GDPR Subject Access Request for the application Tumblr

This report is about the GDPR Subject Access Requests (SAR) I made for the application Tumblr and the website tumblr.com.

# Tumblr

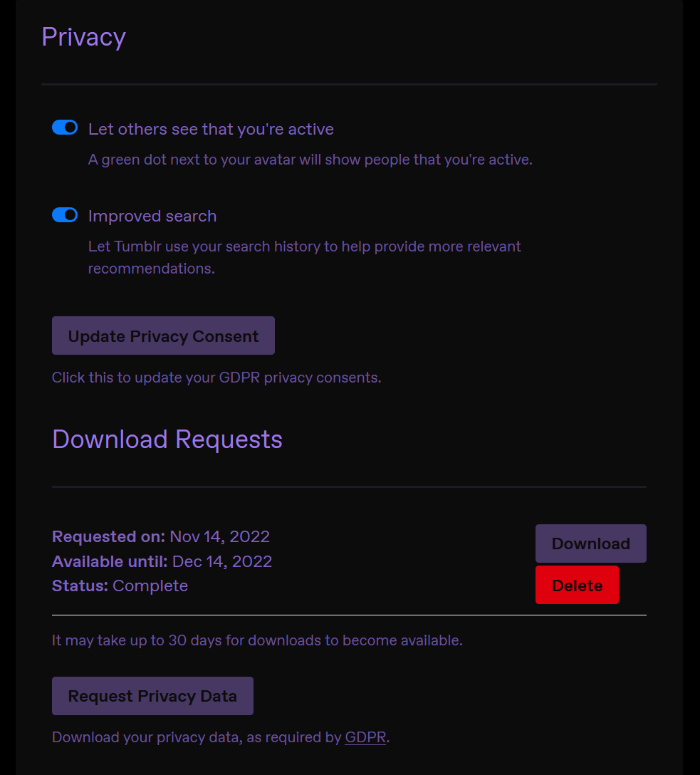


Figure 1: Tumblr's privacy settings page

Tumblr is a blogging platform that allows users to upload text, images, GIFs, and other media to their profile. Users can interact with and follow other users, like and

reblog (repost to your blog) their posts, and privately message one another.

I have had this Tumblr account for slightly over a year, but I began using it more often recently, within the past month or so. I used it irregularly before, but I recently began posting and liking other posts more frequently, interacting with other users and spending more time on the app. However, most of this happened after I made the Subject Access Request, so a lot of it will not appear in the data I will dissect in this report.

## Making the Request

To make the subject access request, I had to navigate to my privacy settings and click on the "Request Privacy Data" button. A message appeared indicating that the request was made, and an email was sent to my address. I was informed that "It may take up to 30 days for downloads to become available".

Since I didn't have much data available on Tumblr due to my lack of use, I received the data on the same day I requested it. The data is available on the website for one month after the date of the request, after which it will be deleted. The user also has the option to delete the request themselves. This request cannot be made from the mobile app.

All this information was made available on their website. [1]

## Surface observations about the data

The data was sent to me in a zip file containing two JSON (JavaScript Object Notation) files. These files were difficult to look through, so I opened them in the Firefox browser, which has a built-in JSON formatter.

One of the files was named "schema.json", and contained the scheme for the other file. The second file was named "payload-0.json" and contained the actual data Tumblr has on my account. schema.json was over 6 times smaller than payload-0.json, as it consisted only of explanations for the fields in its sister file.

payload-0.json has only 3 fields: one labelled "transactionId", containing a string that I assume is simply the ID number for my SAR transaction, one labelled "schema\_identifier", containing a string I again have to assume is the ID for the actual SAR, and a third labelled "data".

## Data

The field labelled "data" contained all the information that Tumblr gave me as part of my GDPR request. There are several fields here with several more fields of data, so I will only highlight some interesting ones I found.

The first few fields show my entered age, my account email, my account language, the time I registered the account, and the most recent time I made a post. The entered age and email are in plaintext.

