly limit the potential for researchers to carry out larger-scale and longitudinal analysis, which by their nature require large volumes of data in order to identify systemic risks.

Accessibility



Programs are not designed for emerging crises or information shocks

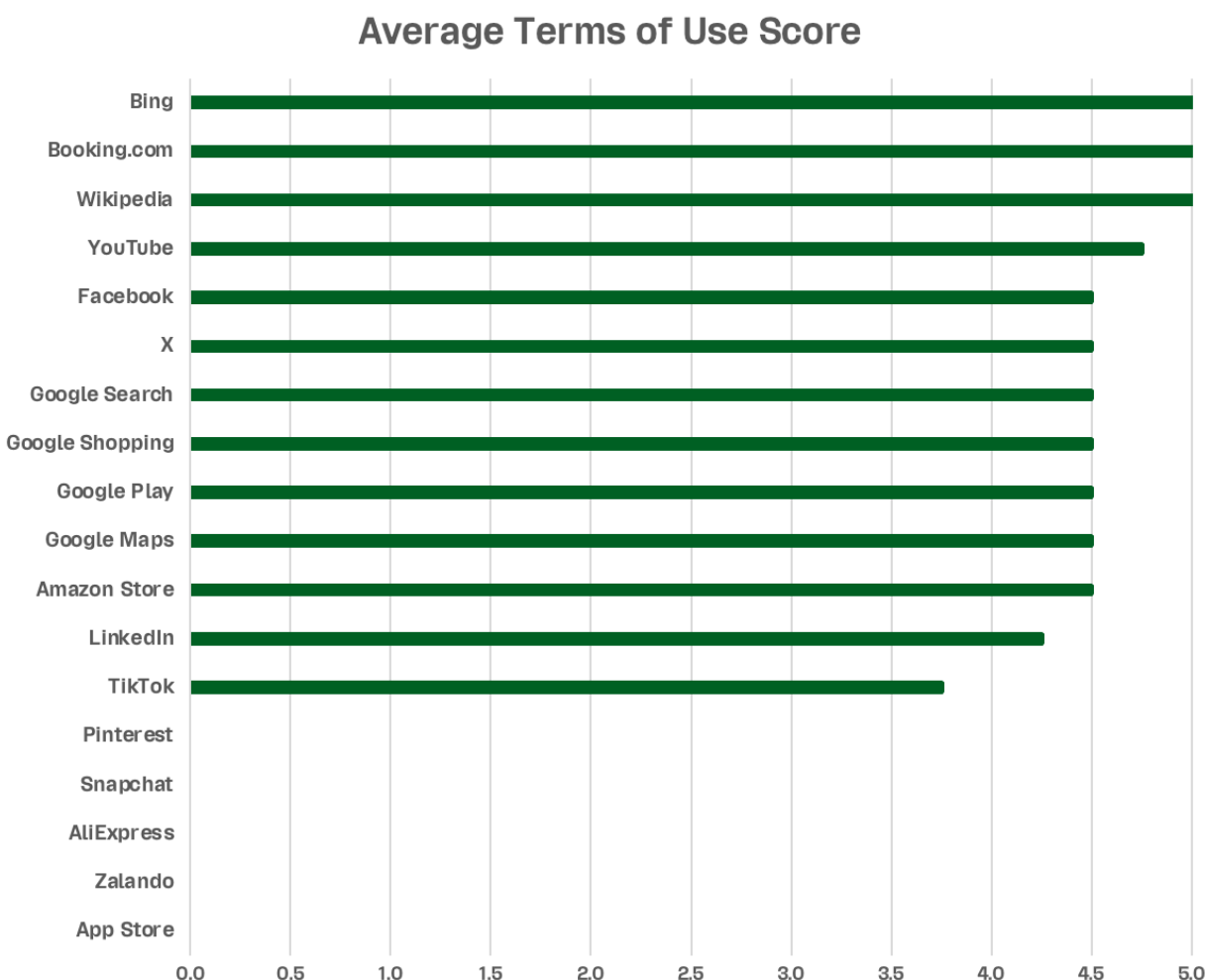
The current structure of the programs do not appear to enable rapid-response research that emerges from current events. For example, if a researcher wants to monitor narratives and trends unfolding in real-time around the conflict in Gaza to identify the prevalence of hate speech⁶, or if they wanted to understand the prevalence of foreign disinformation in the conversation about the European farmers' protests⁷, the research application process would significantly hinder rapid-response monitoring on nearly all

⁶ <https://www.nytimes.com/2023/11/15/technology/hate-speech-israel-gaza-internet.html>

⁷ <https://www.politico.eu/article/europe-farmer-protest-russia-war-propaganda/>

providers. Users with existing access could leverage that access to pivot to rapid response, but might risk violating their agreements which were developed and signed in the context of other specific areas of research.

