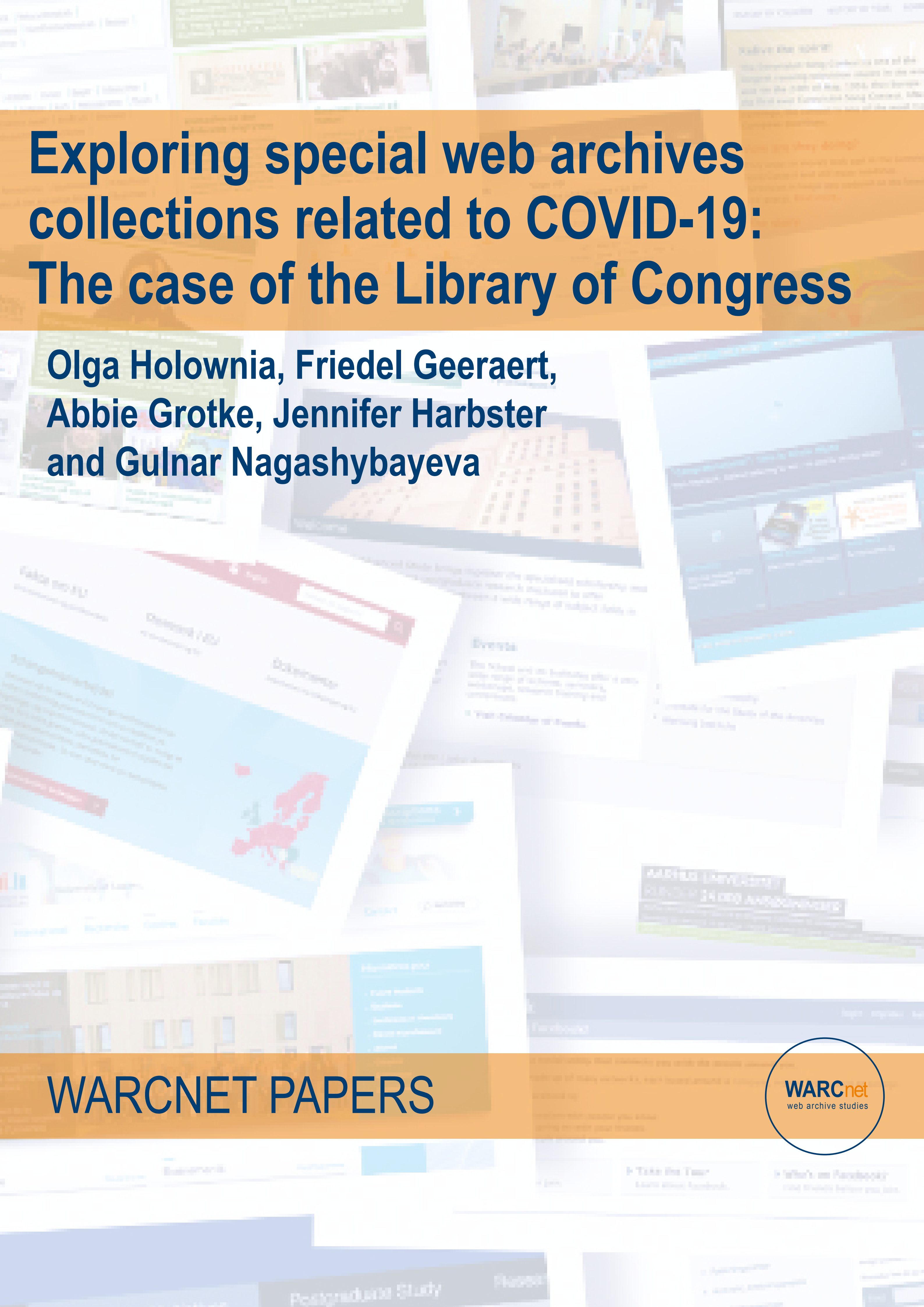


Exploring special web archives collections related to COVID-19: The case of the Library of Congress

Olga Holownia, Friedel Geeraert, Abbie Grotke, Jennifer Harbster and Gulnar Nagashybayeva

WARCNET PAPERS

WARCnet
web archive studies

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

An interview with Abbie Grotke, Jennifer Harbster and Gulnar Nagashybayeva (Library of Congress) conducted by Olga Holownia (IIPC) and Friedel Geeraert (KBR)

olga@netpreserve.org
friedel.geeraert@kbr.be

WARCnet Papers
Aarhus, Denmark 2022

WARCnet Papers ISSN 2597-0615.

Olga Holownia, Friedel Geeraert, Abbie Grotke, Jennifer Harbster and Gulnar Nagashybayeva: *Exploring special web archives collections related to COVID-19: The case of the Library of Congress*

Published by the research network
WARCnet, Aarhus, 2022.

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

Cover design: Julie Brøndum
ISBN: 978-87-94108-07-2

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)

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

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

Niels Brügger: *The WARCnet network: The first year* (Jan 2021)

Susan Aasman, Nicola Bingham, Niels Brügger, Karin de Wild, Sophie Gebeil and Valérie Schafer: Chicken and Egg: *Reporting from a Datathon Exploring Datasets of the COVID-19 Special Collections* (Dec 2021)

Olga Holownia, Friedel Geeraert, Abbie Grotke, Jennifer Harbster and Gulnar Nagashybayeva: *Exploring special web archives collections related to COVID-19: The case of the Library of Congress* (Feb 2022)

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

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

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

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

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

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

Karin de Wild, Ismini Kyritsis, Kees Teszelszky and Peter de Bode: *Exploring special web archive collections related to COVID-19: The Dutch Web archive (KB)* (Nov 2021)

Michael Kurzmeier, Joanna Finegan and Maria Ryan: *Exploring special web archives collections related to COVID-19: The National Library of Ireland* (Feb 2022)

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 Library of Congress

An interview with Abbie Grotke, Jennifer Harbster and Gulnar Nagashybayeva (Library of Congress) conducted by Olga Holownia (IIPC) and Friedel Geeraert (KBR)

Abstract: This WARCnet paper is part of a series of interviews with 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. This interview with Abbie Grotke, Jennifer Harbster and Gulnar Nagashybayeva at the Library of Congress was conducted in collaboration with the International Internet Preservation Consortium (IIPC).

