

Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative collection

Friedel Geeraert and
Nicola Bingham

WARCNET PAPERS

WARCnet
web archive studies

Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative collection

*An interview with Nicola Bingham (British Library) conducted by
Friedel Geeraert (KBR)*

Friedel.Geeraert@kbr.be



WARCnet Papers
Aarhus, Denmark 2020

WARCnet Papers ISSN 2597-0615.

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

© 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: 978-87-972198-8-1

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)

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.

Exploring special web archives collections related to COVID-19: The case of the IIPC Collaborative collection

An interview with Nicola Bingham (British 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, social networks, COVID-19, special collections, IIPC, International Internet Preservation Consortium

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 interview was conducted on 3 September 2020 with Nicola Bingham, Lead Curator Web Archives at the British Library.

The International Internet Preservation Consortium was founded in 2003 by 12 founding members. The mission of the organisation is to 'acquire, preserve and make accessible knowledge and information from the Internet for future generations everywhere, promoting global exchange and international relations'. The IIPC members currently span 45 countries and comprise national, regional and university libraries and archives. (IIPC, 2020a)

One of the Working Groups within IIPC is the Content Development Group (CDG). This group was created in 2014 to create a formal collaborative collection framework for the Consortium, following on previous less formal projects carried out by the predecessor Access Working Group to collect content for the 2010-2014 Olympic games. In 2015 the CDG, led by founding co-chairs Abbie Grotke (Library of Congress) and Alex Thurman (Columbia University Libraries) initiated its subscription with Archive-It and began building thematic collections on World War I Commemoration, the European Refugee Crisis, Intergovernmental Organisations, and the 2016 Rio Olympics. By 2017 Nicola Bingham had succeeded Abbie as a CDG co-chair, and subsequent new collections included Online

News Around the World, Artificial Intelligence, and Climate Change. The Novel Coronavirus (COVID-19) collection, with Nicola and Alex as lead curators, has been the CDG's focus in 2020. All the collaborative collections (including the pre-CDG Olympics collections) can be consulted via [Archive-It](#).

The themes or events that constitute the focus of a collaborative collection need to meet a number of criteria. The theme or event needs to be transnational in scope and broader than the responsibility or mandate of any one member. It also needs to be of high interest to IIPC members and the broader perspective provided by multiple contributing institutions should constitute an added value for research. (IIPC, 2020b)

THE REASONS OF THE SPECIAL COLLECTION

Why did you create a special COVID-19 collection?

Nicola Bingham: Collecting around events of global significance is very much part of the content development policy for the CDG, the Content Development Group. We pursue collaborative collections based on themes or events, if they meet certain criteria: if they are of high interest to the IIPC members, if the global events don't map to any one member's responsibility or mandate and if the subject is going to be of high value to researchers. Basically, the content development group gives members the opportunity to represent perspectives other than their own country or region.

In planning collections, the IIPC Content Development Group has a data budget, which we work under on an annual basis. Coming towards the end of the calendar year, we'll start to plan the next year's collections that the CDG has got to develop. We've got a strong timeline of archiving events, such as the Olympic games. We had plans to archive the 2020 Olympic games in Tokyo, but just at the end of 2019 and early in 2020, when we were starting to confirm the collections to focus on, the pandemic started to emerge. Members expressed an interest in collecting around the pandemic.

Around about the same time in February of this year, Archive-It, who we crawl with, had offered to increase our data budget. We'd agreed at the end of 2019 to a data budget of three terabytes to use for the coming year, but Archive-It offered to further support this collection by allocating us 2 additional terabytes of data for free, so we've had a total of 5 terabytes available in 2020 for our collecting.

The COVID-19 collection was very much in scope of our content development strategy. It fit the criteria of having a global interest. We'd had quite a few expressions of interest from our members in collecting around this event.

THE ORGANISATION OF THE COLLECTING PROCESS

How do you go about organising a transnational collection?

Nicola Bingham: We have an established cohort of very engaged members who actively select content for collections. I co-chair the content development group with Alex Thurman of Columbia University Libraries. The first suggestion of a possible CDG COVID-19 collection came from Jefferson Bailey, following which, Alex and I communicated the idea of the collection to our members' mailing list to gain a consensus to develop the collection.

Because we received a very positive response to this, we went ahead and made available the practical method of collecting the nominations. We've had some experience of this for other collections, but what we do is that we make available a simple Google sheet. This enables anybody who has the access credentials to nominate seeds and to add metadata, according to the fields that we've set out in the spreadsheet.

For this collection, we had a Google sheet for members of the CDG and we had a separate form for public nominations. The members were able to contribute directly to the Google sheet and complete it. They were also able to see nominations that other members had submitted. For the public nominations, we set up a Google form. Members of the public were nominating seeds one by one, meaning one seed per form. The form fed into a Google spreadsheet, which Alex and I picked up and moderated.