Terms of Use



Terms are generally fair

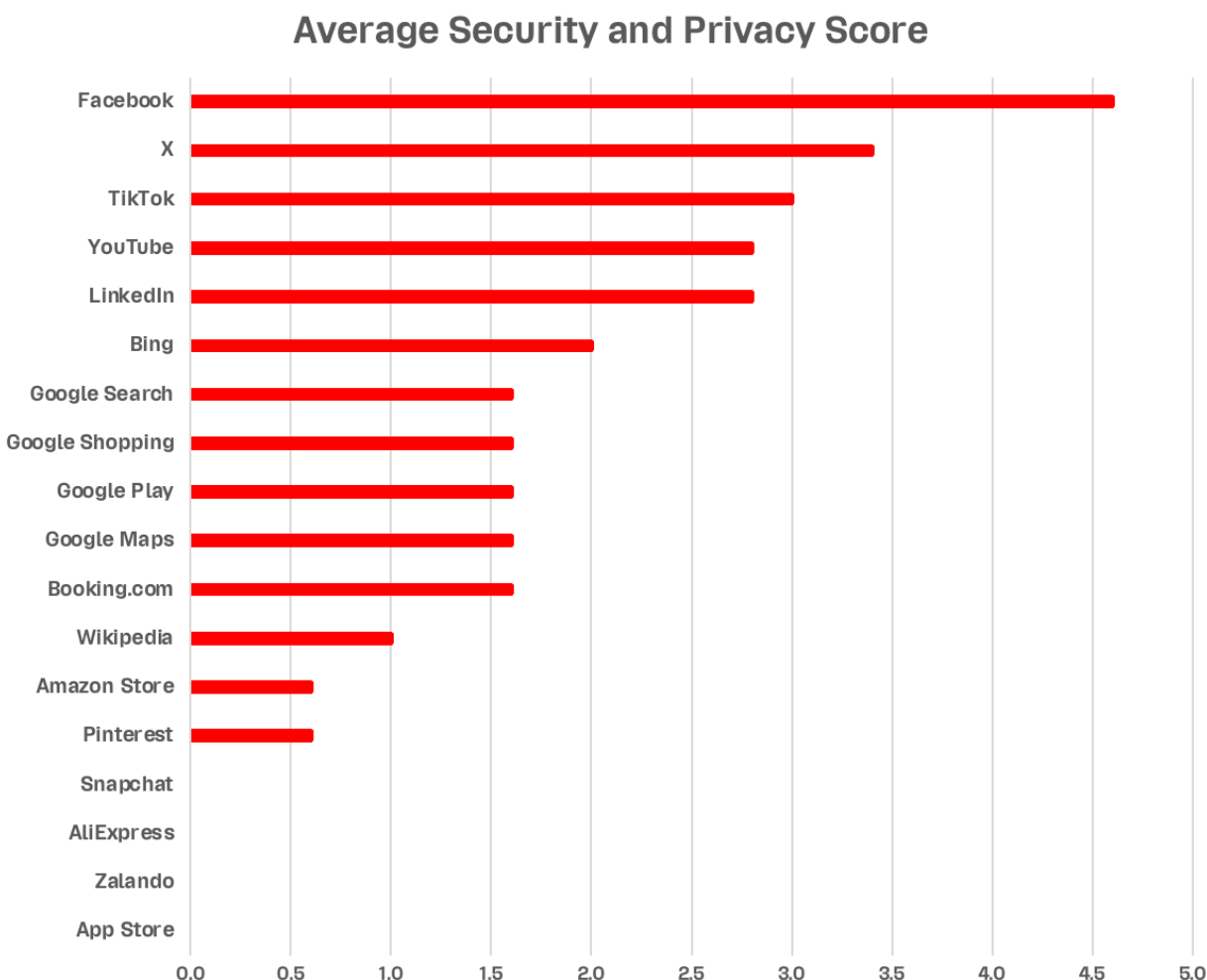
The terms of use for almost all programs—where they exist—are reasonable and straightforward. They do not generally restrict autonomy, with the mild exception of TikTok, which requires pre-publication review.

Research scale may be compromised by technical limits imposed by the terms of use

Technical limitations specified in the terms that relate to data retention could create various issues that impede research, however those limitations are driven at least in part

by legitimate privacy concerns. The practical challenges created by legal terms that limit the volume, speed of access, and scraping of data can significantly impact the quality and diversity of research that can be produced.

Security and Privacy



Privacy appears to be generally protected

Most of the data access programs appear to maintain our inferred user expectations of privacy because they only grant access to data that is publicly available. However, those expectations are somewhat ambiguous when it comes to public data about comments, which is made available from YouTube, TikTok, and Meta.

Security is self-reported and potentially performative

While many of the programs require researchers to provide information about data security and related privacy protections, there appears to be almost no mechanism for

auditing or enforcement, except with YouTube. In addition, Meta maintains control over the research infrastructure which largely eliminates the need for verification.

General

Data Access Programs come in four flavors

Platforms have exercised significantly differing strategies in making public data accessible to researchers. As we discuss below, some programs afford researchers significantly more access than others, and in some cases – as with data requests – it’s challenging to assess precisely how expansive access is. Our work identified four types of access:

	Overview	Platforms
Formal Programs	These platforms have established new, formal programs to service researchers with access to APIs, dashboards, and other resources created for the purpose of research under Section 40.12 of the DSA.	AliExpress Google Search Facebook Instagram TikTok LinkedIn
Existing APIs and Access Programs	These platforms have previously established APIs and data access programs that make public data available. Bing, Wikipedia, X, and YouTube have elected to make these APIs available to eligible researchers under Section 40.12 of the DSA.	Bing Google Maps Wikipedia X (Twitter) YouTube
Permission to Scrape	These platforms explicitly or implicitly allow the scraping of public data by researchers. Google Play, Shopping, and Search permit “limited scraping.” The customer terms of service (Section A14. Intellectual Property Rights) for Booking.com state, “You’re not allowed to monitor, copy, scrape/crawl, download, reproduce, or otherwise use anything on our Platform for any commercial purpose without written permission of Booking.com or its licensors.” Given that the access granted to	Amazon Store⁸ Google Play Google Shopping Booking.com Pinterest⁹

⁸ Amazon has confirmed to our research team via email that “using industry-standard, research-specific tools to collect publicly available data” is not prohibited.

⁹ Pinterest confirmed via email that they “would allow scraping by a researcher who was qualified under DSA Art. 40(12)”

	public data under Article 40 of the Digital Services Act requires researchers to be “independent from commercial interests,” it appears that DSA-eligible researchers are permitted to scrape the platform.	
Data Requests	These platforms provide a way for eligible researchers to make specific requests for public data, which the platform then provides (e.g., in a .csv file). Researchers are not given broad access to public data; rather, they must describe the kind of public data their research requires, and the platform will respond directly to the request. These programs currently lack significant documentation about the format data will be provided in or specifics about what data may be made available.	Booking.com Pinterest Snapchat
None	Two VLOPs do not appear to have made any effort to provide access to public data. As of the time of writing, we have not been able to identify any formal programs, specifically relevant APIs, or other data offerings from Apple’s App Store ¹⁰ or Zalando ¹¹ .	Apple App Store Zalando

Programs are new and untested

Many of the programs and other offerings made available by VLOPs and VLOSEs are new and have not been used extensively by the research community. At least two platforms – Snapchat and Pinterest – said that no researchers had made use of their programs as of early 2024 when asked.

While our work sought to test these programs and offerings, we were only able to gain access to a few of them, namely the Meta Content Library, Bing’s researcher program, the YouTube API, and Wikipedia. Furthermore, many researchers have only just begun using the programs themselves, or are in the process of applying to them. These programs have not faced extensive scrutiny by researchers at this point. The true utility of these programs will become clearer as a larger diversity of researchers test the data access offerings using different methods and exploring a variety of questions and problems.

¹⁰ Apple responded to our inquiry, stating that DSA-eligible researchers can contact the platform to receive publicly accessible data.

¹¹ Zalando’s [legal notice](#) page indicates that researchers can contact them for data access, but provides no information about the offering on their website.

