MDR, Vol. 41, pp. 121–125, December 2012 • Copyright © by Walter de Gruyter • Berlin • Boston. DOI 10.1515/mir-2012-0019

Using DiscoverText for Large Scale Twitter Harvesting

Yngvil Beyer

Yngvil Beyer (Yngvil.Beyer@nb.no) is Ph.D. fellow at the National Library of Norway and University of Oslo, Norway.

Introduction

How can institutions like the National Library of Norway preserve new media content like Twitter for future research and documentation? This question is the topic this short paper, where I am presenting an ongoing research project, focusing on the data collection and some preliminary findings.

The research question builds on two important underlying conditions. Firstly, preserving the nation's national memory is explicitly mentioned in the National Library's mission statement.

> The National Library of Norway is responsible for administering the Act relating to the legal deposit of generally available documents. The purpose of this act is to ensure that documents with generally available information are deposited and made accessible to national collections, thus preserving for posterity a written testimony of Norway's cultural heritage and Norwegian society as well as providing material for original research and documentation. (The National Library of Norway [s.a.])

Secondly, Twitter has, during the last few years turned into an important channel of communication. My research builds on the assumption that at least some of the communication activity on Twitter is a relevant part of our culture, and thus should be regarded cultural heritage. Preserved tweets might provide an insight into our culture for future generations. A recent example of this is when Prime Minister Jens Stoltenberg addressed the Norwegian people[1] and the international community[2] in the evening after the terrorist attack against Norway last summer, and the extensive use of Twitter both among the offended during the terrorist attack and by media and the general public, discussed thoroughly in my forthcoming article «@jensstoltenberg talte til oss på Twitter»[3] (Beyer 2012).

Then, returning to my initial question concerning how such preservation might be carried out. In early 2010 the US Library of Congress announced that they would archive all tweets ever tweeted (Library of Congress 2010). The details about this archive are yet to be revealed. The infrastructure necessary to give access to the collection is yet not in place. Neither is it specified to what extent the collection will be made available, to whom, in which format, under what conditions or including which data.

In addition to this announced archive, several commercial companies, such as Datasift, GNIP and DiscoverText offer their customers access to tweets in order to perform analyses of social media for business purposes. For academic research there are several available tools. yourTwapperKeeper seems to be one of the leading ones, used and described by Larsson and Moe in their recently published article «Studying political microblogging: Twitter users in the 2010 Swedish election campaign» (2011) and by Bruns and Burgess in their «Researching news discussion on Twitter» (2012).

From knowing about this possible access to archived new media objects, I was curious about how to build such an archive, and to using the data to make a collection of new media objects related to a relevant issue. What kind of data would I get, how could I best get them, what obstacles would I meet, and could such data be preserved in a way that makes them valuable for future research on our time?

My ongoing Ph.D. project[4] deals with the making of new media archives. I am exploring the context of new media objects, including judicial issues and the blur of the private and public, as well as technical and practical aspects of establishing such new archival practices. The project is designed as a case study focusing on the terrorist attack against Norway last summer, an important event in Norwegian history, which also turned into a major media event.

Hence, my investigation of Twitter relates to this event. More specifically, in the spring of 2012 I set out to archive the tweets with relation to the trial against the terrorist.

