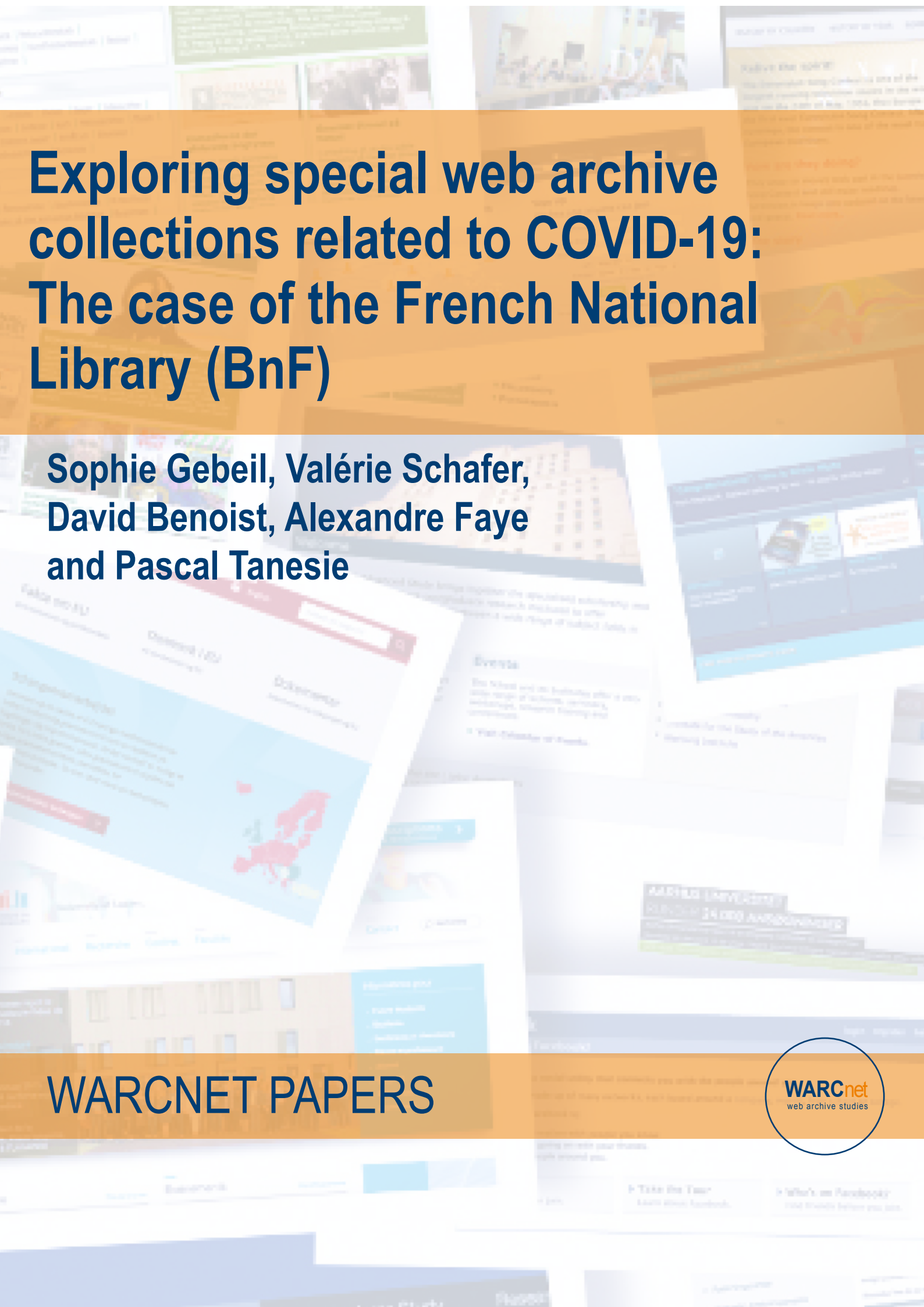
Exploring special web archive collections related to COVID-19: The case of the French National Library (BnF)

