History in the era of massive data: online social media as primary sources for historians

Frédéric Clavert, C2DH, University of Luxembourg

# Abstract

This article draws on the experience of the Luxembourg Centre for Contemporary and Digital History (C2DH) research project, “#ww1 - the centenary of the Great War on Twitter”, to contemplate what it means to use born-digital and big data primary sources, both in terms of methodology and as a contribution to the field of memory studies and, more broadly, to contemporary history. It discusses the notion of distant reading in order to show the interest of multiscale reading in the framework of big data and elaborates upon the use of reflexive thinking in terms of what we really study when using this kind of primary sources (tweets and, more generally, feeds of born-digital sources), i.e. temporalities and information circulation.

# Introduction

“My love, I am writing to tell you that I won’t go back from the war. #11novembre”[[1]](#footnote-2). This very emotional tweet quotes the beginning of a probably fake letter, supposedly written by a Poilu for his wife. It was retweeted 13,437 times from November 11, 2014, to July 13, 2016. It was issued by a little-known account and deleted during 2016. It is extracted from the #ww1 database, which was created during a research project focusing on the echoes of the Great War and its Centenary on Twitter. This project belongs to the interdisciplinary field of memory studies, with a strong emphasis on historical aspects. It is based on born-digital primary sources from social media.

Since the launch of MySpace in 2003, Facebook in 2004 (and its subsequent release to a wider audience in 2006) and Twitter in 2006, online social networks, or social media, have expanded and become more diversified. In 2015, for the first time ever, one billion people accessed their Facebook account in a single day. While MySpace was not able to withstand the competition from Facebookand started to decline at the end of the 2000s, other social media platforms continued to grow or were created (Snapchat, WhatsApp, etc.). Today, Facebook has more than two billion users, Twitter has 326 million users and WhatsApp, a messaging system with social media features, has 1.5 billion users.[[2]](#footnote-3) In the space of a decade, those social media platforms have taken a role in people’s lives that nobody could have anticipated in the middle of the 2000s.

The emergence of social media has led to the production of numerous artefacts – tweets, posts, video links, images, etc. – that document various aspects of our lives. This article explores one way of making use of social media data in historical sources.[[3]](#footnote-4) This approach aims to show how big data inspired methods can bring new elements though with limitations. Can Big Data artefacts be used for historical studies? What are the consequences of this use for the historian’s work? How can these artefacts be studied in more detail?

To answer the question of how to use social media data in historical research,[[4]](#footnote-5) this article is organized into four parts: (I) how to collect and (II) read those sources; (III) how to understand what is being studied when analysing data (i.e. the question of information circulation and temporalities); (IV) how to question, in such research, primary sources and the algorithms used to analyse this data. The conclusion will draw on some thoughts on what this kind of research can bring to memory studies and historical scholarship.

# I. Collecting sources

In this section, the framework of big data (*ie* research based on vast amounts of data) will be discussed. There are several ways to collect data, notably web scraping: extracting data directly from web pages using dedicated software and connecting to an Application Programming Interface. This second case will be developed here, as an example of harvesting of born-digital sources.

Many scholars have discussed the issues encountered by social sciences and humanities in their ability to get big data corpora, as they can first be expensive to collect.[[5]](#footnote-6) Beyond the question of the cost of data harvesting – which, in certain circumstances, is rather low in the case of Twitter - we can – second – collect only certain types of data and we are dependent on the data provider. Getting large amounts of data from Facebook is very hard, if not impossible, at least when we do not work with developers or computer scientists.

Twitter is the only online social network to have opened an Application Programming Interface (API) that partly meets the needs of researchers, whether historians or not. An API is an interface between two types of software - in this case, Twitter and the software that will harvest data - that allows this software to exchange functionalities or data. For instance, when you see a Facebook “like” button on a web page, it means that the webmaster of this page used Facebook’s API. Technically, software on a 24/7 operational server will use a set of commands to provide authentication to Twitter, send a request (obtaining all tweets that contain one or more precise word(s) with or without Boolean expressions) and then retrieve the corresponding data. But there are several sets of commands/APIs that can be used. For instance, the Twitter search API allows up to 3,000 tweets per hour to be retrieved and data from the past seven days on Twitter to be extracted. But 3,000 tweets per hour can be too limited: in that case, we can use the Streaming API, which enables the collection of up to 1% of all tweets published at the time of the data harvesting. This theoretically means that we can harvest up to 5 million tweets per day.

The API is not the only technical mechanism that we should take into consideration. The way the software or scripts are used to collect data should also be understood by the researcher. For example, we started to collect tweets through a server side script called 140dev.[[6]](#footnote-7) The developer behind this piece of software was not interested in collecting some data that he could have collected through the Twitter Streaming API, notably the user’s profile pictures. This information was not of interest for the #ww1 project. Nevertheless, in terms of open data and reuse of data by other researchers, it is indeed a limitation of the #ww1 corpus.

For the needs of our research, we had to use the Streaming API, as we wanted to collect more than 3,000 tweets per hour. Our strategy was not based on the constitution of a corpus of samples of tweets containing hashtags but on the collection of a volume of tweets that was as close as possible to the totality of tweets emitted containing these same keywords. Indeed, in 2014 at least, no sampling approach on twitter seemed viable in the long term. One of the goals of the #ww1 project’s harvesting strategy was to keep the possibility to analyse “weak signals”. But that “big data” strategy required a certain amount of anticipation, as this API does not allow past tweets to be collected. For instance, when starting to collect tweets, it was not possible to predict that the hashtag for the commemoration of the centenary of the battle of the Somme would be #Somme100, nor to anticipate that the “*1 jour - 1 poilu*” (*1 day - 1 fallen soldier*) “challenge” would be so significant in the French-speaking part of this corpus (see below) – and without our “big data” strategy, this hashtag would have been seen later than it was effectively as it started as a “weak signal”. The use of that kind of API therefore means being aware of what is taking place on Twitter.

Anticipation is indeed a key quality when dealing with social media which were not intended to be archived, a big exception being Twitter which concluded an agreement with the Library of Congress (the Twitter archive is not accessible yet). Social media are sometimes very partially being archived by national institutions (archives or libraries). Those archiving actions are often limited to a national sample. What then should be done during major events, such as the 2015 Paris attacks? The *Institut National de l'Audiovisuel*, on this occasion, was fortunate enough to be able to count on personal collection initiatives to complete its archives on this dramatic episode[[7]](#footnote-8). The numerous initiatives of digital artefacts archiving (including tweets) that were born during the Covid-19 crisis seem to show the importance of, as historians, librarians or archivist, being responsive to such an event[[8]](#footnote-9).

