

# Exploring special web archives collections related to COVID-19: The case of the Swiss National Library

Friedel Geeraert  
and Barbara Signori

WARCNET PAPERS

WARCnet  
web archive studies

# Exploring special web archives collections related to COVID-19: The case of the Swiss National Library

*An interview with Barbara Signori (Swiss National Library)  
conducted by Friedel Geeraert (KBR)*

Friedel.Geeraert@kbr.be



WARCnet Papers  
Aarhus, Denmark 2020

WARCnet Papers ISSN 2597-0615.

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

© 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: 978-87-972198-9-8

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* (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)

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)

All WARCnet Papers can be downloaded for free from the project website [warcnet.eu](http://warcnet.eu).

# Exploring special web archives collections related to COVID-19: The case of the Swiss National Library

*An interview with Barbara Signori (Swiss National Library)  
conducted by Friedel Geeraert (KBR)*

*Abstract: This WARCnet paper is part of 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, COVID-19, special collections, Switzerland, Web Archive Switzerland, Swiss National Library*

This WARCnet paper is part of a series of interviews with European web archivists who have been involved in special collections related to COVID-19. This interview was conducted on 14 September 2020 with Barbara Signori, Head of e-Helvetica at the Swiss National Library.

Web archiving at the Swiss National Library started in 2008. Switzerland is one of the few countries without legal deposit legislation. Therefore, the scope of the web archive collections is selective and permission from the owners of websites are sought in advance following the fair use approach, before the content is crawled. The general selection criteria stipulate that websites need to be: published on the web; not protected by passwords; free of charge; refer to the Cantons and Switzerland; treat topics of historical, social, political, cultural, religious, scientific or economic importance to the Cantons or Switzerland; and are of interest to a broad public. Social media are currently excluded from the collections. (Swiss National Library, 2019) The Swiss National Library works with a number of partners comprising specialist libraries and cantonal libraries who nominate websites related to their fields of expertise or Cantons. The collections can be consulted in the reading rooms of the Swiss National Library as well as in the reading rooms of a number of partner institutions across Switzerland.

## THE REASONS OF THE SPECIAL COLLECTION

*Why did you create a special COVID-19 collection?*

Barbara Signori: We have been collecting the Swiss web since 2008. Next to the general collection of websites related to Switzerland we also create web collections for special events. COVID-19 is such a special event. It has an enormous influence on Switzerland as a country and on its inhabitants, that's why it is important to be documented. The special COVID-19 collection is integrated into the general web collection which is called Web Archive Switzerland. Web Archive Switzerland is a collaboration between the Swiss National Library and 30 Swiss institutions, mostly libraries and archives.

Another reason is our engagement with the IIPC - the International Internet Preservation Consortium. The IIPC set up a COVID-19 collaborative collection and that is also why we started to identify and to select websites even before COVID-19 became a threat for Europe. At the beginning when COVID-19 started in China, the IIPC, which is an international consortium with members all over the world and also in Asia, already started to collect websites. Members were asked to contribute websites related to COVID-19 in their respective countries - so we started identifying Swiss websites already in February 2020.

When we realised that we were facing a worldwide pandemic, we also used these websites we had already identified for our own archive. We don't always integrate the websites we contribute to the IIPC for other collaborative collections into our archive, but in this case it was clear that it also would have an impact on Switzerland and that it was important to create our own special collection.

*You already talked about the partnerships you have with the regional libraries and your participation in the IIPC collaborative collection but did you set up any other partnerships for this particular collection?*

Barbara Signori: The partnerships with the cantonal libraries, the other Web Archive Switzerland members and the IIPC are the only ones we have set up for the COVID-19 web collection so far.

## THE SCOPE OF THE COVID-19 COLLECTION

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

Barbara Signori: For the special collection on COVID-19 within the Web Archive Switzerland, we focused on entire websites. For legal reasons we are not allowed to add content from social media platforms to our web archive. We identified two categories of websites we wanted to include in this special collection. After doing some research, we realised that there are websites that have been specifically created because of COVID-19.

They belong to the first category. These websites didn't exist before and we call them the '100% COVID-19 websites'. Obviously we wanted to collect all these websites that have been published in Switzerland or contain Swiss content.

