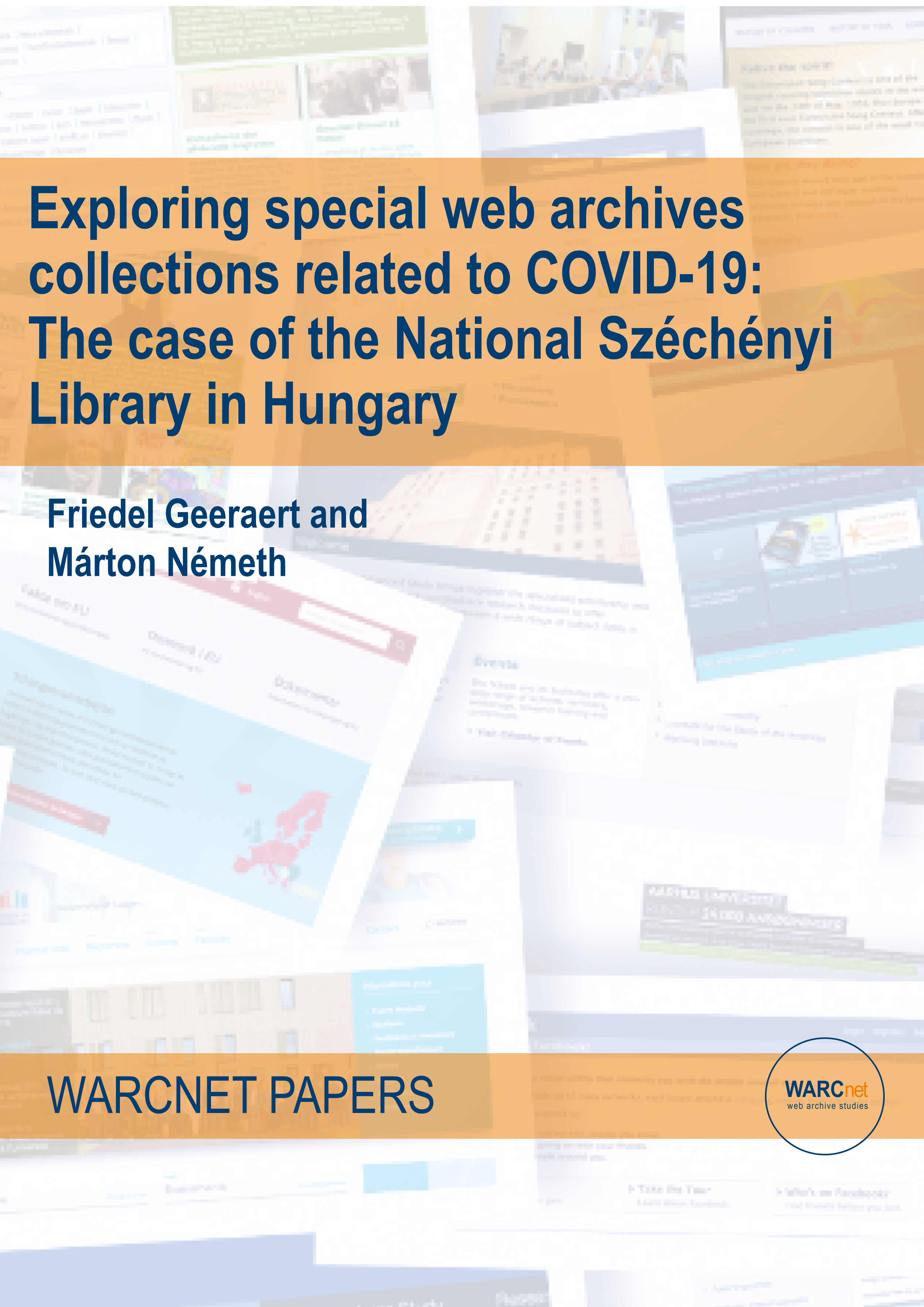


Exploring special web archives collections related to COVID-19: The case of the National Széchényi Library in Hungary

Friedel Geeraert and
Márton Németh

WARCNET PAPERS

WARCnet
web archive studies

Exploring special web archives collections related to COVID-19: The case of the National Széchényi Library in Hungary

An interview with Márton Németh (National Szé-chényi Library) conducted by Friedel Geeraert (KBR)

Friedel.Geeraert@kbr.be

WARCnet Papers
Aarhus, Denmark 2020

WARCnet Papers ISSN 2597-0615.