As mentioned, the basic metadata is also submitted such as title, language, description, top-level domain name and we also ask members to specify the crawl scope as well. The nominations on the sheets are overseen by myself and Alex and we manually review every seed before adding it to Archive-It to be crawled.

In reviewing the seeds, we look at the relevance to the collection and we also look at the compliance of the URL. For example, if the URL is malformed or if it's missing a trailing slash at the end, we edit it so that the URL is formed correctly. We also check for duplication as well. This was quite a big job because we did review every seed and edited and supplemented the metadata as appropriate. From the sheets, we took the seeds and we added them in batches into Archive-It and then crawled the seeds through the Archive-It account.

How did you moderate the collection, knowing that the collection comprised many different languages?

Nicola Bingham: Yes, that was quite a challenge. I was picking up websites in lots of different languages. Particularly in case of websites with non-Roman scripts that I couldn't interpret it was challenging. In some cases dropping a URL into a browser and having a look will tell you if the content is relevant because you can pick up visual clues. You can look at images that tell you that the content is related to COVID-19.

In other cases, I was just a little bit lost. We did make use of Google Translate. So we would drop particular pieces of text into their tool to ensure that a website is relevant. In some cases we'd had to ask colleagues to help us out. Through the IIPC and through

networks in our own institutions, we cover a lot of languages. That was something that we did make use of: to contact a friendly colleague and say: 'Please, can you tell me if this is a website talking about coronavirus?'

Did you impose any limits on participating institutions regarding types of content, maximum number of seeds etc.?

Nicola Bingham: Yes, we do. We've got a general policy of placing scoping rules to the seeds when we crawl them with Archive-It. We can apply data limits at the level of the seeds, depending on the crawl scope. So for example, if we wanted to scope in just one page which might be applicable to an online article or a document such as a PDF, we apply the one-page scope. Against that, we would typically set a limit of a two gigabytes crawl. If we wanted to do what is known as a standard crawl which scopes in the full website or host, we would probably use a slightly bigger limit of three gigabytes or perhaps five gigabytes. The other crawl scope that we use in Archive-It is called 'one page plus'. This allows you to crawl a particular page and then also external links from that page. Again, we might apply a three gigabyte or five gigabyte limit to that. So we apply scoping on a per seed basis. We also apply limits in terms of the number of documents that are retrieved from a particular seed. This enables us to pick up a crawler trap, for example.

We initially didn't apply any limits to the number of seeds that any particular institution or region could nominate, but we did find quite early on that we had overwhelming nominations from some agencies. For example, one of our colleagues had implemented an automatic retrieval of content that was related to coronavirus. This returned a couple of thousand seeds. This is brilliant but due to the data limits and the fact that we want to represent all regions and countries, we did some filtering based on that list of a thousand seeds. We also went back to the person who nominated those seeds and asked if they wouldn't mind being a little bit more selective. It's quite a challenging message really to get the balance between encouraging people to nominate content and then telling people that we need them to focus a little bit more because we don't want to overwhelm the collection with a bias to one particular region or one particular language.

One of the things that we can do relatively easily is if, for example, one member has nominated individual news articles as different rows on the spreadsheet, we can ask them to make the seed reflect the starting point for that subject within the news articles. For example, some newspapers might have a science or medicine subsection to them. In these cases, we would define that as the starting seed, rather than defining every single article underneath.

Something we did review was the relevance of the seeds. In some cases we found that the website that had been nominated, was a generic website. For example, a local authority website, which may at some point have had coronavirus content on its home page but that at the point of archiving doesn't necessarily have that content. In that case, we may have to review the seed and point it to a more specific point in the website with relevant content.

THE SCOPE OF THE COLLECTION

How many seeds are included in the collection?

Nicola Bingham: Over 10,000 seeds have been nominated and crawled, with over 9,200 so far captured well enough to be added to the public collection.

Did you manage to stay within your data budget with this large amount of seeds?

Nicola Bingham: I think at the moment we're at 3.3 terabytes. So we are still within the data budget. We do scope out video content and social media content. This was partly because we had concerns about the data budget.

Were you looking for specific kinds of data in the collection?

Nicola Bingham: We put together guidelines for our curators. We specified that our high priority subtopics would be the origins of the coronavirus, information about the spread of infection, regional or local containment efforts, the medical or scientific aspects, the social, economic and political aspects. We also prioritised published information resources rather than social media feeds or hashtags for this collection.

How many institutions participated in creating this seed list? How many countries and languages are represented?

Nicola Bingham: We had over 30 IIPC members that contributed and we had public nominations from over a hundred individuals or institutions. In terms of languages, we represent 51 languages. This can be viewed on the Archive-It collection page. You can list the languages in the order of the prevalence of the language. Most seeds contain content in English and then Portuguese, Spanish, Japanese, French, German and Dutch. Then there's also a long list of languages going down, some of which I'd not even heard of before. We collected publications from 137 countries.