Keywords: web archives, social networks, COVID-19, special collections, U.S.A., Library of Congress

This WARCnet paper is part of a series of interviews with web archivists who have been involved in special collections related to COVID-19. The interview was conducted on 5 January 2022 with Abbie Grotke, Jennifer Harbster and Gulnar Nagashybayeva at the Library of Congress. Abbie Grotke is Assistant Head of the Digital Content Management Section where she leads the Web Archiving Team. Jennifer Harbster is Head of the Science Reference Section in the Science, Technology and Business Division. She is a subject matter expert and nominator and the majority of her work with the COVID-19 collection is as collection leader. Other collections she leads are the "Science Blogs" and the "Earth Day 2020" Web Archives. Gulnar Nagashybayeva is the Business Reference Specialist in the Science, Technology and Business Division. She also is a subject matter expert and nominator and collection lead of the COVID-19 collection and the ongoing "Economics Blogs Web Archive."

The web archiving programme at the Library of Congress started in 2000. The U.S.A. has no national legal deposit legislation. Mandatory deposit laws exist but web archiving is not part of those. The Library therefore relies on permissions that are based on a fair use analysis led by the Library's Office of General Counsel. Determining factors are the country of publication and the category or type of the website. There are four levels that are used depending on the country and category of website: (1) no notification is required, (2) notification of the intent to crawl and provide access is sent but no permission is required, (3) notification of the intent to crawl is sent and permission is requested to provide offsite access and (4) permission to both crawl and provide offsite access is requested.

Web archive collections at the Library of Congress are based on particular subjects, events or themes and are curated by Recommending Officers who also set the crawl frequency and scope. One exception is the “Single Sites” collection that “allows for the collecting of representative websites in a variety of subject areas” (Library of Congress, 2017). A number of collections are comprehensive, for example the national election campaigns and the U.S. House and Senate offices and committees, but most collections consist of representative samples. The Library of Congress also has an international collection scope and includes foreign websites in their collections if these fit the selection criteria and are not already archived and made publicly available in those respective countries. The web archive comprised 2.8 PB in September 2021. 29.154 web archives were available via loc.gov as of January 2022.

The web archive collections can be consulted online. On the one hand, there is the collection framework where the collections can be consulted and on the other, the descriptive records. The Library of Congress makes the content available after a one-year embargo.

THE REASONS OF THE SPECIAL COLLECTION

Why did you create a special COVID-19 collection?

Abbie Grotke: Why don’t I start with how it came about? We have the Web Archiving Team that I lead, and then there’s folks like Gulnar and Jennifer around the Library who are selecting the content and deciding what collections we do. At the very beginning of when things started to shut down and maybe even a little bit earlier as the news was coming out of Asia, some of the recommending officers that are working on other collections were asking: “Should we be collecting around this topic?”, “What should we do?”, “How should we approach it?”

Some divisions, such as the Asian Division, were nominating things in existing, ongoing collections that were COVID-related. Once we started to collect, we initially thought: “Let’s see where this goes.” I was working with the Collection Development Office who helps guide the policies and the collection development process. We were monitoring it and we had everybody, as they were nominating content, tag in our curatorial tool called Digiboard. They were tagging COVID-19 and we had three different tags that we were using.

People were all over the place and everybody was trying to figure out what to do. We had people tag for about four months. People were nominating content in their own collections or in what we call the “Single Sites” collection [General Collections on loc.gov], which is not a thematic collection but is a place where recommending officers can nominate content that doesn't fit a larger theme or event collection. So we had people tagging content in that, and we were seeing what direction things were taking and also monitoring the news. And then, a few months into that, we generated some reports, and there was a conversation with senior management about this and what the status was.

Our senior manager at that time, Robin Dale, who was Associate Librarian for Library Services, requested that we do this as a formal collection. So Rashi Joshi, who's in the Collection Development Office, worked on a plan for how to approach it. This was outside of our regular collection proposal process. A lot of the collections are done by individual divisions and not library-wide, but this was meant to be a more library-wide initiative. So the plan came together and a project team was identified and then Gulnar and Jennifer graciously took on the role as this collection's leaders. The plan was approved in June, so it took us a number of months to organize.

By that point we had archived thousands of sites related to COVID already. And we were doing a lot of collecting, but it wasn't in an organized way, and we had gaps and there were areas that we weren't covering at all. Those that were adding content were very active in their subject areas, but there wasn't this holistic view. So that was the intent of doing the collection. It got a little bit formalized a few months into the pandemic. Jennifer and Gulnar can say how they took it on from there.

Jennifer Harbster: I should also note that the Library's subject matter experts also participated early on with archiving COVID-19 content for the IIPC's "Novel Coronavirus Web Archive." We were very active – I believe we nominated around 900 URLs.

Abbie Grotke: There was dual activity – the contributions to IIPC and then our collecting, which has a much more rigid permissions process. So that's why we do two approaches in some cases and we're trying to get content in different areas.