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

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: 978-87-972198-6-7

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* (2020)

Ian Milligan: *You shouldn’t Need to be a Web Historian to Use Web Archives* (2020)

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

Niels Brügger, Anders Klindt Myrvoll, Sabine Schostag, and Stephen Hunt: *Exploring special web archive collections related to COVID-19: The case Netarkivet* (2020)

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

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

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

An interview with Márton Németh (National Széchényi Library) conducted by Friedel Geeraert (KBR)

Abstract: This WARCnet paper is part of a series of interviews with European web archivists who have been involved in special collections related to COVID-19. The aim of the series is to provide a general overview of COVID-19 web archives.

Keywords: web archives, COVID-19, special collections, Hungary, National Széchényi 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 13 August 2020 with Márton Németh, Web Librarian at the National Széchényi Library in Hungary.

Web archiving at the National Széchényi Library began in 2017, although the initial idea arose in 2006. A pilot project ran until the end of 2019 during which tools were tested and the necessary infrastructure was set up for web archiving. Since early 2020, a permanent service model has been in place. The legal context for web archiving changed in May 2020 when the Cultural Law was extended to include web archiving. The law entitles the National Library to archive the Hungarian web thereby making it one of the core tasks of the institution. A ministerial decree is currently in development to establish detailed rules regarding rights and obligations of the National Library. (Németh, 2020b)

These web archive collections largely comprise content related to education, culture, public life and science in Hungary. Three types of harvests are done:

- Snapshots of the Hungarian web, comprising content on web servers on the .hu domain and other content related to Hungary;
- Harvests related to events, comprising relevant websites, blogs and specific sections of news portals (the COVID-19 collection falls under the event-related harvests);

- Periodic harvests of selected Hungarian websites based on specific themes, types of institution or genre. (OSZK Webarchívum, 2020a)

The web archive currently contains approximately 40 TB of data. More than 30.000 seeds are collected for the event-related and thematic harvests, and more than 270.000 seeds are collected in the broad crawl of the Hungarian web space. As for social media, currently 700 Instagram profiles are included in the collections. (Németh, 2020a)

The web archive collections are not yet accessible since the library is currently undergoing an infrastructural overhaul, nor are the metadata included in the library catalogue. In the future the archived content due to copyright restrictions will be made available in the reading rooms of the library. However, three collections are (partly) available online to the public: the Francis II. Rákóczi Memorial Year collection, the demo archive and the archive of the National Széchényi Library's websites.

THE REASONS OF THE SPECIAL COLLECTION

Why did you create a special COVID-19 collection?

Márton Németh: At the beginning of this year, we started talking about the kind of event-based harvests we would continue to do and which ones we would like to start when the first news appeared on the media that a global pandemic had perhaps started. Before it appeared in Hungary, we thought that we should perhaps create a thematic collection as part of our annual plan because we expected that it would be a global pandemic. Of course, COVID-19 would also affect Hungary, so we included it in our plan.

THE SCOPE OF THE COVID-19 COLLECTION

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

Márton Németh: We started collecting some websites, some related hashtags on websites and when specialist sections emerged from news outlets, we also collected those. It's also really important to note that we focus on content in Hungarian. It does not necessarily mean that this content is edited in Hungary. Some international resources that have been edited in the neighbouring countries or even further abroad are on the seed list of the collection. For example, the Turkish radio has some news related to the coronavirus in Hungarian.

Could you give some examples of the tags that you mentioned?

