Digital Data Collection through Data Donation

HUN-REN Centre for Social Sciences

https://recens.tk.hu/en/digital-data-collection-through-data-donation

Please do not cite this material without permission!

Contact: kmetty.zoltan@tk.hu

Data Collection Process

Data Collection and Validation

The data collection in the project was started in February 2023, and ended on 18th June 2023.

The data donation process is outlined below.

From the respondents' view, this was a three-phase procedure, that consisted of (1) a short questionnaire with eligibility questions and a consent form, (2) downloading and uploading their data to the project's website, and (3) filling a detailed survey questionnaire.

For each respondent, the process started with an invitation email from the data collection company (NRC). This email contained information about the project's aims and the incentives, along with a direct link to the dedicated webpage of the project.

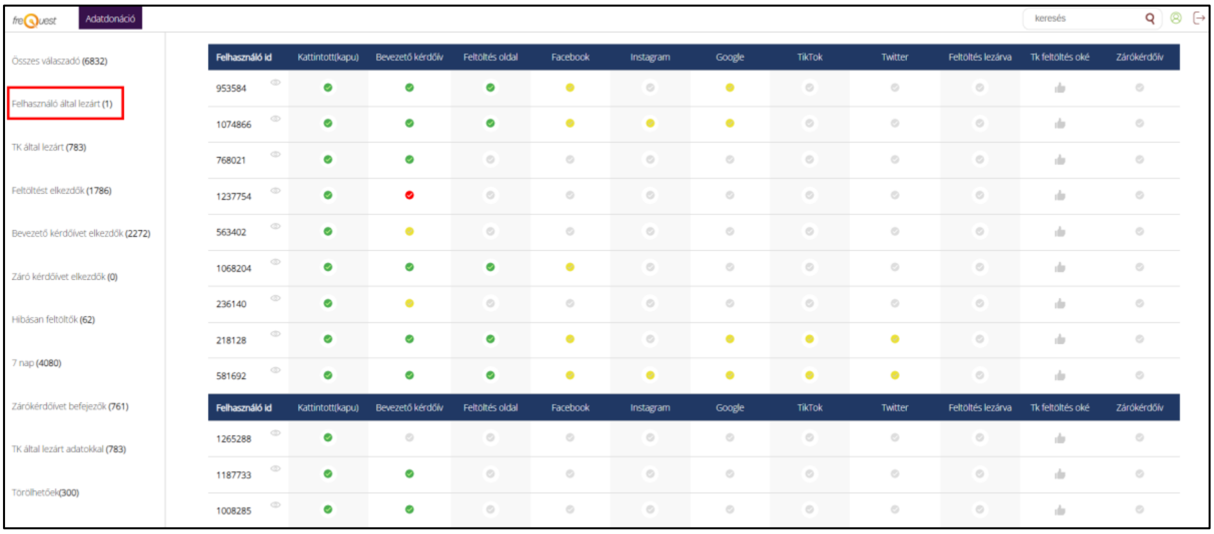
Firstly, (1) respondents filled out a short questionnaire, to filter out respondents who do not have Google and Facebook profiles, or don't use them regularly.

In this step the respondents had to read a detailed research description and accept the consent form. Without expressing consent, the respondents could not move forward.

If the given respondent was eligible to continue, (2) they were presented a webpage that contained detailed guides for each platform (Google, Facebook, Instagram, Twitter and TikTok). These guides contained step by step instructions on how to export and download information from these platforms, along with a video tutorial with the same content.

Once the respondent downloaded their data onto their computer, they had to upload the file(s) for each platform without opening or extracting them to the project's website. Uploading data for Google and Facebook is mandatory, however, those who uploaded their Instagram, TikTok and Twitter profiles, would receive more incentive. After the respondent uploaded the files for Google and Facebook, they were able to indicate that they have finished uploading (meaning that they do not want to upload data from additional platforms), by pressing a button on the website.

At this point, after pressing the button, the given respondent would appear in the "Closed by respondent" ("Felhasználó által lezárt") category on the administrator interface of the project.