Jennifer Harbster**:** Sometime around June or July 2020 Gulnar and I agreed to co-lead the Library's "Coronavirus Web Archive Project." Gulnar and I are in the Science Technology and Business Division. Gulnar represents business and I represent science. We're both pretty seasoned web archivists. Our first step was to form a team to help us because Gulnar and I could not do this alone. We got a core team together that we affectionately called CAT, which is the Coronavirus Archive Team. We love acronyms in D.C. When we selected our core team, our goal was to enlist subject experts for the high-priority and gap areas. We also needed to recruit seasoned web archivists, because this project was going to require a lot of work and we needed experts familiar with our Digiboard tool. We selected 10 experts from across the Library who represented a multitude of subject areas such as law, government publications, performing arts, religion, poetry, psychology, agriculture, economics as well as subject experts that covered Latin America and the Caribbean. We had added an extra business librarian and another science librarian. We even recruited a folklife specialist. Once we got the team together, we reviewed the collection plan that Abbie was talking about and began to outline the actions our team should be taking.

THE SCOPE OF THE COVID-19 COLLECTION

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

Jennifer Harbster: As Abbie mentioned, we had all this existing COVID-19 content that we were already crawling. For example, have been crawling the White House and the CDC [Center for Disease Control and Prevention] for decades. Then there were over 2,000 recently nominated websites tagged with COVID-19 for our "Single Sites" and other web archive collections. It was an impressive number.

Abbie Grotke: It was crazy.

Jennifer Harbster: CAT started off by looking at those 2000 sites. Well, we didn't look at all 2000 seeds, but we started really looking at what was collected and getting a plan together to start reviewing that content and making recommendations to move it over into the "Coronavirus Web Archive" or to another curated collection.

We spent a lot of time working out the scope of the Coronavirus collection and figuring out what to do with the 2,000+ sites that might not fit the scope of the project. It is not that the information was not valuable, rather that the content featured some COVID-19 information and a whole bunch of other non-related topics. Many of these sites were from big news aggregator sites and did not have enough of a focus specifically to COVID-19.

So we made recommendations to add the content to a news collection or just keep them as "Single Sites." What greatly helped us review and make decisions on content was the use of a rubric. Melissa Wertheimer, one of our team members from performing arts with incredible web archiving skills, created an amazing rubric for our team to use. After using the rubric, I feel strongly that every web archivist should use a rubric. It was vital for the selection process and also helped us prioritize the content. We scored the content on things like: "Is this a gap area?" and "What are the permissions?" If it's a "no permission" seed, obviously, it gets rated higher. What I really appreciated were questions like "Does it have a unique value?", "Is this duplicated information?", "What's the evidentiary value?", "Is this talking about a specific event or creative act?", "What's the informational value?", "Are there statistics in this?", "Are there embedded documents?" Because as we all know, a lot of these websites have all this information in there. It's not just text, there's PDFs and so on. We also rate the authenticity of the content. Sites that scored high were added to the collection. For me, that rubric was the key to help manage a massive amount of content – it helped us decide what goes in and what is left out. I thank her all the time for developing the rubric and sing her praises every chance I get. Having both a collection plan and a rubric, helped to focus our efforts.

Abbie Grotke: I would say that the work that Jennifer and Gulnar's team did, narrowed down the actual collection that will be released into a much smaller, tighter collection. We didn't delete any of the content that was selected in those early days but some of it moved into other collections. Some of the content we just collected and stopped crawling, but we're crawling 245 [seeds] currently. It's a much narrower, more focused collection that will be presented to the public. All that other content is still in our archives and we still have it, but we stopped collecting a lot of that, including the social media. Regarding languages, we have material in English, with Burmese, French, German, Greek, Hindi, Indonesian, Italian, Japanese, Khmer, Korean, Lao, Malay, Malayalam, Marathi, Persian, Portuguese, Russian, Sinhala/Sinhalese, Spanish, Tagalog, Thai and Turkish in the formal collection.

Jennifer Harbster: Our team also had a lot of questions about adding social media, so I will defer to Abbie on this topic.

Abbie Grotke: At the time that we started the collection, we were still attempting to get social media. We are currently not trying to capture social media because it failed so miserably. We had people nominating Twitter and Facebook pages, Instagram, and Pinterest. Those are the big ones. Contributors, particularly from the Asian division, were adding a lot of the social media content, a lot of Twitter accounts or Facebook and the crawling tools, as you know, are just not performing well. And we were also crawling with our prior vendor at the time, and they were using Brozzler to get some of that. Sometimes it was successful, but when we tried to get Instagram, for example, we weren't successful. Last year we switched vendors. They were attempting to get social media through their crawling tools, but again, it hasn't been successful because of the way the platforms are trying to keep us out. So as of last summer, we pulled all of the social media out of the crawl. When the edict came from above to do a formal Coronavirus collection, the senior management told us not to focus on social media primarily. The focus really was on websites and just traditional web content. In addition to the technical challenges, we can't do hashtag crawling, for example, because of our permissions approach. We couldn't crawl platforms such as Reddit. Social media is complicated. Jennifer and Gulnar can talk more about the types of sites that ended up in the final collection.

Jennifer Harbster: After reviewing the existing content and using the strategies set forth in the collection plan, we targeted high-priority areas like federal, state, local and indigenous law, public policy, science, business and cultural content. Then we looked and found gap areas that we weren't collecting much of. We really pivoted our focus from: "Let's just collect everything and anything", which was the original approach, to a strategy. We had gap areas of representation from historically excluded groups. African Americans and Asian Americans are one example. We were lacking in what I like to call "the everyday person information" such as religion: what were the churches' and temples' responses. Also, the psychology content, sports and recreation – there were all these other areas that were missing, because, when we initially think of COVID, we're always thinking of these high-level things like science and policy. But this is something that affects everyone's lives, so we asked: "How do we bring it down to the everyday person?", "What was the impact of COVID?" We're still crawling that high-level content from governments, and we also pivoted towards collecting more U.S.-focused content.

Because by that time – summer going into fall of 2020 – there were a lot of established COVID collections going on. And the IIPC was insane! I mean, it's amazing what they've been doing. So we were figuring out how we could enhance or supplement the work of other web archives. Of course, there'll be duplication across web archives, and there is never an absence of web content to archive.