Data harvesting software also has the same flaws as all software, the most significant one being its maintenance or lack of it. The #ww1 project had to deal with several server side harvesting software, as the first one that it used became obsolete in 2017. The transition from one piece of software to another was not easy and required a supplementary step after data harvesting that is not neutral (see below*,* the question of “in between tools”).[[9]](#footnote-10) One example is the merging of different databases that correspond to the two tools we have used to collect data; this changes the structure of the data, with information included in one database and not the other.

The methodological questions posed by the harvesting of data are not only related to the data itself. In fact, in our case, the most difficult issue is the question of the choice of keywords to be collected. Big data, as described in a famous article in Silicon Valley’s journal *WIRED*[[10]](#footnote-11) or by Mayer-Schönberger and Cukier,[[11]](#footnote-12) implies inductive methodology. In this reasoning, data is seen as “neutral” and thus conclusions can be drawn without any hypothesis and, this, therefore would imply the “end of theory”. As danah boyd and Kate Crawford argued,[[12]](#footnote-13) this “end of theory” is a myth as all data is the result of social processes, and the #ww1 database is no exception. In this sense, as R. Kitchin has extensively argued, our database is oligoptic and not panoptic[[13]](#footnote-14). Though collected on a large scale, for instance, many tweets concerning the centenary could not be collected because of the choice of a keyword or hashtag-based method. This method does not allow the harvesting of tweets that answer a harvested tweet without using one of the collected hashtags.[[14]](#footnote-15) This does not mean that a hashtag-based corpus of tweets is not usable, but it does highlight the need for caution and an awareness that all conclusions drawn from analyses of this kind of corpus can only be limited, if not temporary.

Indeed, one of the problems of using hashtags to collect tweets is that it is hardly possible to go back. In a way, it means that historians should remain very humble in the way they “write” history. More generally, it should also remind the profession that Google-style keywording – now a wide practice of historians contrasting with “old-styled” indexing based on the archivist’s know how – should not be the only way to look for information. Unfortunately, in the case of the #ww1 project it was the only path to the desired primary sources.

Constituting the corpus of primary sources is only the first step of the research process. In research based on huge volumes of data, once the harvesting is completed and the data stored[[15]](#footnote-16), it will be read by historians, and this implies having recourse to the notion of distant reading.

# II. From close to distant reading and back again

How can a historian read seven million tweets? How can we read in the era of big historical data? We cannot, but it is possible to ask a computer to carry out the task for us, using what is now called distant reading. This concept was introduced by Franco Moretti, a historian of 18th- and 19th-century European literature.[[16]](#footnote-17) His research examined the following issue: how can we study the history of literature by taking into consideration all literature, not just famous novels? How can we write a history of literature which would embrace forgotten works by famous authors and also forgotten writing by “minor” authors? The answer was developed in three articles written in 2005 and published as a book in 2007: graphs – i.e. applying to literature the quantitative methods used by the School of the Annales in the 1960s -, maps - mapping literature -, and trees - using Darwin’s evolutionary methods to understand literature. This method and the impressive set of data to be analysed (i.e. 200 years of European literature) require the use of computing. In other words, we need to ask a computer to read for us. Moretti called this a “little pact with the devil: we know how to read texts, now let’s learn how *not* to read them.”[[17]](#footnote-18) For Moretti, distancing in this context is “a condition of knowledge” if we are to understand a system in its entirety.[[18]](#footnote-19)

Moretti’s distant reading has been widely criticized. Maurizio Ascari sees distant reading, or at least the way Moretti makes use of it, as a way of bending reality to theory, as a way of neutralizing critics thanks to the mask of a supposed objectivity claimed by the use of a scientific approach.[[19]](#footnote-20) This criticism has also been used by Brennan, who refers to the work of the digital humanist Ted Underwood and adds to this argument that digital humanities (here defined as quantitative approaches to literature) require a large budget for questionable results.[[20]](#footnote-21) Less controversial and more to the point, Sculley reminds us all that data mining - a kind of distant reading - are impacted by implicit hypotheses.[[21]](#footnote-22)

This article nevertheless argues in favour of distant reading, if it is combined with other kinds of reading. Specifically, what does it mean to practice distant reading? In our research, we perform - or rather, our computers calculate for us - several operations that can be considered as distant reading.

Our research began in April 2014, when we started to collect theoretically all tweets published publicly that contained keywords and hashtags linked to the First World War in French and English[[22]](#footnote-23). We then regularly increased the number of collected keywords to 71 in December 2018. As of 1.12.2018, the database consists of 7.2 million tweets. Two thirds of them are retweets and one third are original tweets. The tweets were published by approximately 1.8 million Twitter accounts.

|  |  |  |
| --- | --- | --- |
| Event | Date | Number of tweets |
| Centenary of the armistice | 29 October -14 November 2018 | 1804480 |
| including | 11 November 2018 | 894 978 |
| Centenary of the Battle of the Somme | 1 July 2016 | 262 275 |
| 11 November 2016 | 11 November 2016 | 104 525 |
| 11 November 2014 | 11 November 2014 | 88 518 |
| 11 November 2017 | 11 November 2017 | 88 370 |
| Declaration of war | 4-5 August 2014 | 151230 |
| 11 November 2015 | 11 November 2015 | 70 702 |
| Centenary of the Battle of Vimy Ridge | 9 April 2014 | 66 092 |
| Centenary of Verdun (May 2016) | 29 May 2017 | 60 495 |
| Centenary of Verdun (February 2016) | 21 February 2016 | 24 163 |
| Centenary of Verdun:Black M controversy | 11-13 May 2016 | 62664 |
| Easter Rising Concert for the centenary | 29 March 2016 | 21 893 |

Table 1 - The main events when the centenary of the Great War received the most tweets, grouped by event.