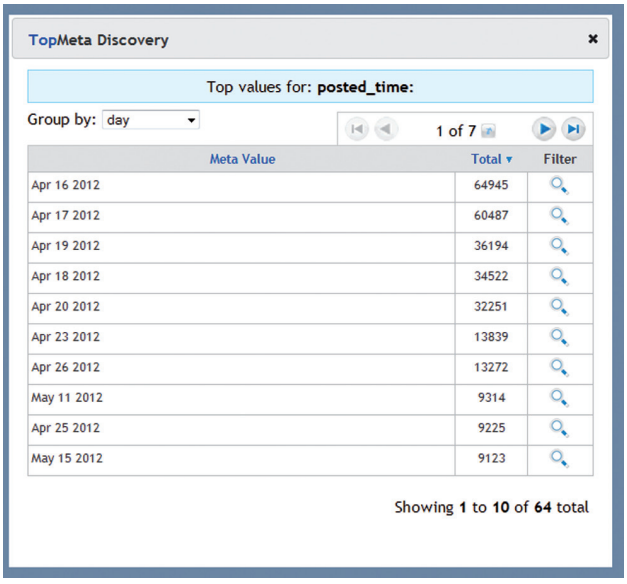
Archiving tweets

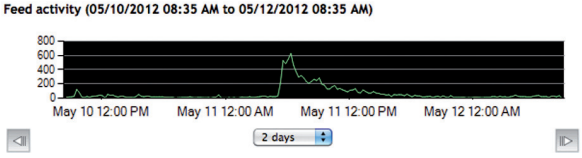
I chose to use a tool and a service provided by DiscoverText[5]. In addition to facilitating for capturing data from social media, the service provides well integrated analytic tools for exploring and coding the data, and the data are easily exportable as text files. At the time I started testing the tool, it was offered as a free beta account including a special service called a GNIP-enabled Power Track for Twitter. Whereas the general Twitter API is letting people harvest tweets from two weeks back, but with strong limitations regarding the amount, regulating the frequency of fetching the data as well as the amount of tweets in each fetch, the Power Track is a so called full fire hose, collecting *all* tweets with the keyword specified by the user, in real-time. Additionally, the Power Track assures a wider range of metadata for the collected tweets.

The weekend before the trial started, I set up 17 feeds. The selection was based on searches exploring activity on Twitter in the period before the trial. I chose to have feeds both filtering by keywords and by accounts. Some of the keywords I chose are misspellings of the word *trial*, which seems to be difficult to spell correctly for many Norwegians. One of the keywords, «Breivik», the name of the terrorist, was set up with an additional feed where I specified the language to be Norwegian. For «Utøya», the name of the island where the massacre was carried out, I added «Utoya» as an additional feed, given that ø is a special Norwegian character. I searched through the Twitter accounts of the major Norwegian media houses, and tried to choose accounts that were used for news reporting issues. ABB, the initials of the name of the terrorist was set up as a feed, the only one in my sample being an abbreviation. Finally, I set up a feed for the Prime Minister Jens Stoltenberg's account, and the account belonging to The Norwegian Labour Party's youth organization, AUF, which was the organization attacked at Utøya last summer.

Exploring the archive

The tweets started ticking into my archive immediately after I started the feeds. From the start I was able to conduct preliminary analyses of my data, both concerning the amount of accumulated tweets, and the content. By selecting the *TopMeta Discovery*, I could for instance see which days that have the most activity, revealing that the activity from the first week dominates the archive. Apart from being the first week of an event that spurred immense interest, this was the week for the terrorist's testimony. We might also take a look at how an incident such as a shoe thrown towards Breivik was reflected almost immediately on the Twitter activity, and thus pointing to Twitter as a hyper-reactive medium, almost hysterical in its mediation of certain events.





Furthermore, we may for instance explore which users are the most active ones, the hashtags used most frequently, or the hyperlinks included in the tweets.

