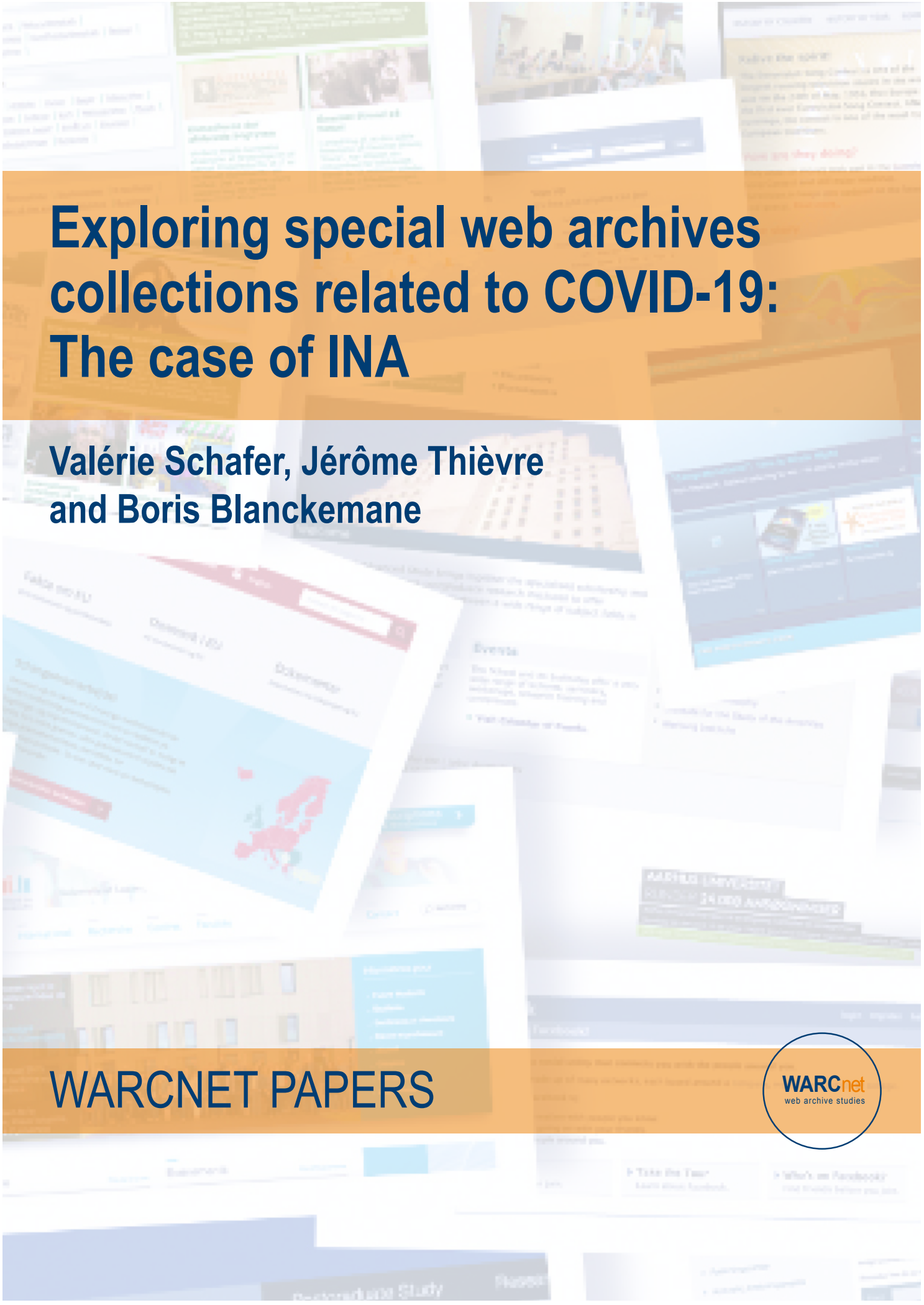
Exploring special web archives collections related to COVID-19: The case of INA

Valérie Schafer, Jérôme Thièvre
and Boris Blanckemane



WARCNET PAPERS

WARCnet
web archive studies

Exploring special web archives collections related to COVID-19: The case of INA

An interview with Jérôme Thièvre and Boris Blanckemane (INA) conducted by Valérie Schafer (C2DH, University of Luxembourg)

valerie.schafer@uni.lu

WARCnet Papers
Aarhus, Denmark 2020

WARCnet Papers ISSN 2597-0615.

