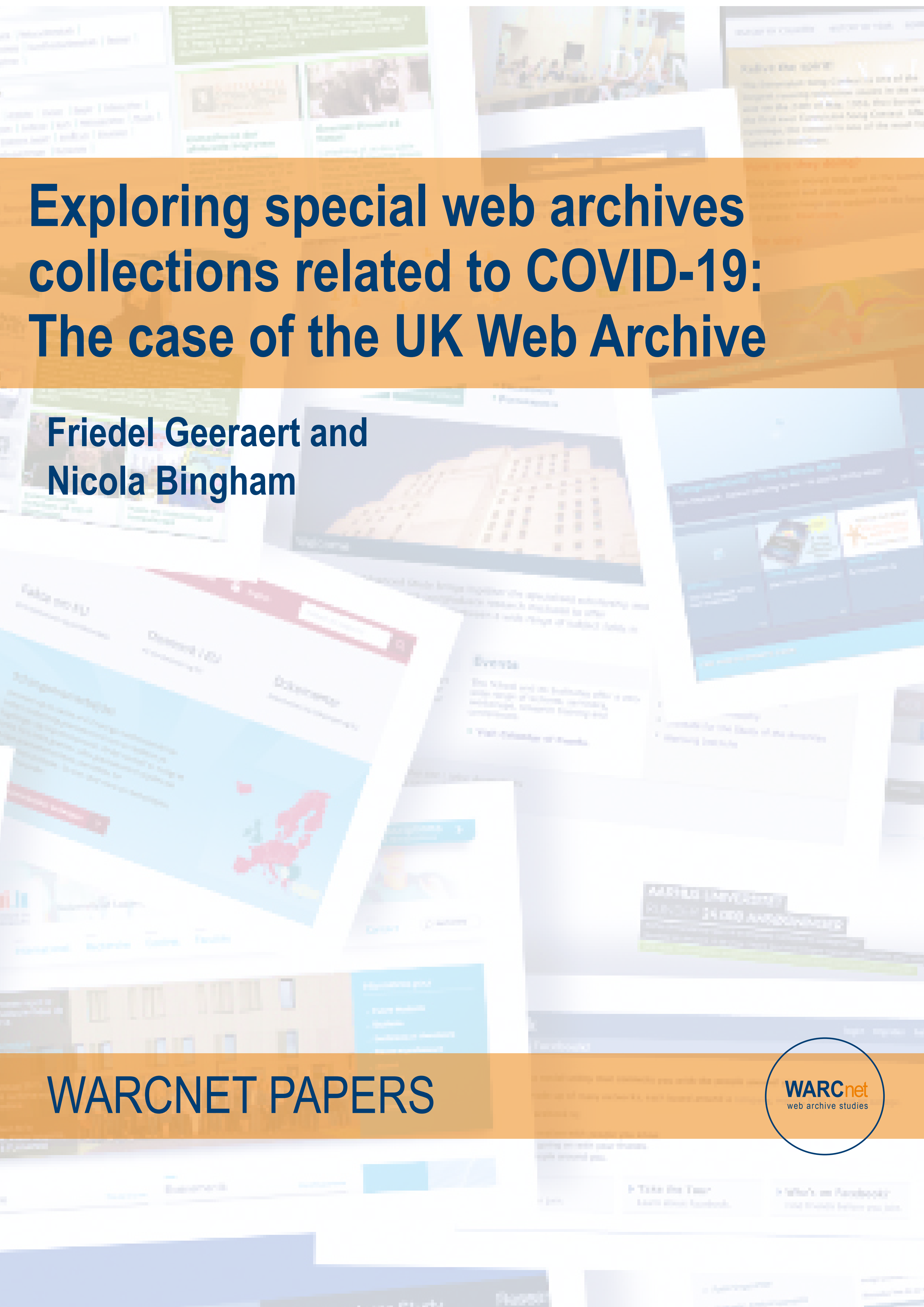

Exploring special web archives collections related to COVID-19: The case of the UK Web Archive

Friedel Geeraert and
Nicola Bingham

WARCNET PAPERS

WARCnet
web archive studies

Exploring special web archives collections related to COVID-19: The case of the UK Web Archive

