**Capture and Preservation of Born Digital News**

*March 5-6*

Universidad Nacional Autónoma de México, México City

Down the rabbit hole: Collecting #metoo through social media and news articles

Pablo Morales Henry

Schlesinger Library, Radcliffe Institute, Harvard University, Cambridge, United States.

E-mail address: pablo\_moraleshenry@radcliffe.harvard.edu

**Jennifer Weintraub**

Schlesinger Library, Radcliffe Institute, Harvard University, Cambridge, United States.

E-mail address: Jennifer\_weintraub@radcliffe.harvard.edu

Copyright © 2020 by Pablo Morales Henry and Jennifer Weintraub. This work is made available under the terms of the Creative Commons Attribution 4.0 International License: <http://creativecommons.org/licenses/by/4.0>

Abstract:

*In July 2018, the Schlesinger Library on the History of Women in America launched a project to comprehensively document social media, news articles, press releases, judicial decisions, and other materials related to the #metoo hashtag that went viral on Twitter during the last quarter of 2017. This challenging project includes acquiring social media content published before the project started, sifting through thousands of online news articles to ensure there is an adequate representation of different industries and topics, reviewing terms of use and copyright issues around usage and distribution of materials, and deploying the appropriate infrastructure and applications to collect data. All this work ensures the material represents all voices around the issue to build a full historical record for researchers.*

*With the guidance of a steering committee composed of Harvard University faculty and librarians, a series of policies and procedures have been developed to strategize the best ways to approach the ethics of providing access to electronic news and social media. These policies cover appropriate platforms to distribute and provide access to the materials for researchers, while being mindful of the terms of use and copyrights surrounding the millions of items included across the dataset.*

*This presentation will engage the audience in the journey down the #metoo archival rabbit hole and will share lessons learned and best practices for documenting social movements on social media.*

**Keywords:** Social Media, Social Movements, metoo, Capture Born-digital.

# Put body of the paper here

**Introduction**

In October 2017, the hashtag #metoo ignited a global discussion around sexual harassment, assault, abuse, consent, and justice (or the lack thereof). This discussion, predominantly led by women, provided a place for victims of sexual misconduct to share their personal stories. Scores of stories were shared, exposing the pervasive nature of sexual misconduct across industries. What started on Twitter, quickly spread to other social media platforms, and resulted in thousands of news articles. In response, many companies have revised their internal policies, sponsorships have been withdrawn, and and in some states, news laws have been proposed as a response to #metoo.

Considering the preeminent role that women have played in the #metoo movement, the Schlesinger Library on the History of Women in America, decided that #metoo falls within the library’s collecting scope. The Library embarked on a project to comprehensively document the movement’s digital footprint.

**Background**

While #metoo became viral after actress Alyssa Milano shared a message with her Twitter followers, advocating the use of “Me Too” as part of their status on October 15th, 2017 (Alyssa Milano on Twitter: “If you’ve been sexually harassed or assaulted write ‘me too’ as a reply to this tweet. Https://t.co/k2oeCiUf9n” / Twitter, n.d.), the idea of using “Me Too” as a way of sharing experiences related to sexual abuse came originally from African-American activist Tarana Burke in 2006 (TIME Person of the Year 2017: The Silence Breakers | Time.com, n.d.). However, it was not until Milano’s tweet and the following accusations against Harvey Weinstein became public that #metoo became a global internet movement. Between Oct/15/2017 – Dec/31/2017 there were 4,416,766 tweets, identified with the hashtag #metoo (the number of tweets is higher when associated hashtags are included). It did not take long to see that #metoo was not a problem exclusive to Hollywood, and similar hashtags like #sciencetoo or #churchtoo were created to better communicate the problem of sexual abuse around specific industries and/or institutions, and eventually, organizations like the TIME’S UP Foundation were established to specifically fight problems like sexual harassment in the workplace (TIME’S UP Now | About, n.d.).