You have already touched upon this but could tell us more about the number of staff involved in nominating for this collection?

Abbie Grotke: Quite a large group of recommending officers were involved in the early stages. About 25 people contributed nominations to the collection that was recently announced, and a number of others contributed related nominations to their existing collections and IIPC efforts.

Gulnar Nagashybayeva: There were 10 on our team.

Jennifer Harbster: Over the past two years we had a team member retire and another one resigned from the Library.

Abbie Grotke: During the start of this collection there were only 4 of us, and now the web archiving team currently counts seven plus me. There are also other people who we work with on the access side. There are people in our IT departments who help with the framework and the infrastructure to get the records online, for all of our web archive collections. So there's a handful of people who we work with over there too.

Jennifer Harbster: And we also receive emails from Library staff that don't do web archiving but came across content that they think would be great for the coronavirus web archive.

You mentioned that you had tried to make the collection more representative in the sense that you tried to include historically excluded groups, but I wonder how you went about that in concrete terms.

Jennifer Harbster: Because there were some subject matter experts that were not on this team, we did a lot of consultations with experts across the Library. I think we even gave them the rubric to help prioritize the content selection. I think everyone at this time, was just ingesting so much information and news that when we saw something mentioned, for example the Asian-American hate group project we were thinking: "Oh my God, are we collecting this?" Then we would go into our Digiboard tool to check. So it was a mixture of having subject matter experts to contribute and serendipity. As Abbie said, during this time in the U.S. and across the world, actually, there were a lot of protests going on against racial injustice. Some of the content they were collecting touched on the content we were collecting. We had a team member that was in both groups so she bridged the two together. There was a lot of that cross-fertilisation going on. I would like to do a deeper analysis of the archive. We did general analysis after a year or so of collecting and were able to identify other gaps. Then we'd say: "This month, or this week, this is the gap area we're going to focus on, or let's focus on these groups." The one group that we don't still have a lot of content on is from the LGBT groups.

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

Gulnar Nagashybayeva: The challenge was when we were told to add 250 new sites (200 were supposed to be U.S.-based and 50 international), we also had those original 2000 sites to go through. The challenge was that at the beginning of telework everybody was archiving as most of the staff did not have many telework-ready projects. Some people were just archiving all day which is how we got all those big numbers. By the time this project with a formal collection plan started, everybody got their work laptops set up and were doing all kinds of other assignments as well. Our task was to sift through 2000 sites and select appropriate ones for this collection from there. I worked with a myriad of spreadsheets and contacted each nominator. Many of them were from the Asian Division and our overseas offices. Some of them were not very active in archiving by then, and it was hard to get responses. Many of the sites were either post-crawled or added to other collections. So this collection became pretty small. I just checked today and we currently have 699 sites in the collection. More than half of the sites are in post-crawl, because the relevant content has been captured.

Abbie Grotke: Post-crawl is a status in our Digiboard tool. It basically means getting it out of the crawl, but we still make it accessible for research use. Just to speak for the resources, the collection came at an interesting time because the frenzy to add so much in the early days led us to hit some capacity limits with the contract crawling by that summer. It ended up affecting all of our collecting across the board because we had to make sure we had enough room to crawl this content. And we were crawling the U.S. elections and the protests hadn't started yet, but there were these high-profile collections that we wanted to make sure we still had capacity to crawl. We ended up pausing all new collecting across the other collections, which was frustrating for people because it was a great telework activity. We had to say: "Wait, we need to make sure we can collect this really important content right now." So we didn't stop collecting things that were already in the crawl but we said: "Don't add anything new to all these other collections that we have." There's about 75 [event and thematic] active collections at any given time so everything else went on hold and we focused on this COVID content. It took us a while (until February 2021) to reopen the collections for business. By that point, people started to go onsite, and the activity level has dwindled greatly. But it becomes an ongoing activity at some point, particularly with something like the pandemic.

And we can't easily say that this collection contains these many gigabytes or terabytes. We don't have the ability to pull that data out right now. There's a total of 699 records in the collection. We are currently crawling 245 seeds and 413 seeds are post-crawl. The rest never got crawled because they were maybe nominated but ultimately weren't selected for archiving.

What was the capture frequency?

Abbie Grotke: All of our collections go into a weekly, a monthly or a quarterly frequency. We don’t do daily collecting. A lot of this content was in the weekly crawls. Some of the government content was in the quarterly crawl, but to get the rapidly changing news, we were trying to get it fast. We have the crawl frequency on the framework: weekly, monthly and some quarterly. So it was a mixture depending on the type of site.

You have already touched upon it, but how did you go about archiving on a national level about an event that is fundamentally global?

Jennifer Harbster: The collection plan really helped us. And the fact that there were other web archiving activities across the globe. The Library of Congress has international collections and specialists. At first it was very much a “Wild West” kind of thing: everything was going in and in large amounts. After we got that collection plan together, that’s when we really started having more of a focus. Whenever I talk about this collection, I have to quote Rashi Joshi, our Digital Collection Coordinator. She sees web archiving COVID-19 as a “digital curation challenge.” So that’s where a lot of the review and analysis of the content we were currently collecting came into play because we could see that we’re really strong in South and Southeast-Asian material, but not so strong in South and Latin America. We also knew there was a lot of collecting going on in Europe. My kind of thinking was: “Where can we fill in gaps and collect content that is not already being collected?” We were thinking about countries that don’t have an existing or robust web archiving program. It was definitely very challenging, but I think since then, there is a Southeast, South Asian web archive.

Abbie Grotke**:** There’s also some work happening around planning, how we approach the global collecting and where there might be gaps across the library. Some of that’s emerging out of a lot of the activity that took place over the last couple of years. There’s renewed interest in making sure that we don’t have gaps or that we’re covering areas of the world that might not be covered.