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 UK Web Archive

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: 978-87-972198-7-4

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

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

DANMARKS FRIE
FORSKNINGSFOND
INDEPENDENT RESEARCH
FUND DENMARK

WARCnet Papers

Niels Brügger: *Welcome to WARCnet* (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)

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 UK Web Archive

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

Web archiving at the British Library started in 2005 on a selective basis. (British Library, 2020) In 2013 the legal deposit legislation was adapted to include non-print works such as websites and other online publications. The UK Legal Deposit Libraries may therefore crawl UK web content without having to ask explicit permission from right-holders. The UK Web Archive is a collaboration between the six legal deposit libraries in the UK: Bodleian Libraries, British Library, Cambridge University Libraries, National Library of Scotland, National Library of Wales and Trinity College in Dublin.

There are two main collection strategies:

- An annual broad crawl comprising all UK websites (starting from April 2013)
- Curated collections about certain themes or events that are captured more frequently (up to daily) (UK Web Archive, 2020)

There are currently over 100 curated collections in the UK Web Archive, one of which is focused on COVID-19. (Webber, 2020) The collections comprise information that is publicly available online. Social media content collection is limited due to technical difficulties in capturing this content. (UK Web Archive, 2020)

The collections can be consulted in the reading rooms of the UK legal deposit libraries although a limited part of the collections is openly available in the UK Web Archive. This concerns web content for which explicit consent was obtained from the website publisher to make it openly available. The metadata for all thematic collections is also made available in the UK Web Archive.

THE REASONS OF THE SPECIAL COLLECTION

Why did you create a special COVID-19 collection?

Nicola Bingham: We have collected around unforeseen events in the past such as pandemics, natural disasters or terrorist attacks as part of our collection development policy. To focus on events such as the COVID-19 pandemic is something that we have a little bit of experience in, meaning this kind of rapid response technique to collecting web material.

We run annual broad domain crawls and they capture the big picture. They're the best way of capturing a snapshot of the web, but they're quite shallow in that we limit each domain to a capture of 500 megabytes. Because the snapshots only run once a year, we naturally we miss a lot of updates to websites. So we miss a lot of UK content from the automated domain crawls. We also only scope in any UK content so that we are compliant with our legislative framework. This means that there's a lot of content, particularly on the .com domain names that is significant and that we miss.

To supplement these crawls, we undertake more frequent in-depth crawls of targeted websites to adequately capture web content in line with our collections development policies and to reflect the UK web space so that it will be a record of our national heritage and be of value to researchers in the future.

We collected around other pandemics, for example the Ebola outbreak in 2014 or the Zika virus outbreak in South Africa. Even though these viruses didn't particularly have a European or a UK perspective to them, we thought that they were of such significance that we focused on the UK response to those disasters. So we did special collections around those. Earlier this year in the first months of 2020, library curators and web archiving staff were becoming more aware of the coronavirus' impact. We monitored the situation because at any one time we have outbreaks of diseases or different natural disasters or unfolding political events that demand our attention as archivists that could potentially form the basis of a special collection.

But it became clear very quickly that due to the global nature of the pandemic, this was going to be something that was going to be of such significance globally, in Western Europe and the UK, that to not collect around it was not an option. So I began to be contacted by curators and external contacts in the library and the archives sector expressing an interest in contributing to this collection. We set it up in March. It started off as a collection in our annotation curation software so that our users could contribute nominations of seeds to the collection.

There’s a bit of an element of a public good in collecting around this subject because it helps to reinforce the value of the national library. We are creating a research collection that can support efforts to understand the social and economic impact of the virus. There is a certainly a role for the Library to play in that researchers of the future might want to mine this collection in managing epidemics in the future.

THE SCOPE OF THE COVID-19 COLLECTION

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

Nicola Bingham: What we collect is in line with our general collection policies. So that is within our legislative framework and the technical limitations. So we predominantly collected websites and the majority of our targets were records at the level of the website, but in some cases we targeted more granular publications. There were some online research articles or news articles that we described as the target belonging to the collection, for example the health or science sections of online newspapers.