Programs are changing

During the course of our research, many VLOP/SEs stood up new programs/data access offerings, or changed their offerings. This necessitated revisiting our grading as programs and applications changed. **Our grades reflect platform offerings as of May 2024.**

Programs don't explain what they offer

Apart from the differences between program offerings, VLOP/SEs vary widely in the types of documentation and support they make available to researchers. In general, programs that offer APIs tend to make more extensive documentation available to researchers to describe the types of data available and how to use it. Platforms that require researchers to make specific data requests offer very little information about what data they will share (beyond "public data") or how the data will be formatted. This results in "cart before the horse" research planning, as researchers cannot understand what is available. This makes it challenging to formulate precise research questions before applying for data access. This may discourage researchers from using the programs.

There is currently no formal definition of "publicly accessible data" which allows platforms to remain ambiguous about how they define the "public data" offered.

Programs lack tools for those without technical skills

Few platforms offer dashboards or other tools that would allow a researcher without any technical skills to view or analyze data. Meta, a notable exception, introduced the [Meta Content Library and API](#) in 2023, which includes tools to search and view data via keywords and "producer" lists. In 2024, Meta added the ability for researchers to download a selection of publicly available data from sources that have significant public followings (pages with more than 15,000 likes and profiles with greater than 25,000 followers).

Loss of CrowdTangle

CrowdTangle, a research platform from Meta that will be discontinued as of August 2024, has served as an example of a relatively robust data access tool that can be used extensively by technical and non-technical researchers alike. Unfortunately, the Meta Content Library (MCL) does not yet offer the same functionality and in some cases level of granularity as CrowdTangle. While Meta's Content Library demonstrates the most extensive effort to provide researcher-specific tooling, and adds several new types of data that were previously unavailable to researchers in CrowdTangle, the system lacks features and data that were previously available.

Recommendations for public data sharing programs

For Regulators

1. **Regulators should offer a formal definition of “public data”** in regulatory guidance that platforms, researchers, and regulators can reference.
2. **Regulators should require platforms to publish specifications and documentation** for their researcher access programs.
3. **Regulators should expand researcher access needs** to support exploratory research without a pre-defined research question or hypothesis, in relation to systemic risks.
4. **Regulators should allow organization-based access** to research programs, rather than project-based access to enable researchers to continually monitor and probe data for new questions/problems.
5. **Regulators should require platforms to capture and publish the number of researcher data access requests they receive**, the number they approve, and provide guidance on their standards for approval.
6. **Regulators should establish a way of including the behavior of algorithmic recommender systems** in the concept of public data so that researchers can request access to the aggregate outputs of algorithmic recommendation systems to understand which public data users see.
7. **Regulators should clarify the legitimacy of scraping** as a means of accessing public data. Several major platforms already permit, either implicitly or explicitly, data scraping as a form of data access for research purposes. Implicit permission to scrape, as we have received for instance over email from some providers, is likely not sufficient reassurance for most researchers. Legal ambiguities remain and privacy concerns in the practice of scraping are therefore not systematically addressed.
8. **Regulators should provide funding to support** and encourage researchers' initial exploration and testing of these programs.

For Platforms

9. **Platforms should publish more extensive documentation** about what data is available to researchers and how that data will be provided.
10. **Platforms should provide some way of auditing what is “publicly available”** within a given platform by distributing data schemas, surface audits, or other means that enable researchers and regulators to have a clear and complete understanding of what the typical user can see on the platform.
11. **Platforms should create offerings for non-technical researchers**, such as dashboards or downloadable datasets for non-programmatic analysis.

12. **Platforms should enable and encourage the creation of third-party tooling** such as dashboards, data-donation repositories, and historical archives to increase accessibility and capacity for collaborative research.
13. **Platforms should provide easy-to-find or direct links to their DSA-specific researcher data access programs.**
14. **Platforms should maintain and enable access to time-series data** about content engagement, account growth, and any other relevant attributes that change over time and are necessary for monitoring systemic risks
15. **Platforms should develop mechanisms for documenting and validating the consistency and accuracy of data over time.**
16. **Platforms should develop more robust researcher resources** for both applying for access and using data. This includes documenting the anticipated wait time for applications, providing examples of data privacy and security protocols to comply with, and providing technical customer support and feedback tools for the data access tools themselves.
17. **Platforms should increase rate-limits** and/or provide mechanisms for securing exceptions to the limits for specific use-cases.
18. **Platforms should create multi-language documentation.**
19. **Platforms should create a path for rapid-response access** during moments that demand action.
20. **Platforms should develop and document processes for reviewing and auditing researchers' security and privacy protocols** in ways that are material and verifiable, but do not feel disproportionately burdensome.

For Researchers

21. **Researchers should apply to and use these public data programs** to improve understanding of offerings, document and report on challenges of applying and using data, and participate in [field-wide surveys](#) to increase shared awareness.
22. **Researchers should share best practices for meeting privacy and security requirements** for data access to reduce the burden on researchers who are not familiar with or don't have capacity to develop these protocols.
23. **Researchers should develop a shared public list of researcher data access use-cases** that represent edge-cases not currently supported by platform programs.
24. **Researchers should identify and document time-sensitive rapid-response moments** that would benefit from researcher data access that are not possible under the current programs.
25. **Non-academically affiliated and non-EU-based researchers should partner with EU-based academic institutions to pursue research** related to systemic risks to exemplify the value of global research collaboration, especially those in Global Majority context.

Challenges and Limitations

We wish to acknowledge some significant challenges we faced in our effort to develop this scorecard. Understanding these limitations is necessary context for how the current set of scores can be used responsibly, and should also guide future efforts to expand this work

Program Access

We were not able to access most of the programs themselves, because of the various limitations imposed by the applications for access, including geography, specific research requirements associated with systemic risks, academic credentials, and technical specifications for data security and privacy. This means that our scores are based primarily on documentation, public statements, feedback from researchers who currently possess or have applied to get access, and in some cases, inferences based on general knowledge of the programs and platforms.

As of May 1, 2024, few other researchers have secured access to these programs, which has limited our ability to get detailed information about their experiences with usability, completeness, consistency, the timing for application reviews, and other factors. We anticipate that forthcoming research, for example as led by the Coalition for Independent Technology Research will provide a much larger sample of researchers' experiences that will shed light on these issues.

Documentation

The current state of researcher access programs generally suffer from limited or non-existent documentation. In addition, for many of these programs, the information and application forms that do exist are buried and difficult to find. As a result, it is impossible to verify if what we found is comprehensive and accurate. In many cases, the information about researcher data access programs linked to the DSA obligations live on separate sites without good SEO or links to the top-level of the terms, trust and safety, or other transparency resources. After we developed draft evaluations and shared their scores for comment, more platforms did respond to provide additional information and context.

Timing

Over the course of this period, several programs and application forms emerged that were not available when we first began. In addition, programs changed over the course of this research, offering new information, or announcing enhanced features. The new releases are often un-dated making it difficult to determine when they were released and/or changed. We have made our best effort to revise our grades and collect additional evidence, but because these programs are actively under development, some findings may be outdated upon publication.

