

Exploring special web archive collections related to COVID-19: The case of the BnL

Valérie Schafer and Ben Els

WARCNET PAPERS

WARCnet
web archive studies

Exploring special web archive collections related to COVID-19: The case of the BnL

*An interview with Ben Els (BnL) conducted by Valérie Schafer
(C2DH, University of Luxembourg)*

valerie.schafer@uni.lu



WARCnet Papers
Aarhus, Denmark 2020

WARCnet Papers ISSN 2597-0615.

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

© The authors, 2020

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: 978-87-972198-3-6

WARCnet

Department of Media and Journalism 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 and Ben Els: *Exploring special web archive collections related to COVID-19: The case of the BnL* (2020)

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

*An interview with Ben Els (BnL) conducted by Valérie Schafer
(C²DH, University of Luxembourg)*

Abstract: This WARCnet paper is the second 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 media, COVID-19, special collections, Luxembourg, Bibliothèque nationale du Luxembourg (BnL)

This WARCnet paper is the second 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 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 of the series dedicated to INA).

Our second interview was conducted on 10 August 2020 with Ben Els, Digital Curator at the National Library of Luxembourg (BnL). The BnL has been harvesting the Luxembourg web under the country's digital legal deposit scheme since 2016 and has performed a special COVID-19 collection, presented on the IIPC blog by Ben Els,¹ which we revisit with him in our discussion. This interview demonstrates the importance of human curation in the choice of which websites to collect and the challenges of selection, whether linked to budgetary constraints (the BnL relies on Archive-It for its crawls), a desire for inclusiveness or the potential for enrichment of the collected URLs.

1. Ben Els, "Luxembourg Web Archive – Coronavirus Response", 23 June 2020, <https://netpreserveblog.wordpress.com/2020/06/23/luxembourg-web-archive-coronavirus-response/>

THE REASONS FOR THE SPECIAL COLLECTION

You decided to perform a special COVID-19 collection. Why?

Ben Els: For the past three to four years, we have generally operated broad crawls twice a year, and additionally performed event-based crawls – mainly because there were four election campaigns in three years. We decided to focus on topics that concern Luxembourg at a national level, events of national importance. We followed local and national elections as well as elections for staff representatives and also other types of events such as the death of Grand Duke Jean or the tornado that struck the south of Luxembourg in August 2019. COVID-19 is another event that really concerns every aspect of life of every citizen in Luxembourg.

THE SCOPE OF THE COVID-19 COLLECTION

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

Ben Els: We tried to concentrate on websites, because social media platforms raise a lot of technical difficulties and they are also more expensive to harvest in term of data budget, with often unreliable results. So we have included some Facebook pages for example, but we haven't saved a lot of social media data; we have archived much more news media, websites and Twitter, which can be archived much more effectively than other social media.

Why does it cost more to archive social media?

Ben Els: Facebook actively tries to block crawler robots and as a result you have to collect much more data to obtain reliable results. So, for the cost of one capture on Facebook you can archive a regular website many times and the capture of Facebook may even be unusable, with videos not working, for example.

How did you select the content you archived, the websites, the profiles?

Ben Els: We tried to be as inclusive as possible. The subjects most closely related to the pandemic include medical care and national protection measures as well as the development of the situation, communication via the official government website, the websites of the different ministries, laboratories, medical research, etc. The next subjects to look at were the lockdown, working from home, home schooling, cultural activities at home and holidays at home. There is also the international situation, neighbouring countries, etc. For example, we archived Luxembourg news media in Portuguese, like *Contacto* (<https://www.wort.lu/pt>) and *Jornal do Luxemburgo*, and we also captured articles related to the COVID-19 situation in Portugal and in our neighbouring countries. Since foreign communities are a large part of the Luxembourg population, it is also important to keep abreast

of developments in the Greater Region and in countries that many residents are going to travel to.

How did you make the selection?