The second category consists of websites that we already have in our archive and that are now also reporting about COVID-19. For example the websites of many of the institutions we already have in our web archive such as hospitals, universities, camping sites, restaurants etc. published specific information about COVID-19. For us, it was important to collect these websites and to take a snapshot of these websites during the (first) lockdown.

The '100% COVID-19 websites' we collected as soon as we identified them. For the other websites that were already part of our web archive collection, we scheduled additional crawls. We usually take a yearly snapshot. It was very important to recrawl the sites during the lockdown independently of their normal crawling frequency. So we checked the crawl frequency of all websites and identified those that were important to capture during the lockdown, for example the website of the Swiss Federal Office of Public Health - where you can find the most important information to the crisis.

Obviously, we had to make a selection since almost all organisations had COVID-19-related information on their websites, even the website of the Swiss National Library for example, and it simply wasn't possible to crawl all these websites again. That's when we came up with specific selection categories. We identified categories such as culture, everyday life in Switzerland, news associations, volunteering, healthcare institutions, employee organisations, administration, experts, sports, education, professional organisations, politics, agriculture etc. All of these themes are important for COVID-19. We tried to match or identify institutions that aligned with these themes and then extended this across the regions of Switzerland in order to get a good geographical representation of important information about COVID-19.

*Do you also harvest national news channels?*

Barbara Signori: We do include some news websites but not specifically for the COVID-19 collection. We focused on key organisations and on initiatives that have been launched because of COVID-19.

We did contribute some Swiss news sites to the IIPC collection, but even the IIPC stopped collecting news pages at a certain point as there was just too much content.

*Since you work with a distributed model of selection, did the partner institutions also nominate the websites they thought would be most pertinent for the special COVID-19 collection for their Canton or was this decided centrally?*

Barbara Signori: We did both. We knew there wasn't much time to react and coordinating a collaborative collection takes time. So we decided to start on our own in order to move forward quickly. At the same time, we asked our partners to provide suggestions for websites in their Cantons or websites related to their specialty that are important and related

to COVID-19. Eleven partners contributed COVID-19 websites. We included all those suggestions. It was really a mix of an external and internal selection. The goal was to have a good overview because it was clear that we couldn't be comprehensive and crawl every single Swiss website. We had to be selective but we also wanted to cover all the languages in Switzerland, all the regions in Switzerland and all the major topics and institutions that are involved in this crisis.

*You work with a web form to collect seed URLs. Is that only used by the partner institutions or may the public also recommend websites?*

Barbara Signori: The form is only used by the partner institutions to suggest websites. Metadata is added on the form: the URL, the title of the website, the language, the Dewey Decimal Classification, the organisation etc. We reuse the metadata to create a bibliographic record. All collected websites, including the COVID-19 websites, are catalogued and can be consulted in our library catalogue. We actually added a specific controlled subject heading for COVID-19 so that the collection can be identified within the library catalogue as well.

*Did you focus on complete websites or were there also cases where you only took a section of a website or a single web page?*

Barbara Signori: We always focus on entire websites. Only if a website is too big to be crawled or archived do we make selections in language or a single section– but these are exceptions.

*How do you archive nationally something which is fundamentally global?*

Barbara Signori: That's what we always do. You could say that web archiving in general is global because there are no barriers on the Internet, there are no frontiers on the Internet. We are crawling content that is related to Switzerland. It's the relation and origin that counts. For books, it would be books by Swiss publishers and Swiss authors. For the web it's similar: for example Swiss organisations and institutions with their headquarters in Switzerland or Swiss researchers who publish something on the web. If a Swiss hospital or researcher has done research on a COVID-19 vaccine, we're going to collect that information. For tracing apps as well, Swiss researchers were very early developers. We assess the relationship to Switzerland.

*So you don't necessarily stay within the .ch domain?*

Barbara Signori: No, we don't because a lot of Swiss content is on other domains as well.

*Do you also crawl content in other languages than German, Italian and French?*

Barbara Signori: Yes, as long as the content is related to Switzerland, we include other languages. It's really the content that is important and not where it's published or in which language it's published. For example, in Geneva, there are lots of international organisations, with websites in many languages which we all collect. We only exclude languages from the collection if the crawls are too big to be archived and we try to down size. But in general we try to collect websites as comprehensively as possible.