Platform Distinctions

The 19 VLOP/SEs vary dramatically from one another. Several are retail marketplaces for physical and virtual goods and services (Amazon, Apple App Store, Booking, Google Play, Zalando) others are mainstream social media platforms (Facebook, Instagram, TikTok, X, YouTube), some are more private social networks (LinkedIn, Snapchat) and others are more complex services (Bing, Google Maps, Google Search). As a result, it is difficult to evaluate what constitutes “public” data on the platforms, and to compare their offerings to one another. While this baseline is a meaningful start, future work in this space should likely separate these distinct types of platforms from one another for analysis.

Lessons Learned

Our initial set of measures and categories of criteria for evaluation we developed with care and feedback from the community of stakeholders who will use this data access. However, after attempting to evaluate the platforms on these measures, it became clear that although we wish to know the answers to these questions, it is not feasible to identify those answers. This was true for various measures, but particularly the Security and Privacy questions, as well as areas where we start from a hypothetical expectation for data access, including the Coverage of Inferred Data, the Rapid Response Feasibility, and Aggregate Data.

Broad Scope

There are some key aspects of these platforms which are beyond the scope of this methodology. Algorithmic recommender systems are not explicitly mentioned under DSA 40.12, but represent a possible factor contributing to systemic risks, for instance to civic discourse, electoral processes, or public security. These features straddle the line between what can be considered “public data” and proprietary systems. The results of algorithmic recommenders are frequently visible as public data, but their risk and impact can only be evaluated in aggregate (e.g., how many people are recommended toxic content based on behavior). Currently, these platforms do not provide researchers access to information about algorithmic recommenders, so this research does not attempt to address this critical area. This will likely be an area for DSA “vetted researchers” to interrogate through Article 40.4 requests, which allow researchers meeting certain qualifications to apply to access non-public data sets. Access to public data via Article 40.12 could be an important exploratory mechanism to help formulate stronger Article 40.4 requests.

Conclusion

This research has sought to take a first look at many of the new or revised public data sharing mechanisms provided by large online platforms, in the context of the Digital Services Act. The DSA is establishing a novel structure for data sharing between platforms and vetted researchers under its Article 40. However, it also establishes a requirement for public data sharing (Article 40.12).

In this research we consider public data sharing by 19 platforms. We developed an assessment rubric to score each program based on criteria developed in consultation with partners. In our 47 criteria, we try to reflect the research needs of the wider public interest research community who currently or would likely make use of this public data.

We developed these criteria in the absence of key regulatory concepts like a definition of public data. Our scores **cannot and should not** be interpreted to assess compliance under the Digital Services Act.

We acknowledge that our scores also represent a snapshot or moment in time look at these programs. Indeed, we hope this report will nourish the continued development of these programs. To that end, we provide our rubric for evaluating these programs based on our detailed criteria along with this first set of scores. We also offer 25 specific recommendations to regulators, platforms, and researchers.

Platform Appendix

Detailed Scores Breakdown

For a complete list of each platform and their respective scores on each measure, we have produced a separate document that can be accessed [here to review, compare, and download the data](#). These scores are accurate as of May 1st, 2024 and do not reflect any changes by the platforms since that date.

AliExpress

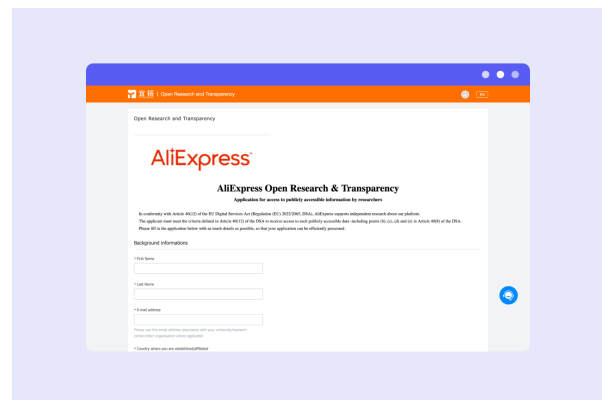
Overview: AliExpress has established the Open Research and Transparency portal, which allows approved researchers to request access to publicly available data. AliExpress [offers](#) “privacy protected datasets and other statistical data” as well as the Dataworks API through an “internal network-based controlled access environment that does not allow data to be exported.” Researchers affiliated with academic institutions and nonprofit organizations in the EU can apply for access.

What we looked at:

- Researcher access overview site
- Application form

Resource Links:

- [EU Digital Services Act](#)
- [Open Research & Transparency](#)
- [AliExpress Open Research & Transparency: Application for access to publicly accessible information by researchers](#)



Outreach:

- We reached out to AliExpress for comment and discussion, but did not receive a response.

Amazon Store

Overview: Amazon Store does not offer a formal program for data access at this time. However, a spokesperson for Amazon Store responded to our inquiry via email, “We can confirm that our Conditions of Use do not prevent researchers who comply with the conditions of the DSA from using industry-standard, research-specific tools to collect publicly available data from our EU Store, for the purposes of Article 40(12) of the DSA.” The feasibility of collecting data in this way remains ambiguous because of the potential for anti-scraping systems to inhibit this activity in the real-world.

What we looked at:

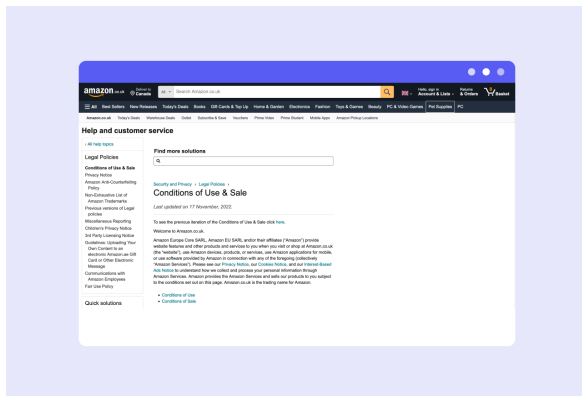
- Website terms and conditions

Resource Links:

- [Conditions of Use & Sale](#)

Outreach:

- Amazon EU responded to our request for comment and discussion, stating that the EU store Conditions of Use permit DSA-eligible researchers to use “industry-standard, research.-specific tools” for public data collection.



Apple's App Store

Overview: Apple's App Store does not offer a formal program for data access at this time. When reached for comment, a member of Apple's DSA Compliance team stated the company has "procedures in place for researchers to request App Store data that is publicly available, in accordance with Article 40(12) of the DSA" and that researchers "interested in obtaining publicly accessible data may contact Apple DSA Compliance."