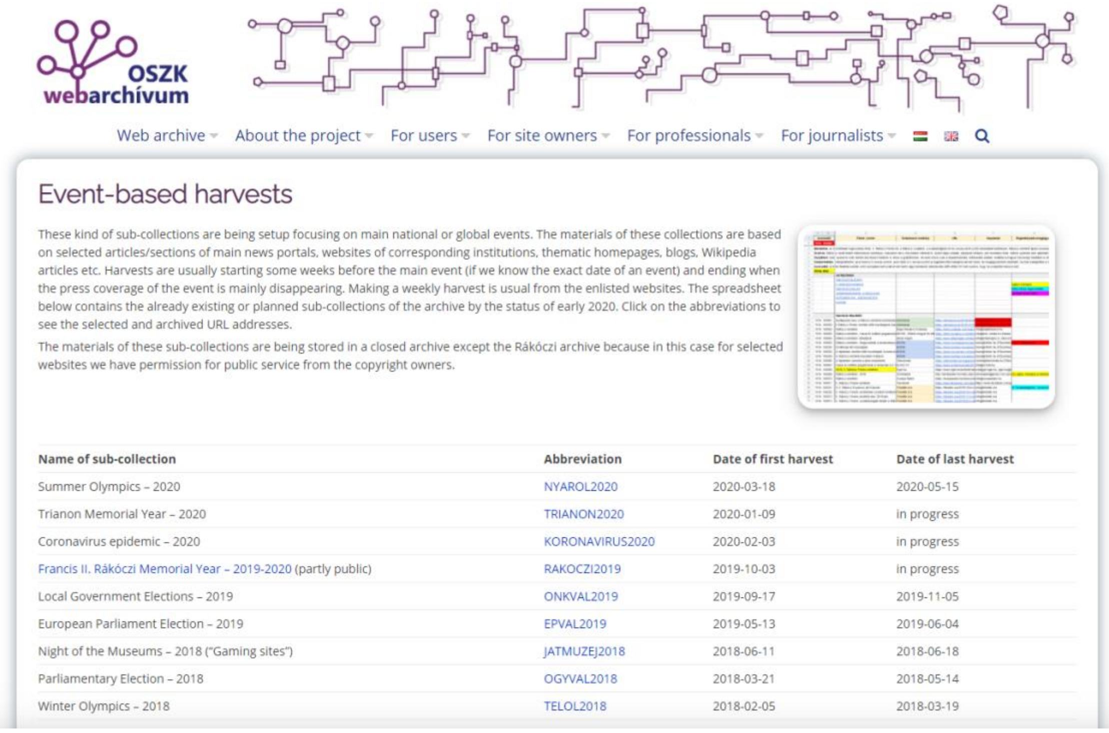
Márton Németh: The tags and the specific search terms we look for can be found in the URLs on the seed list. There are three main types of sources: search terms, sections and hashtags.

There is an example, of a news portal called ‘Index’. They created two specific sections on COVID-19: one about the Hungarian events and one about the global events that started in China and then appeared everywhere. In this case, we included both sections in our seed list. Another news website, ‘HVG’, is using a thematic hashtag *koronavirus* in Hungarian. The news outlet *Hirstart*, meaning new start, is also using thematic hashtags. When you see the term *tematikus címke* in the seed list, it means that there’s a thematic hashtag related to this source.

When you see *tematikus kereső* it means that we could use a search term that allows listing all the relevant resources that are related to the coronavirus. One example is the news service of the Hungarian public media hirado.hu. They don’t use a thematic hashtag nor a thematic section, but it was possible to search for a specific search term to find all the related resources.

In the case of some other resources we could find a special thematic section related to COVID-19, for example on the Euronews website or on the website of the daily newspaper of the Hungarians in Slovakia called *Uj Szo*. Another example is the Office of Statistics in Hungary where they have also created a specific section on COVID-19 on their homepage. On the official website of the city of Budapest, they even created their own subdomain focusing on issues related to COVID-19. The website of the Hungarian railway company also has a specific section on general issues and international travel, which trains are still operated etc.

In all these cases, we could clearly delimit the information to be crawled but as you can see, there is a relatively high level of variety in how this information was found.