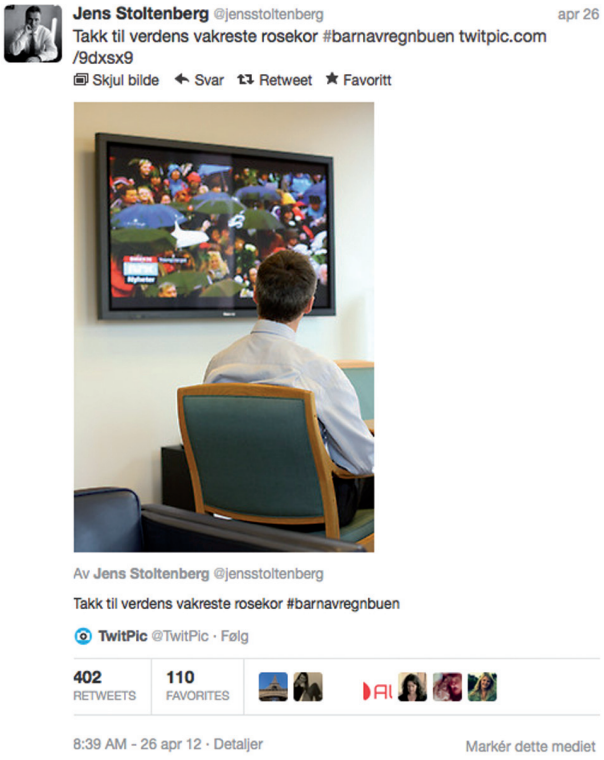
However, there are several challenges as well related to the collection of this material. Let's first take a look at how the selected feeds worked out. The feed selecting all tweets containing the word «Breivik» has been the most numerous throughout the whole trail. After ten weeks of collecting, at the end of the trial, this feed counted 435.641 tweets. The feed where I tried to filter tweets containing «Brevik» written in Norwegian is about 10% the size. However, the language filter does not seem to work, it contains tweets in several languages. The reason for this is yet to be found. The choice to harvest tweets using both correct and incorrect spelling of the word trial in definite and indefinite form[6] seems to have been a good one, given that the incorrect forms contributes with between approximately 2/3 the amount of the correct spelled feeds, and thus augments the collected data substantially. The most problematic feed turned out to be the one selecting tweets containing «abb». Not only is this also the name of an international company[7], which leads to lots of tweets regarding this company, it also turned out that «abb» forms part of several words, other abbreviations and even Twitter user names, making it a lot of work separating the tweets pertaining to the case I am working on, and the tweets that does not have a relation to the Breivik-case.

Dealing with a tool (provider)

Another problem that arose during the collection of data was a more technical one. A couple of days into the second week of the trial, I noticed that there were no accumulated tweets for quite some time. How come? I e-mailed the service support, and after some e-mailing back and forth, an answer pointing to downtime and problems with the code was given[8].

The same thing happened once again a couple of days later, and made the support assure me that this issue would be given top priority. However, a substantial amount of tweets had been lost during the stop. The service is set up as a real-time harvest, which means that tweets are collected immediately after being tweeted by someone. This time I was following Twitter live as well, due to an ongoing event related to the trail – the gathering of forty thousand people singing the Norwegian version of Peter Seeger's *My Rainbow Race* as a reaction to Breivik's uttering that this song was used to brainwash Norwegian children with Marxist ideology.

Jens Stoltenberg's reaction to the event (Stoltenberg 2012) was among the «lost» tweets. I decided to set up an additional feed based on a widely spread hashtag related to this event, #barnavregnbuen[9], that was also used by Stoltenberg.



For this feed I chose to collect the tweets with a service based on the regular API, which is able to include historical tweets. In the context of Twitter, this means tweets tweeted one to two weeks back in time. This feed would give me the tweet from Stoltenberg, and I would also be able to investigate the proliferation of a hashtag. From the stats of this collection, made by selecting tweets using the above mentioned hashtag or the name of the song («barn av regnbuen»), I could see how information about the event spread the days before the happening, and we might also perform a distant reading of the collection building a so-called Tag Cloud.

Yet, not only the tweet by the Prime Minister was lost during the GNIP feed stop, but thousands of other tweets as well. I had been informed that GNIP had launched a so-called 30-days replay in February, and that DiscoverText was planning to include this service in their tool. The 30-days replay facilitates for «replaying history», to quote the GNIP press release (Johnson 2012). On May 11th my feeds stopped for the third time, again it was fixed, but this time I pushed the support harder on how to recover my lost data, and asked them about the replay – to which they replied that this would take some time, given to be, «a whole different beast than the real-time streaming». Our correspondence continued, and I urged them to do something, but it seemed out of reach for the time being[10].

Would it help offering to pay extra for the replay, I asked, and even though this was not the problem, little by little it seemed that they took the issue more seriously[11].

I e-mailed the CEO of the company in order to put some more pressure on the case, and then, a week later it was all fixed, just in time to replay all my lost data[12].