What we looked at:

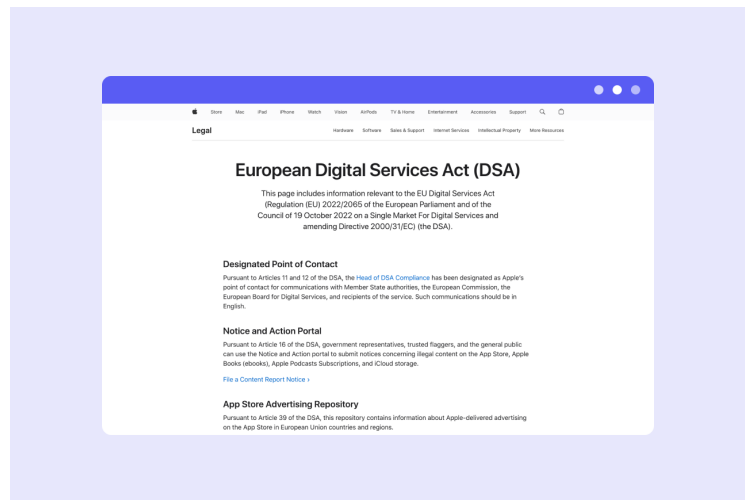
- Digital Services Act compliance page

Resources:

- [European Digital Services Act \(DSA\)](#)

Outreach:

- We reached out to Apple for comment and discussion. We received a response from Apple, which detailed the above.



Bing

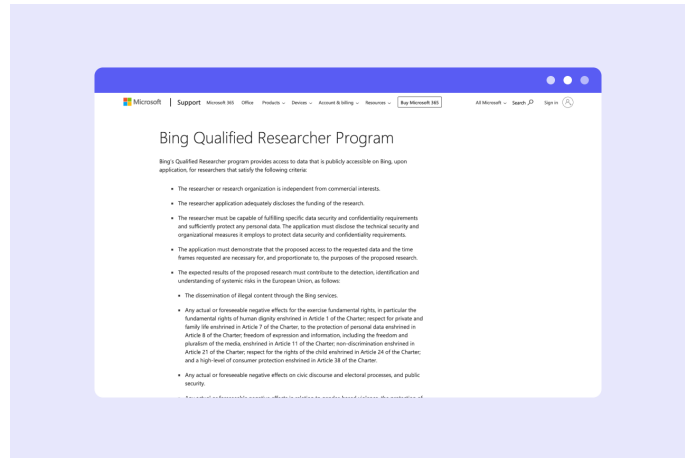
Overview: Bing has established the Qualified Researcher Program to offer researchers access to public data. Researchers free from commercial interests can apply from anywhere in the world. The program makes available the Bing Search API, Bing Webmaster Tools, and datasets to approved researchers.

What we looked at:

- Digital Services Act compliance page
- Researcher access overview site
- Application form
- Tool documentation

Resources:

- [EU Digital Services Act information](#)
- [Bing Qualified Researcher Program](#)
- [Bing Qualified Researcher Program Application](#)
- [Bing Research Resources](#)
- [Bing Web Search API](#)
- [Bing Webmaster Tools](#)



Outreach:

- We reached out to Bing for comment and discussion. We applied for and received access to Bing's program.

Booking.com

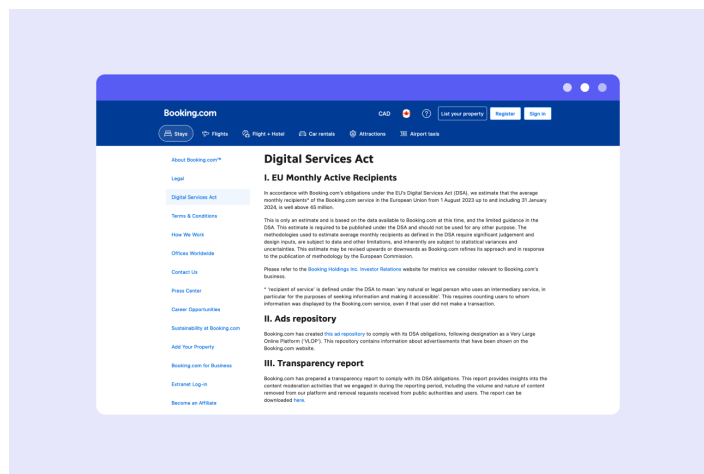
Overview: DSA-eligible researchers can apply for access to public data via Booking.com's Researchers Data Request Portal, though it is not clear in what format the data will be provided. In addition, the [customer terms of service](#) (Section A14. Intellectual Property Rights) for Booking.com state, "You're not allowed to monitor, copy, scrape/crawl, download, reproduce, or otherwise use anything on our Platform for any commercial purpose without written permission of Booking.com or its licensors." Given that the access granted to public data under Article 40.12 of the [Digital Services Act](#) requires researchers to be "independent from commercial interests," it appears that DSA-eligible researchers are likely permitted to scrape the platform.

What we looked at:

- Website terms and conditions
- Digital Services Act compliance page

Resource links:

- [Digital Services Act](#)
- [DSA Researchers Data Request Portal](#)
- [Booking.com Researcher Data Use Policy](#)
- [Customer terms of service](#)



Outreach:

- We reached out to Booking.com for comment and discussion, but did not receive a response in time to incorporate into our report. We received an automated response with an application to determine eligibility for data access.

Meta: Facebook & Instagram

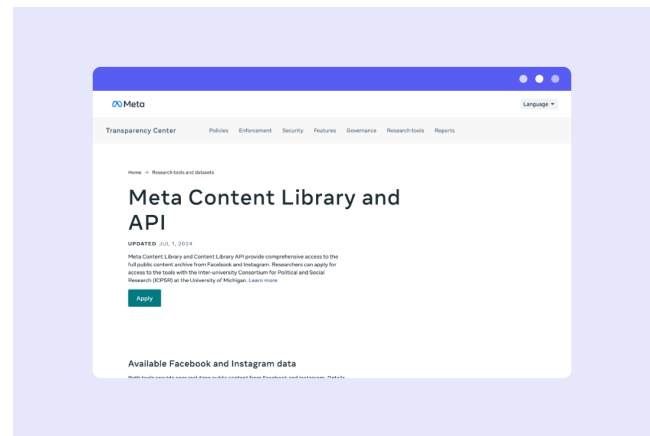
Overview: Meta has established the Meta Content Library and Content Library API – a new formal program offering researchers access to public data on Facebook and Instagram. The MCL offers a dashboard interface for users to query, sort, and filter content. The Content Library API enables researchers to query public data programmatically and analyze it in a clean room. The application process and MCL access are managed by the Inter-university Consortium for Political and Social Research (ICPSR) at the University of Michigan. Researchers can apply from outside of the EU. Meta's existing program, CrowdTangle, is set to be discontinued in August 2024.