In order to learn from similar previous projects, we looked at the Catalonian Independence Referendum (Vicenç Ruiz Gómez, 2017), Cuéntalo (#Cuéntalo, n.d.) in 2008, and of course, Documenting Ferguson (Documenting Ferguson, n.d.) in 2014. While these projects collect social media, personal experiences, and in some cases news media, the scale and unknown timeline of #metoo presented new challenges related to collection scope, preservation, and collection access.

**#metoo Project at the Schlesinger Library**

In June 2018, the Schlesinger Library was awarded by Harvard Library an S.T. Lee Innovation Grant to initiate a large-scale collection and preservation project around the #metoo movement. The original project proposal stated: “The Schlesinger Library, in collaboration with Harvard faculty drawn from five schools (Radcliffe Institute for Advanced Study, Harvard Faculty of Arts and Sciences, Harvard Business School, Harvard Law School, and Harvard School of Public Health) will embark on a large-scale project to comprehensively document the #metoo movement and the accompanying redefinitions of gender-based harassment and sex in the workplace. Through the #metoo Digital Media Collection, Schlesinger Library will provide enduring scholarly access to content including social media conversations, press stories, and multimedia declarations and rebuttals: resources which are now pervasive in our collective consciousness and social media feeds, yet will prove acutely vulnerable in the long term, as proprietary platforms, individual user accounts, and the ever-changing landscape of the web continually transform.” (Innovation Grants, n.d.) In other words, in order to bring this project to fruition, seven items were required: funding, a steering committee, a collection policy, collecting infrastructure, preservation policy, preservation infrastructure, and delivery infrastructure.

The original grant funding provided enough money to hire a part-time assistant for the project, and to fund some of the initial infrastructure expenses, like the domain name and hosting for the project’s website. The steering committee was comprised of seven faculty members across the university, and eight library professionals, also from different schools at Harvard, but primarily from Schlesinger. With the guidance of the steering committee, decisions were made around the collecting scope, suggested language around the Ethics Statement prepared by the core team included in the project’s website (Ethics Statement, n.d.).

As an interdisciplinary project, the collecting scope for the project includes relevant social media platforms, news articles, statements of denial or apology issued in response to accusations, web-forum discussions, legislation originated as response to #metoo related issues, lawsuits documentation, employee/company manuals generated primarily by Fortune 500 companies, statistical studies, and other related materials and projects. Furthermore, because we wanted these materials to represent industries beyond entertainment, we used the North American Industry Classification System (NAICS & SIC Identification Tools, n.d.) codes to define how to approach the search of materials.

As for the project’s date-frame, since #metoo is an extremely fluid event, it is difficult to determine a start and end date. A starting point is perhaps simple, as there is a triggering event that can be identified, but there is no end in sight. In fact, at the moment of this writing, the trial to Harvey Weinstein (one of the initial accusations that served to catalyze the movement) will conclude in a matter of days (Feuer, 2020), which has only helped to revitalize a movement that has only decreased slightly in its activity, but certainly not dissipated, as Figure 1 demonstrates. Additionally, changes around company policies or new laws are not immediate, which provides the project with additional uncertainty in terms of an end date, as these items are part of our collecting scope.



Fig. 1: #metoo interest in news search from October 15, 2027 – January 31, 2020.

In terms of the regional source and language of the materials, while the collecting policy of the Schlesinger Library is limited to the United States of America, it was decided not to limit, at least social media materials, by location or language, as actors involved in the development of these issues could easily move around the globe and express their ideas from different locations. Additionally, because of the Facebook-Cambridge Analytics scandal, social media users’ awareness regarding their privacy and what they share online has changed, providing fewer identifying parameters. (Beck, 2018) As for language, the United States does not have an official language, thus limiting our collecting to English only would impoverish the historical record this project is trying to build.