Sophie Gebeil, Valérie Schafer, David Benoist, Alexandre Faye and Pascal Tanesie



WARCNET PAPERS

WARCnet
web archive studies

Exploring special web archive collections related to COVID-19: The case of the French National Library (BnF)

An interview with David Benoist, Alexandre Faye and Pascal Tanesie (BnF) conducted by Sophie Gebeil (Aix-Marseille University) and Valérie Schafer (C²DH, University of Luxembourg)

sophie.gebeil@univ-amu.fr and valerie.schafer@uni.lu

WARCnet Papers
Aarhus, Denmark 2020

WARCnet Papers ISSN 2597-0615.

Sophie Gebeil, Valérie Schafer, David Benoist, Alexandre Faye and Pascal Tanesie: Exploring special web archive collections related to COVID-19: The case of the French National Library (BnF)

Published by the research network WARCnet, Aarhus, 2020.

Editors of WARCnet Papers: Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Peter Webster, Michael Kurzmeier.

Cover design: Julie Brøndum
ISBN: 978-87-94108-02-7

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

DANMARKS FRIE FORSKNINGSFOND
INDEPENDENT RESEARCH FUND DENMARK

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

WARCnet Papers

Niels Brügger: *Welcome to WARCnet* (May 2020)

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

Valérie Schafer and Ben Els: *Exploring special web archive collections related to COVID-19: The case of the BnL* (Sep 2020)

Niels Brügger, Anders Klindt Myrvoll, Sabine Schostag, and Stephen Hunt: *Exploring special web archive collections related to COVID-19: The case Netarkivet* (Oct 2020)

Friedel Geeraert and Nicola Bingham: *Exploring special web archives collections related to COVID-19: The case of the UK Web Archive* (Nov 2020)

Friedel Geeraert and Barbara Signori: *Exploring special web archives collections related to COVID-19: The case of the Swiss National Library* (Nov 2020)

Matthew S. Weber: *Web Archives: A Critical Method for the Future of Digital Research* (Nov 2020)

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

Niels Brügger, Valérie Schafer, Jane Winters (Eds.): *Perspectives on web archive studies: Taking stock, new ideas, next steps* (Sep 2020)

Friedel Geeraert and Márton Németh: *Exploring special web archives collections related to COVID-19: The case of the National Széchényi Library in Hungary* (Oct 2020)

Friedel Geeraert and Nicola Bingham: *Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative collection* (Nov 2020)

Caroline Nyvang and Kristjana Mjöll Jónsdóttir Hjörvar: *Exploring special web archives collections related to COVID-19: The Case of the Icelandic web archive* (Nov 2020)

Sophie Gebeil, Valérie Schafer, David Benoist, Alexandre Faye and Pascal Tanesie: *Exploring special web archive collections related to COVID-19: The case of the French National Library (BnF)* (Dec 2020)

All WARCnet Papers can be downloaded for free from the project website warcnet.eu.

Exploring special web archive collections related to COVID-19: The case of the French National Library (BnF)

An interview with David Benoist, Alexandre Faye and Pascal Tanesie (BnF) conducted by Sophie Gebeil (Aix-Marseille University) and Valérie Schafer (C²DH, University of Luxembourg)

Abstract: This WARCnet paper is part of a series of interviews with European web archivists who have been involved in collections related to COVID-19. This interview enlightens practices, processes, challenges and uses at the French National Library (BnF).

Keywords: Web archives, COVID-19, special collections, France, French National Library (BnF)

This WARCnet paper is one of a series of interviews with European web archivists who have been involved in collections related to COVID-19. The working group for transnational events within the WARCnet project has decided to focus part of its research on web archiving of the COVID-19 crisis and we felt the need to start our research by giving the floor to those responsible for these special collections (see the first WARCnet paper in the series about INA).