We collect a small amount of social media, generally. We're limited, of course, by the technical difficulties in harvesting content, such as Facebook, which is quite locked down. We exclude Facebook content, not from a policy point of view, because we would like to archive selected Facebook content, if it is publicly facing. But because of the technical limitations, we've scoped it out. In terms of Twitter hashtags, we rarely collect hashtags because in asserting the provenance of tweets it's easier for us to focus on an account that clearly belongs to an organisation or an individual. If we scope in hashtags, then there's some uncertainty around who is contributing to the hashtag, for example where the responses are coming from. Because we can't say with any certainty that they're all going to be within our UK scope, we very rarely collect hashtags, although we do make exceptions when a hashtag is clearly related to the UK such as #Brexit.

The UK web archive is made up of the six UK legal deposit libraries. So that's the British Library along with the national libraries of Wales and Scotland, Trinity College in Dublin, Bodleian Library, Oxford and Cambridge University Libraries. We assign the responsibility for collecting the constituent parts of the UK to the different national libraries. So for example, the national library of Scotland will select content that's related to Scotland et cetera.

Our collecting is predominantly focused on English language resources. Within the UK, we do try to represent larger non-English language speaking communities. The Polish community is I think the biggest non-native community in the UK. But to be honest, we are limited in the expertise that we have in selecting non-English languages. So we rely on our network of curators to voluntarily submit content that is not in the English language.

Do you collect any other social media content from Instagram or other platforms?

Nicola Bingham: No, we haven't done it for this collection. In general, we have started using the Webrecorder or Conifer tool to target selected Facebook and Instagram websites but this collection around the COVID-outbreak has been very much a rapid response collection. Using Webrecorder or Conifer is quite resource-intensive. So we haven't had the staff to actually use Webrecorder as well for this collection.

You mentioned that the social media profiles that you collect are linked to individuals and organisations. Could you elaborate on that?

Nicola Bingham: We're lucky in that we have quite a large cohort of selectors. As I said, we work in partnership with the six UK legal deposit libraries and we have quite a few selectors across the British library who select content in line with their own areas of interest or expertise. We have, for example, a curator who is responsible for official publications, we have a news curator and we have staff who work in archives and manuscripts. Within their own areas of expertise, they all have different types of publications or organisations that they are interested in.

Our official publications curator takes responsibility for government publications and national health service publications. We'll have a list of, for example, government departments that she's aware of and she will focus on scanning those websites or those organisations for any relevant content. Our news curator, for example, has contacts with journalists and has relationships with the main news publishers. We have distributed selecting responsibility amongst those curators who have the expertise and the knowledge in those particular areas, as well as partnerships with external agencies such as the Royal College of Nursing.

It's not perfect because we do have gaps and we're aware that we have them. At the British Library, we have a web archiving team of nine people who work full time on web archiving, but it's not our main role, as we see it, to select content. We see our role as facilitating the people who are the content experts to archive content. But this means that anybody who does select web content for us is doing so pretty much on a voluntary basis. So it relies on us engaging with them and them having a level of interest in web archiving.

We know, for example, that traditionally STEM subjects, Science, Technology, Engineering and Medicine are not as well represented in the web archive because we haven’t always had that resource in the Library. As I mentioned, non-English language websites that are based in the UK, also constitute a slight gap for us as well. I think this is something that we talk about a bit in that within archives there’s always been selector bias. But I think being aware of it and trying to be transparent about this to the users is something that we're working on.

We're also doing a lot in trying to contract out selecting. For example, we know that with our, ’BAME’, Black Asian Minority Ethnic communities in the UK, we need to reach out much more to those communities to be informed of the type of content that we should be selecting to properly represent those communities.

You already talked about the partnerships you have with the other UK legal deposit libraries and content experts but did you have any other partnerships with local stakeholders during the collecting process?

Nicola Bingham: Apart from the other legal deposit libraries that you mentioned, we worked with the Royal College of Nursing in the UK. Their archivists have selected content and they've also coordinated with the other Royal Collages of Health and Medicine. The Royal Colleges of Health and Medicine in the UK all have their own specialism. Almost each of the colleges has an archivist, which is great news. Through our contacts, the archivists of the Royal Colleges have been asking their members to submit content to us. That's been really helpful because that way we've been able to commission scientific and medical information. Whereas I have to say the British Library has traditionally had collecting strengths more in the arts and humanities rather than STEM subjects. That subject expertise has been very welcome.