Table 1 shows a simple distribution of the #ww1 corpus, by simply looking at the events of the centenary of the Great War that were the most tweeted. It provides an insight into the hierarchy of the commemorations: the event on which the largest number of tweets about the Great War were published was on the centenary of the Armistice -11 November 2018 (around 900,000 tweets), with a sharp increase in the daily number of tweets from the end of October 2018. The commemoration of the Armistice, on 11 November, is also the only recurring event in this database. Every 11 November is marked by quite a large number of tweets.

The day with the second highest number of tweets was 1 July 2016, the centenary of the beginning of the Battle of the Somme. Further analysis of the tweets reveals that the organizers of the centenary of the Battle of the Somme included online social networks in the commemoration by providing instructions on site regarding which hashtag (#Somme100) to use on social media. The Battle of Verdun and the French (February 2016) and Franco-German (May 2016) commemorations were also present but for different and very controversial reasons (see below), particularly in the case of the Franco-German commemoration.

Some of the entries are more related to the British dominions: the centenary of the Battle of Vimy Ridge which is particularly important to Canada and the centenary of the Easter Rising in Ireland. The commemoration of the landing at Gallipoli on 25 April 1915, which became ANZAC day, the national day of remembrance in Australia and New Zealand, is not on this list but still remains among the most tweeted events (25 April 2015).

This list of the fifty most active days should not prevent us from reflecting on other peaks in activity, admittedly less remarkable but that can be due to regional interest. Among these smaller peaks, we can mention the centenary of women’s suffrage in Great Britain (2018). In the list above, we can also find the declaration of war, especially 4 August (UK). Other notable dates are conspicuous by their absence, such as the centenary of the assassination of Archduke Franz Ferdinand. This commemoration was supposed to be international in its scope but was not met with great success.

Some other events are missing, especially the First and Second Battle of the Marne. This is linked to another element: the manner in which French memory institutions tweet themselves. In other words, there were no clearly defined hashtags for these events. Another event which is absent is the declaration of war by the United States of America. Furthermore, apart from ANZAC day, colonial troops, especially from Africa (for example, the Senegalese Tirailleurs) and India, are not really present in our database. Nevertheless, a closer look at the autumn 2018 sequence shows that the Twitter discussions around colonial troops depend on a commemorative vector. For example, on 11 November 2018, the Beninese singer Angélique Kidjo sang in memory of the Senagalese Tirailleurs. Many tweets were then published, some of them openly racist. In addition, the analysis of the French-speaking corpus of tweets is highly dependent on the activity that has taken place around the database of the Dead for France and the records in the database concerning colonial soldiers were much more difficult to transcribe: birthplaces have since then disappeared or their names have changed, family names were misspelled, etc. These records were nevertheless indexed but were less visible on digital social networks. In a way, this low visibility is part of the complex memory of the Tirailleurs[[23]](#footnote-24).

The table 1 allows us to establish a sort of hierarchy of the commemorations (what is there and what is missing) and, more generally, the global temporalities of the centenary on Twitter. However, it does not help us understand the proportion of languages (mainly English and French) and the temporality of this ratio.

 Though the English language is largely dominant, on certain days French is in the majority: this is the case on 11 November, with the notable exception of 11 November 2018, and the commemoration of the Battle of Verdun (February and May 2016). There are some striking elements revealed by the data: the commemoration of the start of the war is marked by more English-language tweets whereas the end of the conflict on 11 November attracted more French-language tweets. Obviously, with the exception of 11 November 2018, the commemorations were more national than international in their focus. This is all the more interesting as it shows a strong difference when compared with the historiography of the Great War, which has become manifestly international, at least since the 1990s, as have argued Jay Winter and Antoine Prost[[24]](#footnote-25), though they do recognize that the national framework of the Great War narratives are still strong. This internationalization has not, according to this research on Twitter, percolated into the collective memories of the event.

Concerning the content of the tweets, we proceed to a clustering of tweets (Figure 1) and project it through time. The aim of this clustering is to find themes using diverse statistical methods, which are essentially based on co-occurrences of words. This is what is revealed in Figure 1, for the French-speaking part of the corpus.



Figure 1 - Clustering of French-speaking tweets 2014-2018

French-speaking accounts published tweets about the *poilus* (French soldiers): classes (or clusters) 1 and 2 consist of tweets that point to precise individuals, usually declared *Mort pour la France* (Died for France). Related to those classes are the clusters (9 and 14) that are related to an initiative by amateur historians (the “*1 jour - 1 poilu”* challenge) to index the *Morts pour la France* database in 2013. This database, published by the French Ministry of the Armies,[[25]](#footnote-26) contains images of administrative documents with some metadata. The ministry opened an indexation module in 2013, meaning that users could create an account and index the database, i.e. transform an image into proper text (text seen as text by the computer), allowing a more comprehensive search of the database. This action was seen as an individual homage to each *poilu* who died during the war and, at the same time, as a service to future historians.[[26]](#footnote-27)

Some other clusters reveal what was organized (for instance, exhibitions) or published (books, articles) during the centenary (classes 3, 4). Days of major commemorations are also evident (classes 9 & 12) and are concomitant with tweets paying a general homage to soldiers (class 11). Sometimes, the events which took place during the war are written about, including references to Germany (classes 13, 8). The nature of 11 November as a public holiday is mentioned too (class 5). Lastly, some classes are more specifically linked to the centenary of Verdun (classes 7, 10 and 6). This particularity will be explained later in the article.

These clusters can be projected through time. What can be derived from this projection is the fact that, on major days of commemoration, the words used to pay homage to soldiers are more general. On the other days of the centenary, the words used are more personal (clusters 1 & 2), focusing on specific soldiers, their names, origins, regiment or the cause of their death. Furthermore, the clusters containing tweets with contestation-related words are present during all great commemoration events. It seems that the French collective memory of the Great War is not that consensual after all.

What we can learn generally from these two analyses is the following: *poilus* who died for France have been at the center of the French collective memory since 1919. In a way, this means that the act of commemoration on Twitter is merely a continuation of the broad lines of the French collective memory. What has changed, however, is practices, as is shown by the collective indexation of the *Morts pour la France* database.

Some of the conclusions that were drawn here could not have been reached without the long-term collection effort undertaken by the #ww1 project. Sampling would have minimized some of the elements that could be analysed, first and foremost the tribute tweets to individual soldiers, which were primarily issued on days without official commemoration.

