FAQ social media archiving tools

Document version

- 0.1 - After online session 9-4-2020
- 0.2 - After online session 16-4-2020
- 0.3 - After online session 30-4-2020

Before reading this FAQ (and please do)

When asking questions about digital archiving and digital preservation, answers infamously begin with 'it depends'. Asking about social media archiving tools is no exception. Because the first question, before talking about archival tools, really is: what is the audience who's going to use the archived material and what do they expect from the archived material? Talking about social media archiving tools they roughly speaking come in two forms: tools that harvest the look and feel of the social platform, and tools that capture the structured data. The latter will satisfy the more statistical-minded researcher and the former the researcher or user who has a more qualitative approach. And one approach of course is not better than the other, they are just different.

In the Preliminary Social Media Archival Tool Report - April 1 2020 (a first product of the NDE sponsored report about social media archiving tools, written by Zefi Kavvadia from the IISH) this distinction is clearly made.

This FAQ is made on the basis of different online NDE Q&A sessions about social media archiving and will grow alongside with these sessions.

The questions

Is there one tool who can harvest all the different social media platforms?

No. From the viewpoint of the 'it depends' observation above this already is not possible. First you will have to choose if you want a more structured data or look and feel approach. Once you have decided that concerning the latter the Webrecorder and Brozzler tools come closest to being a universal tool. The first one is very easy in use but time consuming doing the actual harvesting. The latter is harder to deploy and use (being a command line tool) but can yield much more harvested material. In the Preliminary Social Media Archival Tool Report - April 1 2020 you can read more in detail about these tools. Looking at the structured data approach there is not one tool who can do it all. For each platform you will have to find the right tool, even though Social Feed Manager is an example of a tool which combines capturing capabilities for Twitter, Tumbler, Flickr and Sina Weibo; the choice of whether to choose a dedicated tool or a more inclusive one will depend on a combination of available time, resources, and expertise.

What can Webrecorder do for you?

As said above, Webrecorder is easy in use, but time-consuming when doing the actual harvesting since in many cases it must be done manually. Also it is a tool that harvests the look and feel, so if you want to create big harvests and/or you are more interested in statistical analysis material this is not the right tool. But for relatively small harvests Webrecorder can work very well. It is advisable to use the desktop version (which can be used on Windows, Mac and Linux) and not the online version as the former is limited to 5GB of storage. The new Autopilot functionality does automate some of the formally manual harvesting tasks. More detail in Preliminary Social Media Archival Tool Report - April 1 2020.

What can Twarc do for you?

Twarc is a tool for harvesting Twitter and yields structured data material. It is a command-line tool, so it has something of a learning curve. But much of it is really well documented. This tool was developed by Documenting the Now, and has an avid community of researchers and activists behind it. More detail in Preliminary Social Media Archival Tool Report - April 1 2020.

An alternative to twarc is TAGS, which is used via Google Spreadsheets. This is very good for a novice user and allows you to harvest API data from Twitter in a similar fashion with twarc, only with a little less flexibility.

How can you choose specific combinations of hashtags, accounts, keywords, etc. to archive (e.g. all the posts with a specific keyword from a Facebook user or all the tweets with a specific hashtag from a specific Twitter user)?

If you are using Webrecorder, it is currently not possible to filter the material that you want to archive in this way, since Webrecorder records everything that is served on your browser; thus if you record a hashtag feed, you will have recorded all tweets with that hashtag independently of which user made the tweets. The same is true for Facebook posts.

To deal with this issue, one could use twarc or a similar API-based harvesting tool, which generally allows for far more fine-grained search and capture capabilities but differs in the output it offers (structured data instead of a WARC file (see below)). Such data are also much more malleable even after they are harvested, as they can be queried and filtered with various data cleaning, analysis, and visualization pieces of software so that they result in different sub-datasets, for example focused on specific locations, or users.

How can you replay the harvested material?

Most look and feel tools produce WARC files. WARC a is web archiving file format used for long term preservation and giving access to the harvested material. To replay WARC's you can use the Wayback machine or the easy to use Webrecorder Player. Most structured data tools produce JSON or CSV files. These can be read by most statistical analysis programs and code editors.

How can you harvest YouTube videos?

You can use youtube-dl for this purpose. This is a command line tool (so has a learning curve) that is pretty reliable. More detail on this tool will follow.

Can you remove unwanted content from inside a WARC?

Yes you can, for example with a command-line tool called warc-extractor. By providing some keywords, this tool will search the URLs stored inside the WARC, filter out what you do not want, and create a new WARC.

How difficult is the harvesting of the different social media archiving platforms?

In general none of the platforms make archiving easy because this activity doesn't really fit their business models. Also the rules of use and software behind the platform will change so tools that work now can be obsolete tomorrow.

Looking at the biggest platforms:

- Twitter doesn't make it easy but with some trouble a lot can be done. You will need to create a developer API account when using tools that follow the structured data approach.
- Instagram is definitely archivable with Webrecorder and tools that work similarly - some more research is needed to look into the API harvesting possibilities for it (there are tools available, such as Instaphyte, etc.)
- Facebook makes harvesting very difficult and in may cases harvest will be incomplete or not work at all.

Not strictly questions about tools but important to answer anyway

Should I use a institutional account when harvesting material?

This is advisable because otherwise personal information concerning the one doing the harvesting will end up in the archive. Also from an ethical perspective it is better to be clear who is doing the harvesting.

Are there ways to archive social media apart from using tools to harvest material?

Using tools to capture social media content can often result in lower quality captures or rights violations, and that is an accepted reality for this task. An alternative to this is to contact the archival creator/donor and ask them for a copy of the data you want to preserve. Most if not all social media platforms offer this functionality of downloading one's own data, although the output differs in form, structure, and richness. Doing this is of course only possible when you have access to the social media users or creators you want to archive and are able to have/build such a relationship with them, which can often prove to be difficult.

Is information about the tools themselves important to document and preserve together with captured materials?

Yes it is. In social media archiving, and in web archiving in general, the collected materials are essentially created by the tools as we use them. The WARCs, JSONs and other files we harvest and capture do not exist before we start the harvesting process, like a DOC or JPEG that is selected to be archived does. Because of this, and in order to be able to preserve the provenance of social media records, recording information about specific tools used, their versions, operating systems, specific configurations, commands, etc. could be invaluable for the users of the material in the future.