We're also part of the Welcome Trust Network as well so we've had submissions that way. We also work with colleagues at the Public Record Office in Northern Ireland, who've selected content related to Northern Ireland. As part of the UK Web Archive, we also work with Trinity College Dublin who are a legal deposit library but for Northern Ireland we're getting this coverage from the Public Record Office of Northern Ireland as well.

Apart from that, members of the public have contributed nominations. We have a 'save a website' form on our website, which is available for anybody to submit a nomination. These nominations are moderated by the web archiving team and added the collection if they are in scope.

How do you archive nationally something which is fundamentally global?

Nicola Bingham: We focus on how the events have affected the UK by archiving UK-based organisations and individuals. With a collection such as this, each target is assessed on a case by case basis. A curator will justify its inclusion in the collection by determining that it is a UK website. We can define a website as pertaining to the UK if it has a top level domain name that relates to the UK, such as .uk, .scot or .wales. We can also link into a geo-IP database to determine if a website is hosted physically in the UK. We can do those two tests automatically, but there is quite a bit of manual effort to scan a website and determine whether or not it is the website of a UK organisation or if there's a postal address in which case it will be scoped in.

At the same time, we recognise that the researcher might not just have the UK as a focus for their research question. From their point of view, it would be impossible to understand a global event, just from the point of view of one nation or region. We very much intend our collections to complement other collections that we know have been undertaken by heritage institutions and national libraries globally.

We hope that we can point our users to these collections. The Memento protocol is one of the tools that facilitate a federated search across different archives' collections. It is something that is quite challenging to just focus a global event onto the UK. There'll be a lot of global organisations, the Red Cross, the World Health Organisation that have an entirely global focus but are still working in the UK. It’s important for us to represent those websites, in which case we'll probably archive a subsection of that website that's relevant to the UK, or at least highlight that relevant content in the collection.

THE FRAME OF THE COLLECTION

Could you provide more information with regards to the amount of data collected?

Nicola Bingham: We’ve collected over 6,000 targets, predominantly websites, but in some cases we've specified a directory level further down than the main host. Of those 6,000 targets, nearly 600 were Twitter accounts. We haven't archived very much video for reasons related to technical limitations. So unless the video is actually embedded and linked to in an explicit way we’re unable to crawl video without quite a bit of intervention. So we don't have much video content in the collection, but certainly images, HTML pages and documents published on the web such as PDF and word documents.

In other departments, such as oral history, there's been a lot of work collecting around COVID: collecting TV and radio recordings and also in producing content as well. We've had a team of people that have been working on a project to interview health care workers about their experience of the outbreak. There are different formats of content that have been archived elsewhere in the library.

As it stands, one of the challenges that we have is to federate the searching across our different format types. We do have our main online library catalogue, which is called ‘Explore’ which is searchable online. But of course there are backlogs in creating catalogue records for exposing content. There's still one or two different catalogues that are not available within the main catalogue as well. So we're working towards a more federated search.

You already mentioned that you started collecting in early March. When do you plan to stop? What was the capture frequency?

Nicola Bingham: We'll will continue certainly until the end of this calendar year. Of course, the pandemic has been an unfolding event. We can't say when we will stop collecting. It's not like a planned event, such as a general election where you can plan your collecting period. We're certainly going to continue collecting whilst the virus is still active. Of course there are second waves of the virus at various points in the UK and the social and economic effects of the virus and of lockdown are still very relevant. We are concerned to close a collection before it becomes too unfocused. If we are to continue our collecting periods for too long, I think it becomes less useful to the researchers. If it starts to lose focus, because you have events that are happening of which you can say with less certainty that they are related to the outbreak. But of course we archive news publications on an ongoing basis. The main online news publications in the UK are archived on a daily basis. That relevant content is captured anyway.

Do you only take the home pages of these newspapers or do you search for a specific 'coronavirus' tag and then only capture this content?

Nicola Bingham: Yes, that's right. We archive online news publications in their entirety as part of our general collecting activity but for a special collection, a curator might want to surface a particular news article or a section of an online newspaper. In our curation software, they can point the crawler to that particular relevant section of the website and add metadata to it, so that it is tagged as part of the COVID-19/Corona virus collection.