Ben Els: Between mid-March and mid-July I selected articles individually from all the Luxembourg news websites. This constitutes a collection of just over 26,000 articles. Then from the beginning of June we requested the daily capture of each news media website from the Internet Archive. Previously this was only done twice a year, but now we will continue to crawl these 55 sites on a daily basis for at least a year. Obviously, between my first selection involving precise captures of targeted articles and the daily collection of media news websites, the results are different. The possibilities for analysis are different depending on whether I have a WARC that covers the whole website or one per article. I also added live feeds, such as the radio station 100.7, the *Luxemburger Wort* and RTL.lu. Since these feeds were displaying new content in a dynamic manner, they had to be crawled manually – it could take up to an hour to scroll and click through 8 different news feeds each day. At the end we moved to just once a week. We also have other languages, Italian websites, a Serbian radio channel, but it's harder to read and select what's relevant. Media websites are the central part of the collection and by following the media every day we discovered new initiatives, a new solidarity platform or a site to register for testing, etc. We also worked with the Government IT Centre (Centre des technologies de l'information de l'Etat – CTIE) to identify the government's main channels, because they changed all the usual information and communication channels: normally everything is on gouvernement.lu but they put it all on guichet.lu.

We also launched a call for participation at the end of March and there too we got a very good response from the media and a few communities that we would not have otherwise thought of suggested their websites.

Can you give me an example?

Ben Els: For example, the Muslim community and shoura.lu. I hadn't thought of looking for religious communities. The Muslim community posted online information and recommendations for its members about services in mosques, religious holidays, etc. Based on this suggestion, we then looked more closely at other religious communities. It is important to capture the minority viewpoint as well and I have also included sites for border residents, for example the site frontaliers.lu, because they face different problems from residents. I have also included a lot of information from trade unions about remote working and home schooling, teachers' unions which are opposed to the policy of the Ministry of Education, parents who did not agree to the reopening of schools, etc., and the views of political parties and the opposition.



Bibliothèque nationale du Luxembourg
Nationalbibliothék

WHAT WE DO
HOW IT WORKS
WHAT WE HAVE
FAQ
DICTIONARY
CONTACT


Collection du Covid-19 : La Bibliothèque nationale a besoin de vous tous pour compléter sa collection



La BnL est en train de rassembler une collection de toutes les informations pertinentes publiées sur la propagation et l'impact du Coronavirus au Luxembourg. Pour cette raison, dans le cadre de sa mission d'archivage du web luxembourgeois, la BnL aimerait demander votre support, afin de capturer tous les aspects de l'épidémie Covid-19 dans notre pays.

La liste des URLs de tous les articles moissonnés sera mise à disposition en ligne, de la même façon que nous présentons déjà d'autres collections ciblées sur webarchive.lu.


Dans la situation sans précédent à laquelle la pandémie Covid 19 nous confronte, de nouvelles initiatives de solidarité et le besoin d'information publique ont mené à la création de nouveaux sites web et plateformes en ligne, répondant aux besoins d'une société cherchant à s'adapter à cette crise. Afin de documenter ces développements pour les générations futures, nous vous demandons de nous communiquer les références des sites web informatifs, pages Facebook et Twitter traitant la thématique du Covid-19 dont vous avez connaissance.





Bibliothèque nationale du Luxembourg
Nationalbibliothék

Période Covid-19
Mardi à vendredi: 10:00 - 20:00
Samedi: 10:00 - 18:00

RECHERCHER
INFOS PRATIQUES
DÉCOUVRIR LA BNL
SERVICES AUX PROFESSIONNELS




Actualités


Recherche sur a-z.lu

Webarchive of Covid-19 : the National Library needs your help to complete its collection

30/03/2020

The National Library of Luxembourg is building a collection on all relevant information on the spread of the Coronavirus in Luxembourg. Naturally, online news media are an important part of this collection. Therefore, we would like to ask for your help in capturing all aspects in relation with the Covid-19 outbreak:

Version française

RENSEIGNEMENTS ET CONSEILS À DISTANCE
Informations sur l'inscription en ligne:
inscription@bnl.etat.lu

Figure 1: Web-archive of COVID-19, a call for participation on the website of the BnL. ©BnL

THE FRAME OF THIS SPECIAL COLLECTION

That represents a lot of websites...

Ben Els: There are 540 websites; 145 were collected weekly and 60 were collected daily, including the government website, the Statec website for statistics and hospital websites. In parallel I tried to find their Twitter accounts and we also carried out a very broad crawl at the beginning of April 2020, and we will start again at the beginning of August. We will do a final crawl in December, so that represents a capture about every 4 months during the crisis to complete the picture obtained via the special collection. For example, I chose not to harvest all the websites of the individual communes, as there are 120 of them and it would have blown the data budget, but the broad crawls give us these websites too. Sometimes a clothing brand launches a corona-themed garment and I collect their website once. The whole thing represents 1.3 terabytes on Archive-It, or 11 terabytes including the broad crawls, because all websites have been affected in one way or another by the epidemic.