Valérie Schafer, Jérôme Thièvre and Boris Blanckemane: Exploring special web archives collections related to COVID-19: The case of INA

Published by the research network WARCnet, Aarhus, 2020.
Editors of WARCnet Papers: Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Sofie Flensburg, Peter Webster, Michael Kurzmeier.

Cover design: Julie Brøndum
ISBN: 978-87-972198-2-9

WARCnet
Department of Media and Journlism Studies
School of Communication and Culture
Aarhus University
Helsingforsgade 14
8200 Aarhus N
Denmark
warcnet.eu

The WARCnet network is funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

DANMARKS FRIE FORSKNINGSFOND
INDEPENDENT RESEARCH FUND DENMARK

WARCnet Papers

Niels Brügger: *Welcome to WARCnet* (2020)

Ian Milligan: *You shouldn't Need to be a Web Historian to Use Web Archives* (2020)

Valérie Schafer, Jérôme Thièvre and Boris Blanckemane: *Exploring special web archives collections related to COVID-19: The case of INA* (2020)

Exploring special web archives collections related to COVID-19: The case of INA

An interview with Jérôme Thièvre and Boris Blanckemane (INA) conducted by Valérie Schafer (C²DH, University of Luxembourg)

Abstract: This WARCnet paper is the first in a series of interviews with European web archivists who have been involved in special collections related to COVID-19. The aim of the series is to provide a general overview of COVID-19 web archives.

Keywords: web archives, social networks, COVID-19, special collections, France, National Audiovisual Institute (INA)

This WARCnet paper is the first in a series of interviews with European web archivists who have been involved in special collections related to COVID-19. The working group for transnational events within the WARCnet project decided to focus part of its research on the web archiving of the COVID-19 crisis. This was clearly not planned at the beginning of our project, but given the circumstances which arose during the first half of 2020 (which was also the start of our research project), it was decided for several reasons:

- COVID-19 and lockdown are transnational events which have had a huge impact on every country in Europe (and the world) in health, economic, societal and cultural terms. The current context can therefore provide us with a comprehensive picture of unexpected transnational events in web archives.
- The crisis represents an opportunity to analyse an event with significant consequences on our digital lives, as the lockdown emphasised the need for remote communication and content and rapid digital adaptation of every aspect of human activities.
- Recent events have also had an impact on web archives; many web archiving institutions decided to launch emergency collections dedicated to COVID-19, and those that did not still have vast reams of content that reflect recent months.

Although the working group will focus on subtopics within this wider topic of web archives and COVID-19 – two in particular, the first on cultural institutions and COVID-19 and the second on women, gender and COVID-19 (more details and results to come in the next few

months) –, we felt the need to start our research by giving the floor to those who conducted these special collections. Once again, there are several reasons for this choice:

- The launch of a special collection is clearly not something new; there have already been similar initiatives in the past. During the wave of terrorist attacks in France in 2015 and 2016, the French National Library (BnF) and National Audiovisual Institute (INA) launched special collections and other institutions and individuals also collected testimonies and digital traces, both on websites and on social media platforms (see for example Schafer, 2018). But times of emergency are always of real interest to investigate web archiving processes as they reveal the choices, negotiations and debates that may occur within institutions and the challenges facing web archivists (for example restricting the scope of crawling, identifying which hashtags should be followed, etc.).
- To analyse these COVID-19 web archives, we also need to understand the archiving process itself, especially when moving from one European collection to another and exploring realms of the web that may be less familiar than the national context. It is very difficult to conduct an effective qualitative or quantitative analysis without a clear idea of the web spaces involved, the selection made by web archivists, their practices, methods, tools and curation procedures. Let's take just one very simple example: for the French National Audiovisual Institute (INA), the first respondent in this series of interviews, there is limited scope for web archiving, as INA and the BnF share the mission of archiving the "French Web". INA therefore focuses on audiovisual websites, and this of course imposes particular characteristics on their collection. Not all countries archive the same social media platforms, invest the same efforts in preserving audiovisual content, have comparable website samples, etc.

