Platform Governance Archive (PGA): Dataset PGA v2 [Data Paper]

Christian Katzenbach[1,2], Daria Dergachava[1], Alex Fischer[1], Adrian Kopps[1,2], Sergei Kolesnikov[1], Dennis Redeker[1] , Paloma Viejo Otero[1]

[1] *Centre for Media, Communication and Information Research (ZeMKI), University of Bremen*
[2] *Alexander von Humboldt Institut for Internet and Society, Berlin*

Keywords: Platform Governance, Platform Policies, Terms of Service, Community Guidelines, Privacy Policies, Platform Law, Content Moderation, Digital Constitutionalism, Platform History, Archive

Dataset PGA v2
URL: https://www.platformgovernancearchive.org/data/dataset-pga-v2-ongoing-collection/

Recommended Citation
Katzenbach, C., Dergachava, D., Fischer, A., Kopps, A., Kolesnikov, S., Redeker. D., Viejo Otero, P. (2023). Platform Governance Archive (PGA): Dataset PGA v2. https://doi.org/10.26092/elib/2373.

Cite a Single Policy Document
Name of platform. (Date of version). Name of policy. *Platform Governance Archive*. Direct URL.

Bremen / Berlin, July 2023

Lab Platform Governance, Media and Technology
Centre for Media, Communication and Information Research (ZeMKI)
University of Bremen

Platform Governance Archive (PGA): Dataset PGA v2 [Data Paper]

This dataset PGA v2 is part of the Platform Governance Archive (PGA). It is the result of a continuous data collection of policies of up to 18 (social) media platforms since April 2022, performed with the Open Terms Archive engine, and hosted and curated by the Lab Platform Governance, Media And Technology (PGMT) at ZeMKI, University of Bremen. The PGA v2 dataset contains a selected set of policies for each platform that usually includes community guidelines, privacy policies and terms of service agreements. A number of platforms have more complex structures for laying out their rules, so PGA v2, for instance, tracks seven specific policies for TikTok and nine for Twitter.

Platforms: Facebook, Instagram, LINE, LinkedIn, Parler, Pintest, Quora, Reddit, Snapchat, Spotify, Telegram, TikTok, Tumblr, Twitch, Twitter, WeChat, WhatsApp, YouTube

Time Frame: since April 2022 (most platforms)

Project Website: https://platformgovernancearchive.org

Background

This dataset PGA v2 is part of the Platform Governance Archive (PGA), a data repository and platform that collects and curates policies of major social media platforms in a long-term perspective. In addition to PGA v2, there is also the historical dataset PGA v1 that provides a long-term collection of platform policies of four major platforms from 2005 to 2021. The PGA v2 picks up this work and writes it into the future, while also broadening the scope of data collection. Instead of looking backward, the PGA v2 entails data collected in-real time as changes are made to platform policies. This way, the database remains always up to date allowing access to the data through the Platform Governance Archive even for the most current versions of platform policies.

Data Collection

The data collection is performed with the Open Terms Archive engine. Open Terms Archive (OTA) is a French NGO focussed on making company terms and their changes available to consumers and the public, and offering an almost real-time alert service for such changes. The dataset PGA v2 utilizes the open source software, the Open Terms Archive Engine to download and archive changes of 18 platforms. Two aspects for data collection are important and linked to the OTA. The OTA Engine automatically updates the archive of the English language version of selected platform policies on a regular basis, capturing all meaningful edits in terms and policies. This is done by using an automated web scraper that tracks the changes of each web URL. The relevant parts of platform policies are selected here with the use of CSS selectors and Javascript filters, to select the correct content, remove insignificant content (e.g. ads, illustrative pictures, internal navigation links…), and filter out noise (e.g. tracker identifiers in links, relative dates…). The engine scrapes through the selected policies multiple times a day, and keeps all HTML snapshots here. If changes are detected within the HTML snapshots, new versions of the policy are populated to PGA v2 dataset. Please consult the documentation for more information on the Open Terms Archive Engine.

Even when taking an automated approach, we could not archive all platforms this way. Going beyond the four platforms of PGA v1, we selected a total of 18 platforms to monitor for PGA v2. The logic for archiving decisions has largely been based on scale i.e. selecting the platforms with the largest possible usership. Since we were interested in social media platforms and not in commercial retail platforms such as Amazon or news websites that allow for comments under published articles. A number of these platforms can be best described as “chat services” although they might have a number of additional features. The sampling decision was made in early 2022 and usership numbers vary over time. For instance, the use of Parler has since strongly declined. Another sampling criterion

is that the location of the platforms must be in the jurisdiction of the European Union (at least through a form of representation or agent).

Using the Data

We are more than happy if you want to use our dataset in your research, reporting, and explorations. If you do:

1. Consult the respective data documentation;
2. reference this project and the actual dataset;
3. send us a note so that we include you in our research and output page.

PGA v2 is made available under the Open Data Commons Attribution License (that means what we say above: use it, but reference us).

Dataset PGA v2
URL: https://www.platformgovernancearchive.org/data/dataset-pga-v2-ongoing-collection/
Licence**:** ODC-BY v1.0 https://opendatacommons.org/licenses/by/1-0/

Recommended Citation
Katzenbach, C., Dergachava, D., Fischer, A., Kopps, A., Kolesnikov, S., Redeker. D., Viejo Otero, P. (2023). Platform Governance Archive (PGA): Dataset PGA v2. https://doi.org/10.26092/elib/2373.

Recommended Citation for a Single Policy Document
Name of platform. (Date of version). Name of policy. *Platform Governance Archive*. Direct URL.