If I understood correctly, social media content and video were excluded, so the focus was primarily on websites?

Nicola Bingham: Indeed, that's correct.

THE FRAME OF THE COLLECTION

Can you give more insight into the size of the collection?

Nicola Bingham: As said we have 3.3 terabyte for now. It comprises 21.600.104 documents such as HTML, Word files, PDFs etc.

When did you start crawling and what was the capture frequency?

Nicola Bingham: This collection started a little bit earlier than the UK Web Archive collection. We started in the first week of February calling for content and promoting our intention to crawl around this topic. The first crawl was on the 21st of February. At this point in time, the collection is still ongoing. We do have data budget left, but the number of members regularly nominating sites has dropped off a bit during the summer

I think we're up to date with the nominated content. So it might be that we decide to re-crawl some of the seeds that we've already crawled that may have been updated, or we do a final call for content.

The crawl scope is indicated by the person who nominates the content. This describes the way in which we'll treat the seeds: whether it's a one-page crawl or whether it's a whole domain crawl. The crawl frequency on the other hand is determined by the people doing the work in Archive-It. In this case, that was myself and Alex. It was quite a difficult balancing act. The situation is evolving constantly, we had a data budget and we also didn't know how many nominations we might receive.

How did you carry out quality control on the collection?

Nicola Bingham: The Archive-It tool has quite a sophisticated QA module. So you can do an official check of individual seeds because there's a link to Wayback that is embedded in the tool. You can see if that seed renders well. But Archive-It also gives detailed crawl reports. You have a high-level summary of whether the crawl has finished. For example, due to limits of documents on the data, you can see how much content has been acquired. You can review error codes to see why a seed might not have been archived. You can review specific problems with individual seeds. It might be that the crawler has pulled in too much data that's outside of your tolerance. In our case, for example, if too much video content has been archived, that might be something that we might want to adjust the crop parameters for.

For content that has been missed by Archive-It, we can run patch crawls to pick up particular missing seeds or items. You can also invoke the Brozzler crawler in Archive-It. We would do this for more dynamic websites that the Heritrix crawler had struggled to get.

We also had support from staff at Cornell University in doing quality assurance. We actually produced a list of about 300 seeds for them, which had known issues. They were assigned to staff at Cornell to look at in-depth and re-crawl the seeds if necessary. That's what I like about these kinds of collaborations: it really makes the web archiving feel like a community and you can easily reach out for help. There's expertise in many different areas.

ACCESSIBILITY AND SEARCHABILITY

How can the collection be accessed or searched?

Nicola Bingham: The collection is already available on the [Archive-It](#) website and the metadata facets can be browsed so you can explore the collection by language or website type. We've categorised websites into different types: media article, government agency, non-profits, universities, medical research, etc. You can also filter by top-level domain or country of publication.

We also promoted the collection on the World Pandemic Research Network. It's a federated network of global agencies that had collected resources about coronavirus. So the Archive-It collection is also linked to from that network.

Have researchers already shown interest in this collection?

Nicola Bingham: So far, we haven't had any requests from specific researchers, just general enquires. We're not making the collection available that way just yet, because we want to finish the collection. We will be able to offer the collection to researchers in that way, but only when we've closed the collection.

There also is a research agreement that somebody would have to complete if they wanted to use the WARC files that we've collected. A series of guidelines is available on the [IIPC website](#).

How do you communicate about this special collection?

Nicola Bingham: We promoted it on the IIPC website and through our main collecting page. We were updating the collection statistics weekly there. We've also used the usual netpreserve Twitter accounts and the social media accounts of our individual member organisations. We've also promoted it at the IIPC virtual General Assembly which we had a couple of months ago for our members. Olga Holownia, the IIPC Programme and Communications Officer, already promoted the collection at a library conference in South-America. We think that there will be lots for us to use in conferences going forward. Certainly within next year's IIPC conference and potentially other journals or conferences as well.

BEST PRACTICES FOR COLLABORATIVE COLLECTIONS

Was there anything in particular that surprised you when working on this collection or anything that you found particularly interesting?

Nicola Bingham: One of the interesting challenges that I came across was an archival challenge. I was reviewing a seed that had been nominated from a member of the public. It was an Armenian website and I couldn't interpret what the website said. So I did a bit of research around the website and put some of the text into Google Translate. I discovered

that this website was purporting the point of view of COVID-denial. It was quite an extreme right-wing website that had an anti-vaccination policy. It was putting forward the viewpoints that the coronavirus was a made up pandemic. The information was presented in a way that was kind of couched in the terms of a news article. It looked like it was quite factual and verified information.