As for content copyright, in previous projects, an effort was made to notify content creators before crawling their websites if authorization was not already available. After assessing how impractical it would be to attempt a similar approach, and consulting with our library copyright expert, it was decided to proceed with the crawling of websites, as there would be a transformative implementation of the material for researchers, and the potential profit received by each site should not be affected.

Once the collecting policy and scope was defined, it was necessary to establish what existing tools could be used to collect these materials, how they could be integrated into a larger collection workflow, and what type of infrastructure was required to deploy these applications. Upon initial exploration, experimentation, and suggestions coming from the steering committee, different tools were selected for different areas, as there is not yet a suite of applications encompassing all the projects requirements.

Social Media

To collect social media, and particularly Twitter, it was clear that reviewing the incredible work done by the team behind Documenting the Now (https://www.docnow.io/) was the best option. Based on the set of tools made available for the community and other suggested applications, the team decided to use Social Feed Manager(Media Cloud, n.d.) (https://gwu-libraries.github.io/sfm-ui/), based on system requirements, ease of configuration, and friendliness of its interface, the application proved to be the best choice for our needs.

News Articles

To collect news articles, members of our steering committee suggesting using Media Cloud, a joint project between MIT’s Center for Civic Media and Harvard’s Berkman Klein Center for Internet and Society (Media Cloud, n.d.). Using Media Cloud turned out to be an excellent choice due to the possibilities of content the application offers, in this case selecting content at a national and regional level, but also due to our affiliation with Harvard University, it was not necessary for us to setup an instance of the application.

Other Online Content

To collect online content, most capture could be automated using Archive it (https://archive-it.org/), a subscription service developed by the Internet Archive to crawl, capture, and preserve websites. Since Harvard University already has a subscription to this service, and Schlesinger staff already had experience using the tool, it appeared to be a logical choice as it was easy to feed lists of URLs to the application and schedule the capture of the sites. However, because due to the dynamic nature of some of the content found online, not all content is easily capture through Archive-it, Webrecorder (https://webrecorder.io), another well-known tool to preserve online content, was also used.

Since most of the infrastructure to collect data for this project was already available to us, the only requirement was to setup Social Feed Manager on a dedicated workstation using a Linux virtual machine, as we initially encounter issues during our testing/experimentation stage using the application directly on the Windows operating system. Once the application was running, backups were guaranteed, and the corresponding credentials were configured, Social Feed Manager was scheduled to collect Twitter data on a weekly basis.

As for a preservation policy and infrastructure, we are fortunate to use our own Digital Repository Service at Harvard University, a system that ensures long-term preservation of the objects, and that already has adequate preservation policies in place. (Digital Repository Service (DRS)—Harvard Library Digital Preservation—Harvard Wiki, n.d.)

Finally, when it comes to a content delivery infrastructure, there are two venues. One is the project website at https://www.schlesinger-metooproject-radcliffe.org/, which contains information about the project and links to the Archive-it site in which the capture websites reside. The other venue is Harvard’s Dataverse (https://dataverse.harvard.edu/), which is an excellent tool to provide access to datasets, and where the Twitter data which we can make available to users for research will be available.

**Collecting #metoo**

The project started just as consequences of the Facebook-Cambridge Analytica data scandal (“Facebook–Cambridge Analytica data scandal,” 2020) were starting to emerge, which for our purposes, meant changes in policies for data access and usage, and of course, further scrutiny to projects planning to collect data categorized as potentially sensitive. These changes delayed data collection for at least three months.

Despite the aforementioned delay, there was still a much larger backlog of Tweets which we were interested in acquiring. While Twitter is currently planning (More on restricted use cases – Twitter Developers, n.d.) to create a platform to provide access to Twitter data for academic/research purposes, the lack of such platform and the difficulties presented to contact a sales representative, made obtaining a quote on the required data rather difficult. The original values presented through the automated responses were prohibitive for any academic institution. It was not until August 2019, after months of emails and further research through our steering committee network, that we were able to negotiate licensing the content of the twitter data for a calendar year.