This interview was conducted by Sophie Gebeil and Valérie Schafer on October 15, 2020 with David Benoist, Alexandre Faye and Pascal Tanesie from the French National Library (BnF).

Pascal Tanesie is Deputy Head of the Digital Legal Deposit (DLN) Department at the BnF and was responsible for the COVID-19 collection during the lockdown period. He allocated the work and coordinated the collection process.

David Benoist is a Collections Officer in the BnF Audiovisual Department, where he is responsible for curating video game collections. He is the Digital Legal Deposit correspondent in the Audiovisual Department.

Alexandre Faye was involved in the collection as a member of the DLN Department. He led a collection that was opened up to wider participation than usual, on the basis of the existing Topical News collection, which follows news stories, especially on social media.

The BnF has been harvesting the French web under France's digital legal deposit scheme since 2006. It recently carried out a special COVID-19 collection, presented on both the Web Corpora blog (in French) and the IIPC blog (in English) by Alexandre Faye, which we revisit with him and his colleagues in our discussion. This interview once again demonstrates the importance of human curation, while also highlighting many years of experience and a strong network of correspondents, discussing issues related to web technologies and social media, space, scales and timeframes, and exploring current and future collaboration with users and research projects.

THE REASONS OF THE SPECIAL COLLECTION

Why did you conduct a special COVID collection?

Pascal Tanesie: We don't use the term "collection"; we prefer to talk about a "series to be collected". This year, for example, we created two new series, "Artificial Intelligence" and "Environmental Challenges". We also created a series called "Topical News", with the aim of collecting material from websites related to specific events (festivals, deaths of figures in the public eye, events, etc.). We used the Topical News series when the outbreak was announced in China in January, with the keyword "coronavirus". When we entered lockdown, we stepped up the collection process and it became more systematic. We continued using the keyword "coronavirus". Initially, the collection was largely based on the contribution of correspondents and the Digital Legal Deposit (DLN). At the beginning we didn't think in terms of a collection, but we have ended up with a collection that has been indexed to some extent, which researchers can work with.

As for the reasons behind the collection, the BnF was very keen for all staff who could do so to contribute to this collection during the lockdown period. So, there was a logistical effort required to make sure that all those who wanted to contribute to the collection were able to, especially in legal deposit libraries. We didn't systematically give these correspondents access to the "BC Web" application; they generally used spreadsheets to record any relevant sites they had identified, and we could then incorporate these into our application.

David, what was your role in the process?

David Benoist: As far as I remember, the request came from above — a memo written jointly by the Collections Department and the DLN asked the various BnF departments to organise the collection in conjunction with the DLN, a task that I was already doing for Topical News. We also had to involve as many people as possible in the collection process, including those without any experience in the area, who weren't familiar with BC Web. During the lockdown, everyone was monitoring new web content, and the idea was to make

good use of this monitoring work being carried out by colleagues. It was also a way of providing tasks for people working from home who were unable to perform their usual job remotely, for example warehouse workers. And it helped prevent isolation and created links between departments, giving people a shared task and helping boost the morale of colleagues who were feeling isolated. In the Audiovisual Department, participation was decided on a voluntary basis, after the project was presented to staff by their line managers. We then held an online meeting with the volunteers. The DLN also proposed a methodological guide, a four-page PDF. I produced a special Audiovisual version of the guide which emphasised the aspects that were most important for us: the impact of the crisis on cinema, culture, content distribution platforms, YouTube channels linked to COVID, technical points. Usually around fifteen of us take part in the DLN selection process; this time there were around twenty additional colleagues involved, but with no experience in this type of collection and no access to the interface. So, I set up a collaborative spreadsheet on Framacalc and I sent the link to my colleagues. Those who already had access to BC Web carried on creating records; my job was mainly to help out colleagues who weren't used to the selection process by sending examples to show them sites that would be worth exploring. The shared table generated a sort of competitive rivalry — as it filled up, it spurred others on. It also provided examples of websites or content that had been collected. Every day I took stock of what had been added. The table was a simplified version (so that it would be easier to use) of the DLN table used for imports. To start with some colleagues didn't dare enter their websites in the table — they were worried about getting it wrong or proposing resources that were irrelevant. I wanted them to feel comfortable and not to censor themselves.

