Tormented ways

Frédéric Clavert

9/7/2022

Thanks for the invitation, etc

Who am I

I like to present a bit myself before starting a talk. I'm a historian initially trained in political sciences and international history.

What I am going to talk about today is linked to chance. I have started collecting tweets in 2009 for a conference I was organising, just to know what the conference participants were discussing on Twitter. I then learned – first for subjects of personal interests – to collect tweets massively, thanks to the streaming API.

When the Centenary of the First World War started, I was ready to collect tweets massively and easily, and I did it, first to see what would happen, after some discussions with historians of the Great War. And it worked so well, that it became my main research project until 2019.

In March 2020, when the first lockdowns where initiated in Europe, I had my server ready as the #ww1 project had reached an end for a few months. So I started collecting tweets – it's still under way, with more than 62 millions tweets, mostly french-speaking, stored somewhere in a database, on a university server.

That's the two projects I will base my keynote on.

So, « api or archives? tormented ways to transform tweets into historical sources ».

Defining an *Application Programming Interface* (API) is easier than defining *archives*. An API is a socio-technical device - a piece of software - that allows two apps to exchange features for instance. Do you see on a webpage a facebook "like" button? That's a possible use of the facebook API.

Concerning what we are discussing today, I am - like many developers and researchers - using the Twitter API to get information. Concretly, a piece of software on a server is connecting to the Twitter API, send it some information (what we want to collect, ie all tweets containing a word or hashtag from a list of words or hahstags), and the API is returning, if the request meets some condition that are defined by Twitter, information – in this case tweets and metadata about those tweets and their authors.

archives ?

Archives is much more complex to define.

I need to precise here that I am a historian and speaking as such – y considerations on archives might be contested by archivists who are far more rigourous than us!

"It's"Archives" is a word and concept that is much more polysemic than API, all the more that the first digital humanities projects where often called 'archives' and, in a way, blured the definition. Furthermore, even if most european languages are using more or less the same word (the French "archive(s)", the English "Archives", the German "Archiv" or "Archivalien" to give some examples), they don't cover the same concepts, and even when they do, traditions to concretely collect and sort archives are not the same from one country to the other.

So as it is complex to define, I asked the system that simplifies everything to define it: Google (define:archive).

Here is the definition of "archives" from Goole, in English. Some elements are here - though they could all be defined better. 'collection', 'historical documents', 'records', 'store'. But this definition lacks something.

archives? (in French)

And this something is present in the definition in French. 'classés' means 'sorted'. It's an important point, because, behind this word lays most of the work of archivists. Archivists have to decide what will be collected, what will be kept, how it will be sorted, and where it will be stored. They then index documents that are stored in archive centers.

archiving as the setting of historians' work

We remain in a too simple definition of 'archives', but let's say that archivists (and librarians) work can be summed up in three words: preserving, sorting, indexing. I'll let aside storage for today, I'll speak about accessibility a bit later.

Those three words are basically the setting of historians' work. With no preservation, there's no history. If documents or records are not sorted, they're not findable, and there's no academic history. If they are not indexed, searching the right document might still be too hard – and history is biased.

Just one example taken from my previous researcher's life. When I first went to the Bundesarchiv in Berlin, I had to look for documents that used for some of them to be stored in the former West Germany (Koblenz) and some in the former East Germany (Potsdam). In Koblenz, they worked on indexing. I could do very efficient search, sometimes at the level of the document, rather than at the level of the box. Archives that used to be stored in East Germany were preserved, sorted, but not indexed. I had to work very differently, less efficiently, and so much slower.

So archiving is the basis of almost all work of history. Archives are 'mediated' thanks to the archivists' work.

primary sources and archives

But there's a difference between archives and primary sources. All archives are primary sources, at least potentially. But all primary sources are net necessarily archives.

Indeed, in some cases, historians have to get their primary sources without archivists. For instance, oral history often implies a method where the historian is creating, together with interviewees, their own primary sources.

Some 'private archives' – family archives for instance – can sometimes be included in primary sources but not in archives: when they are preserved, but not sorted and even less indexed. In that case, historians will have to sort and index archives themselves.

Tweets, data – I am saying data and not archives, you all noticed it, though I am not going to speak about that today – from social media can also be primary sources but are most of the time not archives - in the sense that they are not mediated by archivists.