Összes válaszadó (6832)
Felhasználó által lezárt (1)
TK által lezárt (783)
Feltöltést elkezdők (1786)
Bevezető kérdőívet elkezdők (2272)
Záró kérdőívet elkezdők (0)
Hibásan feltöltők (62)
7 nap (4080)
Zárókérdőívet befejezők (761)
TK által lezárt adatokkal (783)
Törölhetőek(300)

Felhasználó id	Kattintott(kapu)	Bevezető kérdőív	Feltöltés oldal	Facebook	Instagram	Google	TikTok	Twitter	Feltöltés lezárva	Tk feltöltés oké	Zárókérdőív
953584	🟢	🟢	🟢	🟡	⚪	🟡	⚪	⚪	⚪	👍	⚪
1074866	🟢	🟢	🟢	🟡	🟡	🟡	⚪	⚪	⚪	👍	⚪
768021	🟢	🟢	⚪	⚪	⚪	⚪	⚪	⚪	⚪	👍	⚪
1237754	🟢	🔴	⚪	⚪	⚪	⚪	⚪	⚪	⚪	👍	⚪
563402	🟢	🟡	⚪	⚪	⚪	⚪	⚪	⚪	⚪	👍	⚪
1068204	🟢	🟢	🟢	🟡	⚪	⚪	⚪	⚪	⚪	👍	⚪
236140	🟢	🟡	⚪	⚪	⚪	⚪	⚪	⚪	⚪	👍	⚪
218128	🟢	🟢	🟢	🟡	⚪	🟡	🟡	🟡	⚪	👍	⚪
581692	🟢	🟢	🟢	🟡	🟡	🟡	🟡	🟡	⚪	👍	⚪
Felhasználó id	Kattintott(kapu)	Bevezető kérdőív	Feltöltés oldal	Facebook	Instagram	Google	TikTok	Twitter	Feltöltés lezárva	Tk feltöltés oké	Zárókérdőív
1265288	🟢	⚪	⚪	⚪	⚪	⚪	⚪	⚪	⚪	👍	⚪
1187733	🟢	🟢	⚪	⚪	⚪	⚪	⚪	⚪	⚪	👍	⚪
1008285	🟢	🟢	⚪	⚪	⚪	⚪	⚪	⚪	⚪	👍	⚪

At this point, the data uploaded by the respondent was on the data collection company's server. Approximately every two days, the data of all respondents in this category was validated using a Python script in the following way.

- Uploaded files were automatically renamed after uploading to the following format: <respondent_id>_<platform_name>_<timestamp>_<file_number>
- Our script, running on our server, connects to the server of the data collection company, and downloads all files that were uploaded by the given respondent.
- After that, all files are extracted, and their content is compared against a pre-compiled list of files and folders that are mandatory for the given platform. Additionally, the following information is extracted to make sure that, e.g., the profile is not empty, not newly registered, or that the profiles for the different platforms belong to the same person:
- All the extracted information, along with the list of files and folders and whether they exist in the respondent's uploaded archive, was saved in a text file on the server, in the individual folder of the given respondent, that, at this point, contained their data (zipped, and extracted), and the report file.

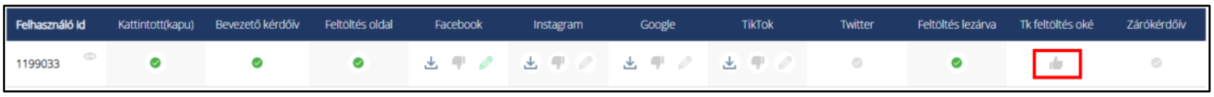
Using the report file, we determined for each respondent whether (1) each archive was exported in the proper (JSON) format, (2) the date range setting was correct, and (3) the profiles belong to the same person. If something seemed questionable, we contacted the data collection company, to provide more information about the given person, according to what was saved in the web panel.

If, for some reason, **the respondent's data was not acceptable**, they would receive detailed feedback, which contained all issues that we found with their uploaded archive. This feedback was given separately for each platform; therefore, it was possible to only mark one platform as invalid. This triggered another automated email, which informed the respondent that we found an issue with their data, and directed them to the data collection website, where they could read our feedback, and re-upload their data.

After re-uploading, they were once again able to finalize their data, after which they would appear on the administrator website, again, in the “Closed by respondent” category. This cycle continued until the respondent managed to upload acceptable data for each platform, or, in some cases, until they gave up participating.