*How much data was collected and what was the nature of the data?*

Barbara Signori: We have a relatively small web archiving system and our approach is very selective. For this special collection we have identified almost 500 websites up to date. More than 100 of these websites were just created for this event. These are the '100% COVID-19 websites'. The rest are the websites that are already part of our web archive and we identified as important to be captured again during the lockdown. We will do a second crawl of these '100% COVID-19 websites' later this autumn as well. We will break our rule to have only one crawl per year, and crawl these websites more often, as long as they still exist. It also depends on the content of course: some of these COVID-19 websites are maybe not being updated anymore or might not even exist anymore. So we have to check the collection and then decide which websites we are going to harvest again. This is manual work we are doing. And as long as the crisis goes on we will continue selecting and identifying new websites.

As for formats, we have the regular formats in the web archive: mostly HTML, PDFs, JPEG, TIFF, CSS, JavaScript, etc. We try to harvest everything that's on a website as long as it is technically and legally possible. We don't have a legal deposit in Switzerland. Before we harvest any website, and this was also the case for the COVID-19 websites, we ask the website owners if we are allowed to collect the website following the fair use approach. If the owners tell us that they don't want us to archive their websites, we don't collect them. It's an opt-out approach. We also remove snapshots from the web archive when a website owner objects only after the crawling and archiving have already taken place.

*Do you have any idea of the size of the collection in TB?*

Barbara Signori: The collection is about 1.75 TB.

## **THE FRAME OF THIS SPECIAL COLLECTION**

*You already mentioned that you started contributing to the IIPC collaborative collection in February, but when did you start collecting in Switzerland?*

Barbara Signori: Indeed, we started contributing to the IIPC collection mid-February. On March 6<sup>th</sup>, we started to create our own collection and the peak was in March, April and May during the (first) lockdown. The web archiving team was focusing on searching pertinent websites, nominating those websites on our web form and harvesting them.



*Do you know when you will stop collecting?*

Barbara Signori: No, not yet. We are monitoring the situation because the crisis is still ongoing and there are still websites that are important to be harvested. For example if a second wave happens, we are going to collect those websites as well and will continue to do so. I think we will only stop once the crisis is over.

It's important to document what is happening. It's definitely important for the IIPC to do so because the collaborative collection really allows you to obtain a worldwide view. In addition to that, every country is creating their own collection and that's important as well. We try to do our best for Switzerland.

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

Barbara Signori: We do carry out quality control on the collection. This is a manual task and takes a lot of effort and time. Most of it has been done whilst we were working from home during March, April and May.

We have a semi-automated approach. We're using a tool that allows us to compare the home page of the live website with the home page of the harvested website. If we already have a copy in the web archive, we also compare it to the home page of the archived snapshot. The tool checks if the home pages are similar based on a mathematical comparison between pre-selected indicators. The tool shows a green light when a home page looks the same as on the live web. As mentioned, it's only the home page that's compared, not everything that lies underneath. Every website that is lighting up in green, we don't have to check. For every website that is lighting up in orange or red, we know that there are important changes and we need to check manually. We have a list of criteria to check. For example, you try to click through all the levels of the website, check if the images are captured and many other things.

*Do you sometimes crawl websites again?*

Barbara Signori: When the quality is acceptable, the website is added to the archive. If the quality is not acceptable, the person who is doing the check adds a note and it goes back to the crawl engineer who tries to make a better capture. If we manage to do so, we then integrate it in our archive. If the quality stays poor it's the curators who decide if the website is added to the web archive anyhow.

Sometimes we see that there is still content that is worth being archived and that it's better to have this than nothing at all. Especially with the COVID-19 collection that was often the case. We know that it's just impossible for us to technically crawl and archive this website in a decent way. But in order to not lose the information, we archive it once at least. It allows us to keep a trace of history, to show that it existed even if we weren't able to harvest it properly. It may not look like the original website, but at least the content is preserved. It still gives an idea of what existed.

This is what Web Archive Switzerland is all about: to give an idea of how the Swiss web is used right now. At the moment we don't aim to comprehensively collect the Swiss web. We want to give an idea of how the web has been used, how the web was presented and how institutions and people have presented themselves on the web so that future generations can understand how we lived with the web because maybe in the future it won't exist anymore.