Our first interview was conducted on 29 July 2020 with Jérôme Thièvre and Boris Blanckemane, who work at INA. INA was given responsibility for the legal deposit of some Internet content by the French Act of 1 August 2006 on Copyright and Related Rights in the Information Society (known as the "DADVSI" Act), whose implementing decrees were published at the end of 2011. INA is responsible for the legal deposit of online audiovisual content (approximately 14,000 websites are collected), while the French National Library (BnF) archives other "French" websites (approximately 4.5 million websites). Websites whose content is preserved by INA include:

- The websites of audiovisual media services (public and private channels), including on-demand services
- Online TV and radio services
- Websites about radio and TV programmes as well as websites dedicated to feeds and series and fan websites
- The websites of professional and institutional bodies in the audiovisual communication sector.

With more than 14,000 websites preserved at INA, the collection has been extended to online social media (more than 20,000 media-related user accounts from social media and video publishing platforms such as Twitter, YouTube and Dailymotion, and continuous streams from 30 web radio stations) (https://institut.ina.fr/collections/le-web-media). As

mentioned earlier, this is not the first time INA has conducted an emergency collection. Let's find out more about the reasons, uses, choices and limitations of this special collection.

THE REASONS OF THE SPECIAL COLLECTION

You conducted a special COVID-19 collection. Why?

Jérôme Thièvre: The first collection linked to a media event took place after the 2015 terrorist attacks in Paris. It was the first time that we had made a special collection and departed from the broad scope of our audiovisual collections. Since then, we have decided to try to carry out collections for media events with a high media impact on social media platforms, in particular Twitter. We have done this for major societal crises like the Me Too movement. So that's why we also launched a collection on COVID-19. It is a massive event that reflects our decision to monitor major media events. For us this collection was not so special as we had already had previous experience with other collections, but what did make it stand out was the sheer volume that was circulating on social media compared to other crises. We had a higher volume of hashtags too. Hashtag tracking was also more complicated because there was a lot of online communication.

Boris Blanckemane: I agree that one difference compared to other events is the higher number of hashtags. For the Me Too movement, for example, we had just under ten hashtags. During the pandemic there were a lot more. Some developed over a long period of time, some were used for one day only, such as #deconfinementJ1, J2, etc., which appeared and disappeared every day.

In terms of volume, are there more tweets than the number you collected during the terrorist attacks – Charlie Hebdo, Bataclan, Nice, etc. – in 2015-2016? You collected more than 40 million tweets at that time if I remember rightly?

Boris Blanckemane: For COVID-19 we have around 120 million tweets at the moment.

Jérôme Thièvre: Indeed, it's clear with the peaks we're getting that it's on a much larger scale, and obviously we're only talking about the tweets that we collected; in reality there are even more.

THE SCOPE OF THE COVID-19 COLLECTION

What exactly did you collect? Websites, social media? Which specific platforms, hashtags, profiles, languages?

Jérôme Thièvre: On Twitter, we have allowed ourselves to go slightly beyond the scope of the legal deposit of web content for which we are responsible, to make sure that we could

follow these important media events properly. The BnF makes general captures of French websites, so we tend to stay within the scope of our field, namely audiovisual content. Of course, there are documents that talk about the COVID-19 on TV and radio websites, but we have not added any specific websites on the coronavirus. In terms of video channels, we added...

Boris Blanckemane: Yes, a few YouTube channels, but less than a dozen. And in those cases we are really collecting videos.

The main focus of this COVID-19 collection is social media platforms.

Jérôme Thièvre: Mainly Twitter because on Facebook we are still not able to collect data properly via the API.

And Instagram?

Jérôme Thièvre: It's like Facebook, it's blocked in the same way.