The harvesting effort to collect these natively digital sources also made it possible to capture a certain number of new memory-based practices and, in particular, that which consisted of organizing collaborative work around the database of the *Morts pour la France* which became a *lieu de mémoire* during the Centenary. Certain analyses that have not yet been carried out should provide more information on these practices, for example the way emojis are used when evoking elements of collective memory. The identification of automated (bots) or semi-automated (cyborgs) accounts and an analysis of the influence of (malicious) automation of content publishing on collective memory are among the elements that the project has not yet undertaken.

We can also try to understand interactions between Twitter accounts thanks to an analysis of the links between Twitter accounts mentioned in our database. It reveals two groups of Twitter accounts that correspond to the English-speaking part and the French-speaking part of our corpus. Investigating the nature of the Twitter accounts (collective or individual; institutions; historians; historical projects, etc.) is of major interest: there are more museums, for instance, in the English group, starting with the Imperial War Museum (@I\_W\_M); there are also a few accounts created for the promotion of public history projects; the Poppy Legion (@PoppyLegion) is also very active, as are amateur historians (@HistoryNeedsYou, for instance). In the French-speaking part, the most important account is the one of the *Mission du Centenaire* (@Mission1418), the inter-ministerial mission that was in charge of the organization of the commemoration. A closer look reveals that the @1J1Poilu account - a Twitter account that encourages people to transcribe the *Morts pour la France* database - is quite popular among the French-speaking community. The Twitter account of a highly active journalist, Stéphanie Trouillard (@stbslam), is also quite prominent. There are few signs of historians, and almost no trace of French museums, despite the fact that there are many Great War museums in France. For instance, the *Musée de la Grande Guerre* in Meaux does not often use hashtags related to the Great War. This museum tweets without taking part in a more general discussion by using the corresponding hashtags: it is a top-down communication policy. The Twitter account of the *Historial de la Grande Guerre* (in Péronne, Somme) is also fairly active, but it only started to use hashtags after the beginning of the centenary.

These few ways of using distant reading methodologies demonstrate that they are practical for gathering a large corpus in its entirety and for understanding the information stored in a database. But distant reading has a consequence; by using it, we lose something, one might lose details, lose his/her link to the specific and individual primary source as well as his/her link to those that have produced the tweets or are quoted in tweets.

A multiscale approach should be encouraged, mixing distant and close readings (i.e. the traditional approach used by historians to read primary sources), which involves carefully choosing the software that is used to perform distant-reading operations. Some software will produce results that, once obtained, cannot be linked to original primary sources. The software we chose allowed us to go back to every single tweet, where necessary.

In the case of the Franco-German closing ceremony to commemorate the Battle of Verdun in Douaumont, which triggered huge controversy because the commemoration was staged in the military cemetery in front of the ossuary, the ability to mix distant and close reading has proved essential to understanding what was at the origin of the controversy. Distant-reading methodologies must be used with caution and we must be aware of their real benefits for research.

Nevertheless, some aspects of distant reading bring strong elements to the research. In the case of the #ww1 project, distant reading has allowed an analysis of the relationship to temporality, history and memory in the age of social media. Any approach that would have been too "micro" would have been difficult to generalize. The collection strategy adopted allowed analyses at different levels - from singular tweets to the corpus as a whole, from weak to strong signal.

# III. Understanding what is being studied: information circulation and temporalities

Practicing distant reading on a large number of tweets or of other kinds of born-digital or digitized sources requires us to think about what we are actually studying when we use large born-digital corpora. In the case of tweets, we are basically studying information circulation and temporalities, as argued by the sociologist and linguist Dominique Boullier.[[27]](#footnote-28) Boullier points out that big data helps us to think about the circulation of information through what he calls “vibrations” - those short but very intense moments of information circulation. In the case of Twitter, these are the moments when tweets are retweeted on a massive scale, when news (whether “fake” or not) circulates in a few seconds through countless “replications” (retweets, mentions, etc.) of the same information.

The centenary event to commemorate the Battle of Verdun is a good example.[[28]](#footnote-29) Two controversies took place: the first controversy occurred a few days before the commemoration itself when a far-right website revealed on 10 May 2016 that a concert by the French rapper Black M would be organized by the city of Verdun to thank young German and French people for their participation in the commemoration; the second controversy occurred when those young German and French people participated in the event by running through the military cemetery in Douaumont (29.5.2016). Those two controversies led to two different types of dissemination of information: the first (the Black M controversy) came from a far-right website, with an article that was then heavily tweeted. Those tweets were then used by the French media, on- and off-line, during the weekend. Those articles written by “regular” journalists were then quoted in numerous tweets again on the following Monday. The second controversy (the Douaumont controversy) was the result of an interaction between TV (the broadcast of the commemoration) and Twitter (people commenting live on the commemoration): a few seconds after the young people from both countries started running, the number of tweets that mentioned the ceremony doubled, most of them not coming from people in Douaumont but from people watching the ceremony on TV or people just retweeting tweets about what they perceived as a scandal. These kind of controversies are typical of the way information is disseminated, the way the “vibrations” described by Boullier have arisen with online social networks. It is not yet clear how these vibrations influence collective memory and its temporality.

The second element, related to the first, that can be studied with such a corpus coupled with distant reading tools is temporalities, and, more precisely, the interlacing of temporalities. In the case of our corpus, it is easy to distinguish at least four kinds of temporality: the temporality of each user (why are tweets sent at a precise moment?), the temporality of the commemoration of the First World War (which depends mainly on national agendas), the temporality of the Great War (similar, in part, to the previous temporality but a hundred years earlier), and the temporality of Twitter (the Western, open, linear view of time that implies a never-ending flow of tweets on innumerable subjects).[[29]](#footnote-30) It is precisely this interlacing of temporalities that can be studied and that will provide an insight into the influence (or lack of influence) of Twitter and other online social networks on collective memories.