What we looked at:

- Researcher access overview site
- Application form and resources
- Tool documentation
- Terms and conditions

Resource links:

- [Research tools: Meta Content Library and API](#)
- [Meta for Developers: Meta Content Library and API](#)
- [SOMAR InfoReady Application Guide](#)
- [Other research tools and datasets](#)
- [Product Terms for Meta Research Tools](#)
- [Frequently asked questions](#)



Outreach:

- We reached out to Meta for comment and discussion, and met with members of the research partnerships and policy teams.

Google Maps

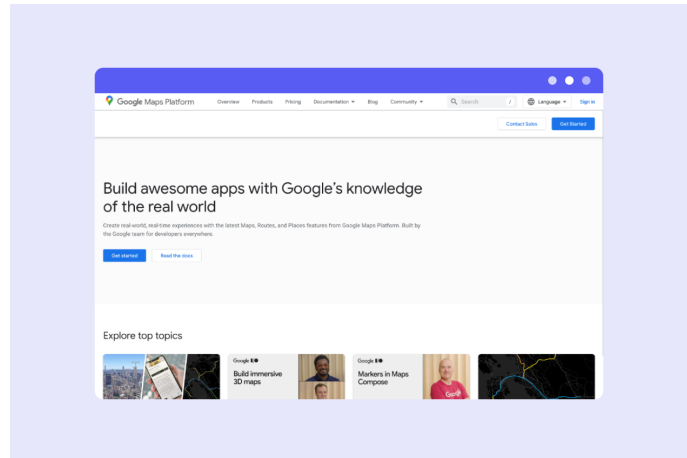
Overview: Alphabet has established a formal program and application process for accessing public data. To access Maps data, Alphabet [offers](#) "access to public data through a cloud-based solution." Only researchers in the EU are eligible for access to the Google Researcher Program.

What we looked at:

- Researcher access overview site
- Application form
- Tool documentation

Resource links:

- [Google Researcher Program](#)
- [Google Researcher Program Application](#)
- [Google Transparency Center: Researcher Engagement](#)
- [Google Researcher Program Acceptable Use Policy](#)
- [Google Maps for developers](#)



Outreach:

- We reached out to contacts at Alphabet for comment and discussion, but ultimately only met with a representative from YouTube's policy team.

Google Play

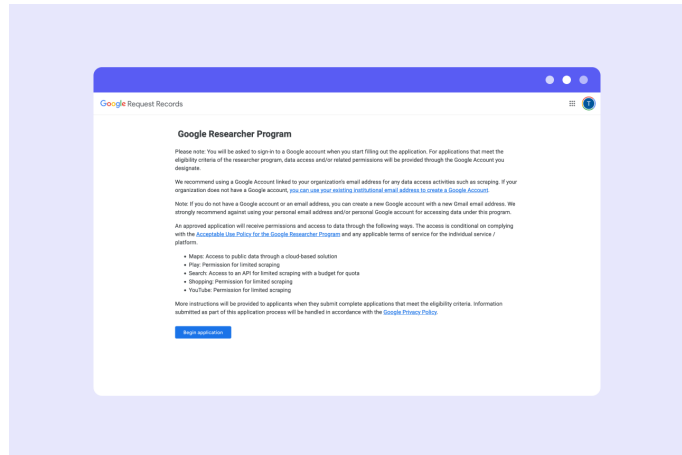
Overview: Alphabet has established a formal program and application process for accessing public data. To access Play data, Alphabet [offers](#) “permission for limited scraping.” Only researchers in the EU are eligible for access to the Google Researcher Program.

What we looked at:

- Researcher access overview site
- Application form

Resource links:

- [Google Researcher Program](#)
- [Google Researcher Program Application](#)
- [Google Transparency Center: Researcher Engagement](#)
- [Google Researcher Program Acceptable Use Policy](#)



Outreach:

- We reached out to contacts at Alphabet for comment and discussion, but ultimately only met with a representative from YouTube's policy team.

Google Search

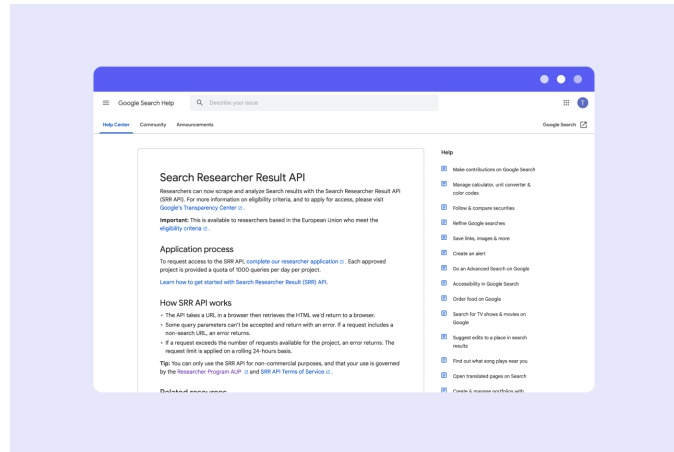
Overview: Alphabet has established a formal program and application process for accessing public data. To access Search data, Alphabet [offers](#) an “API for limited scraping with a budget for quota.” Only researchers in the EU are eligible for access to the Google Researcher Program.

What we looked at:

- Researcher access overview site
- Application form

Resource links:

- [Google Researcher Program](#)
- [Google Researcher Program Application](#)
- [Google Transparency Center: Researcher Engagement](#)
- [Google Researcher Program Acceptable Use Policy](#)



Outreach:

- We reached out to contacts at Alphabet for comment and discussion, but ultimately only met with a representative from YouTube's policy team.

Google Shopping

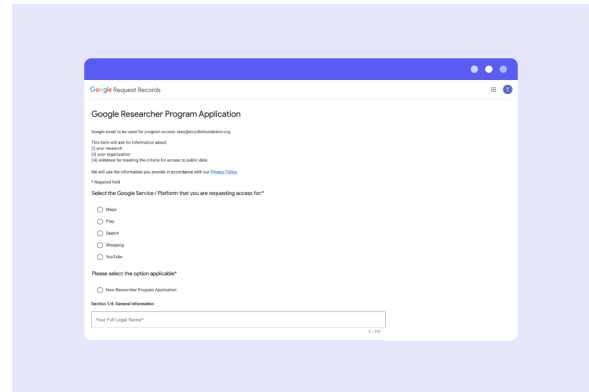
Overview: Alphabet has established a formal program and application process for accessing public data. To access Shopping data, Alphabet [offers](#) a “permission for limited scraping.” Only researchers in the EU are eligible for access to the Google Researcher Program.