Things become a little more interesting the further down I go. The "dashboard" field contains, according to the schema, "Everything you have been served on your dashboard in the order of when you saw it". Every post that had been sent to my dashboard, seen or not, was under this field. This does not contain every single post that has been sent; rather, it seems to be a list of all posts sent to my dashboard since July 22nd, 2022. I am not sure what the significance of this is, but it's possible that Tumblr simply discards information about posts sent to my dashboard every 115 days, or maybe every 152 posts due to the amount of actual data in this field. This information isn't ordered as seen in Figure 2, but it can be ordered through a Python script (see order\_dashboard.py in Other). The "dashboard" field also contained a field describing the served post. Each post was one of three things: a post, an advertisement, or "other". Counting each of these, there ended up being 40 advertisements, 39 posts, and 73 "other" elements (152 in total). The schema does not explain what these "other" elements are, nor can I find anything on the website. Each field also contains a field labelled "content\_url". Interestingly, only fields with a "post" element contained links to the element. The advertisements and "other" elements contained only the string "\\N". Finally, the "platform" field has only two values, always either "android" or "redpop". Using the sorted view, all values for the platform field contained "android" and inexplicably changed to "redpop" on November 1st, 2022. The string "redpop" does not appear in either schema.json or payload-0.json outside of the "dashboard" field, nor does it appear anywhere on the Tumblr website (the only reference to "redpop" I could find on the internet was about a cream soda product). The schema defines the platform field as "the device or platform you were using when served this content element", which implies it refers to a desktop session, which doesn't make sense given that I use both the desktop website and the mobile app interchangeably.



Figure 3: Four fields within the file "new\_data.json" (see order\_dashboard.py). The "platform" field changes from "android" to "redpop" after November 1st.

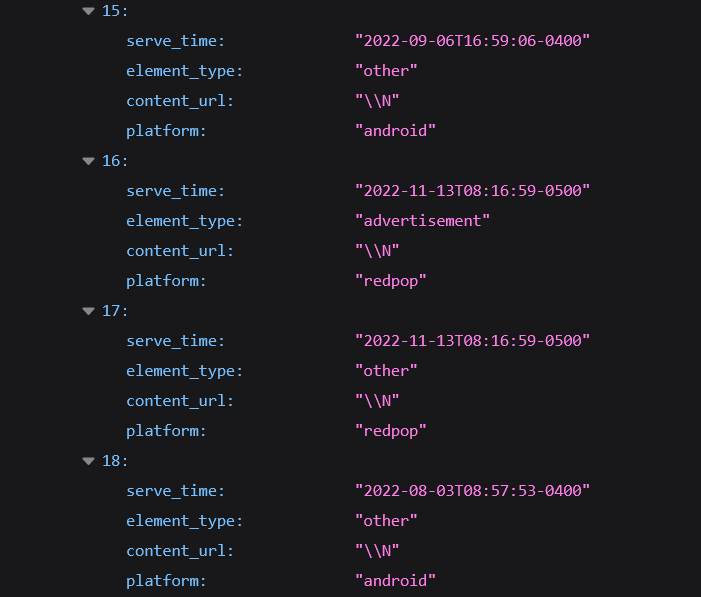


Figure 2: Four fields within the "data" field in payload-0.json

The "dashboard" generally shows each post served to my dashboard, with the time it was sent and the type of post it was. If this list went on long enough, Tumblr would have access to a trend of recommended data for my account across a more expansive timeframe. This data could show how my habits, tastes, actions, and reactions change according to what I they recommend to me, and they could recommend tailored posts to match.

The data also included a field labelled "ads\_analytics". The schema says this field contains "user events on posts created by Tumblr premium partners", which seems to imply that these are posts shared by Blaze. Tumblr Blaze is a relatively new service Tumblr offers that lets users pay to promote their posts. [2] Given that Tumblr also blocks these with their Ad-Free Browsing policy [3], it seems that these are Blaze posts, however there is evidence that disproves this. Also, this field is different to a later field, "client\_side\_ads\_analytics", which I will discuss later.

This field contains a "serve\_time" field, keeping track of when it was served to my dashboard. Sorting this, I can see that the earliest post in the package I received was on October 25th, 2021, the date I created my Tumblr account. Tumblr Blaze was started on April 20th, 2022, so these can't be Blaze posts. Still, they are posted by "Tumblr premium partners" and are functionally indistinguishable from regular ads.