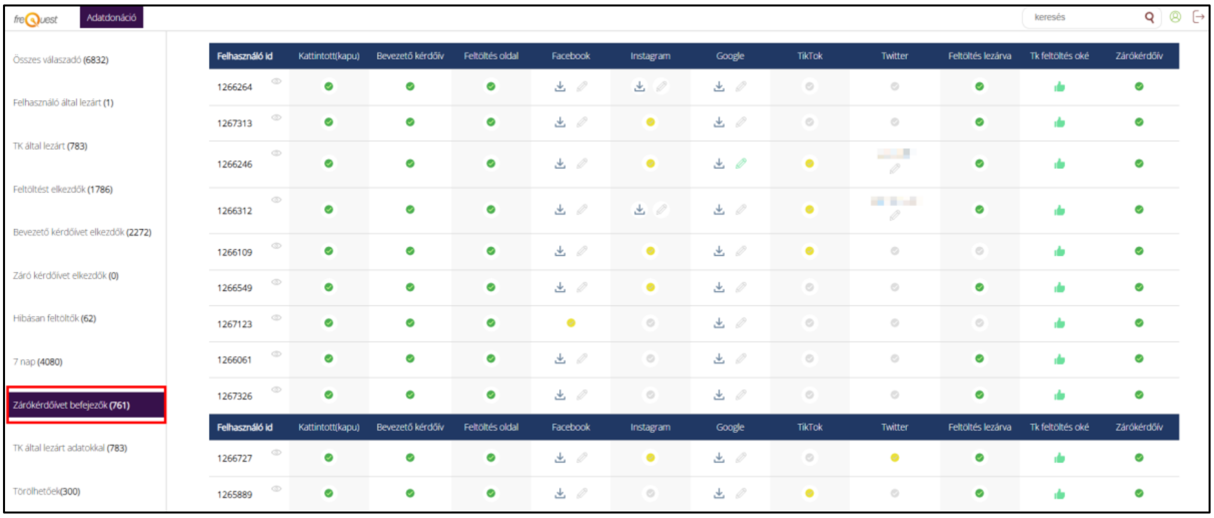
Additionally, if the respondent decided not to re-upload their data for any of the optional platforms (Instagram, TikTok and Twitter), they could just click on the button to close the uploading process, letting the researchers know not to process their data from those platforms that we marked as invalid.

At this point, **if the uploaded archives were acceptable**, we marked the respondent as “uploaded data is OK” (“TK feltöltés oké”).



Felhasználó id	Kattintott(kapu)	Bevezető kérdőív	Feltöltés oldal	Facebook	Instagram	Google	TikTok	Twitter	Feltöltés lezárva	Tk feltöltés oké	Zárókérdőív
1199033	✓	✓	✓	⤓ 👎 ✎	⤓ 👎 ✎	⤓ 👎 ✎	⤓ 👎 ✎	○	✓	👍	○

After this, finally, the (3) respondent received an automated email, that invited them to finish the data donation process by filling out the final questionnaire. After filling out the survey, they would appear in the “Finished the final questionnaire” (“Zárókérdőívet befejezők”) category on the administrator website, as shown on the screenshot below.


freQuest
Adatdonáció
keresés
Összes válaszadó (6832)
Felhasználó által lezárt (1)
TK által lezárt (783)
Feltöltést elkezdők (1786)
Bevezető kérdőívet elkezdők (2272)
Záró kérdőívet elkezdők (0)
Hibásan feltöltők (62)
7 nap (4080)
Zárókérdőívet befejezők (761)
TK által lezárt adatokkal (783)
Törölhetőek(300)

Felhasználó id	Kattintott(kapu)	Bevezető kérdőív	Feltöltés oldal	Facebook	Instagram	Google	TikTok	Twitter	Feltöltés lezárva	Tk feltöltés oké	Zárókérdőív
1266264	✓	✓	✓	⤓ ✎	⤓ ✎	⤓ ✎	○	○	✓	👍	✓
1267313	✓	✓	✓	⤓ ✎	●	⤓ ✎	○	○	✓	👍	✓
1266246	✓	✓	✓	⤓ ✎	●	⤓ ✎	●	[redacted] ✎	✓	👍	✓
1266312	✓	✓	✓	⤓ ✎	⤓ ✎	⤓ ✎	●	[redacted] ✎	✓	👍	✓
1266109	✓	✓	✓	⤓ ✎	●	⤓ ✎	●	○	○	👍	✓
1266549	✓	✓	✓	⤓ ✎	●	⤓ ✎	○	○	✓	👍	✓
1267123	✓	✓	✓	●	○	⤓ ✎	○	○	○	👍	✓
1266061	✓	✓	✓	⤓ ✎	○	⤓ ✎	○	○	✓	👍	✓
1267326	✓	✓	✓	⤓ ✎	○	⤓ ✎	○	○	✓	👍	✓
Felhasználó id	Kattintott(kapu)	Bevezető kérdőív	Feltöltés oldal	Facebook	Instagram	Google	TikTok	Twitter	Feltöltés lezárva	Tk feltöltés oké	Zárókérdőív
1266727	✓	✓	✓	⤓ ✎	●	⤓ ✎	○	●	✓	👍	✓
1265889	✓	✓	✓	⤓ ✎	○	⤓ ✎	●	○	✓	👍	✓