Which hashtags or profiles did you select?

Boris Blanckemane: For the hashtags we used various methods, but we carried out daily monitoring anyway, because some hashtags regularly appeared during the night or during the day. Some of them were also retrieved from websites that publish trends of the main hashtags being used. We particularly used Trends24. This helped us identify hashtags linked to the epidemic. We also closely monitored some programmes, especially TV programmes, which might use hashtags.

Can you give me some examples of hashtags you collected?

Boris Blanckemane: They can be grouped thematically. There are all those that directly concern the COVID-19 outbreak: COVID19, CoronavirusFrance, etc. Then there are those related to containment, such as Restezchezvous [Stay indoors], and then to easing of lockdown measures, including those I mentioned above: Deconfinement [Easing of lockdown] followed by days J1, J2, J3, etc. Then there is a whole category related to the health aspects of the crisis with hashtags on masks, thanking care workers, on Didier Raoult [a French scientist whose controversial views on the COVID-19 crisis made the headlines on several occasions], hydroxychloroquine, etc. Finally, there is a category that is a bit of a catch-all that touches on all the other facets of the crisis – political, environmental and other aspects. These are hashtags like Macronavirus, those related to speeches by politicians like Macron20h02, etc.

Confinement (Lockdown)	Déconfine-ment (Easing of lockdown)	Epidémie (Epidemic)	Soignants (Care Workers)	Hors-périmètre (Other facets)
ALaMaison	11Mai	CoronaPande-mie	Darmanin	Allocution
attestationde-deplacement	2juin	coronapocalypse	EnsembleA-vecNosSoig-nants	Anneeblanche
attestationsurl-honneur	deconfine-ment	coronapocolypse	Ensemble-SauvonslHopital	Bergame
confinement	Deconfine-ment11Mai	coronavirus	masques	Blanquer
CONFINE-MENTJOUR1	deconfine-mentjour1	CORONAVIRU-SENFRANCE	masquesFFP2	BuzynGate
Confinement-Jour35	Deconfine-mentJour10	coronavirusfr	MerciAuxSoig-nants	carlabruni
Confinement-Jour36	Deconfine-mentJour11	Covid_19	OnApplaudit	CohnBendit
Confinement-Jour37	Deconfine-mentJour12	Covid_19fr	onnoublierapas	ConseildEtat
Confinement-Jour38	Deconfine-mentJour13	COVID19	ParolesDeSoig-nant	conseilscienti-fique
Confinement-Jour39	Deconfine-mentJour14	COVID19france	respirateurs	criseeconomique
Confinement-Jour40	Deconfine-mentJour15	COVID19Pande-mic	Ricard	CymesDegage
Confinement-Jour41	Deconfine-mentJour17	CO-VID2019france	ScandaleSani-taire	DIS-COURSMACRO N
Confinement-Jour42	Deconfine-mentJour2	COVID—19	SOSFrance	ElizabethII
Confinement-Jour43	Deconfine-mentJour3	DeuxiemeVague	Chloroquine	Enten-donsLeursCris
Confinement-Jour44	Deconfine-mentJour4	etatdurgencesa-nitaire		EtApres
Confinement-Jour45	deconfine-mentjour5	JeNeSuisPasUn-Virus		Eurovision2020
Confinement-Jour46	Deconfine-mentJour8	Stade3		FranceUnie
Confinement-Jour47	Deconfine-mentJour9	VirusChinois		Hantavirus

Confinement (Lockdown)	Déconfine-ment (Easing of lockdown)	Epidémie (Epidemic)	Soignants (Care Workers)	Hors-périmètre (Other facets)
Confinement-Jour48	Ecole			JeanPierre-Pernaut
Confinement-Jour49	Ren-dezNousNos Plages			JermeSalomon
Confinement-Jour50	resterprudent			JeVeuxAider
Confinement-Jour51	RestezPru-dents			KarineLacombe
Confinement-Jour52	StopCovid			Kawasaki
Confinement-Jour53				LCIVousDonne-LaParole
Confinement-Jour54				macron20h
Confinement-Jour55				Macron20h02
CONFINE-MENTJOUR8				Macronavirus
Confinement-Total				MacronDestitu-tion
couvrefeu				montparnasse
culture-cheznous				Municipale2020
DistanciationS-ociale				NationAppre-nante
Irresponsables				Necker
jeniraipasvoter				Nicotine
JeSauve-DesVies				OnVousRepond
maisonlumni				PatrickCohen
PessahChez-Vous				Penicaud
QuarantineLife				Raoult_didier
restecheztoi				sanofi
resterchezsoi				soiree2linfo
ResterChez-Vous				SoutenonsNo-sEntreprises