The "post\_url" field contains the URL to the post served. This usually is a valid link, except for two occasions: some fields have the string "\\N" in place of an actual URL, and links starting with the string "house-ad.tumblr.com" always have an error occur when visiting. Figure 4 (below) is one such error.

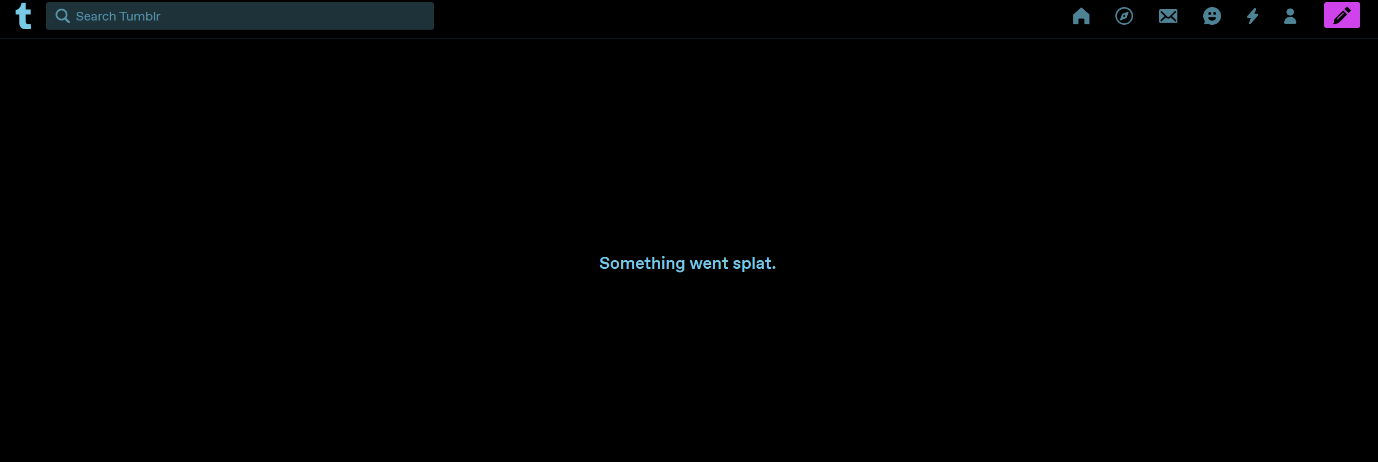


Figure 4: An error page after vising the URL <http://house-ad.tumblr.com/post/700537564490350592>

Of the 79 of these links, only 26 were unique. Most other URLs link to valid Tumblr pages.

Finally, the last two fields, "viewed" and "interacted" refer to if the user sees and interacts with the ad, respectively. Of all 178 ads served to my dashboard, 0 were only viewed, 8 were only interacted with, and 13 were both viewed and interacted with. This leaves 157 ads ignored entirely. The distinction between "viewed" and "interacted" is odd. Neither field is described further than "if you [action] with the post" in the schema.

In any case, this shows that Tumblr keeps track of which posts the user takes note of and which they don't, which can be used to further refine ad recommendations. These ads are dated from a little over a year ago, so Tumblr would have that long to figure out what ads to recommend to me.

A few smaller fields are the "user\_interest\_profiles" and "geo\_profiles" fields.

The "user\_interest\_profiles" field consists of "Your interests based on historial data such as post engagements, posts you made, blogs you follow, and more", according to schema.json. It seems that this field, which only contains 29 items, is a list of tags that Tumblr uses to recommend posts to me. A tag is a phrase users can append to their posts to boost recognition for those following those specific tags. Some of these tags are ones I appended to my posts or are on posts I've liked or reblogged. Given that these are tags that I have either intentionally used or have purposefully interacted with in the past, it is unlikely that Tumblr could use these to influence me. However, they could use it to track me and my usage of Tumblr.

The "geo\_profiles" field consists of the general login locations Tumblr has detected for me. These were stored for "product and performance analysis" according to the schema. The only fields within it log my city (county), state (province), and country. What effect using a VPN while on Tumblr would have, I don't know. It is obvious how this might be used to track me, but with how unspecific the location is, tracking me would still be difficult.