Preservation, sorting and indexing are dealt differently. Either they are archived through processes such as web archiving, or they are not archived. They will be preserved, or not, by social media firm themselves. And you will access them through either possibly not legal ways (scraping), or through an API. If there is one.

So, to access tweets, there is a mediation through the API – not through the work of archivists.

Tweets as primary sources

So Twitter has an API and it is relatively open – that's why we are all working on Twitter. Facebook API is much more restricted and some services (whatsapp) have no API or at least no API that allows to collect data, even for research (which is not a bad thing).

Using Twitter API is a convenient way to get primary sources over events, phenomenon, information circulation since 2006. But it's not archived in a proper sense – there's no archivist's work of selection, preservation, sorting and indexing, but at the Library of Congress, but with no means to access it.

So instead of having the archives and the archivists' work as setting of the historians' work, we have an API. The mediation between historians and their primary sources is here totally different than in archives centre.

This API is:

- changing over time, sometimes dramatically (think about Facebook, that closed many features of its API in 2016);
- forcing you to make choices (stream or search, that is the question. Or pay. Or get the validation as a researcher since 2021);
- giving you the opportunity, but also the responsibility, to create your own primary sources, your own corpus, your own dataset.

api as the setting of the historians' work

More generally, when you work with social media (or lots of web based / internet accessible online services) and their APIs, there are many elements you should pay attention to:

- Is there an api? As I said, we are many to work on Twitter, because it's feasible. Because there's an API. It's a strong source of biases. Of course there are other ways (scraping) and in some cases, there will be enough data or primary sources in web archives.
- What are the conditions to use the API? Before 2021, to use twitter's API with no budget, you had to choose between 1) streaming (1% of the firehose, meaning 5 to 6 millions tweets per day theoretically) that implied anticipating hashtags (for instance) or 2) search but with strong limitations – you could only go back up to 7 days in the past.
- What is the sustainability of the API? usually it's not sustainable, it will change, it might be brutally closed and that will destroy at the same time your research project. It happened to the french project algopol based on facebook in 2016 – hopefully, they had enough data to work on.

All those elements will orient your research and influence the way you work. The disadvantage is that the work archivists are usually doing with lots of historical documents will be yours to do: it's time consuming, it also requires to learn technical (not only computing) skills. The advantage is that you can taylor your corpus as you wish and you do this very explicitly.

researching echoes of the past on twitter

So good research – I hope so – is possible with tweets as primary sources. I will here speak about two research projects based on massive corpora of tweets: #ww1 and #covid. Of course, my method is quantitative, sometimes qualitative. Some research can also be performed qualitatively.

The interest of those two projects for today is that the first one was entirely based on Twitter API 1.1 (streaming, more precisely) and the second one is based on this same API and the new Twitter APIv2 (ie access to 10 millions tweets per months if you are recognized by Twitter as a researcher, which I am).

#ww1 and the echoes of the centenary of the first world war online

I have started to collect tweets about the Centenary of the First World War in April 2014 – before the commemorations started – and stopped in December 2019 to have one 11th of November after the end of the Centenary.

I collected a bit more than 9 million tweets, published by around 1.5 million Twitter accounts. Around two third are retweets. I have used hashtags or keywords to collect those tweets, mostly French and English. 85 % or the tweets are in english, 10% are in French, the 5% remaining gather other languages.

I have tried to explore this corpus with distant reading tools. Distant reading is usually a reference to Franco Moretti's book, *graphs, maps and trees* published in 2007. It is a set of quantitative methods that allows a researcher to ask the computer to read for them, when there is too much data to be read humanly. Since 2007, most researchers are integrating distant reading in scalable reading, where distant reading and close reading are used at the same time (Fickers, Clavert).

Distant reading might be necessary as tweets are primary sources, but massively collected, they are primary sources that you cannot read yourself. Close reading and other kinds of readings might be necessary to get to the details, to get to precise events.

api as a framework

I have used the Twitter streaming API, such as available in its 1.1 version. That implied constraints:

- 1% of the firehose (access to all tweets being published) – technically 5 to 6 millions, but the 1% is based on quarters of hour – so I did, particularly on the 11th of November 2018, sometimes go over the 1% – which means that some tweets have not been collected on those occasions;
- anticipation – at the beginning of the research project, I could not guess that the 2016 commemoration of the battle of the Somme on the first of July 2016 would use the #somme100 hashtag.
- no past tweets available: errors in anticipation could hardly be corrected. One example in the #ww1 project case is the assassination of Jean Jaurès. We should also remember that, sometimes, there's no efficient ways to collect tweets for one event: in the case of the centenary of the first battle of the Marne for instance.

some results (1, 2, 3)

to be done

covid19 on twitter

I will not detail too much the results of this research project. I have started it while obvious that France, Belgium and Luxembourg and other european countries would be placed under lockdown, mid-March 2020. This research and harvesting is still going on, with now a bit more than 62 millions tweets collected and the ambition to let it go as long as possible.