Pascal Tanesie: David has described the process very well. That was exactly how it worked for us with the departments. We worked very closely with the Audiovisual Department and most of the other departments. The staff network was very well coordinated and it showed. We had never had as many people working on a specific topic before.

Alexandre Faye: There were 52 correspondents who contributed directly to BC Web but, overall, there were more people participating, around 80.

Pascal Tanesie: In BC Web we set up a table template that we could use to import records directly from the table, which made it easier to carry out checks before importing. In the application there are two parts: documentary and technical information (frequency of collection, depth, budget, etc.). The technical information can be confusing for those who are not used to it. The first stage was to prepare a document to explain how to create records and define themes, and also how to integrate various aspects, such as education.

David Benoist: That's why I proposed a simplified version of the eight-column DLN document. I then added the technical information afterwards — I didn't want people to feel intimidated.

The monitoring process was a good activity for the lockdown period. Some colleagues were living in difficult conditions during lockdown. This was an easy activity that could be spread throughout the day; it was well suited to lockdown.

Alexandre Faye: At a broader level, with the BnF so willing to get involved and the positive response from BnF departments and from libraries in other French regions, we had a good distribution in terms of disciplines and geographical areas — and that resulted in a quality collection, which we are rather proud of. The role of the DLN was also to fill any gaps in terms of topics that might have been overlooked or were gaining increasing media attention such as domestic violence, diabetics, etc. To that extent the collection kept that "topical news" quality — unlike targeted collections, it remained flexible, which meant that we could adapt to a changing situation. We had selections that we could monitor over time, like media sections on websites, and more one-off content that was just collected once. It was also the first time we had such an extensive video collection. Overall, we obtained good coverage of the online reaction to the health crisis, with a wide variety of thematic websites and media types.

THE SCOPE OF THE COVID-19 COLLECTION

What did you collect? Websites, social media platforms (if so, what DSNs, specific hashtags, profiles, languages)?

Pascal Tanesie: We collected pretty much everything you have just mentioned. In terms of social media, we mainly collected from Twitter and lots of hashtags. We had different hashtags for each day of the lockdown. We drew on the News collection, but we also collected content from institutional and local authority websites, lots of blogs about the epidemic, and we collected YouTube videos.

Alexandre Faye: Our mission is to harvest the French web, which limits our scope. It's helpful for us to have the Afnic list of .fr domain names. During lockdown, websites were set up and we were able to identify them with the lists provided by Afnic and pinpoint these new domain names, which included mutual support platforms, for example.

Pascal Tanesie: As Alexandre mentioned, we have a partnership with Afnic and we identified COVID domain names that were created, sometimes just to host advertising. We looked for domain names where pages had actually been created — some domains were purchased but not actually activated.

Alexandre Faye: At the same time the press collection continued, and we also developed a video collection. The video collection began in July with the identification of 200 YouTube channels linked with the coronavirus. That's an approximate figure, it was an initial selection, and we then had to make a second selection because there were too many channels.

David Benoist: I was particularly concerned about representativeness, choosing sites that reflected all viewpoints. That was all the more important because there were highly controversial topics like chloroquine and Professor Raoult. We chose French hashtags, although the term "coronavirus" is international so it's difficult to target just French content. So where possible I selected hashtags that were specific to France, even if sometimes that spread over into other French-speaking countries. For audiovisual material, we tended to take the subjects we usually monitor, in the fields of software and video games for example, with gaming experiencing a massive surge. Colleagues looking for audio content targeted one-off concerts, and when it came to videos/animated images we suggested focusing on video creations and performances. One example was the website "Par ma fenêtre" [From my window — https://www.parmafenetre.fr], where individuals and some celebrities film what they see from their window. We found that very interesting. Colleagues from the Animated Image Department even contacted the webmaster of that site to retrieve the videos. And we also collected some local news, but that was the case for all French departments.