Response rate

We can divide the data collection into two periods. In the first period (until 2023.05.04), the field company invited active panel members who regularly participated in their survey. After that, they started to invite those more passive panel members who do not regularly fill out surveys. The final response rate was 3.6 percent for the active and 0.4 percent for the passive panels. However, the main difference was between the ratio of those who clicked on the platform after getting the invitation. If someone clicked on the platform, the probability of going one step further was the same for the active and passive panel members.

We have some additional information about those panel members who have filled out the preliminary survey, which was fielded in 2022 March and April. They are all active panel members and were the first to be invited to the study. In the preliminary survey, they had to answer hypothetical questions about data donations. In their case, the final response rate was 17 percent, which is exceptionally high compared to other groups. We can also divide those screened out and those who did not accept the consent form. Based on the preliminary sample, 40 percent of those who filtered out this level did so because of the screening questions, and 60 percent was who did not consent.

<table>
<tr><th colspan="2"></th><th>Filled out preliminary survey</th><th>Active panel members (20230504)</th><th>Passive panel members</th><th>Full panel</th></tr>
<tr><td colspan="2">Deleted</td><td>40</td><td></td><td></td><td></td></tr>
<tr><td colspan="2">All Invitations</td><td>960</td><td>13958</td><td>66042</td><td>80000</td></tr>
<tr><td colspan="2">Clicked on the platform</td><td>799</td><td>4931</td><td>1901</td><td>6832</td></tr>
<tr><td>Viewed the consent form</td><td rowspan="2">Started the filter questionnaire</td><td>733</td><td rowspan="2">3509</td><td rowspan="2">1332</td><td rowspan="2">4841</td></tr>
<tr><td>Screen out</td><td>601</td></tr>
<tr><td colspan="2">Started the uploading process</td><td>518</td><td>1878</td><td>691</td><td>2569</td></tr>
<tr><td colspan="2">Finished the uploading process</td><td>169</td><td>508</td><td>275</td><td>783</td></tr>
<tr><td colspan="2">Finished the final survey</td><td>169</td><td>505</td><td>256</td><td>761</td></tr>
<tr><td colspan="2">Final sample</td><td></td><td></td><td></td><td>758</td></tr>
</table>

Data Fields and Information Content

Respondents were asked to export only certain parts of their social media data, where this was possible. In the case of Facebook and Google, one can choose what data fields / products to export, but in case of Instagram and TikTok, all profile information is exported. In case of Twitter, we asked respondents to enter their Twitter handle.

In the case of Facebook, the following 20 data categories were collected.

- Your activities on Facebook
    - Posts
    - Pages and profiles
    - Events
    - Comments and reactions

    - Stories
    - Reels
    - Groups
    - Reviews
    - Other activity
- Personal data
    - Profile information
    - Other personal information
- Contacts
    - Friends and followers
- Logged data
    - Your topics
    - Location
    - Music recommendations
    - Search
    - Your interactions on Facebook
- Settings
    - Feed
    - Preferences
- Advertising information
    - Advertising information

In the case of Google, the following data categories were collected.

- Location History
- My Activity: Search
- YouTube and YouTube Music: all, except uploaded videos

Storing and Anonymizing Data

All the data validation procedures, error checking and the general handling of the respondents' data took place on a dedicated server computer in TK. During the data collection phase, only dedicated researchers had access to the data on the server.

After the data collection, a thorough anonymization take place, along with restructuring, which include the following:

- Removing all uploaded data that is not subject of our analysis, and which we did not ask consent to process,
- Masking the name of the respondents' and their social media friends using hash functions,
- Parsing, sanitizing, converting data, and uploading them to a highly integrated SQL database.

Using this final database, it will be possible to create (1) smaller sub-samples of respondents, and (2) to export data files with restricted information content, containing the minimal amount of data that is necessary for the project's researchers to conduct their analyses.

Final Sample Data Distribution

The final sample consists of 758 respondents. These are the respondents who not only uploaded all their data correctly, but they also managed to fill out the final questionnaire on time.

The platform distribution of these respondents is displayed in the table below.

Facebook	758
Google	758
Google Search	707
Google Location History	395
YouTube Search History	692
YouTube Watch History	703
Instagram	247
TikTok	78