Understanding temporalities in online social networks is not a small task: for approximately the past 15 years, many researchers have questioned the concept of time, from its acceleration[[30]](#footnote-31) to the changing regimes of “historicity.”[[31]](#footnote-32) Gumbrecht notably questioned a broadening of our present.[[32]](#footnote-33)

Linking those considerations on present and time to Twitter studies is not an easy task. Nevertheless, Jan Assmann’s definition of cultural memory helps: “We need a term to describe these processes [of collective remembering and forgetting adds Marek Tamm] and to relate them to historical changes in the technology of storage systems, in the sociology of the groups concerned, in the media and in the structures of storage, tradition, and the circulation of cultural meaning in short, to encompass all such functional concepts as tradition forming, past reference, and political identity or imagination. That term is cultural memory”.[[33]](#footnote-34) Assmann considers in this definition that technology (and the way it stores information) has an impact on memory. While we cannot at present give precise indications of how Twitter affects collective memory, it seems to us that a phenomenon is emerging, that is linked to our current time information and storage technologies: the constant updating of the way we relate to the past in the present, a memory confronted with the circulation of information through high-frequency replication.

The study of Twitter as a born-digital feed or archive shows that a short event, such as the Verdun controversies that we described above, can be understood and analysed second by second. In a sense, this is not a broader present but a constantly updated present[[34]](#footnote-35) and hence a constantly updated presence of the past within our present as well as a constantly projected present to the past. Constantly updating our present necessarily has consequences on how historians will see the past and its memories. The ways information is circulating through social media is already influencing the way we question some more ancient primary sources: research projects based on digitized newspapers are looking closely at how information was memeticaly copied and circulated[[35]](#footnote-36). The ways they are analysing this information circulation is inspired by how social media functions.

Another element to consider is how, by choosing software to collect, prepare and analyse our data, the embedded algorithms in software are changing our relationship with our primary sources, thus affecting the “allure” of the born-digital archive. In her famous 1989 book[[36]](#footnote-37), Arlette Farge sees the intimacy that she created with her primary sources as a way to give back a voice to historically ignored actors – in the case of her research, imprisoned French women of the 18th century. She sees the manipulation of paper archive as a reminiscence of the past that seems necessary to understand it. Within a digital framework, with born digital primary sources that are published by accounts rather than by people, will historians see in the future the today’s voiceless historical agents the same way?

# IV. The historian, primary sources and algorithms: the changing allure of the (born-digital) archive

Studying information circulation and temporalities is not necessarily new.[[37]](#footnote-38) Nevertheless, understanding what is at stake for a (future) historian when studying online social networks implies knowledge of what it means to use online social networks as primary sources and an awareness of their limitations.

Informally, an algorithm is any well-defined computational procedure that takes a value, or set of values, as input and produces a value, or set of values, as output. An algorithm is thus a sequence of computational steps that transforms the input into the output.[[38]](#footnote-39)

With this definition of an algorithm, the distant reading of a text is the same as having this text as an input and clusters/themes as the output, or having a set of Twitter accounts and links between them as an input and the calculation of a network (the degree of nodes, for instance) and a visualization as the output. But what happens between input and output? That is where the notion of algorithm can help. The simple definition above implies that we identify a “class of problems” and a way to solve them, which indicates that defining an algorithm will rely on specific theoretical framework. It means that those apparatus, whatever they are, will then be translated into software that implements this algorithm. As a consequence, choosing software means choosing an algorithm and a specific framework of analysis. Two examples of software are here presented: gephi[[39]](#footnote-40) and IRaMuTeQ.[[40]](#footnote-41)

Gephi is an application that allows its users to calculate and visualize networks and perform social network analysis. It is widely used in the field of humanities and social sciences and beyond[[41]](#footnote-42). It is based on the sociology of social networks, which in turn draws on graph theory in mathematics. Furthermore, gephi was at least partly developed within SciencePo’s medialab, founded by Bruno Latour, the sociologist and philosopher who defined actor-network theory.[[42]](#footnote-43) The first versions of gephi were used within the eDiaspora project (led by Dana Diminescu), which uses a sociology of networks.[[43]](#footnote-44) In other words, this tool should be chosen only if a historian intends to implement the tools/methods provided by the sociology of networks. It might be more problematic either to ignore that gephi is a computer tool implementing algorithms derived from this sociology and graph theory, or to try to use it to analyse a phenomenon based, for example, on the methods provided by Bourdieu’s fields theory.[[44]](#footnote-45)

IRaMuTeQ, a software for text analysis developed in Toulouse, is based on the work of the French school of data analysis, and, in terms of clustering, on research carried out by Max Reinert in the 1980s.[[45]](#footnote-46) Max Reinert developed a set of statistical algorithms based on co-occurrences of words in segments of texts, which he called “Theory of lexical worlds” (“*Théorie des mondes lexicaux*”). This theory starts from the hypothesis that if the same kind of vocabulary is used in two texts or segments of text (a segment of text is roughly a sentence), then those two or more texts or segments of texts are considered somehow similar and share the same point of view.

The text analysis embedded in *IraMuTeQ* can be compared to topic modelling tools that “take[s] a single text (or corpus) and looks for patterns in the use of words; it is an attempt to inject semantic meaning into vocabulary”[[46]](#footnote-47). Topic modelling tools are today widely used in several fields, such as litterature[[47]](#footnote-48), sociology[[48]](#footnote-49) or history[[49]](#footnote-50) for instance. As Graham, Milligan and Weingart put it, it should be remembered that “The topic model generates hypotheses, new perspectives, and new questions: not simple answers, let alone some sort of mythical “truth.” The act of visualization of the results too is as much art as it is science, introducing new layers of interpretation and engagement”[[50]](#footnote-51). They emphasize that tools that are implementing topic models implement it differently and their outputs can be different from one tool to another. Our own choice of IRaMuTeQ depended on three factors: it deals well with english and french corpora; it can handle rather large data; by experience it does give interpretable outputs with corpora of tweets. It is worth stating that MALLET, a machine learning based software, is considered standard for topic modelling in digital humanities[[51]](#footnote-52).

For our research, although gephi and IRamuTeQ, their algorithms and hence the sociological or linguistic logic that inspired them provide the most visible results, we also used a wide range of software that is more discreet but nonetheless just as useful as the aforementioned applications. Our project is not an exception: many researchers in digital history/digital humanities use software that we might call “in between” tools, situated somewhere between those tools that help collect data and those that perform analyses on which the interpretation of primary sources will be based.

“In-between” tools are all those tools that help prepare data before it is analysed: spreadsheet software like *Excel* or *Calc*, data-cleaning tools like *OpenRefine*, large dataset tools such as *Dataiku Data Science Studio*, etc. These tools allow researchers to carry out many tasks on data that remain “invisible” (usually publishers are not interested in details on how your corpus was prepared). For instance, if we want an interpretable analysis by IRaMuTeQ, we have to remove from tweets all hashtags, Twitter account names and URLs. Those interventions on our corpus, specifically for IRaMuTeQ and with the constant concern regarding the ability to trace back any operations, and the fact that analyses can only be performed on original tweets (and not retweets), obviously change the analyses but are necessary in order to obtain meaningful results.