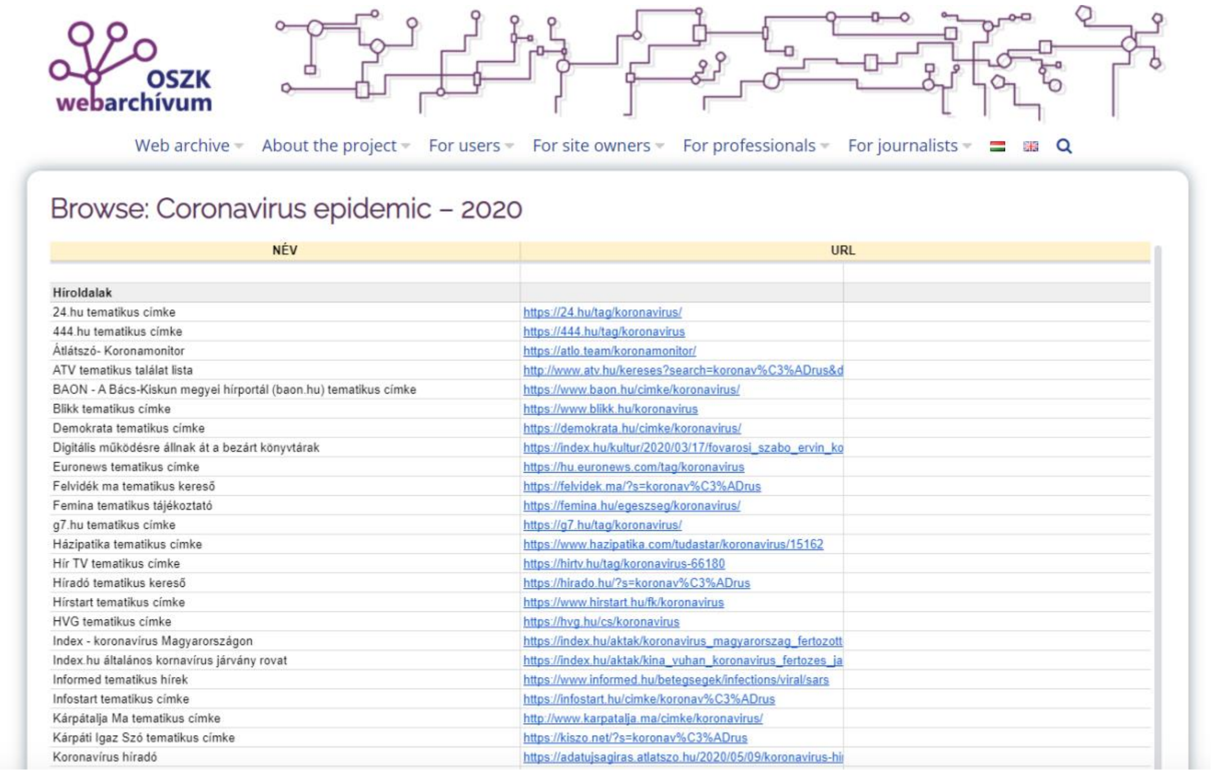
Event-based harvests

These kind of sub-collections are being setup focusing on main national or global events. The materials of these collections are based on selected articles/sections of main news portals, websites of corresponding institutions, thematic homepages, blogs, Wikipedia articles etc. Harvests are usually starting some weeks before the main event (if we know the exact date of an event) and ending when the press coverage of the event is mainly disappearing. Making a weekly harvest is usual from the enlisted websites. The spreadsheet below contains the already existing or planned sub-collections of the archive by the status of early 2020. Click on the abbreviations to see the selected and archived URL addresses.

The materials of these sub-collections are being stored in a closed archive except the Rákóczi archive because in this case for selected websites we have permission for public service from the copyright owners.

Name of sub-collection	Abbreviation	Date of first harvest	Date of last harvest
Summer Olympics – 2020	NYAROL2020	2020-03-18	2020-05-15
Trianon Memorial Year – 2020	TRIANON2020	2020-01-09	in progress
Coronavirus epidemic – 2020	KORONAVIRUS2020	2020-02-03	in progress
Francis II. Rákóczi Memorial Year – 2019-2020 (partly public)	RAKOCZI2019	2019-10-03	in progress
Local Government Elections – 2019	ONKVAL2019	2019-09-17	2019-11-05
European Parliament Election – 2019	EPVAL2019	2019-05-13	2019-06-04
Night of the Museums – 2018 ("Gaming sites")	JATMUZEJ2018	2018-06-11	2018-06-18
Parliamentary Election – 2018	OGYVAL2018	2018-03-21	2018-05-14
Winter Olympics – 2018	TELOL2018	2018-02-05	2018-03-19

Figure 1: Screenshot of the overview of event-based harvests included in the collections of the Hungarian National Library (OSZK Webarchívum, 2020b).