You mentioned working with the IIPC, but we were wondering if you have any existing collaborations, such as with the State Libraries. While covering the pandemic nationally, was there any focus on getting information from different States?

Abbie Grotke: I don’t think so. We already have a “State Government Web Archive” that was already underway for a few years before that. We were already collecting a lot of the state government websites and our Acquisitions and Bibliographic Access (ABA) division is leading that effort. Our colleague Rick Fitzgerald from ABA went through and added all the

COVID portals to the IIPC collection, and we made sure that the US States were represented there. For this collection, we weren't coordinating really, but I know that some of the Asian division folks were talking to the Ivy plus consortium about some of it and perhaps nominating some things. I'm not always on top of who was doing what with those external groups. And I don't know if Melissa was talking to performing arts societies to get information about what to nominate. There may have been independent recommending officers talking to their communities about topics, but maybe you all know more about that.

Jennifer Harbster: We were also going to webinars to learn about the focus of other COVID-19 collections. And we also were reading things that were coming out on blogs about web archiving and coronavirus. We really wanted to understand the landscape out there too.

Abbie Grotke: That reminds me: we are members of a federal government web archiving interest group which has met sporadically. It's kind of re-invigorated in the last couple of months, but there was a meeting mid-pandemic, like a year in or six months in where folks from the National Library of Medicine National Archives and GPO [Government Publishing Office], the Smithsonian [Institution], they all talked about the types of collecting they were doing. That was another touch point with some of the agencies in the local DC area, at least at other federal institutions. We were trying to keep on top of what they were all up to.

So you were aware of what they were doing, but was that to make sure that you're not duplicating the content or you did have any formal agreements on who is collecting what kind of content?

Abbie Grotke: It was more just to keep on top of what others were doing, but we were still going our own way.

For the IIPC collection, you mentioned that you could nominate more websites than you would normally do for your local collection. Could you elaborate on that?

Abbie Grotke: It was interesting because our team spread the word about the IIPC collecting, and initially the collection development office was collecting, standardizing, and inputting suggested URLs on behalf of LC staff. But once IIPC created a Google form where anyone could input, staff were just told to use the IIPC form directly, so we didn't monitor what people were doing after that. At some point, I got some reports from Alex [Thurman, Content Development Group Co-Chair] about LC nominations and some people didn't always identify themselves as Library of Congress, but went in and nominated content. A lot of the nominations were from our overseas offices. We have offices in Rio and Indonesia and in other locations, and some of those folks who had already been active in the IIPC collaborative collections were nominating to the IIPC COVID-19 collection. We didn't really promote it, but people knew about it and they were adding to the collection in addition to adding to our web archives.

THE FRAME OF THIS SPECIAL COLLECTION

You mentioned that you started collecting quite early on. What was the exact timeline?

Abbie Grotke: We started in March. As always with this kind of thing, people start saying: “Are we collecting?”, “Should we be collecting?” There were a couple of weeks of that. And then suddenly people started adding things. It was probably as soon as we shut down so March 16th or the week before that we started collecting in some of the existing collections. March 24, we sent a message to all of our nominators saying we had been receiving inquiries about COVID-19 related web archiving, and provided some guidance there which said to continue to add content in their retrospective collections when in scope, or to use the “Single Sites Web Archive.” We also mentioned the IIPC collection as an option. Since everyone was home, we also provided instructions about accessing Digiboard remotely for staff who were not used to teleworking. The formal collection began in July officially. The project was approved in June.

Gulnar Nagashybayeva: June 30th is when we proposed it to team members.

Abbie Grotke: July 20th is when I activated the collection in Digiboard. This is the crawl start date. The formal collection started up then.

Do you have any stop date in mind? Will you continue crawling the selected number of URLs you mention or add new ones as well?

Abbie Grotke: I have a question for Jennifer and Gulnar. Are you adding things around Omicron right now? Are we already picking up the newest updates?

Gulnar Nagashybayeva: Melissa has been adding new content.

How did you carry out quality control on the collection?

Abbie Grotke: On our side, once things were in the collection and being crawled our team is responsible for most of the QA, of reviewing the archived content itself. During the early parts, when Internet Archive was our crawling agent, we had access for the team to look at the captures and to do some review, but it’s typically not something people have much time to do. Our team does the sort of high-level QA across all of our collections and this [COVID-19 collection] was just one of them. We get various reports from the crawler and we can see where there are problems and then do spot checking as we go. It was integrated into all of our QA processes across the team, but we weren’t focused specifically on this collection. Individually there wasn’t somebody assigned to it, but we worked it into our process. Now, under the current vendor, there’s no access yet for recommending officers to look at their content. We’re working on that still. Our team is responsible for reviewing all the content and making sure that things are still healthy and in the crawl. If we find a problem, we sometimes interact with the person that has nominated the content and make some decisions about whether or not to keep crawling or take it out of the crawl.

Gulnar Nagashybayeva: Currently we are doing the annual collection review, so I sent spreadsheets to all the nominators to look at their seeds to see if they are still active, if there is new content being added. If it's a static website, then they need to send them to post-crawl. This review lasts two or three months so we're supposed to finish that by the end of this month. We'll see how many will stay in the collection.

Jennifer Harbster: As Gulnar mentioned, we were doing our yearly assessment and I think that was a little shocking for some people that they have to maintain this content.

Abbie Grotke: This is new. As of last year, we're making collection leaders look at all of the sites in their active collections to determine whether or not we should still be crawling it all.

Jennifer Harbster: You don't just fill out a form and say: "Bye!" like you do with a book here at the library. Instead, you become the owner of that record and of that content. I think that became a challenge for some people and especially for people that nominated a lot of content. The maintenance for me has always been a really important thing with web archiving and trying to communicate that information with our subject matter experts. And explaining that you need to be committed to the entire life cycle of the content.