Confinement (Lockdown)	Déconfine-ment (Easing of lockdown)	Epidémie (Epidemic)	Soignants (Care Workers)	Hors-périmètre (Other facets)
restezàlamai-son				StopProduction-NonEssentielle
RESTEZCHEZ VOUS				Tocilizumab
RestezChez-VousBordel				tokyo2021
stayhome				VacancesAppre-nantes
stayhomechal-lenge				VincentLindon
StayTheF-Home				ViolencesConju-gales
TousAlaMai-son				YvesCalvi

Table 1: 149 hashtags were identified and collected. They have been divided into 5 categories. ©INA

As part of the WARCnet project, we are planning to conduct research on women, gender and COVID-19, exploring topics such as feminist impulses, the status of carers, domestic violence and aesthetic issues.

Boris Blanckemane: Domestic violence is a hashtag that we collected.

THE FRAME OF THIS SPECIAL COLLECTION

When did you start? When did/do you plan to stop? What was the capture frequency?

Jérôme Thièvre: We started on 13 March. And we are still collecting.

Is it a daily practice, with a dedicated person?

Boris Blanckemane: We have a TV/radio archivist who selects hashtags to collect, but it doesn't take up all of his time. Anyone in the team can also suggest hashtags that are growing and major trends. Everyone is involved.

And the technical part?

Jérôme Thièvre: We have a well-established procedure. All the selected hashtags go into databases. We collect tweets using Twitter's APIs. The problem is not the collection process but the fact that we can't collect everything, due to the APIs limits. We can only collect some of what is tweeted; real-time APIs are limited to 50 tweets per second. We are constantly reaching the limit.

Do you always collect limit messages when the limit is reached, like you did for the collection on the terrorist attacks?

Jérôme Thièvre: Yes, but it's becoming increasingly complicated. We use several accounts linked to the APIs, but with each account we don't only collect COVID-19 tweets. So if I receive a limit message telling me that I missed 10,000 tweets, I'm not sure that all those tweets are directly related to COVID-19. We're sticking with the idea of a large sample. If we are careful about how we make use of it then it can provide us with information, but any quantitative analysis needs to be interpreted with caution and is limited. Moreover, hashtags have to be popular for us to spot them. And sometimes we had to retrospectively search for important hashtags that we missed, using an API that allows to retrieve tweets over seven days. We have ended up with a collection of tweets for which it is difficult to evaluate what proportion of the general movement it represents. So we don't have a technical problem; it's more that we are not doing what we thought we could do at the beginning: we had a very quantitative-oriented interface and a lot of work is still needed to rethink how best to perform quantitative analysis. When it comes to the data collection itself, on the other hand, it remains easy to collect, store, etc.

I imagine that there is a slowdown at the moment?

Boris Blanckemane: Yes, indeed.

Jérôme Thièvre: By 1 July we had already noticed a significant decrease in the volume of tweets and also the identification of new hashtags...

Boris Blanckemane: ...Yes, there are not really any new hashtags anymore.

Jérôme Thièvre: But we are letting the current collection run because numbers may go up again.

How did you carry out quality control on the collection (if applicable)?

Jérôme Thièvre: We have monitoring tools to ensure the collect is working on a day-to-day basis. So, it's a partial control given the volume of data, but every month we check that what we have collected can be visualised in the interfaces. We do not have the same problem as with websites, where we have to ensure the quality of the collection. With Twitter, we collect data and we either have them or not, it's as simple as that.