While the software that is mentioned here is so far relatively understandable by historians (though it requires a lot of time), will this be always the case? Nowadays, software that integrates algorithms based on machine and deep learning research - branches of artificial intelligence - is being used.[[52]](#footnote-53) This software uses algorithms which are “black boxes”: although the algorithms on which the software is based are perfectly well known, the reasons why it produces specific outputs are not, as yet, really understood.[[53]](#footnote-54)

Software is not the only thing that a digital historian should look at closely. Corpora from big data platforms such as Twitter should be understood as they are: as the product of a specific social, political and economic configuration. Although the discussions surrounding big data, as mentioned above, suggest that it can help a researcher access all the available data on a given subject, this premise is anything but true. Taking the example of the use of Canadian 19th-century press, Ian Milligan describes an “illusionary order”:[[54]](#footnote-55) the conviction that everything is online led Canadian historians, in their articles published in the *Canadian Historical Review*, to use mostly digitized newspapers. At the time of the writing of Milligan’s article, most major national English-speaking newspapers had been digitized but French-speaking newspapers and the myriads of small newspapers had not. The result, if we follow Ian Milligan, was that gigantic databases of newspapers available online gave the impression that it was possible to browse “all the available data”, when only a biased part of it was available. We must remain constantly critical of this “illusionary order” when practicing history with digital tools and databases.

# Conclusion

By using new born-digital primary sources, by distant reading them while keeping a multi-scaled approach through digital methods, the #ww1 project and its team hope they could contribute to the field of memory studies. These new primary sources can help us to understand commemoration practices and, in some cases - as in the example of the *Morts pour la France* database - everyday collaborative practices. A second, relatively novel, element is the analysis of the circulation of information (in a broad sense) and practices among citizens through their Twitter accounts. The example of the commemoration of the Battle of Verdun is paradigmatic: how disputed aspects of collective memory are debated and how they can lead to controversies, often very quickly, creating intense debates through a high-frequency propagation of information (“vibrations”).

More generally, working with vast born-digital archives that are constantly changing requires historians to be aware of the constraints regarding the harvesting of those archives, such as the way that computers will read them for us. It is therefore vital to understand the software used (and the algorithms it is based on), while not succumbing to the illusion that big data provides a complete picture. These rather new constraints to the historian’s work are not, in this case, laid bare to discourage the use of digital tools and born-digital archives. We are simply, as historians, at a stage where criticism of data and algorithms is a day-to-day necessity. While it goes without saying that critically assessing our primary sources has always been a concern, whether we practice digital history or not, we would argue that it is all the more necessary to cast a critical eye over our tools, data and methods today, as Peter Haber stated in 2011.[[55]](#footnote-56) What is sure is that, as future historians will probably use Twitter and other born-digital archives differently from the way we do, perhaps because they will be able to use a tweet archive, for example, such as the Library of Congress Twitter archive (all tweets 2006-2017, sampling since 2018, not yet accessible)[[56]](#footnote-57) or the French *Institut national de l’audiovisuel* archive on the Bataclan attacks (accessible on-site only),[[57]](#footnote-58) the necessity to critically assess born-digital sources and software will always remain very strong.

This digital and academic development has changed the allure of the archive, at least in our experience as digital migrants. Arlette Farge[[58]](#footnote-59) described the intimacy that historians establish with their sources, through reading but also by handling them and physically touching them when in an archive center. What happens when your main archive is a database? Some of us will increasingly find ourselves handling not documents but programming interfaces, establishing a kind of intimacy not with a physical object but with information feeds and passing from one modelized social aspect of a person or institution (a Twitter account) to another.

What the #ww1 experience seems to demonstrate is that in this digital era that has gradually emerged since the 1950s, historians are still, at present, in a stage of development, a time when they are able to take potentially disruptive methods from computer sciences and adapt them to their practice of history: it is bricolage in the sense of Levi-Strauss.[[59]](#footnote-60)