Browse: Coronavirus epidemic – 2020

NÉV	URL
Híroldalak	
24.hu tematikus címke	https://24.hu/tag/koronavirus/
444.hu tematikus címke	https://444.hu/tag/koronavirus
Átlátszó- Koronamonitor	https://atlo.team/koronamonitor/
ATV tematikus találat lista	http://www.atv.hu/kereses?search=koronav%C3%ADrus&d
BAON - A Bács-Kiskun megyei hírportál (baon.hu) tematikus címke	https://www.baon.hu/cimke/koronavirus/
Blikk tematikus címke	https://www.blikk.hu/koronavirus
Demokrata tematikus címke	https://demokrata.hu/cimke/koronavirus/
Digitális működésre állnak át a bezárt könyvtárak	https://index.hu/kultur/2020/03/17/fovarosi_szabo_ervin_ko
Euronews tematikus címke	https://hu.euronews.com/tag/koronavirus
Felvidék ma tematikus kereső	https://felvidek.ma/?s=koronav%C3%ADrus
Femina tematikus tájékoztató	https://femina.hu/egeszseg/koronavirus/
g7.hu tematikus címke	https://g7.hu/tag/koronavirus/
Házipatika tematikus címke	https://www.hazipatika.com/tudastar/koronavirus/15162
Hír TV tematikus címke	https://hirtv.hu/tag/koronavirus-66180
Híradó tematikus kereső	https://hirado.hu/?s=koronav%C3%ADrus
Hírstart tematikus címke	https://www.hirstart.hu/fk/koronavirus
HVG tematikus címke	https://hvg.hu/cs/koronavirus
Index - koronavírus Magyarországon	https://index.hu/aktak/koronavirus_magyarorszag_fertozott
Index.hu általános kornavírus járvány rovat	https://index.hu/aktak/kina_vuhan_koronavirus_fertozes_ja
Informed tematikus hírek	https://www.informed.hu/betegsegek/infections/viral/sars
Infostart tematikus címke	https://infostart.hu/cimke/koronav%C3%ADrus
Kárpátalja Ma tematikus címke	http://www.karpatalja.ma/cimke/koronavirus/
Kárpáti Igaz Szó tematikus címke	https://kiszo.net/?s=koronav%C3%ADrus
Koronavírus híradó	https://adatujsagiras.atlatszo.hu/2020/05/09/koronavirus-hi

Figure 2: Screenshot of the public seed list of the Coronavirus epidemic 2020 (OSZK Webarchívum, 2020c).

Is any social media content included in the collection?

Márton Németh: No, no social media is included in the collection. The problem is that currently we don't have good tools to archive social media sites. We did some experiments with Instagram, but Instagram is not relevant in the COVID-19 context. Twitter is not frequently used in Hungary. Facebook on the other hand is frequently used, but the only tool we found to crawl this content was Webrecorder. However, there are often problems with the scripts. So, we realised that we had to exclude social media resources for now. These can be important, but we don't have the technical capacities to collect this type of content.

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

Márton Németh: At this stage, it is hard to estimate the exact size of the collection because occasionally other non-COVID-19-related content appears in the WARC files. We have around 120 seed URLs on our list.

With regards to the nature of the content, we exclude videos since we're focusing on textual content. As mentioned, we also exclude social media for practical reasons. We focus on three aspects. We try to collect news on national, regional and international news portals and other resources. We also focus on official resources such as the official communication of the various bodies of State, for example the official bodies of the Ministry of Health that are responsible for managing the pandemic. It is interesting to note that official resources appeared in Hungarian, not just in Hungary but also in neighbouring Slovakia as part of the government websites in that area. In Romania, the party of the Hungarian minority, called the Democratic Alliance of Hungarians in Romania, also started an official information web page. They translated the most important elements of official State communication from Romanian into Hungarian. There are some parts of the country in which Hungarians are the local majority and their language competencies in Romanian are often rather weak because they are living in and staying within the local community. Of course, this can be critical in a pandemic situation. So, we also focused on finding these resources.

As for the tags, we collect some sites that are not technically news outlets but focus on health issues and of course these also focused on COVID-19. Whenever we could do data mining on a special section for a hashtag, we pointed the crawler to that specific content and included those sites in our collection.

To conclude, we can say that there are three main elements: the general international, national and regional news portals in Hungarian, the official State communication and other information resources that are available in the Hungarian segment of the web.

How do you archive nationally something which is fundamentally global?

Márton Németh: The most important criterion from this point of view is the language. We are focusing on the appearance of global events in Hungarian. Of course, a global pandemic is a fundamentally global phenomenon, but on the other hand, I think it's important to try to archive the appearance of this global phenomenon in the Hungarian public life. This collection in my view is a major element in this effort. I think that the global and the national level complement each other including different viewpoints.