All this collection is based on meticulous human curation. This would have been difficult to achieve without precise knowledge of the Luxembourg web...

Ben Els: Yes, indeed. It would have been interesting to launch a second call for participation, and maybe we'll see in the autumn. The Luxemburger Wort published an article on the COVID memory initiative launched at the University of Luxembourg (covidmemory.lu) and in the article a small insert also mentioned the BnL. People have sent us their poems, short stories and blogs. People like to share what they have created during the lockdown. For the poems we redirected them to COVID memory or the National Literature Centre (CNL) in Mersch.

When did you start? When do you plan to stop?

Ben Els: We started on 16 March 2020, and it is difficult to say when we will stop. It is more of a budgetary issue.

Do you have an annual budget?

Ben Els: Yes. I have a lot of terabytes at my disposal, but we won't necessarily put 5 terabytes more into it. I was hoping that it would calm down in June, but it has started again and we are at the same pace as in March. With the individual articles that I collected from media websites we can compare the number of articles and relate it to the number of cases recorded.

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

Ben Els: We don't really have precise quality control, the only control in Archive-It is when I run a crawl test and if there are too many or too few results, I know something strange is going on. But I have no intention of checking all the results. The crawls are done by Archive-It; you can run and set up crawls, and quality control is also performed remotely.

What are the issues, challenges and limits of the collection process?

Ben Els: We really need more staff to get better coverage. As the choice of what to collect is decided by one person, the results remain subjective. If we had three people, for example, we would probably have three times as many ideas. If the web archive was better known, there would also be more responses from editors and bloggers, there would be more submissions, without us having to ask. Here too, more efforts are needed to raise awareness about web archives.

Another important point concerns the budget: as we don't know when this collection will be finished, there are questions about the distribution of the data budget. It was hard to know whether I was spending too much at the beginning.

And what do you particularly take pride in?

Ben Els: I think the large number of news media articles archived represents a lot of work and a major achievement. I think we also did a good job selecting websites and this is our best list because we were able to build on the lists from previous years. It seems to me that we have achieved a very good overview of sites of national importance.

In WARCnet we plan to work on Women, Gender and COVID.

Ben Els: On that topic, the head of "CID Fraen an Gender"² called me and we had a useful discussion. I asked her to send me a list of websites to collect, which will also come in handy for next time if we launch a gender collection in our web archives.

This COVID-19 collection experience is also useful for the future: we plan to find national partners for thematic collections in order to expand our lists and collections. The list is still short, but we are receiving help from the CTIE for government websites and from the CNA (National Audiovisual Centre) and the CNL on Luxembourg authors and publishers, and we intend to go deeper.

We have always followed events; in years to come we intend to focus more on thematic collections, but events help to create the thematic map and COVID-19 has touched on so

2. An organisation dedicated to feminism, gender issues, etc. <https://cid-fg.lu>

many different topics! We have found many associations of interest to the community. This will help provide us with a good basis for our thematic collections, which will grow over time. This reminds me of another interesting aspect that I have not yet mentioned: the case of public petitions submitted to the Chamber of Deputies. Anyone can submit a public petition with a subject to be discussed in the Chamber, and if the petition gets enough signatures, members must invite the petitioner to present the subject. It is rare for petitions to reach this stage but during the crisis there was a surge in proposals. About twice as many petitions as usual were proposed during the early months of the crisis, almost all of them related to the crisis (opposing the government, seeking a bonus for healthcare workers, etc.). The composer and musician Serge Tonnar, for example, did a lot to encourage support for the cultural sector, to change the political direction, etc. I conducted an interview with him that I will put on our website. We discussed how he experienced the lockdown, and the consequences for the cultural sector. He created a concert channel on Facebook, which became kuk.lu. He's a prominent voice on social media.

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

Ben Els: It didn't really involve any changes compared to before. Video conferencing worked well. Our team is very small and for crawls the Archive-It platform is accessible from home. A lot of things that seemed impossible before have now become possible.