Alexandre Faye: In the settings, we indicate a depth, which might be the domain if we want to retrieve all the pages from a newly created website (e.g. ethique-pandemie.com). But for a media section within a website, we would opt for less depth (page+2 click) and instead increase the frequency, collecting the site more often. That enabled us to monitor a media section and all the articles published over time. We already had a system for archiving national and regional press, and we supplemented this by including specialist press outlets and specialist blogs like the one run by the Syndicat National des Jeunes Médecins Généralistes [French Union for Young General Practitioners, snjmg.org/blog/].

For social media platforms, did you also use Heritrix?

Alexandre Faye: Yes. On that topic, we are weighing up whether we need to continue to crawl Facebook. For the time being we are continuing to do so despite a high failure rate; the success rate is 10 to 15% but with a bit of effort we are able to retrieve captures. There are several research projects on Twitter that are less complicated to collect in technical terms. We have realised that we are the only ones doing it, so we think it's a good idea to carry on. For Twitter we use Heritrix, and also for videos.

Pascal Tanesie: Even if we adapt Heritrix, it's still the same system. Sometimes we vary our methods but it's always Heritrix, conventional methods.

Alexandre Faye: We changed to a new version of Heritrix at the end of the collection process. The new version is quicker and is better at retrieving images, which was the case for the July crawl.

We will make available all the data collected from 1 February to 31 July. We intend to index the content collected during that period to enable keyword searches.

Could you provide more information about the volume of data collected and the nature of the collected data?

Alexandre Faye: We had 5,000 URLs to start with, and if we add video collections and online press, the total collection amounts to 15 TB of data or 275 million URLs. That represents 1,014 collection tasks and a final output of 15,504 WARC files.

When you talk about URLs, a URL can refer to a single image or part of a page, that's right isn't it?

Alexandre Faye: Yes, absolutely.

Pascal Tanesie: That's why we also provide information about the type of content.

What quantity does that represent compared to the collection on the Gilets Jaunes [the "Yellow Vest" protest movement], for example?

Pascal Tanesie: It's hard to say, but it's definitely more. It has really been a highlight in terms of our collections. There are topics that we have monitored like the Gilets Jaunes, but I don't remember ever having done such a large collection. The Topical News collection usually involves around ten correspondents; this is the first time such a large number of people have been involved.

Alexandre Faye: The lockdown really gave new momentum to the movement. So, 2020 has been a strong year with the Coronavirus collection, and even for the French municipal elections the network continued its work. New colleagues contributed to the work, and we want to carry on working with them and maintain the momentum, even though we realise that of course it can't be as strong as it was during lockdown.

When did you start? When did you (do you plan to) stop? What is the capture frequency?

Alexandre Faye: The first content was collected at the end of January, a few hashtags and web pages, especially humorous content, the hashtag "jenesuispasunvirus" [I am not a virus], accounts like the one belonging to the French Medical Board [Ordre des médecins] which we are continuing to monitor, but to start with it came under the Topical News collection so it only involved around ten people. When the lockdown started (16 March), the network grew significantly. We entered a second phase in the collection process, with a huge number of selections made in April and May. The number of bytes increased threefold between March and June — we were collecting three times more material.

David Benoist: There was understandably a time lag for colleagues who were learning something new; it took a few weeks to train up a group. We held a virtual coordination meeting on 20 April to give us time to put everything in place for these additional staff.

Pascal Tanesie: Some captures occur almost instantly and bots crawl several times every day or week. Other sites are captured once a week or once a month, like institutional websites such as Inserm. In mid-July we asked to stop the collection, we needed to bring it to a close, people were going back to their usual work and we also needed to do a quality check. We had in mind that if there was a second wave, we would consider starting a second collection, to round off the period. To include the press and other content, we need a timeline. We also wanted to carry out some checks, for example on the AP-HP platform, to improve the way the bot handles PDF content.