Abbie Grotke: The maintenance is less about the quality of the archive. It's wondering if this thing is still in scope, if the URL has changed and if it is still up. Because we catch some of that when we're doing QA, but our team is not looking at it in terms of the scope and the content. If we get a 404, then we'll maybe address taking it out of the crawl. Another element is checking that the content is still active.

Jennifer Harbster: Or just how things were very headline news and then started waning. We would review the crawls or the live site and discover that there is no COVID-related content on the site now, so we would send it to post-crawl.

Did you encounter any issues, challenges, or limits related to the collecting activity? You already mentioned that you ran into the limit for the capacity with the vendor, but were there any other?

Abbie Grotke: I think social media was a big one in the early days when we were trying to get it. And then also that being a limiting factor when the formal collection started up. I think it was because we're so used to adding social media to the collections that a lot of folks were saying: "Why can't we? We're doing it in all these other collections?" But it was something we had to follow for this one. And I think from my perspective, I don't think it was really a challenge, it's more like an opportunity – this was one of the first big cross-library collections. I think having so many people from around the library working on one collection maybe brought some new challenges. I don’t know, Jennifer and Gulnar, if you agree.

Jennifer Harbster: There were sort of those normal challenges you encounter with web archiving. Permissions obviously are a big one for us, and making sure that we were following that rubric. The social media, that definitely was an issue for us because so much communication is through the social media channels and it’s our nature. That’s how we’re ingesting information now, through the social media channels, maybe not so much via websites per se. When we were getting together a list of new content to nominate, so much of it was social media and we had to say: “We’re just not going there.” That was hard for us to try to not nominate social media, because we were finding such great content and because people were using social media as a platform to create content. The permissions are always difficult for us because there is such wonderful content out there but if we needed permissions and the content owners don’t respond back, we then would email them directly. So that takes time out of your day.

Abbie Grotke: That was a lot of the creative content. The things that have the most restrictive permissions are the personal things that Jennifer was talking about.

Jennifer Harbster: Anything that was creative or reflective of the performing arts would have been in that “needs permission to collect” category. That can be frustrating too. There were hard decisions to make. Some of that content is still existing in other collections or in the “Single Sites” collection but deciding what goes in and what goes out is difficult because, realistically, we can’t have everything in there. Telling people “It’s not going to be added to this collection” is also hard. But we had a lot of the data and the evidence to say: “This is why, but it does still exist.” We were assuring our subject matter experts saying: “The content is still being crawled, it’s still there, you could find it. If you tag it, you could search and find it. It’s just not going to go into this collection.”

Abbie Grotke: The front end for the collection, once it goes live, does point to some other collections. And it does say: “You may also find coronavirus in other collections.” At one point we looked across a lot of the collections and even in the “United States Elections Web Archive”, the coronavirus is all over. It was mixed in with all the collections. So it’s not that this is the only thing that will document coronavirus. That was kind of tough to then weed out the content. I didn’t have to do it, luckily.

Gulnar Nagashybayeva: I mentioned that when we were reviewing the large number of sites from the initial archiving effort, many of them were added to other existing collections as appropriate. For example, the state governments sites were sent to the “State Government Websites of the U.S.” collection and were tagged as COVID-19 / coronavirus. And we have the “Business in America’ Web Archive” and “Economics Blogs Web Archive”, so they would be all covering COVID-19's impact on business and economics for this time period. And there is more COVID-19-related content outside this collection. And I also want to add to the challenges that there are some really cool websites with tons of data, these data dashboards we couldn't collect. They are dynamic and changing weekly and daily.

Abbie Grotke: You all know about other tools out there that might have been more successful. Our scale is such that we can't be running Conifer or something and try to manually archive the dashboard. So that was very frustrating for our team as well, to see those beautiful visualisations containing such valuable information but not be able to get it. I think that was frustrating for everybody.

As I mentioned above, at the start of the pandemic, our web archiving team was down by two members. We only had four of us in those early days. We were in survival mode in terms of what we could or could not do and had to make some really tough decisions across the board. And then everybody started to web archive and we just couldn't do these boutique crawls, as we say, and we still can't really because of our scale, but we now have more staff thankfully. It was kind of a wild ride for a couple of months there while we were trying to figure out how we could proceed without Gina [Jones] and Chase [Dooley].

Jennifer Harbster: This project has definitely been very interesting, but there's a lot of reflection that's been going on too. Just in terms of tracking it. In the United States, a new administration came in during this time and there was also vaccine development. It's so interesting to look at the content from 2020.

I remember hearing about COVID in January 2020. I was in San Francisco and people were talking about it. When reflecting back on the main hot topics, one thing that was a challenge was how we deal with misinformation. Because that is a very large part of the story. I don't understand how these really wild ideas of how to cure COVID gain so much traction. We decided on collecting a big thing. It was an aggregator that put together all of this misinformation content. We were wondering how to do this because this is so much a part of this experience, and we cannot neglect it, but we also don't want to give authority to these misinformation websites by nominating them because they receive a permission letter that says that the Library of Congress accepted their website to its collection. We're not collecting them because of their authority; we are collecting them because they were an example of misinformation. The solution was to go to one of these big aggregators that listed them all – NewsGuard – so that is in the archive and so there is content that represents the misinformation.

I feel like the project will never end because we're always thinking of ways to connect, build upon or identify gap areas that are missing in order to give a more holistic picture. We could not have created this archive without a team and each one of the members brings in his or her own expertise and perspective. We have somebody from the law library who is developing an indigenous law collection. While she was developing that collection, she was also helping us develop the coronavirus collection and adding the content created from the indigenous peoples in the U.S. She was cross-collecting, if that's a word. That was really wonderful too. We also have content that focuses on food. As we all know, there were a lot of quarantine kitchens springing up. One of our other web archivists leads the "Food and Foodway Web Archive" collection and Gulnar is part of that too. She was building up both collections (Foodway and Coronavirus) at the same time. The input from our subject matter experts was vital to this project.