Because the delay we experience through our own collection using the free tier for Twitter data, the backlog licensed included 24,289,671 tweets from October 15, 2017 to December 15, 2018. Through the free tier, as represented in Figure 1, we have collected 4,572,921 tweets from December 2018 until January 31, 2020. However, our collection through the free tier also included hashtags such as #timesup, #churchtoo, and #gamani. As for other hashtags, as long as they include “metoo” as part of the string, they are also collected. E.g. #metoomedicine, #metooliars, and #mcdonalsmetoo.



Fig. 2: Twitter data collection. January 2019- January 2020 including all hashtags

As for collecting news articles, two minor issues were recognized: regional outlets mostly repeated content produced by news agencies, and with the exception of the news cycle around the confirmation of Brett Kavanaugh to the Supreme Court of the United States, much of the content was focused on the entertainment industry. In order to decrease the impact of these issues on the end result, the number of news outlets dedicated exclusively to the entertainment industry was reduced to a minimum, and only a sample was submitted for collection by Archive-it from the original lists of URLs provided by Media Cloud.

Besides the automated processes above, there was a fair amount of manual collecting of data, particularly when it comes to blogs, videos, portals, documents, and other smaller sites providing valuable data which is not necessarily found in social media or news articles.

**Accessing #metoo data**

Some of the data collected for this project is already available online (https://archive-it.org/collections/10866), mainly data which is not dependent on terms of agreement or licensed. Providing access to Twitter data has provided many challenges, not only because Twitter sets limits to the number of tweets that can be shared within a 30 day period (Twitter Terms of Service, n.d.), but also because the preferred method of distribution is through tweet IDs, meaning that researchers are responsible for rehydrating the tweets, in other words, query Twitter and retrieve the tweet contents if they are still available. Since distributing data identified only by tweets IDs is such a poor method of providing information, the team is currently researching how to associate metadata to the tweet IDs or at least group tweets by common elements.

Another minor inconvenience has been the fact that the structure of the tweets provided through Twitter’s historical track and the tweets collected through Social Feed Manager is slightly different. While both can be easily represented as JSON data, the key-value representation is different enough that requires different parsers.

It is expected that Twitter data will be available to researchers through Harvard’s Dataverse in the summer of 2020. Additionally, it is likely that more traditional access points to the materials will be made available to researchers; a record in Harvard’s library catalog, a finding aid providing archival description to the materials, and a LibGuide to support the collection.

**Lessons learned and conclusions**

The project presented a series of unexpected challenges, not just related to collecting data associated with events occurring in real time, but also for events that do not appear to have an end. As it can be seen in Figure 2, between the moment in which #metoo exploded, until January 31, 2020, news related to #metoo keep surfacing and the trend appears to peak again in the near future. Although, technically, the project is funded until the end of the fiscal year 2020, we see an imperative to continue collecting materials, preserving, and making materials available to researchers.

When deciding where to limit our collection scope, many difficult decisions had to be made. For example, the hashtag #gamani was included, as it is the form in which members of the American-Jewish community are sharing their stories of sexual harassment, but hashtags like #niunamenos were not included, mostly because they were mainly focused on issues of violence against women, and particularly, femicide. Similarly, events like Brett Kavanough’s confirmation to the United States’ Supreme Court were included, as it was highly associated to #metoo, and both supporters and detractors were particularly active during that time. Fortunately, other events like the trial to Harvey Weinstein are so intrinsically associated to #metoo, that the event it and of itself, has server as an agent to revitalize the activity around #metoo in the press and social media.

Another challenging issue was to deal with the different APIs, policies, and the constantly changing terms of service coming from the different social media platforms. Following the Facebook-Cambridge Analytica scandal, many platforms (particularly Facebook) have restricted further the access to their data, either by changing policies or limiting APIs to selected areas of content. Since there is no academic key to access the data, a multidisciplinary research project like this becomes rather difficult.