THE FRAME OF THIS SPECIAL COLLECTION

When did you start? When did/do you plan to stop? What was the capture frequency?

Márton Németh: We started collecting in February and we are doing weekly harvests. For the 120 seed URLs we only include the relevant sections of the websites. It's a very important distinction. When we couldn't locate a particular section or when we couldn't browse by hashtags focusing on COVID-19, we don't harvest complete websites. When we only had the option to crawl the entire website, we abandoned the site. We used this strategy before in other collections so it has become our main policy because at the beginning we only had a restricted amount of storage space on our server.

In general, when you are creating a collection, you have to state the limits of the collection very clearly. When it is not possible to find the exact elements, it is not worth archiving it for us. Often this relates to large general news outlets, where it is simply not possible, technically, to crawl everything. Of course, if an entire website is focusing on COVID-19, we crawl it entirely, but these are exceptions. There are only two or three such cases. The central library of the Semmelweis University, the Hungarian Medical University in Budapest, for example, started a special information service for the general public about COVID-19. Since it's entirely about this topic we started harvesting the entire website.

We would like to constantly update this collection while COVID-19 is an important topic in the public life. It's impossible to tell how long we will continue these weekly harvests. I think that a second wave can appear and we still don't have a solution for how we can prevent the appearance of COVID-19 nor do we have a vaccine. I think that when the official emergency situation will be called off, we can start thinking about how long we will continue to crawl this collection. Currently we can't estimate the exact end date.

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

Márton Németh: Another major problem that is also important for the event-based collections in general is that we only have two full-time staff members. We also have a project coordinator and the head of our directorate, but they are both working on other projects as well. Currently we do not have the human capacity for doing quality control. It's simply not possible.

ACCESSIBILITY AND SEARCHABILITY

What about access to and searching in this collection?

Márton Németh: One major restriction is that due to copyright reasons, we cannot offer access to the collection. Another important issue is that the infrastructural background of the service environment is not ready yet. We hope that dedicated terminals will be made available in the library so that we can provide access to these closed collections. These are plans for the future. It's a strange situation that we started the web archiving project at the same time as a huge infrastructural re-establishment of the library. That's how this special situation came to be. So we are currently crawling this information for future use but we can't offer access to the archived websites yet.

PARTNERSHIPS AND USES

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

Márton Németh: Researchers haven't shown interest in the collection yet. When researchers visit our reading rooms, we can show them some of the work we are doing. For other collections, we talked with researchers and shared our experiences with them. They sometimes run small-scale crawl projects of their own so we could compare those with our results. But in the case of the COVID-19 collection, no requests have been submitted yet.

How do you communicate about this special collection?

Márton Németh: The seed list itself is public. It's important to us that the people know what kind of sources are included in the archive. It can also be a relevant information resource for them if they want information about COVID-19. These are the websites they can trust. Luckily these websites have also been collected by the library of the Semmelweis University. It's very useful that people can use two collections of information resources from different origins and that they can compare them.

The reference to the collection is also included in the news section of our new home page. Every time we have a media appearance, we also mention it. For example, the project coordinator last spoke about the project and the head of our department participated in an interview for the Hungarian public. We also wrote some professional articles about the project. Every time we mention that we have this special collection that reflects the current social challenges and the public life.

It was also interesting that after the first wave of COVID-19 appeared, an online event was organised by the technological section of the Hungarian Library Association about the impact of COVID-19 on library workflows and services. I shared our experiences and viewpoints during this event as well. Many interesting aspects were mentioned, for example the fact that the home use of various full-text resources increased exponentially.

It's also important to mention in this context that some other departments of the library were also focusing on different aspects of COVID-19. For example the special library and information science collection of the National Library started to spread information about the availability of library services in a national sense. They collected the information and the public could find the relevant information in one place. They also started to collect the experiences in several European countries and in the U.S. and compiled a weekly online review of these experiences. This is also very useful content for library users. So together with our colleagues, we created a small information service portfolio about this special event considering different viewpoints and different kinds of resources. So, our efforts in the web archiving field do not stand alone. We can work together with other departments as well. They are not really focusing on these issues now, but I think that if COVID-19 will get serious again, they will start the regular information services again. It's a bit different for them since they are making materials available for the general public, whereas we are crawling this information mainly for the future. We continue our weekly harvests, but our colleagues in charge of the Library and Information Science collection need to reflect about the actual needs of the users.