Twitter, and social media more generally, obviously differs from traditional media in many ways. My investigations into the archiving of this material, however, makes me aware, like the negotiation summarized above shows, that the archiving of them is not so much connected to how the various media *are* as to how they are dealt with. Developers, tools, researchers like I, APIs, various restrictions, licenses, and so on are all parts in the process of archiving the tweets.

Data/metadata

Let's now take a glimpse into the archived material and ask: What is the data and what is the metadata for tweets? Tweets are limited to include a maximum of 140 characters. However, as we have seen earlier, apart from these characters, each tweet contains a substantial amount of metadata, such as the sender and the time it was posted. The DiscoverText tool extracts information from the Twitter accounts as well, adding these data to each of the archived tweets. Examples of such metadata is the url of the twitter profile, the count of followers and those being followed, the users estimated influence in the realm of social media, that is the so called Klout score, and if shortened urls are included in the tweets, these are expanded into their original format.

The act of re-tweeting is widespread on Twitter. In the early days of Twitter, this was done by starting a tweet with RT and then the «@» and the username of the original sender of the tweet. More recently, this functionality has been implemented in the service as a button. This means that while in the beginning this activity was part of the *data*, and hence the user spent some of the 140 characters on the semantic code defining this act, this is no longer the case, re-tweeting is now made part of the metadata. However, in DiscoverText, the information of the re-tweeting is still considered part of the data. Let me show an example from my material.

User A re-tweeted user B. In the archived version of the re-tweet the text is «RT @RagHolmas: Breivik: I learnt how to fly a plane on YouTube. I'm good at technical things. Prosecutor: In July you needed help to rev ...». Because the RT and the username of user B is included in the text, the same amount of characters is cut off at the end of the tweet, and in this case the tweet loses its point. Consulting the Twitter account of user A, we can read the whole tweet as she published it: «Breivik: I learnt how to fly a plane on YouTube. I'm good at technical things. Prosecutor: In July you needed help to reverse your own car.» Then the tweet makes sense. This consultation of the live account could not have been done if this was research performed on this archive in the future.

Conclusion

The analysis of the material is still in the first phase. In the continuation I will problematize aspects such as considerate selection in order to avoid messy data such as the ABB-feed, and I will look into the practice, possibilities and implications of distant reading as a means to research this kind of mediations.

Then more work will have to be done both theoretically and methodologically in order to answer my

research question concerning how institutions like the National Library of Norway can preserve new media content like Twitter for future research and documentation. Inspiration will be drawn from the recent publications of new media researchers such as David Berry (2011, 2012), Wendy Hui Kyong Chun (2011) and Lev Manovich (2001, 2012), among others. Additionally, perspectives from ANT and Bruno Latour (2005) will be explored in order to see how they might shed light upon the networks that these new media objects all are parts of.

Notes

1 «Today, we have been hit by two savage and cowardly attacks. Tonight we all stand together, taking care of each other.» (Stoltenberg 2011b)

2 «Thank you for the solidarity with the Norwegian people. The support from the international community is a great strength.» (Stoltenberg 2011a)

3 «@jensstoltenberg talked to us on Twitter»

4 A case study within the research project *The Archive in Motion* (AiM) http://www.hf.uio.no/ifikk/english/research/projects/archive-in-motion/. AiM investigates the ways in which archival concepts and practices have been transformed under the impact of the radical changes in writing and recording technologies that have taken place over the last century, and particularly with the introduction of digital technologies.

5 http://discovertext.com/

6 In Norwegian: «rettssaken», «rettssak» (correct spelling) and «rettsaken», «rettsak» (incorrect spelling)

7 http://www.abb.com/

8 «Back up and running now... give it about 15 minutes before the statistics get refreshed on the site. Sometimes GNIP has to close the feed for maintenance, and for whatever reason - it seems like 25% of the time they do this, our code can't catch the stream being closed, so, it just waits... and waits... and waits...» (Personal communication with support, April 24).