What we looked at:

- Researcher access overview site
- Application form

Resource links:

- [Google Researcher Program](#)
- [Google Researcher Program Application](#)
- [Google Transparency Center: Researcher Engagement](#)
- [Google Researcher Program Acceptable Use Policy](#)



Outreach:

- We reached out to contacts at Alphabet for comment and discussion, but ultimately only met with a representative from YouTube’s policy team.

LinkedIn

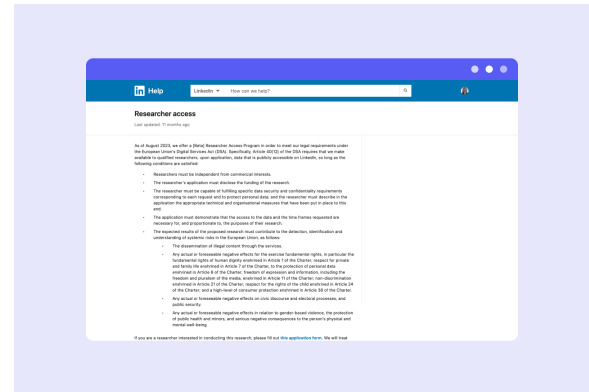
Overview: LinkedIn has established a Researcher Access Program and application process for accessing public data. Global researchers can apply for access to the platform's public data, but the format the data is provided in is unclear.

What we looked at:

- Researcher access overview site
- Application form
- Terms and conditions

Resource Links:

- [LinkedIn: Researcher access](#)
- [LinkedIn Researcher Access Program Application](#)
- [Additional Terms for the LinkedIn Research Tools Program](#)



Outreach:

- We reached out to LinkedIn for comment and discussion and received a response from a member of the platform's legal team. We requested access to LinkedIn's program, but were denied.

Pinterest

Overview: Pinterest has established an application form through which global researchers can request access to the platform's public data. It is not clear what format the data is provided in.

What we looked at:

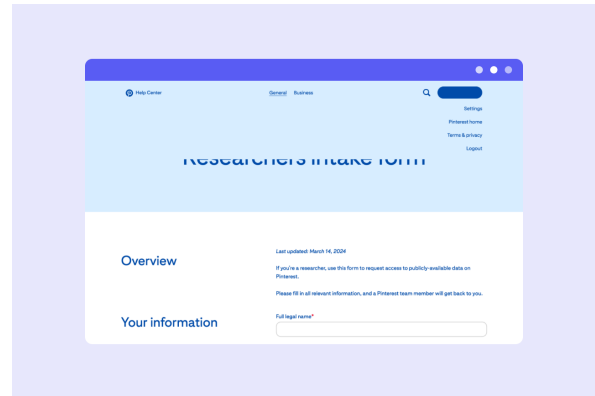
- Digital Services Act compliance page
- Application form

Resource links:

- [Digital Services Act](#)
- [Researchers intake form](#)

Outreach:

- We reached out to Pinterest for comment and discussion. We met with a member of the platform's legal team, who confirmed researchers can request approval to automatically collect publicly accessible data.



Snapchat

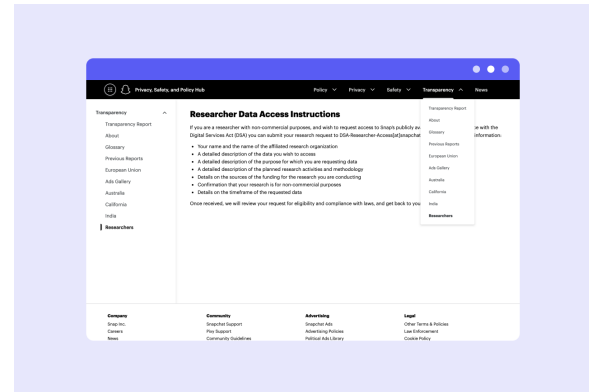
Overview: Snapchat has established guidelines and an email contact through which global researchers can request access to the platform's public data. According to Snapchat, data will be provided to the researcher in a .csv file.

What we looked at:

- Digital Services Act compliance page
- Researcher access overview site

Resource links:

- [Privacy and Security: European Digital Services Act \(DSA\)](#)
- [Researcher and Data Access Instructions](#)



Outreach:

- We reached out to Snapchat for comment and discussion and met with members of the platform's legal and regulatory teams.

TikTok

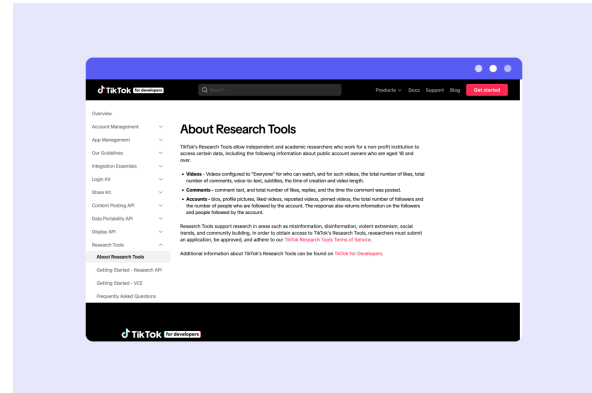
Overview: TikTok offers a Research API that allows access to publicly available data related to accounts and content on the platform. U.S. and E.U. researchers affiliated with non-profit academic and research institutions are eligible to apply.

What we looked at:

- Researcher access overview site
- Application form
- Tool documentation

Resource links:

- [Expanding TikTok's Research API and Commercial Content Library](#)
- [Research API](#)
- [About Research API](#)
- [Research API Getting Started](#)
- [Research API FAQ](#)
- [Research API codebook](#)
- [Research API Terms](#)



Outreach:

- We reached out to TikTok for comment and discussion, but did not receive a response.

Wikipedia

Overview: The Wikimedia Foundation has made its public data available in a variety of ways, since before the enactment of the DSA. Researchers do not need to request or apply for access to the data; researchers can use existing tools to programmatically collect and analyze the data.