You already mentioned that you collected 120 million tweets. What type of data does the collection represent (videos, texts, etc.)?

Jérôme Thièvre: We collected 145,000 videos hosted by Twitter in the 120 million tweets mentioned and by the beginning of July we had 250,000 videos related to covid-19 from all sources.

So, you don't have YouTube videos that are shared in tweets? Or even the content of shared hyperlinks?

Jérôme Thièvre: That's right. At the moment we are not collecting those pages and assets, our tools are yet to be adapted. Moreover, robots are website-oriented, but here we are talking about pages and web pages which may be outside the scope of our legal deposit mission for the Web. Many publications are indeed just links pointing to other content. So, it would be relevant but there are technical, legal and other problems here.

ACCESSIBILITY AND SEARCHABILITY

Can we talk about the accessibility and searchability of these data?

Jérôme Thièvre: Concerning Twitter data, videos and web pages on COVID-19 – because many audiovisual websites made online publications –, our collection is accessible in the INA reading rooms in Paris and in French regions and it can be consulted via interfaces that enable complex queries, with aggregated Twitter searches on the most collected hashtags and searches by emoji, timeline, etc. The interfaces for videos are more simplistic, but there is still the functionality to display results and timelines.

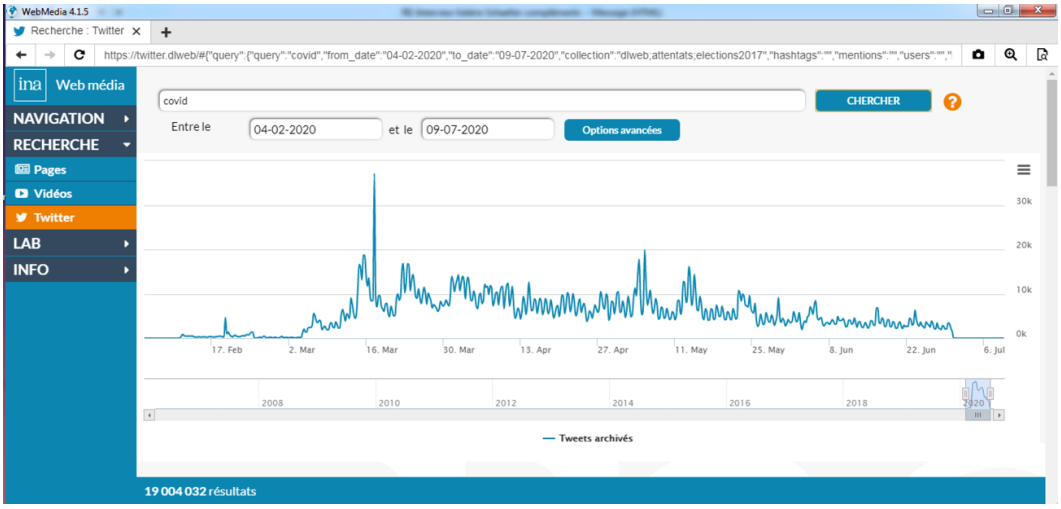


Figure 1: Timeline based on the request COVID between February and July 2020. ©INA