We do impose limitations on selecting news articles for a particular collection because what we've found in the past is that we've overwhelmed our collection with tagging individual news articles. What this does is that we may have a collection that has a total of 4,000 targets and over 3,000 of those targets are individual newspaper articles. What we try to keep in mind is how useful or not this is from the end users' point of view. Once you get to tagging many hundreds of newspaper articles, this becomes quite unwieldy from the end user's point of view.

What about the crawl frequency of for example, websites of organisations? Are these also captured daily?

Nicola Bingham: We tend to base the crawling frequency on a case by case basis. We archive news publications on a daily basis. However, if a website is what we would call self-archiving, we revisit it less frequently. For example, if you have a news article, quite often that content will be edited much more frequently than on a daily basis. A news article might overwritten on an hourly basis. On the other hand, if a website is fairly stable and the legacy content is still searchable on that website, we consider that we have to visit that website less frequently. For example a blog sometimes has an archive that you can search on the website.

It's not a particularly scientific method and it does take a bit of experience to gage how frequently a website will change and how often we should revisit them. But the aim is to get a balance between capturing updated content and not overloading our infrastructure with too much data because there's a cost associated with that. Even though we have deduplication technology embedded in our crawler, the crawler would still detect a small change on a website, for example, a change in the date stamp. This means that the crawler could potentially be getting a lot of duplicated content.

Who sets the capture frequency: the curators or the UK Web Archive team?

Nicola Bingham: We do quite a bit of training with our users about crawl frequencies, but we are starting to feel it may be something that we should handle within the web archiving team to ensure consistency. Because with the best will in the world, though, there is quite a bit of subjectivity involved in setting a crawl frequency. We realise that sometimes our guidelines aren't always clear on this. We're just at the point now in thinking that perhaps the role of the curators is more about selecting the content and being experts on the subject matter and that we (the web archiving team) have a role to play more on the technical side. At the moment our users determine the content that is to be selected and they also determine the crawl frequency.

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

Nicola Bingham: Quality control in this collection has mainly taken the form of service monitoring rather than reviewing individual harvested websites. Because this is a rapid response collection, the effort is focused on crawling targets because there's a time-sensitive nature to this project. We do have alert systems in place to notify us of failures in particular crawl jobs, or, if for example, the crawl had not been written to the storage cluster, although that's not something that has happened. We have service monitoring in real time so this allows us to see the rate of crawling and how much data has been downloaded so we're alerted to any anomalies. Apart from that we do within our curation software have links to Wayback so that curators can check individual web content that they're aware of. We also educate our curators to be particularly aware of technically difficult sites such as rich media websites or dynamic websites. If they are selecting quite a few websites, they may want to prioritise certain websites to review those websites first. Again, this relies on the volunteer efforts of our curators.

In terms of remedial action we can do a few things if we spot problems, for example, adjust the crawl parameters. We can add supplementary seeds or increase the crawl depth or ignore robots.txt. We don't have really high fidelity QA functionality such as patch crawling that is available in the Archive-It tool, but we do some quality assurance work and prioritise those websites that we know are technically difficult to capture. Because we're collecting a lot of content, and we have limited resources we can't investigate every website. It's more the case of if we're notified of a problem by one of our curators, we can investigate that issue. We have been doing a little bit of quality assurance. Hopefully we'll be able to come back to look at the quality of the collection when we've finished the crawl. With previous collections, we've even been able to commission extra resources to do quality control on a particular collection. We've had, for example, a student placement or an intern who's been looking at quality assurance. The only thing about this is if we do spot problems, if, for example we've missed some content in our crawling, quite often there's not very much we can do about it after the event. If that content is still online, then that's fine but as I said, we don't have a very sophisticated tool for things like patch crawling.

ACCESSIBILITY AND SEARCHABILITY

Can the collection already be accessed or searched?

Nicola Bingham: No, it's not currently publicly available on our website because we want to do a bit more quality assurance work. We also want to complete and standardise the collection metadata and apply for permissions for open access for as much of the content as we can. The websites are all available individually in the reading rooms of the legal deposit libraries, but not yet as a discreet collection. So if a user came onsite to one of our reading rooms and they know the URL to look for, they'd be able to see that website in Wayback.

There's a little bit of work to do in terms of applying for permission to make websites available. But when we do make the collection available, it will be browseable as a special collection on the topics and themes page of our website. When the user comes to our website, they can see a list of all of the targets in that collection with a little bit of metadata about each target.