What we looked at:

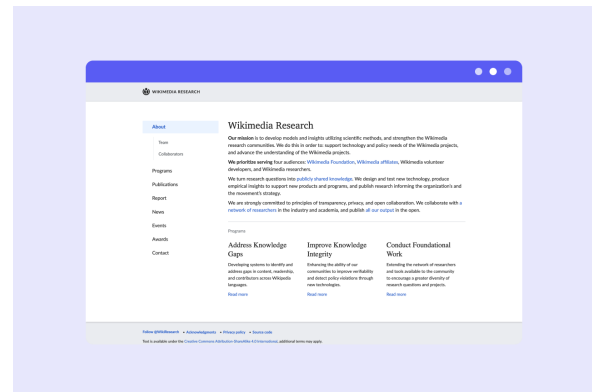
- Tool documentation
- Researcher access overview site
- Terms and conditions

Resource links:

- [Wikipedia is now a Very Large Online Platform \(VLOP\) under new European Union rules: Here's what that means for Wikimedians and readers](#)
- [Wikipedia API Parsed Infobox. Introducing Structured Contents](#)
- [Research:Data](#)
- [Wikimedia Research](#)
- [Wikimedia Foundation Open Access Policy](#)
- [Wikipedia Page History](#)
- [Terms: Content Licensing](#)
- [Wikimedia Downloads](#)
- [Research Data FAQs](#)

Outreach:

- We reached out to Wikipedia for comment and discussion and met with members of their legal and global advocacy teams.

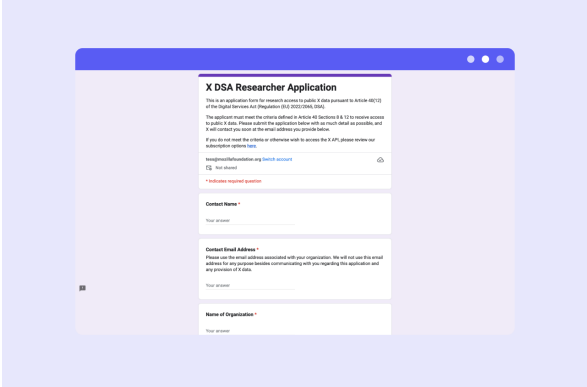


X/Twitter

Overview: X has an application for DSA-eligible researchers through which they can request access to public data via the X API. This pre-existing API is available to other researchers for a fee.

What we looked at:

- Researcher access overview site
- Application form
- Terms and conditions
- Digital Services Act compliance page
- Tool documentation

A screenshot of the 'X DSA Researcher Application' form. The form is titled 'X DSA Researcher Application' and includes a disclaimer about the application being for research access to public data pursuant to Article 40(2) of the Digital Services Act (DSA). It contains several input fields: 'Organization Name', 'Contact Name', 'Contact Email Address', and 'Name of Organization'. There are also checkboxes for 'I am a researcher' and 'I am a developer', and a 'Submit' button. The form is displayed within a browser window with a blue header bar.

Resource links:

- [Country-specific Resources: European Union](#)
- [Developer Terms of Service](#)
- [X Developer Platform: Academic research](#)
- [X DSA Researcher Application](#)
- [Dev community: academics](#)
- [Difference between commercial and noncommercial usage](#)

Outreach:

- We reached out to X for comment and discussion, but did not receive a response.

YouTube

Overview: YouTube has established a Researcher Program to grant academic researchers access to public data on the platform. The program gives access to YouTube's Data API after approval of the researcher's application. A program also allows access for limited scraping for DSA-eligible researchers.

What we looked at:

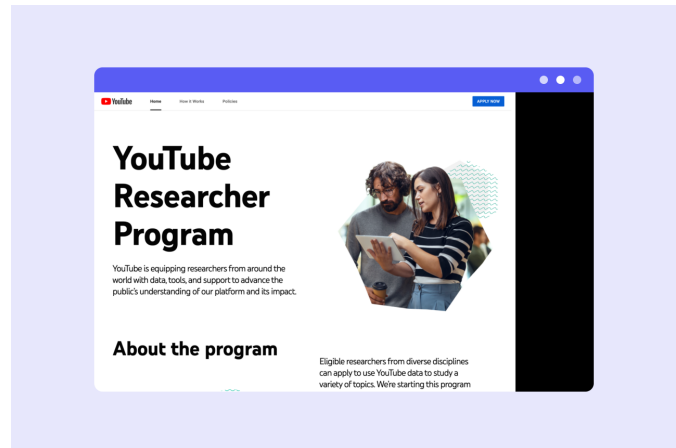
- Researcher access overview site
- Application form
- Terms and conditions
- Tool documentation

Resource links:

- [YouTube Program Policies](#)
- [YouTube Researcher Program Application](#)
- [YouTube Program Terms and Conditions](#)
- [YouTube Data API Overview](#)
- [YouTube API Services - Developer Policies](#)
- [YouTube API Services Terms of Service](#)

Outreach:

- We reached out to contacts at Alphabet for comment and discussion and met with a representative from YouTube's policy team.



Zalando

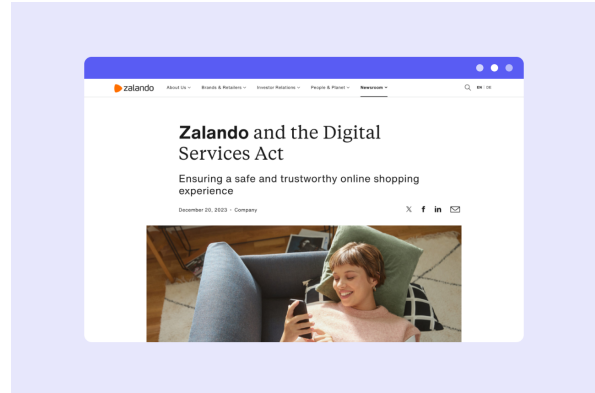
Overview: Zalando states that researchers can contact the platform through the DSA Single Point of Contact to receive data "in accordance with Article 40 DSA."

What we looked at:

- Terms and conditions

Resource links:

- [Zalando and the Digital Services Act](#)
- [Zalando files legal action against the European Commission to contest its designation as a "Very Large Online Platform" as defined by the Digital Services Act](#)
- [Legal Notice](#)



Outreach:

- We reached out to Zalando for comment and discussion, but did not receive a response.

This report was produced by:

Cameron Hickey
Kaitlyn Dowling
Isabella Navia
Claire Pershan
Design by: Shannon Zepeda and Tess Heinrichs

With thanks to:

Becca Ricks
Nicholas Piachaud
Julian Jaursch
Martin Degeling
Brandon Silverman
Svea Windwehr
Anna Lenhart
Henry Tuck

Also a special thanks to the participants of Mozilla's workshop held in April 2023 and everyone who contributed to and endorsed the subsequent [recommendations](#) for public data sharing for public interest research.



Mozilla



ALGORITHMIC
TRANSPARENCY
INSTITUTE



National Conference on Citizenship