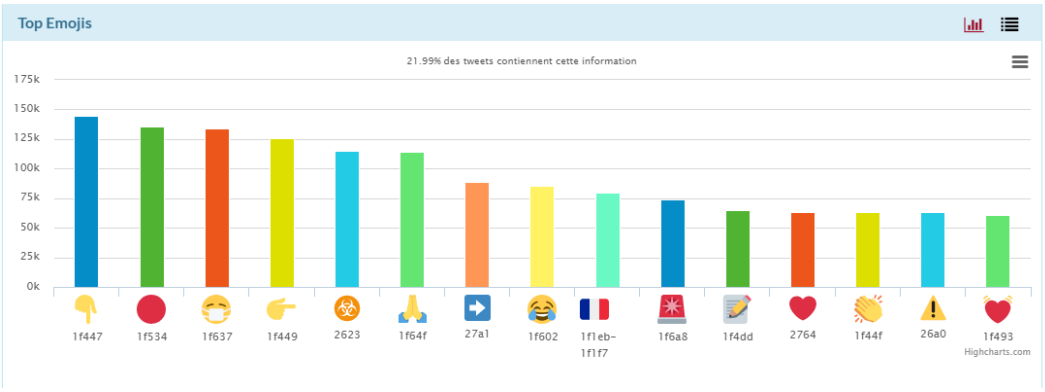


Figure 2: Top emojis between February and July 2020 (request: COVID). ©INA

Boris Blanckemane: There's also another feature to report, as we've added the list of items we are collecting. Before, you had to know the hashtags you wanted to search for, but now the list is accessible and we have documented the hashtags, there is a detailed description and descriptors from the INA thesaurus. You can easily retrieve the list.

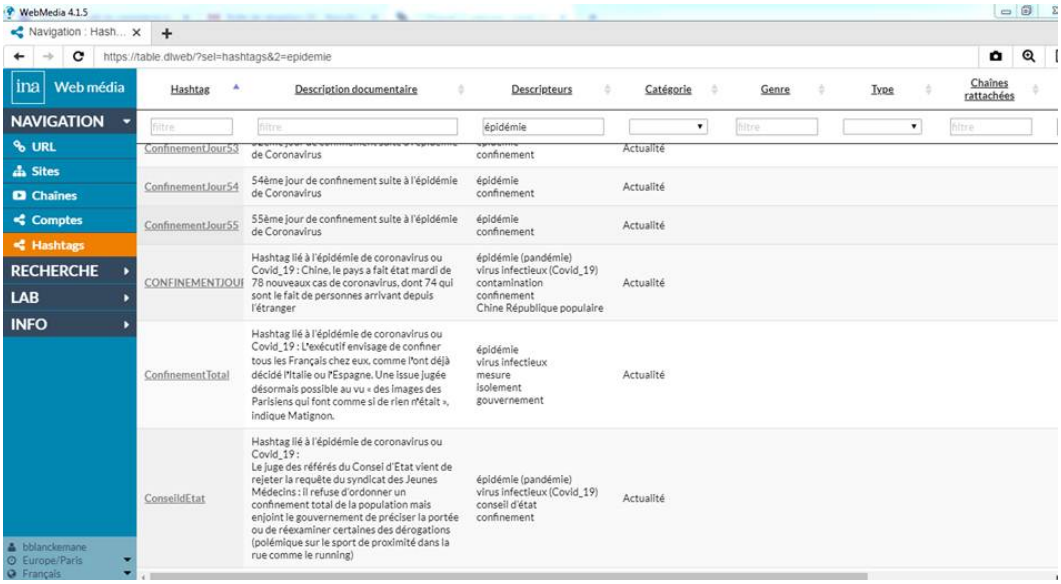


Figure 3: Detailed description and descriptors of hashtags. ©INA

Excellent! This could be very useful if we imagine young researchers in 40 years' time wanting to carry out research but with no memory of these hashtags because they did not live through the pandemic. Did you also do the same thing for events such as the terrorist attacks?

Boris Blanckemane: Yes, we redocumented everything.

PARTNERSHIPS AND USES

Are researchers already asking you about the COVID-19 collection, wanting to analyse it?

Jérôme Thièvre: Consultation was closed until the end of June because of the lockdown and we only communicated internally about this collection. I think that at the beginning of the academic year there will be more interest and more researchers, and we'll work with Inathèque, the lecture rooms department, to coordinate our communication.

Did you have any partnerships – with local stakeholders, Archive-It, the IIPC, etc. – during the collection process?

Jérôme Thièvre: No, not really, we were focused on the Twitter collection, which is not really central to the IIPC's approach, but we will communicate with the IIPC and we are willing to share our experience.

How easy was it to manage the process of collection from home?

Boris Blanckemane: It was not a problem. All the tools are scaled and designed to be used remotely. There is no problem when it comes to managing them from a distance.

Jérôme Thièvre: I would add that remote working has been in place at INA for several years now. It has worked well. Everyone already had the equipment and so on. All the collection and consultation tools are accessible through the remote working system and communication within the team was good.