ACCESSIBILITY AND SEARCHABILITY

How can users access and search in this collection?

Abbie Grotke: Well, all the content should eventually be released. The initial release will have 450 web archives; others will be released as content comes out of our one-year embargo. A lot of the early captures will be already out of embargo when we announce the collection, but there may be some things in the last year that were added that are not there like vaccines.

Gulnar Nagashybayeva: Some of the content related to the vaccine I think might be under embargo.

Abbie Grotke: The records are actually up right now. We basically release the records that are available out of embargo every month. Every month there's new content coming out into all of the web archives. Then, as we build enough content up front for a particular collection, we add what we call the "collection framework", which is the descriptive information and then some featured items. This provides the contextual information for the collection. While this one has been ready for a while, we are trying to coordinate the launch with our communications office and do press releases and such.

Jennifer Harbster: We are doing an official launch of the framework and the archive in February. There'll be a press release and a social media campaign because a lot of the team members manage various social media platforms like Facebook, blogs, and Twitter.

In terms of access to and searching in the collection, I feel that, as we get more experienced in web archiving, our metadata becomes better behind those pages or the content. In terms of searching, we were really trying to make sure we're giving good metadata for those records, like having subjects and tagging and such. The search can be a lot better. Our team was discussing creating a sort of short video or something else instructional about how to use this archive, giving tips, and search strategies, that sort of thing. I think there's so much we can do to make the collection discoverable on the back end with the metadata, but also with the outreach, either webinars or blogs posts. I know Melissa has been presenting on this at a couple of professional meetings. I feel like there's still a lot of work to do, and there's a lot of content that we want to add, as well as existing content that we want to go back to and add into the web archive.

Abbie Grotke: Items in the web archive can be related to multiple collections. We can have something being crawled in the "State government website archive" and also have it associated with the coronavirus archive. There's some of that, I think, that will emerge over the next few years probably.

Jennifer Harbster: The one thing that really helped me when the Coronavirus became a pandemic was learning how people discussed the 1918 influenza pandemic. I started looking at what these historians and scholars were using to describe the 1918 flu and wondering what they will be researching 50 years from now about the Coronavirus. What are they going to need? Obviously, images are huge and responses from governments. Comparing how scholars were using 1918 flu archives and seeing how that could be applied to what we're doing. I guess there can be an end to this Coronavirus because there was an end to the 1918 flu.

Abbie Grotke: I remember when you were adding things related to masks, because we didn't have mask imagery. So it was like: "Oh, we have got to get some content about masks in there."

Jennifer Harbster: Yes, the face mask and Corona couture. But some of that was hard to capture because it fell under' commercial permission'. We had to collect information about facial masks. The importance of face mask was also important during the 1918 flu. I went down a lot of rabbit holes locating content for the "Coronavirus Web Archive." I found myself reading a lot of newspaper articles from 1918. You definitely can get lost in all of that content. It's good to get lost sometimes because it makes you discover things and it gives you inspiration.

Have researchers already expressed interest in using the COVID-19 collection?

Gulnar Nagashybayeva: Speaking of masks, I wanted to share something. We got the comment through the 'Ask a Librarian' service in our division. The researcher or the patron was asking if the library collects the annual reports of companies. In some company annual reports there are images of how they're dealing with COVID. There are pictures of people wearing masks and such. The library does not collect annual reports in print, so I told him that we have the 'Business in America Web Archive' with about 600 companies' websites, which includes the annual reports. And we have electronic resources that have annual reports and SEC [Securities and Exchange Commission] filings that we subscribe to and told him about the upcoming release of the "Coronavirus Web Archive." This patron was very impressed that the Library was doing this collection.

Jennifer Harbster: We definitely have seen interest in terms of people asking: "Are you collecting this?" or "Will you collect my website?" That has happened.

Abbie Grotke: There were a couple of blog posts on the main blog since the beginning of the pandemic that talked about coronavirus collecting and that, I assume, sparked some interest or awareness about it. But it'll be great once we have the framework up because then we can actually start pointing towards it. It's hard to point people to the list of records. It's better when the framework is up.

Jennifer Harbster: We dove in, and Gulnar could attest to that, and we've been getting people interested in what we're doing. And the library collects all sorts of material related to COVID-19, not just websites. People are definitely wanting to recommend things for us to take a look at and sometimes they're really good and sometimes they don't fit and they don't pass the rubric. Once we start our social media campaign and send out the press releases, we'll definitely, hopefully, get interest.

Abbie Grotke: Melissa recently did a Lib guide, or research guide, around her performing arts collections and, I think, cross-promotion across the various blogs to cover the different subject areas – we have the Signal blog that the web archiving team can post something to from the perspective of the creation of the archive and then the different areas. I think we'll be able to roll out some nice announcements about it and get some interest.

You've already partly covered this question but is there anything you would like to add regarding communicating about this special collection?

Abbie Grotke: I think the big thing with this one is that we're doing a press release. Mostly we just announce the collections through other channels or just send out a tweet. There's internal communication about the releases. I think the fact that the communications office is assisting with this is a big deal. I think part of that's because it was an initiative pushed by the senior management so there's this interest in doing a nice release for it and publicizing it. Of course, there's all the cross-collecting that we're doing across the library otherwise. So that's really exciting. We haven't had a press release in a while for web archives.

REFERENCES

Library of Congress (2017). *Collections Development Policy Statements. Supplementary Guidelines.* Retrieved from https://www.loc.gov/acq/devpol/webarchive.pdf.

The Web Archiving Program at the Library of Congress. Retrieved from https://www.loc.gov/programs/web-archiving/about-this-program.