Pascal Tanesie: We are still using the keyword "coronavirus" in our selection application, because we are continuing our selections. But we have fixed the scope for the first series as coronavirus web archives for the period from February to July as representative of this phase of the epidemic, which was particularly marked by the first lockdown.

How was quality control done on the collection?

Alexandre Faye: In July, we were able to carry out some checks on the Wayback Machine in particular for sites that had published a significant amount of documentation about the epidemic online (for example covid-documentation.aphp.fr). When there were gaps, it gave us a chance to rethink the depth, to add secondary URLs to guide the bot and improve the quality of the archiving. But quality is primarily something that has to be done in advance at the point of selection — what David does for audiovisual material —, and then the DLN, especially Pascal, monitors the jobs under way.

David Benoist: When it came to audiovisual content, we had to review the suggestions that had been made. Out of 700 suggestions in the spreadsheet, we kept 320 that closely matched the criteria identified. The review of the URLs was related to several aspects: non-compliance with guidelines (a single video instead of a channel, for example — there had to be several interesting videos and not just one), sometimes there were several URLs for a single site, and some sites were too wide-ranging and went beyond the topic of the coronavirus. We also had to avoid duplication with other collections like the press and comply with the guidance of INA [the French National Audiovisual Institute] — we removed anything related to radio and TV, which INA collects. Sometimes there were things that went off topic, but that wasn't a problem; we discussed them and were able to incorporate them into our other collections. Quality control took place during the collection phase, before the documents were sent to the DLN.

Pascal Tanesie: During the crawls there is of course also a technical dimension. And another aspect that Alexandre mentioned was checking whether a given site has stopped changing, and if so we may decide to stop collecting it.

ACCESSIBILITY AND SEARCHABILITY

How accessible and searchable will the data be?

Alexandre Faye: The content collected, including videos, can already be accessed using the BnF interactive terminals in the reading room, in the Internet Archives. We are preparing to publish a list of our selections, in theory by late November 2020, after a clean-up stage. The list will be a first step that we want to publish quickly. We can then offer indicators about the number of captures, the first archiving date, etc. Given the huge volume of archived content, we want to improve data visualisation and facilitate the work of researchers. We also performed a first keyword indexing test on 10% of the collection, which enabled us to incorporate n-gram tools to analyse distribution over time. The next aim is to apply these tools to the entire period. We need to facilitate exploration of this huge collection.

Have researchers already asked you about these collections, or already analysed them?

Alexandre Faye: There has been a lot of interest in the collection but not many specific projects for the time being, which is quite natural. Sometimes laboratories are in the project development phase and they contact us to find out more about the collection and how it might be useful for their topic, for example in economics when it comes to issues related to relocation. We did a presentation for the SCAI [Sorbonne Center for Artificial Intelligence], with the aim of bringing together those interested in questions related to fake news. We sometimes also get requests for information from colleagues in other BnF departments who want to make use of these archives and get to know them better. We also take a proactive approach, for example we took the initiative of contacting Mines ParisTech. In the short term, I think that the projects in which we are already involved with partners could undoubtedly benefit from this collection. One example is BodyCapital, a European project led by the University of Strasbourg.

Pascal Tanesie: There was also a documentary maker who was interested in producing a documentary on the topic.

How have you spread the word about this special collection?

Alexandre Faye: We are planning publications on our blog, Web Corpora, or like here for the WARCnet papers, where we can share the results of the collection and also describe the process.

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

Alexandre Faye: In February, the IIPC [International Internet Preservation Consortium] contacted the BnF and other institutions. We were immediately interested in taking part

because it was clear that the pandemic was becoming an unprecedented global event. We sent two Excel files in March and April. It was a practical approach that enabled all institutions to take part because all that was needed was to identify target URLs and technical parameters. The IIPC made a specific request, to which we were able to respond. In our selection, we promoted geographical diversity and emphasized scientific and medical content (academic and more general-public content). The results can be consulted on Archive-It.