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

Márton Németh: We haven't received more help in Hungary, beyond that which I have already mentioned. We have shared our list with the IIPC when the IIPC content working group started to create the collaborative collection. We copied the URLs and other important metadata from one Google table to another. Of course, the structure of the tables is different so we had to make some changes. I must check again if it is still up to date because the last update was about two months ago. I think that 95% of the URLs are on the IIPC seed list as well. This kind of international collaboration shows us that the local and global elements are working together. So, you can find the local resources very effectively, but it's also important to share these resources with the international community because in this way some comparative research projects can be done.

I know that it's difficult in all European countries to grant access to the archived web resources, but even if you can only compare the differences between collection policies, it can be an interesting point of view for a comparative project.

So, you do not work with contributions from other partners in Hungary nor the general public for this collection?

Márton Németh: That is correct. We haven't been in touch with any local stakeholders. No one was looking for any kind of collaboration, but we informed the Ministry responsible for culture about our collection and we sent them the link to the seed list. They were appreciative that we are working on it. They also gather links to resources that are relevant for COVID-19. So, this kind of collaboration appeared, but no other collaborations were initiated.

You know, especially in the U.S., many institutions are not really managing the technical processes since they have an agreement with Archive-It or another institution. In Europe on the one hand, we don't have the financial capacity for this and on the other hand, we found that it is much better to manage these technical issues and make decisions about the collection policy by ourselves. I know that these stakeholders have the technical expertise and capacity, but I think this European model is much more suitable even though we have some problems related to capacity and even though we can't do the quality control of the collection. The new law in Hungary also states that it's a core part of the national library portfolio to work on these issues.

Is there anything else you would like to discuss that we haven't talked about yet?

Márton Németh: With regards to archiving social media, there are serious problems. We can develop a strategy for what we would like to archive but we simply don't have any specific tools that we can use effectively. Archiving Facebook is especially challenging. Webrecorder now also has a problem with Twitter as well but they are working on solving that problem based on community-based collaboration. I'm not sure if there will be enough capacity to manage these issues. The IIPC has to be an umbrella in order to organise the management of these automatic scripts because, by itself, this community will not be established. An umbrella organisation is definitely needed.

REFERENCES

Németh, Márton. (2020a, August 8). Personal communication with Márton Németh.

Németh, Márton. (2020b, August 26). Personal communication with Márton Németh.

OSZK Webarchívum. (2020a). *Basic information and data*. Retrieved from https://webarchivum.oszk.hu/en/for-journalists/basic-information-and-data/.

OSZK Webarchívum. (2020b). *Event-based harvests*. Retrieved from https://webarchivum.oszk.hu/en/webarchive/sub-collections/event-based-harvests/.

OSZK Webarchívum. (2020c). *Browse: Coronavirus epidemic - 2020*. Retrieved from https://webarchivum.oszk.hu/en/webarchive/browse/browsing-in-the-event-based-subcollections/browse-coronavirus-epidemic-2020/.

We would like to thank Julie M. Birkholz (KBR and Ghent University) for her help in proofreading this interview.

WARCnet Papers is a series of papers related to the activities of the WARCnet network. WARCnet Papers publishes keynotes, interviews, round table discussions, presentations, extended minutes, reports, white papers, status reports, and similar. To ensure the relevance of the publications, WARCnet Papers strives to publish with a rapid turnover. The WARCnet Papers series is edited by Niels Brügger, Jane Winters, Valérie Schafer, Kees Teszelszky, Peter Webster and Michael Kurzmeier. In cases where a WARCnet Paper has gone through a process of single blind review, this is mentioned in the individual publication.

The aim of the WARCnet network is to promote high-quality national and transnational research that will help us to understand the history of (trans)national web domains and of transnational events on the web, drawing on the increasingly important digital cultural heritage held in national web archives. The network activities run in 2020-22, hosted by the School of Communication and Culture at Aarhus University, and are funded by the Independent Research Fund Denmark | Humanities (grant no 9055-00005B).

WARCNET PAPERS

INDEPENDENT
RESEARCH FUND
DENMARK

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