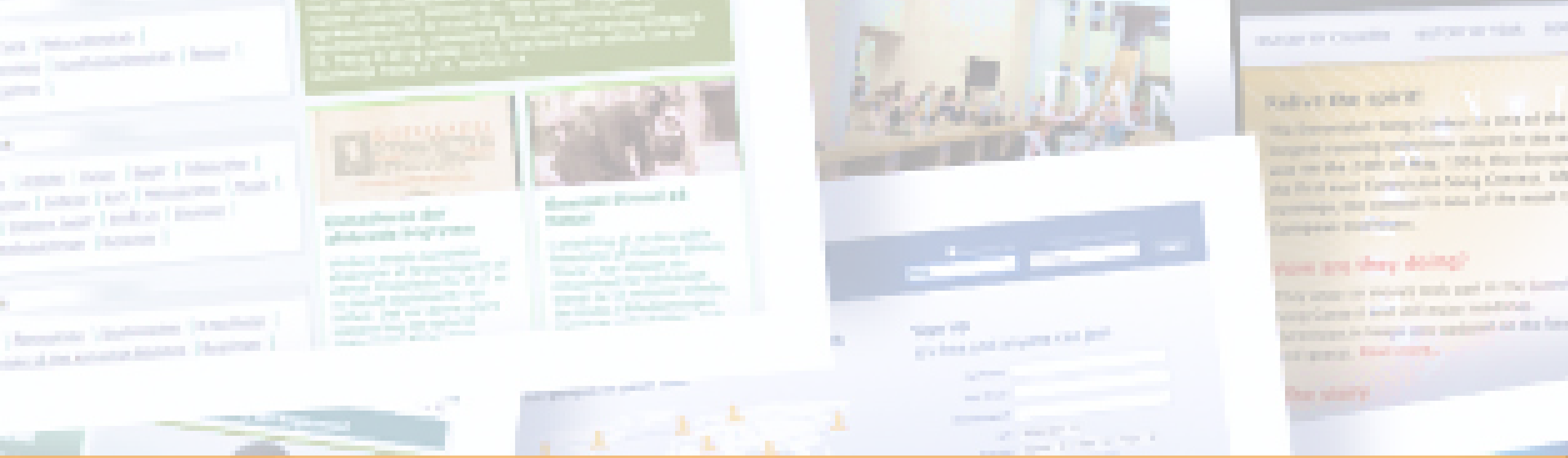
Mandatory Deposit. U.S. Copyright Office. Retrieved from https://www.copyright.gov/help/faq/mandatory_deposit.html.

Library of Congress COVID-19 Collection. Retrieved from https://www.loc.gov/collections/coronavirus-web-archive/about-this-collection.

Library of Congress Research Guides. Retrieved from https://guides.loc.gov.

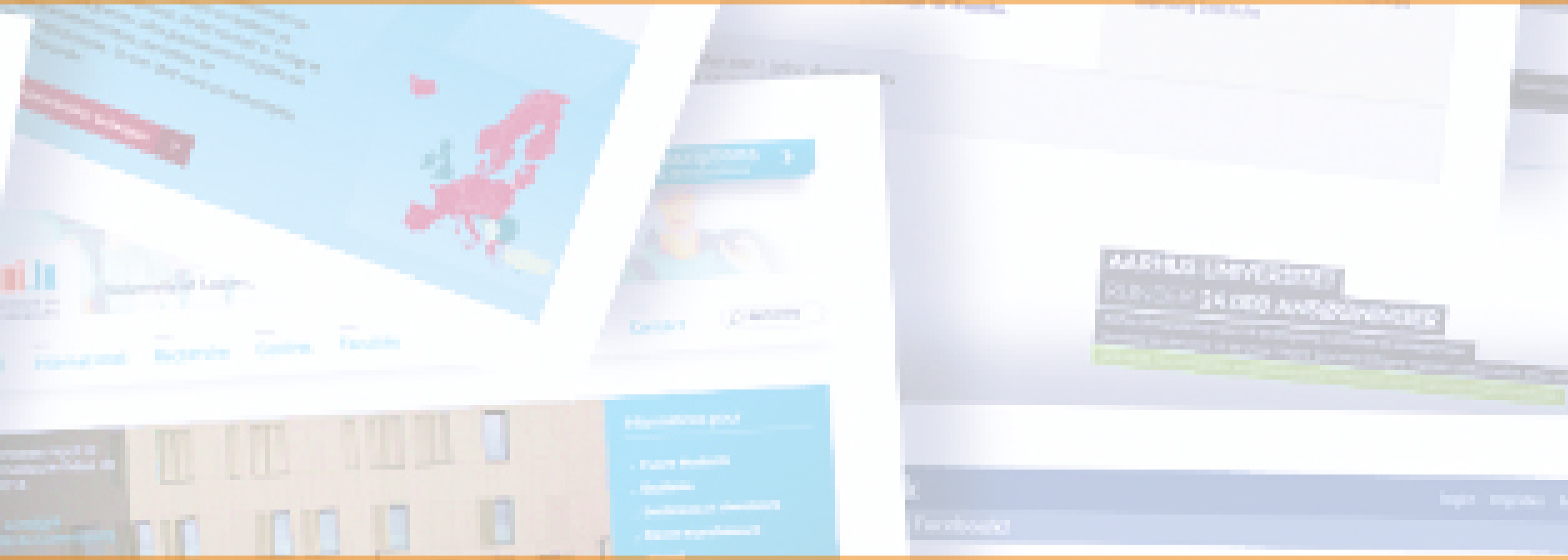
Library of Congress Web Archive. Retrieved from https://www.loc.gov/web-archives.

We would like to thank Abbie Grotke, Jennifer Harbster and Gulnar Nagashybayeva for their 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

INDEPENDENT
RESEARCH FUND
DENMARK

warcnet.eu   warcnet@cc.au.dk   twitter: @WARC_net   facebook: WARCnet
youtube: WARCnet Web Archive Studies   slideshare: WARCnetWebArchiveStu