ACCESSIBILITY AND SEARCHABILITY

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

Ben Els: The archive is accessible at the BnL, but on the webarchive.lu website we try to give information and starting points, to outline the background of the collection and explain the processes to people who are interested.

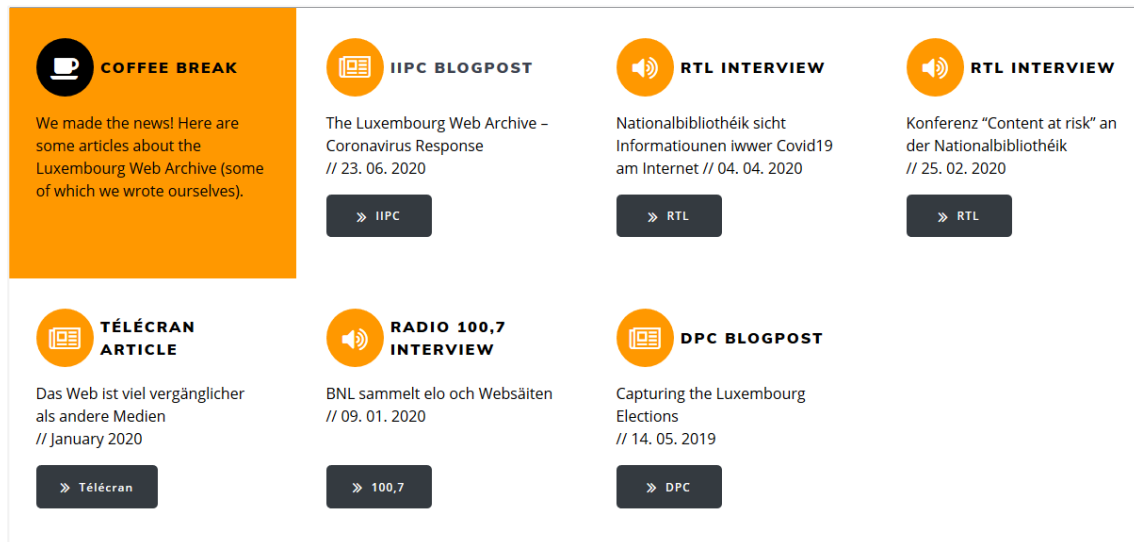


Figure 2: Information on the COVID-19 collection on webarchive.lu. ©BnL

We also intend to provide lists that can be used as a starting point. I would also like to make interactive statistics available for this collection and other event collections. We will be adding metadata, such as the title of the website, the area of interest and a few keywords. Often websites are identified by an acronym followed by .lu and this will allow users to find information more easily and to see the different aspects that are represented in our choice of websites.

PARTNERSHIPS AND USES

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

Ben Els: WARCnet and your team is the first academic partnership on the horizon!

How did you raise awareness about this special collection?

Ben Els: Through the website webarchive.lu, which was also used as a platform for the call for participation. We have received a lot of interest from the media. We have been invited by several radio stations and the Luxemburger Wort will also feature a story about the Luxembourg Web Archive in the coming months. In general, when we explain the principle of the web archive people understand it, but the question still remains: why do we have to archive all this material? And in the context of COVID-19 it becomes much more obvious to everyone that this is a moment in history that has changed everything, changed our online life, affected the web, and it helps us get the message across.

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

Ben Els: We are a member of the IIPC. The IIPC launched a collaborative collection and we submitted our URLs. They recaptured them. I don't know the exact number of websites they collected, but there are tens of thousands. So, we have a good picture from a global point of view.




Figure 3: The BnL COVID-19 collection on Archive-It.

That's an excellent transition to my last question. How do you archive nationally something which is fundamentally global?

Ben Els: Basically, our work is linked to the Luxembourg legal deposit. This legal basis gives us the right to archive websites published in Luxembourg, but we are also interested in websites published by Luxembourgers abroad or in relation to Luxembourg. They are harder to find but when Luxembourg became an "at-risk area" a few weeks ago, I archived the pages of neighbouring countries like Germany with their restrictions for Luxembourgers, or the website of the Luxembourg Embassy in the USA.

We would like to thank Ben Els and Yves Maurer (BnL), as well as 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, 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