The last field I want to discuss is the "client\_side\_ads\_analytics" field. There are 358 fields here, slightly more than double what was in "ads\_analytics". I mentioned previously that this field is distinct from the "ads\_analytics" field and upon looking at the data stored within, it seems to be.

Both fields had a "placement\_id" field that I neglected to mention as it doesn't reveal any legible information. However, what the client-side ads have that the others don't is an "ad\_type" field. This field contains six distinct types: "VERIZON\_NATIVE", "SMART\_BANNER", "DISPLAY\_IO\_INTERSCROLLER", "DISPLAY\_IO", "DISPLAY\_IO\_HEADLINE\_VIDEO", and "GOOGLE\_NATIVE".

"VERIZON\_NATIVE" presumably refers to Yahoo's Native Advertising service [4], where as "GOOGLE\_NATIVE" refers to Google's Native Advertising service [5]. "DISPLAY\_IO", "DISPLAY\_IO\_INTERSCROLLER", AND "DISPLAY\_IO\_HEADLINE\_VIDEO" are likely all advertising services produced by display.io, and are three distinct types of ads [6][7][8]. Finally, "SMART\_BANNER" seems to also be a Google ad, being a dynamic banner ad meant for mobile devices [9].

This field also has a "serve\_time", which also spans from the date of my account registration to the date I made the SAR, and the "viewed" and "interacted" fields. Only 1 ad was ever viewed, and it was also interacted with. None were only viewed or only interacted with.

The fact that there are six different advertising services that Tumblr employs is slightly alarming to me. Each ad has a different way of being displayed, from a simple banner to an obtrusive full-screen display. Some ads, such as the interscroller, can be very obnoxious and in-your-face, and others like the headline video don't fully go away until the video has finished. These ads are likely no different from those covered in "ads\_analytics", and any point I would make about them has already been stated in that section.

# Conclusion

There are several fields I neglected to mention in this report, both because of the sheer density of data in the files I received, and because some are just not as interesting as others. For example, there is a field labelled "mobile\_device\_oses" that only contains the OS of my phone. Tumblr somehow failed to get its actual OS name (Android), though it did manage to figure out the fact that it runs version 11.

In general, there is much evidence to suggest that Tumblr has certain ways of tracking me and trying to keep me on the app, but not much in the way of actually influencing me to do anything – there is nothing particularly inflammatory or overtly political to be seen on the app unless I choose to see it. The information I gleaned from this Subject Access Request shows a lot about what data Tumblr is keeping on my account, and it shows what they either aren't keeping or what they chose not to send. Tumblr is based in New York, but they still must comply with GDPR, as the app is available in EU countries. The GDPR is important for user data protection and access, and the information that Tumblr is willing to share to its users is certainly expansive, if not a little dense. But their compliance with GDPR still shows the depths of their large network, and all the little bits behind how they function.

# References

1. <https://help.tumblr.com/hc/en-us/articles/360003604974-Access-and-manage-data-associated-with-your-account>
2. <https://help.tumblr.com/hc/en-us/articles/5298854249367-Tumblr-Blaze-FAQ>
3. <https://help.tumblr.com/hc/en-us/articles/4418605293975-Ad-Free-Browsing>
4. <https://gemini.yahoo.com/advertiser/home>
5. <https://support.google.com/admanager/answer/6366845?hl=en>
6. <https://www.display.io/>
7. <https://www.display.io/products/product-interscroller/>
8. <https://www.display.io/products/product-headline-video/>
9. <https://developers.google.com/admob/android/banner/smart>

# Other

order\_dashboard.py

|  |
| --- |
| import json  t\_full\_data = json.load(open("payload-0.json"))  t\_data = t\_full\_data[0]['data']  t\_dash:list = t\_data['dashboard']  t\_dash = sorted(t\_dash, key=lambda d: d["serve\_time"]) # sort data based on serve time  with open ('new\_data.json', 'w') as f:  f.write(json.dumps(t\_dash, indent=4)) # write to a json file |