1. “Ma chérie, je t’écris pour te dire que je ne reviendrai pas de la guerre”. Tweet, 11/11/2014. Translation by the author. The tweet is today unavailable. [↑](#footnote-ref-2)
2. Statista, *Global social media ranking 2018 | Statistic*, webpage, <<https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/>>. [20.12.2018] [↑](#footnote-ref-3)
3. For more information on this research: Clavert, Frédéric. “Face au passé : la Grande Guerre sur Twitter”. *Le Temps des médias. Revue d’histoire*, nᵒ 31 (décembre 2018): pp. 173‑186. [↑](#footnote-ref-4)
4. We will not write about the ethical and legal aspects regarding the use of big data and personal data in this article. This topic deserves a full article on its own. [↑](#footnote-ref-5)
5. Others have also discussed the relevance of using big data instead of more traditional techniques like sampling: Rob Kitchin, *The Data Revolution: Big Data, Open Data, Data Infrastructures and Their Consequences*, SAGE Publicatinos Ltd.: Thousand Oaks, CA, 2014, 240 p. Though we will not examine the sampling/big data debate in this article, it is a worthy topic. [↑](#footnote-ref-6)
6. <http://140dev.com/>. [↑](#footnote-ref-7)
7. See the website of the ASAP (Archives sauvegarde Attentats Paris) project: <https://asap.hypotheses.org/>. The key in this project was the cooperation between archivists, librarians and historians. [↑](#footnote-ref-8)
8. There are numerous examples of data harvesting initiatives related to the covid crisis. For instance, the author’s research center launched a tweet harvesting initiative, an oral history project and a platform to collect other kind of digital artefacts (texts, photographies, videos, comics, etc). See *History in the making: #covidmemory* (<https://www.c2dh.uni.lu/projects/history-making-covidmemory>) and *#covid19fr – a confined country on Twitter* (<https://www.c2dh.uni.lu/projects/covid19fr-confined-country-twitter>). [↑](#footnote-ref-9)
9. We are now using DMI-TCAT: E. Borra and B. Rieder, “Programmed method: developing a toolset for capturing and analysing tweets”, in: Aslib Journal of Information Management, Vol. 66 Iss: 3, 2014, pp.262 - 278. <http://dx.doi.org/10.1108/AJIM-09-2013-0094>. Code available on <https://github.com/digitalmethodsinitiative/dmi-tcat>. [↑](#footnote-ref-10)
10. Chris Anderson, “The End of Theory: The Data Deluge Makes the Scientific Method Obsolete”, in: WIRED, June 2008. Online: https://www.wired.com/2008/06/pb-theory/. [↑](#footnote-ref-11)
11. Viktor Mayer-Schönberger and Kenneth Cukier, *Big data: A revolution that will transform how we live, work, and think*, Boston: Houghton Mifflin Harcourt, 2013, 256 p. [↑](#footnote-ref-12)
12. danah boyd and Kate Crawford, “CRITICAL QUESTIONS FOR BIG DATA: Provocations for a cultural, technological, and scholarly phenomenon”, in: Information, Communication & Society 15 (2012), 5, pp. 662–679. [↑](#footnote-ref-13)
13. Robert Kitchin, *The Data Revolution... op. cit.*,p. 133. [↑](#footnote-ref-14)
14. Evelien D’heer et al., “What are we missing? An empirical exploration in the structural biases of hashtag-based sampling on Twitter”, in: First Monday 22 (2017), 2. Online: https://firstmonday.org/ojs/index.php/fm/article/view/6353/5758. [↑](#footnote-ref-15)
15. This article intentionaly does not adress the long term archiving of digital media, that would require alone a complete issue of *Geschichte und Gesellschaft*. [↑](#footnote-ref-16)
16. Franco Moretti, *Graphs, Maps, Trees: Abstract Models for Literary History,* London, New York: Verso, 2007, 119 p. [↑](#footnote-ref-17)
17. Franco Moretti, *Distant Reading*, London, New York: Verso, 2013, kindle edition, here kindle location 796. [↑](#footnote-ref-18)
18. *Idem*. [↑](#footnote-ref-19)
19. Maurizio Ascari, “The Dangers of Distant Reading: Reassessing Moretti’s Approach to Literary Genres”, in: Genre, Vol. 47, Issue 1, 2014, pp. 1–19. [↑](#footnote-ref-20)
20. See Timothy Brennan, “The Digital-Humanities Bust: After a decade of investment and hype, what has the field accomplished? Not much”, in: *Chronicle of Higher Education, vol.* 64, issue 8, 2017. Online: https://www.chronicle.com/article/the-digital-humanities-bust/. Other criticisms have been addressed to distant reading, including in the *New Left Review* where Franco Moretti regularly publishes. A good summary can be read here: Rachel Serlen, “The Distant Future? Reading Franco Moretti”, in: *Literature Compass*, volume 7, issue 3, 2010, pp. 214-225. [↑](#footnote-ref-21)
21. D. Sculley and Bradley M. Pasanek, “Meaning and mining: The impact of implicit assumptions in data mining for the humanities”, in: *Literary and Linguistic Computing,* Volume23, issue 4, 2008, pp. 409-424. [↑](#footnote-ref-22)
22. We also collected tweets using certain German keywords, but the rather low numbers of German tweets did not inspire us to further develop our research in this language. [↑](#footnote-ref-23)
23. Saletes, Jean-Loup. “Les tirailleurs sénégalais dans la Grande Guerre et la codification d’un racisme ordinaire”. *Guerres mondiales et conflits contemporains,* n° 244, 2011, pp. 129-40. [↑](#footnote-ref-24)
24. Jay Winter, Antoine Prost. *The Great War in History: Debates and Controversies, 1914 to the Present,* Cambridge: Cambridge University Press, 2005, 250 p. [↑](#footnote-ref-25)
25. Ministère des Armées, “Base de données des Morts pour la France”, *Mémoire des Hommes*, online: <https://www.memoiredeshommes.sga.defense.gouv.fr/fr/article.php?larub=24>. [↑](#footnote-ref-26)
26. For more information on this initiative: Jean-Michel Gilot et al., “1914-1918 : quand la commémoration devient participative”, in: *Le Temps des medias,* issue 31, 2018, pp. 219–229. The “#1j1p” initiative has a website: “Le défi 1 jour - 1 Poilu”, <https://www.1jour1poilu.com/>. [↑](#footnote-ref-27)
27. Dominique Boullier, “Les sciences sociales face aux traces du big data”, in: *Revue française de science politique,* volume 65, issue 5-6, 2015, pp. 805-828. An english version is available: “Big data challenges for the social sciences: from society and opinion to replications”, 2016, 40 p., online: https://arxiv.org/abs/1607.05034. [↑](#footnote-ref-28)
28. For a detailed analysis of the two Verdun controversies: Clavert, Frédéric. “Commémorations, scandale et circulation de l’information: le Centenaire de la bataille de Verdun sur Twitter”, in: *French Journal For Media Research,* issue10, 2018. Online: http://frenchjournalformediaresearch.com/lodel-1.0/main/index.php?id=1620. [↑](#footnote-ref-29)