Perhaps, the greatest and most challenging issue has to do with the infrastructure and funding required to take on a project like this.

The work involved communicating with Twitter, contacting different people, and finding the adequate channels to negotiate data licensing would have been impossible without the assistance of a part time employee funded by the grant. Similarly, all the work involved in writing the content for the project’s website, manually selecting materials to collect, while including additional metadata and parsing news articles, required the attention of a person whose sole work is to be dedicated to the project. The project would have worked much slower and not efficiently if a full-time staff would have to take on the previous tasks on top of their normal obligations.

Also related to funding was the licensing of the data backlog from Twitter. While our library was fortunate to acquire the additional funds to move forward with the transaction, the costs and terms of usage would be prohibitive for many research institutions.

This project builds on the effort and experience of preceding works focused on collection of social media and related events, as well as the fantastic work done by organizations like Documenting the Now, George Washington University Libraries, MIT Center for Civics and Media, and the Berkman Klein Center for Internet and Security. Certainly, a project of this magnitude could not be possible without the availability of open source tools and other resources made available to the community.

Acknowledgments

The authors would like to thank Amanda Strauss and Jane Kelly for their work on this project.

References

Alyssa Milano on Twitter: “If you’ve been sexually harassed or assaulted write ‘me too’ as a reply to this tweet. Https://t.co/k2oeCiUf9n” / Twitter. (n.d.). Retrieved February 18, 2020, from https://twitter.com/alyssa\_milano/status/919659438700670976?lang=en

Beck, J. (2018, June 7). People Are Changing the Way They Use Social Media. The Atlantic. https://www.theatlantic.com/technology/archive/2018/06/did-cambridge-analytica-actually-change-facebook-users-behavior/562154/

#Cuéntalo. (n.d.). Retrieved February 18, 2020, from http://proyectocuentalo.org/

Digital Repository Service (DRS)—Harvard Library Digital Preservation—Harvard Wiki. (n.d.). Retrieved February 19, 2020, from https://wiki.harvard.edu/confluence/pages/viewpage.action?pageId=204385879

Documenting Ferguson. (n.d.). Retrieved February 18, 2020, from http://digital.wustl.edu/ferguson/

Ethics Statement. (n.d.). #metoo Project Schlesinger Library. Retrieved February 18, 2020, from https://www.schlesinger-metooproject-radcliffe.org/ethics-statement

Facebook–Cambridge Analytica data scandal. (2020). In Wikipedia. https://en.wikipedia.org/w/index.php?title=Facebook%E2%80%93Cambridge\_Analytica\_data\_scandal&oldid=939718385

Feuer, A. (2020, January 20). Harvey Weinstein’s Dark Days. The New York Times. https://www.nytimes.com/2020/01/20/nyregion/harvey-weinstein-trial.html

Innovation Grants. (n.d.). Harvard Library. Retrieved February 18, 2020, from https://library.harvard.edu/innovation-grants

Media Cloud. (n.d.). Media Cloud. Retrieved February 19, 2020, from https://mediacloud.org

More on restricted use cases – Twitter Developers. (n.d.). Retrieved February 19, 2020, from https://developer.twitter.com/en/developer-terms/more-on-restricted-use-cases

NAICS & SIC Identification Tools. (n.d.). NAICS Association. Retrieved February 18, 2020, from https://www.naics.com/search/

TIME Person of the Year 2017: The Silence Breakers | Time.com. (n.d.). Retrieved February 18, 2020, from https://time.com/time-person-of-the-year-2017-silence-breakers/

TIME’S UP Now | About. (n.d.). TIME’S UP Now. Retrieved February 18, 2020, from https://timesupnow.org/about/

Twitter Terms of Service. (n.d.). Retrieved February 21, 2020, from https://twitter.com/content/twitter-com/legal/en/tos.html

Vicenç Ruiz Gómez, A. M. (2017). Catalan Referendum Tweets. http://archive.org/details/catalan-tweet-ids