How do you archive nationally something which is fundamentally global?

Alexandre Faye: We are used to working at national level. The IIPC initiative was great, since it enabled us to focus on the national framework and the scope of our missions. The boundaries are of course rather blurred sometimes, because we broaden subjects to the French-speaking world, including on social media.

The BnF's vision is to remain rooted in this national and regional perspective. We managed to cover various scales and it was very rewarding.

David Benoist: Yes, we are used to working at national level. Some of the suggestions that we couldn't follow up included Belgian websites and channels, for example. The national scope helps guide us; sometimes in the Topical News collection we monitor international events from a French perspective. With the fire in Notre-Dame Cathedral we asked the international community whether there were websites worth including but when it came to COVID, itself a transnational event, we were happy with the national framework.

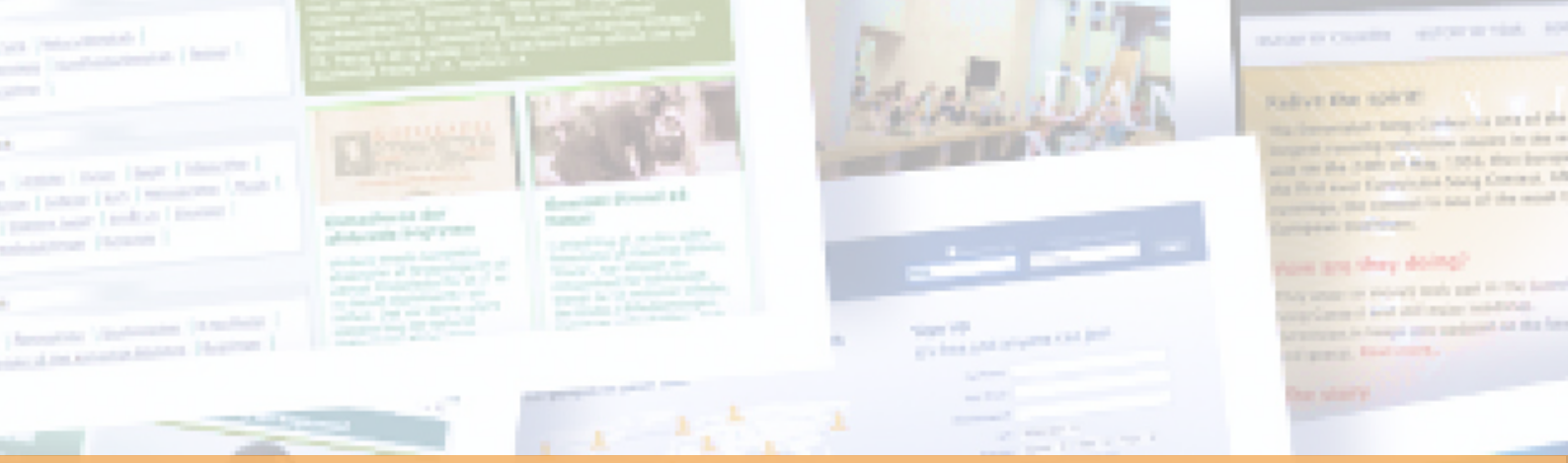
Do you have anything you would like to add?

Alexandre Faye: This is an event that everyone lived through, one that was documented in real time. What happened needs to be discussed, debated and accessible to citizens so that we can analyse the legacy of the event. Web archiving contributes to the memory of the event, including from the perspective of the general public and society at large.

READ MORE

Web Corpora
https://webcorpora.hypotheses.org

BnF list of URLs of the Covid-19 collection (carried out between 1 February and 31 July 2020)
http://api.bnf.fr/node/144



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, Peter Webster and Michael Kurzmeier. In cases where a WARCnet Paper has gone through a process of single blind review, this is mentioned in the individual publication.

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 2020-22, 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