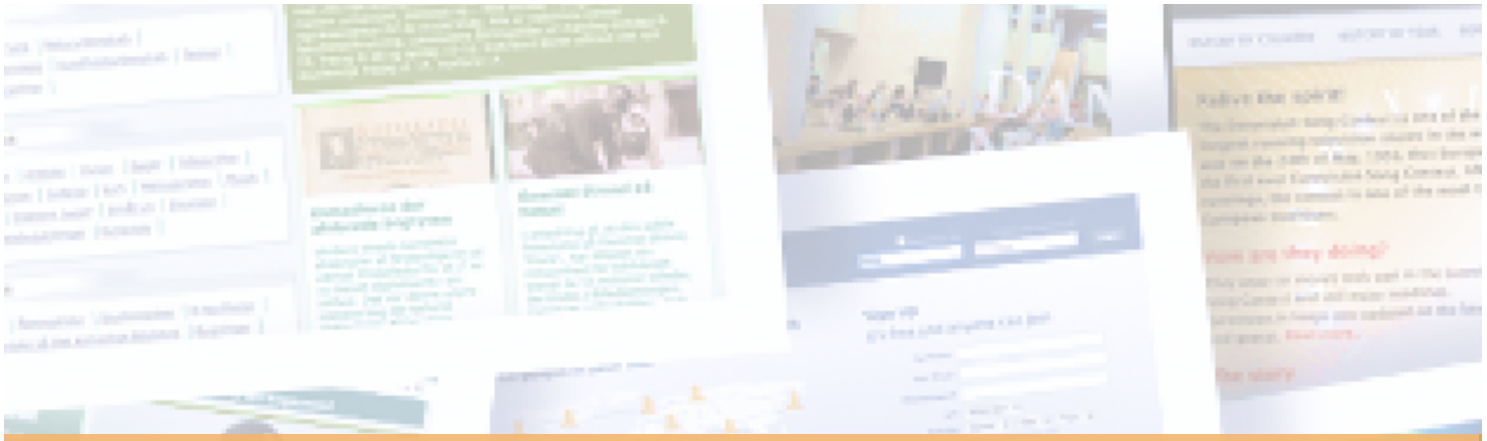
How do you archive nationally something which is fundamentally global?

Jérôme Thièvre: Since the beginning of these media collections on Twitter, the choice was made not to filter by language but by hashtag, so we get tweets that are not in French and we accept this as part of the process. It's difficult to formally identify what is French or not and we have some foreign publications, especially with very general hashtags like COVID19, but I think that our collection mainly reflects the reactions on Twitter in France. However, it is also useful to have publications from abroad, some of which talk about the situation in France.

REFERENCES

Schafer, V. (2018). Le patrimoine nativement numérique des attentats en Europe: regards croisés. *La Gazette des Archives*, 250, 115-130. https://orbilu.uni.lu/handle/10993/36698

We would like to thank Sarah Cooper (University of Luxembourg) for her help in proofreading this interview.



WARCnet Papers is a series of papers related to the activities of the WARCnet network. WARCnet Papers publishes keynotes, interviews, round table discussions, presentations, extended minutes, reports, white papers, status reports, and similar. To ensure the relevance of the publications, WARCnet Papers strives to publish with a rapid turnover. The WARCnet Papers series is edited by Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Sofie Flensburg, Peter Webster and Michael Kurzmeier. WARCnet Papers have gone through a process of single blind review.

The aim of the WARCnet network is to promote high-quality national and transnational research that will help us to understand the history of (trans)national web domains and of transnational events on the web, drawing on the increasingly important digital cultural heritage held in national web archives. The network activities run in 2020-21, hosted by the School of Communication and Culture at Aarhus University, and are funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

WARCNET PAPERS

INDEPENDENT RESEARCH FUND DENMARK

warcnet.eu warcnet@cc.au.dk twitter: @WARC_net facebook: WARCnet
youtube: WARCnet Web Archive Studies slideshare: WARCnetWebArchiveStu