The dilemma was that from the collection managers' point of view, we don't want to filter information or comment on information when we want to represent the whole spectrum of opinions and viewpoints. Because we believe it's for the researcher to interpret the resources, but we do also have a position of responsibility when it comes to public safety. So the decision was taken to not add this website to the collection, because there is a risk that somebody might look at that article and perceive it as a scientific fact and therefore it could potentially cause danger to health. That for me was quite an interesting archivist's challenge. We kept a record that the website had been nominated. We recorded what the content of the website was and why we decided to not include it in the collection.

Is this kind of records accessible for researchers?

Nicola Bingham: It's not made public at the moment. We do create a record within the IIPC. Within the content development group, we do retain the manifest of everything that had been nominated. Researchers could potentially get access to this metadata.

Do you have any advice for people who also wish to set up a transnational collection?

Nicola Bingham: I don't think this is anything that one institution can do on their own. So I think my advice would be to make use of contacts and explore the possibility of building a federated collection in partnership with other institutions, rather than taking all the responsibility on yourself. I don't think you could build a collection this big in isolation. I think my advice is also to think about strategies for sourcing content from areas that are underrepresented in web archiving. In our experience, China, Russia and Africa for example, are underrepresented in web archiving.

What we do in the IIPC is to try and exploit our contacts by reaching out to those individuals or agencies, who might be able to help us to nominate content. For example, the Library of Congress Overseas Office in Rio have made a fantastic contribution to the collection, contributing over 1,000 seeds in total, drawn from all South American countries. I think it's important to ensure that we represent a balance in the collection holistically.

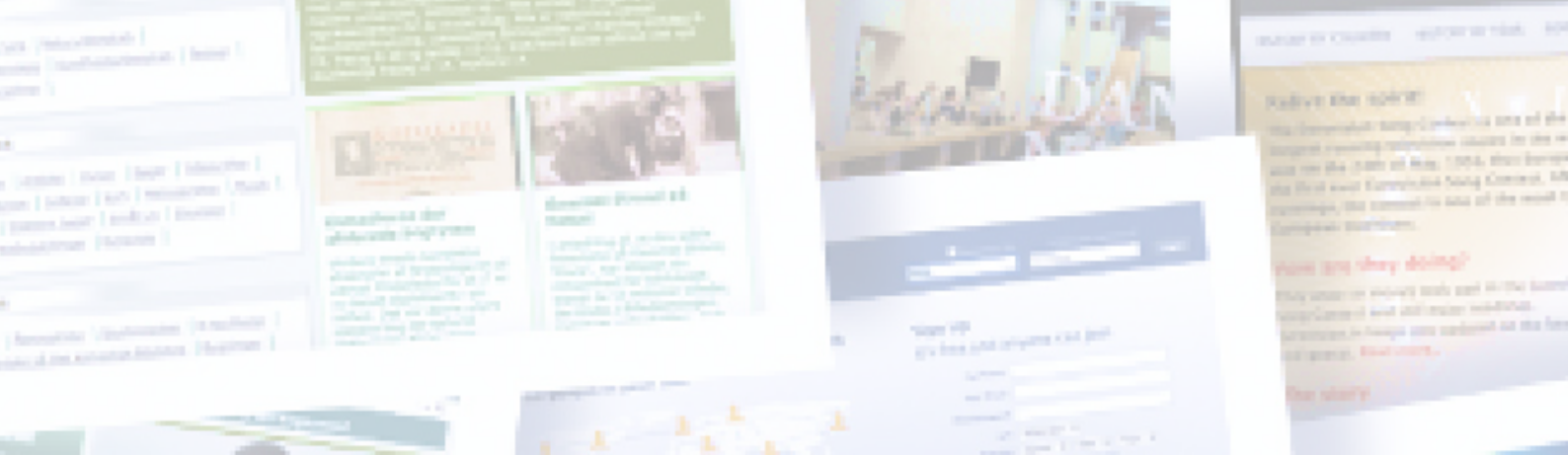
Perhaps something else to think about is, if you want people to nominate seeds for a collection, make this process quite easy and make your expectations as clear as you can. With the COVID-19 collection, we received a lot of nominations and a lot of the nominations were at the level of the website and didn't necessarily contain content relevant to the theme.

REFERENCES

International Internet Preservation Consortium (IIPC). (2020a). *Who is the IIPC?* Retrieved from <https://netpreserve.org/about-us/>.

International Internet Preservation Consortium (IIPC). (2020b). *Collaborative Collections*. Retrieved from <https://netpreserve.org/projects/collaborative-collections/> .

We would like to thank Nicola Bingham (British Library) for her help in proofreading this interview in addition to agreeing to be interviewed. Comments from Alex Thurman (Columbia University Libraries) have also been included.



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



warcnet.eu

warcnet@cc.au.dk

twitter: @WARC_net

facebook: WARCnet

youtube: WARCnet Web Archive Studies

slideshare: WARCnetWebArchiveStu