Depending on whether we have an open access permission for that website, they can be taken through, to view the archived copies. This comes back to our legislative framework in that to make archival copies of websites publicly available we need this additional permission from the content owners. We can make all of our web archive collection available in the reading rooms, but only a subset is made available through our publicly facing website. Having said this, from the publicly facing website, users can still view the full manifest of all the targets that we've harvested. They would be able to see the full 6,000 targets, for example. Being able to view that target depends on the permission form.

We also create a catalogue record in our main library catalogue, 'Explore'. This is only at the collection level, so it's just the one record that describes the coronavirus collection. We don't create catalogue records for individual target records, just because we don't have the resources to do that. There's an advantage for us in that the users of the library catalogue might not necessarily know about the web archive. They might be doing a subject-based search, but they could be directed to web archive content that way. This raises our profile a little bit more.

The other thing is that our archive is all full-text searchable. A user could come across relevant content through a keyword search. Although one of the limitations of the full-text searching is that it quite often returns so many results that it becomes not very useful. So I imagine if a user was to search for COVID-19 or coronavirus, they would have many millions of search results returned to them.

Did you ask for open access permission for all the seeds included in the COVID-19 collection or only for a selection of them?

Nicola Bingham: We generally apply for permissions for everything that we have selected through our curation software. There are well over 100,000 target records; not for this collection, but the totality of our selective collections is over 100,000 records. Generally we will apply for permission at the point of adding the selection to the curation software. There's a module within the system, which allows us to generate an email to a website publisher, and it will send a generic letter saying: 'We are the UK web archive. We'd like to seek your permission to make archival copies of your website publicly available.' We also explain the legal deposit regulations and why their website has been archived.

It's very challenging and we don’t have a fantastic success rate with this process. Firstly sometimes identifying the publisher of a website is quite difficult. Quite often a website might not publish contacts details. Even if we do manage to identify the person that might be responsible for the website content within a large organisation, quite often, they won't know themselves who would be responsible for that kind of a query. We do endeavour to seek open access permissions for all our selected content, but we have probably only about a 25% or 30% success rate. It’s a very resource-intensive process with quite limited gains.

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

Nicola Bingham: Not as far as I'm aware in terms of analysing the collection. It's probably still quite close to the event at the moment and lot of it will still be online. I guess that will be the researchers’ first point of call. We have had a lot of interest in contributing to the collection from other heritage institutions wanting to know about our activities and from members of the public who are becoming more aware of web archiving and wanting to notify us of websites that we might want to archive. It's very positive. I think it speaks to genuinely the international efforts of the web archiving community in that there is more awareness of web archiving. We've had a few topical events over the past couple of years, and we've been contacted by members of the public to say: ‘I understand that you are archiving websites. Have you considered archiving this website?’ or ‘This is my own website I'd like you to archive’.

How do you communicate about this special collection?

Nicola Bingham: We've written a few blog posts that have been published on places such as the Digital Preservation Coalition, which has promoted the collection. I've written on mailing lists. For example, the UK JISC mail web archiving list for the archives and records management community and within the UK, the Welcome Trust has coordinated a professional network of heritage organisations who are undertaking collecting related to the outbreak. This network is relatively informal at the moment. I'm not sure that there are any concrete outputs, but I suspect that they will be conferences and journal articles that we will be able to contribute to.

We’ve spoken internally about the collection at departmental meetings with colleagues. We’ve also used our British Library press and publicity to promote the collection as well on our website and social media. The UK Web Archive has a Twitter account and blog as well. We'll promote the collection in that way. I'm hopeful that this activity will form the basis of papers, conference papers and journal articles potentially in the near future.

REFERENCES

Webber, Jason (2020, 29 July). 15 years of UKWA – Looking back at our first collections. *UK Web Archive blog post.* Retrieved from https://blogs.bl.uk/webarchive/2020/07/15-years-of-ukwa-looking-back-at-our-first-collections.html.

UK Web Archive. (2020). *About us.* Retrieved from https://www.webarchive.org.uk/en/ukwa/info/about.

British Library. (2020). Collection guides. UK Web Archive. Retrieved from https://www.bl.uk/collection-guides/uk-web-archive.

We would like to thank Nicola Bingham (British Library) for her help in proofreading this interview in addition to agreeing to be interviewed.

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