29. For more on temporalities in our #ww1 corpus, see Clavert, Frédéric. « Temporalités du Centenaire de la Grande Guerre sur Twitter ». In Valérie Schafer (dir.), *Temps et temporalités du web,* Nanterre: Presses universitaires de Paris Nanterre, 2018, 318 p., here pp. 113‑134 [↑](#footnote-ref-30)
30. Hartmut Rosa, *Beschleunigung: die Veränderung der Zeitstrukturen in der Moderne,* Frankfurt am Main: Suhrkamp, 2005, 537 p. [↑](#footnote-ref-31)
31. François Hartog, *Régimes d’historicité. Présentisme et expériences du temps*, Paris : Le Seuil, 2003, 272 p. [↑](#footnote-ref-32)
32. See Hans Ulrich Gumbrecht, *Our Broad Present: Time and Contemporary Culture,* New York: Columbia University Press, 2014, 144 p. [↑](#footnote-ref-33)
33. Assmann, J., Assmann, Jan. *Cultural Memory and Early Civilization*. Cambridge ; New York: Cambridge University Press, 2012, 424 p., p. 9, quoted by Tamm, Marek. « Beyond history and Memory: new Perspectives in Memory Studies ». History Compass 11, nᵒ 6 (2013): 458‑73. [↑](#footnote-ref-34)
34. About “updatism”, Mateus Pereira, Valdei Araujo, “Updatism: Gumbrecht’s broad present, Hartog’s Presentism and beyond”, in *Diacronie. Studi di Storia Contemporanea*, issue 43, 2020(3), online: https://www.studistorici.com/2020/10/29/pereira-araujo\_numero\_43/. [↑](#footnote-ref-35)
35. Oceanic exchanges (<https://oceanicexchanges.org/>), Numapresse (<http://www.numapresse.org/>) or the impresso project (<https://impresso-project.ch/>) are some examples. [↑](#footnote-ref-36)
36. Arlette Farge, Thomas Scott-Railton (translation), *The allure of the archives*, New Haven: Yale University Pres, 2015, 131 p. [↑](#footnote-ref-37)
37. Frédéric Clavert, Martin grandjean, Cécile Méadel, “Le temps long des réseaux sociaux numériques, une introduction”, in: *Le Temps des médias,* issue 32, 2018, p. 6-11. [↑](#footnote-ref-38)
38. Thomas H. Cormen et al., Introduction to Algorithms, Cambridge, USA: The MIT Press, 2009, 1292 p., p. 5. [↑](#footnote-ref-39)
39. Bastian M et al., “Gephi: an open source software for exploring and manipulating networks”, *International AAAI Conference on Weblogs and Social Media,* 2009. Online: <https://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>. Software available at <http://www.gephi.org/>. [↑](#footnote-ref-40)
40. Pierre Ratinaud and S. Dejean, “IRaMuTeQ : Implémentation de la méthode ALCESTE d’analyse de texte dans un logiciel libre”, *Modélisation Appliquée aux Sciences Humaines et Sociales*, Toulouse, 2009, pp. 89. Software available at <http://www.iramuteq.org/>. [↑](#footnote-ref-41)
41. For instance, a search on Google Scholar for articles quoting Bastian M et al., “Gephi...”, op. cit., in 2020 returns around 1200 results. [↑](#footnote-ref-42)
42. See Madeleine Akrich, Michel Callon, Bruno Latour, *Sociologie de la traduction: textes fondateurs*, Paris : Presses des Mines, 2006, 304 pages. [↑](#footnote-ref-43)
43. <http://www.e-diasporas.fr/>. [↑](#footnote-ref-44)
44. See Pierre Bourdieu, La distinction: Critique sociale du jugement, Paris 1979 (Le Sens commun). [↑](#footnote-ref-45)
45. See Max Reinert, Une méthode de classification descendante hiérarchique: Application à l’analyse lexicale par contexte, in: Les cahiers de l’analyse des données 8 (1983), 2, pp. 187–198, here pp and, for the implementation into IRamuTeQ: Ratinaud, Dejean, IRaMuTeQ (unpublished). [↑](#footnote-ref-46)
46. Shawn Graham, Scott Weingart, Ian Milligan. “Getting Started with Topic Modeling and MALLET”, in: *Programming Historian*, 2 septembre 2012. Online: <https://programminghistorian.org/en/lessons/topic-modeling-and-mallet>. [↑](#footnote-ref-47)
47. Christof Schöch, “Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama”, *Digital Humanities Quarterly,* volume11, issue 2, 2017. [↑](#footnote-ref-48)
48. Nathan C. Lindstedt, “Structural Topic Modeling For Social Scientists: A Brief Case Study with Social Movement Studies Literature, 2005-2017”, *Social Currents, volume* 6, issue 4, 2019, pp. 307-318. Online: https://doi.org/10.1177/2329496519846505. [↑](#footnote-ref-49)
49. Graham, Shawn, Ian Milligan, et Scott Weingart. Exploring Big Historical Data: The Historian’s Macroscope. London: Imperial College Press, 2016, 282 p., here p. 113ff. [↑](#footnote-ref-50)
50. Ibid p. 157 [↑](#footnote-ref-51)
51. Andrew Kachites McCallum, “MALLET: A Machine Learning for Language Toolkit”, http://mallet.cs.umass.edu, 2002. A comparison between different topic modeling tools is done in: Mark BelfordBrian Mac Namee, et Derek Greene. « Stability of Topic Modeling via Matrix Factorization ». *Expert Systems with Applications* 91 (1 janvier 2018): 159‑69. <https://doi.org/10.1016/j.eswa.2017.08.047>. [↑](#footnote-ref-52)
52. Melvin Wevers et al., “Seeing history: Analyzing large-scale historical visual datasets using deep neural networks”, *DH Benelux 2018 abstracts,* 2018, online: [https://pure.knaw.nl/portal/en/publications/seeing-history(5eb92738-4f9a-43c1-888e-65888ce73238)/export.html](https://pure.knaw.nl/portal/en/publications/seeing-history%285eb92738-4f9a-43c1-888e-65888ce73238%29/export.html) [↑](#footnote-ref-53)
53. For some thoughts on deep learning: Gary Marcus, *Deep Learning: A Critical Appraisal*, arXiv:1801.00631, 2018, 27p. [↑](#footnote-ref-54)
54. Ian Milligan, “Illusionary Order: Online Databases, Optical Character Recognition, and Canadian History”, in: *Canadian Historical Review,* volume 94, issue 4, 2013, pp. 540-569. [↑](#footnote-ref-55)
55. Peter Haber, Digital Past. Geschichtswissenschaften im digitalen Zeitalter, München: Oldenbourg Wissenschaftsverlag, 2011, 184 p. [↑](#footnote-ref-56)
56. Library of Congress, White Paper: Update on the Twitter Archive at the Library of Congress, 12/2017, <https://blogs.loc.gov/loc/files/2017/12/2017dec_twitter_white-paper.pdf>. [↑](#footnote-ref-57)
57. An entire issue of *La Gazette des archives* is dedicated to the archiving of reactions to the attacks in France: *La Gazette des Archives*, issue 250, 2018. [↑](#footnote-ref-58)
58. See Arlette Farge, *The allure... op. cit.* To answer this question, the author of this article is co-leading a project on the allure of the archive in the digital era, available at https://www.gout-numerique.net/. [↑](#footnote-ref-59)
59. Claude Levi-Strauss, *La pensée sauvage*, Paris: Plon, 1962, 347 p. [↑](#footnote-ref-60)