My colleague Deborah Paci (today at the university of Modena) is co-leading this research with me. We could access an italian corpus, for comparison: what happened on twitter during lockdowns in french-speaking countries (moslty France, let's be honest) and in Italy? We are currently writing about the results.

The interesting thing here is that, while still usig the API 1.1 for this project, Twitter released its APIv2. In the mid-term, this version of the API will replace the 1.1, but for the moment both versions can still be used.

new api, new framework...

There's a huge innovation with this new API for researchers: if recognized as researcher by Twitter, then we can collect up to 10 millions tweets a month in the full hsitory of Twitter. It's obviously addictive all the more that you can go back in time, up to 2006. That's long-term (well, from a social media point of view) brought (legally) to twitter research.

It's still not archives or archiving. But it starts to be a bit better than the APIv1.1 and its streaming set of commands, because you have a right to make a mistake with the API v2.

...new limitations

But there's a price to that. You can make a mistake, sure. But there are tweets, while searching the full history of twitter, that you will never find: deleted tweets, tweets from account that switched to a private

status. Those tweets can be captured if you are using the streaming sets of command (that still exists with the v2, but with more restrictions).

That makes the search archive feature a problem if you wish to study some kind of research questions: controversies, the spread of misinformation for instance.

Another limitation is the process that allows you to get recognized by Twitter as a researcher. And if you get rejected? It happened to Nick Ruest who asked for a research access for the Archive Unleashed project. His 'use case' was not elligible, probably because he mentioned web archiving, but it's a guess.

working on archived tweets

Considering what I just developped, would proper archives of tweets be a better solution? For lots of research questions, it is. It would be also a strong gain of time. The problem is rather to define what would be a proper archiving of tweets – that what besocial is about, to.

What would be the point

Just some thoughts, more or less as a conclusion, on the interest of proper archives of tweets.

I am working with CSVs, JSONs, SQL databases. Though in the metadata there are elements about the pieces of software users have used to consult twitter – there are lots of metadata when you collect tweets – there is something I lack: the historicization of the tweet.

an example: Obama's four more years

Let's take as an example a tweet that has long been the most favorited / retweeted tweet of twitter's history: Obama's '4 more years' tweet. If you look at it today, Twitter's interface, Obama's picture, answers, retweets, are today's elements. If you collect it with the API, you will have some elements that historicize it (its publication date), and some that are today's (retweets, favourites, all the metadata about the account who published it). Tweets' metadata are not time consistent, in the sense that there is no concept of history in its metadata (and this would be hard to implement anyway).

Obama's four more years tweet in 2012

In the case of this very famous tweet, we can have a look at how it looked in 2012, thanks to the Wayback machine. But in most cases – probably 99% of the tweets – it's just not possible. Furthermore, it's not that simple to collect archived tweets massively within the Wayback Machine, it's probably even impossible. Using tweets archived in web archives is hence probably limited to qualitative methods.

We are here at the heart of what could bring to researcher a proper archiving of tweets: historicization.

how to really historicize twitter?

I'll let the besocial team, in a way, answer this question. For me, the ideal situation is to have at the same time the API and archives, the best of two worlds. But archivists and librarians have choices to make, decisions to take, and it's not always possible to have the best of the two worlds.

The Twitter archive at the Library of Congress is a good example of the choices archivists and librarians may have to do. An agreement between Twitter and the Library of Congress allows the later to archive Twitter. Up to the end of 2017, this archiving was integral. All tweets from 2006 until December 2017 are there. But the Library of Congress lacks the technology to search its twitter archive. So from the 1st of January 2018, the LoC started to archive Twitter "on a selective basis". It's been depicted as a failure – but archivists know in the end that they must do choices (even if it bothers historians) – clearly I would have prefered a full archiving of all tweets! But archiving everything bears also a risk in itself, which is, basically, losing everything.

It's, in the end, an old question asked to archivists, librarians and historians.