9 #barnavregnbuen is a hashtag derived from the Norwegian title of the song My Rainbow Race, Barn av regnbuen, recreated by the Norwegian singer-songwriter Lillebjørn Nilsen in 1973.

10 «we probably won't be able to have that in place until the end of May to mid-June at the moment, as it does require completely separate code to process and ingest the items as well as additional costs on our end for the additional connector... we'll try and get this up as soon as we can» (support, May 13)

11 «I'll see what sort of schedule this next week holds and try and push for this to get in place within the next week or 2 at the most (hopefully within before the 22nd so we can get back to the April 23rd data) - no guarantees, but we'll try.» (support, May 13)

12 «Yngvil - Just to keep you in the loop on the development of the 30-day replay functionality in DiscoverText - we are entering the final phases of development and will be testing and debugging shortly - there is a chance this will go up tonight, but more than likely tomorrow night (Monday night). Either way, we are certainly going to be up and running before our cut-off date of the 22nd to ensure we are able to grab the data from April 24th.» (support, May 21)

References

Berry, David M. 2011. *Philosophy of Software: Code and Mediation in the Digital Age*. Basingstroke: Palgrave Macmillan.

Berry, David M. 2012. *Understanding Digital Humanities*: Palgrave Macmillan.

Beyer, Yngvil. 2012. "@jensstoltenberg talte til oss på Twitter: Digitalt fødte objekter som kulturarv." In *Viden i Spil*, edited by H. Høyrup, B. Hjørland and H. J. Nielsen. Århus: Samfundslitteratur. Forthcoming.

Bruns, Axel, and Jean Burgess. 2012. "Researching news discussions on Twitter: New methodologies." *Journalism Studies*:1-14. doi: 10.1080/1461670x.2012.664428.

Chun, Wendy Hui Kyong. 2011. *Programmed visions : software and memory*. Cambridge, Mass.: MIT Press.

Johnson, Rob. 2012. "20/20 Hindsight: 30-Day Replay for Twitter Now Available". [Blog post] Published February 16, 2012, cited May 5, 2012. Available from http://blog.gnip.com/historical-twitter-data/.

Larsson, Anders Olof, and Hallvard Moe. 2011. "Studying political microblogging: Twitter users in the 2010 Swedish election campaign." *New Media & Society* (Published online before print: November 21, 2011). doi: 10.1177/1461444811422894.

Latour, Bruno. 2005. *Reassembling the social : an introduction to actor-network-theory*. Oxford: Oxford University Press.

Library of Congress. 2011. "Library to acquire ENTIRE Twitter archive -- ALL public tweets, ever, since March 2006! Details to follow.". [Tweet] Published April 14 at 1.36 PM, 2010, cited September 11, 2011. Available from https://twitter.com/#!/librarycongress/status/12169442690.

Manovich, Lev. 2001. *The language of new media*. Cambridge, Mass.: MIT Press.

Manovich, Lev. 2012. "Trending: The Promises and the Challenges of Big Social Data." In *Debates in the Digital Humanities*, edited by M. K. Gold. The University of Minnesota Press.

Stoltenberg, Jens. 2011. "Thank you for the solidarity with the Norwegian people. The support from the international community is a great strength.". [tweet] Published July 23 at 10.37 PM, 2011a, cited September 9, 2011. Available from http://twitter.com/#!/jensstoltenberg.

Stoltenberg, Jens. "Today, we have been hit by two savage and cowardly attacks. Tonight we all stand together, taking care of each other". Published July 23 at 1.10 AM, 2011b, cited September 9, 2011. Available from http://twitter.com/#!/jensstoltenberg.

Stoltenberg, Jens.2012. "Takk til verdens vakreste rosekor #barnavregnbuen http://twitpic.com/9dxsx9 ". [tweet] Published April 26 at 12:39 PM, 2012, cited May 15 2012. Available from http://twitter.com/#!/jensstoltenberg.

The National Library of Norway. 2012. "Mission statement". [Web page] Published, [s.a.], cited 03.01 2012. Available from http://www.nb.no/english/mission-statement.