## ACCESSIBILITY AND SEARCHABILITY

*Other than consulting the descriptive metadata via the catalogue, can the collection already be accessed or searched?*

Barbara Signori: As for access, we use the e-Helvetica access platform, which is our digital library. The full-text of the websites are indexed, so full-text search is available.

Web Archive Switzerland is only accessible on-site at the Swiss National Library and at our partner libraries on dedicated work stations in the reading rooms where no reproduction is allowed. This is due to the legal reasons that have already been mentioned before. Thanks to the distributed access no one has to travel to Bern to search the Web Archive Switzerland but can go to his or her nearest library or archive. There's a nice geographical spread between the partner libraries: the Northern, Southern, Western and Eastern parts of the countries are covered.

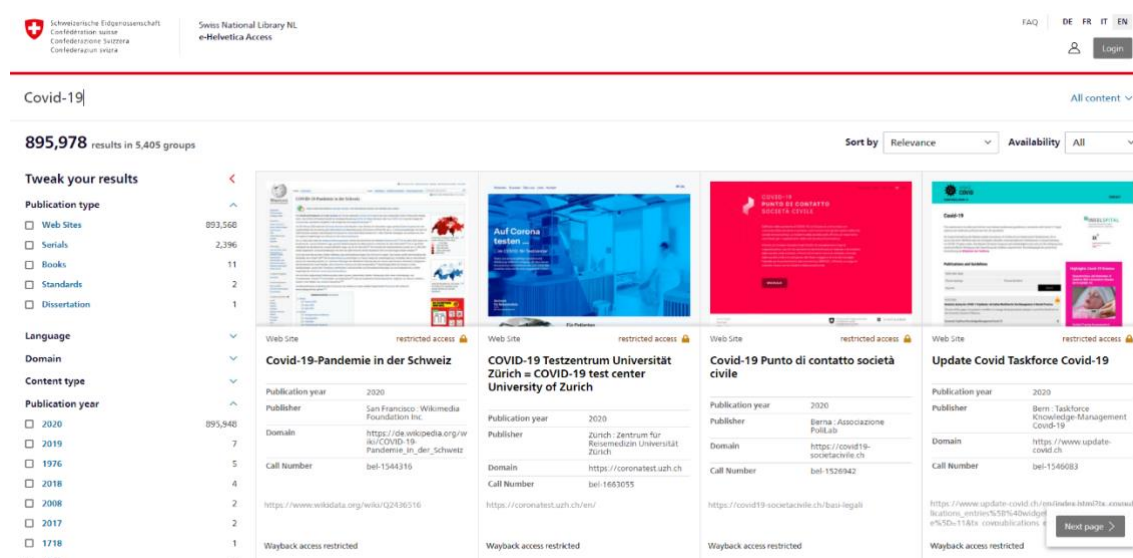


Figure 1: Screenshot of the e-Helvetica interface after query for 'COVID-19'. ©Swiss National Library

*Have you already been approached by researchers who want to use the COVID-19 collection?*

Barbara Signori: No, unfortunately not. At least not that I'm aware of. Everybody can use our collections whether it is the COVID-19 collection or any other digital collection. They don't necessarily need to ask or tell us. Even for Web Archive Switzerland, if someone comes to the library and consults the archive or does research, I don't necessarily know about it. Of course we hope that our web collections are used for research. But we also have to be realistic. Web archives will be important or will be getting more important the older they get. A lot of information, especially COVID-19 information, is still live on the web. With web archiving, we are doing something today that will be very useful and requested in the future.

*How do you communicate about this special collection?*


Barbara Signori: We have communicated about it internally, but we haven't communicated about it externally yet. This interview is the first time where we tell the world about it, which is important and I thank you for this opportunity. COVID-19 was and still is a very challenging time also for libraries. We had to set priorities in external marketing and communication issues.

There are other institutions in Switzerland that have created special platforms for COVID-19. There is for example a website that is collecting stories and memories about COVID-19, which is called Corona-Memory. The special COVID-19 collection of Web Archive Switzerland is one amongst many other collections that are being created around COVID-19.

## REFERENCES

Swiss National Library. (2019). *Merkblatt Sammeln*. Retrieved from <https://www.nb.admin.ch/snl/en/home/information-professionals/e-helvetica/web-archive-switzerland.html>.

*We would like to thank Julie M. Birkholz (KBR and Ghent University) 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

