IFLA Section
News Media

IFLA International News Media Conference 2024

Collecting Online Newspapers and Bypassing Paywalls

Jari Heikkinen
Legal Deposit Office, National Library of Finland, Helsinki, Finland.
jari.heikkinen@helsinki.fi

Topi Chamchoon
Legal Deposit Office, National Library of Finland, Helsinki, Finland.
topi.chamchoon@helsinki.fi

Samuli Sairanen
Legal Deposit Office, National Library of Finland, Helsinki, Finland.

Joel Nieminen
Legal Deposit Office, National Library of Finland, Helsinki, Finland.

Sanna Haukkala
Legal Deposit Office, National Library of Finland, Helsinki, Finland.

Abstract

The Legal Deposit Office of the National Library of Finland has been systematically collecting articles from online newspaper sites and media platforms, as well as other web materials, since 2007. Currently, the initiative extends to around 800 Finnish newspapers and journals, engaging in an ongoing process of article harvesting. This project does not encompass the harvesting of digital editions of periodicals; therefore, it is imperative to select periodicals that provide article content on their websites.

Although numerous online newspapers offer open access, the challenge persists with many being subscription-based, with articles concealed behind paywalls. Consequently, the web crawler is limited to retrieving merely images and snippets of text from the article's commencement.

Confronting this impediment, the National Library of Finland has conceived a methodology for accessing articles behind paywalls. There are two primary strategies for harvesting paywalled articles: one involves IP address recognition; and the other entails obtaining login credentials directly from the newspaper publishers. These credentials are then integrated into the collection tool, facilitating the harvest. This approach necessitates a sustained partnership with publishers, especially as they frequently revise their login procedures, which in turn requires the harvesting tool to be updated with new protocols. Presently, the Library successfully collects articles behind the paywalls of approximately 100 online newspapers.

Acknowledging that the endeavour to harvest paywalled articles is an ongoing task in the face of evolving technical landscapes, it is essential to remain continuously adaptable and vigilant. Nonetheless, the endeavour is useful, considering the discrepancies that may exist between the content, illustrations, and headlines in online newspapers compared to their printed counterparts. Through this paywall project, the National Library of Finland diligently addresses the complexities involved in archiving the evolving landscape of online media.

Keywords: Paywalls, Newspapers, Web Harvest, Legal Deposit, Web Archive

---

1.1 Introduction

One of the strategic goals of the National Library of Finland (2021b) states: "We will be a world leader in the preservation of the diverse and changing materials included in the born-digital cultural heritage." An important part of achieving this goal is to collect articles and other content from news and media platforms, as they are at the core of online news reporting today, with the role of printed newspapers evolving into a collection of articles that have already been published online. With news constantly being updated, it is interesting to preserve even the ongoing reporting process.

Although many newspapers provide digital editions on their platforms, we will not focus on collecting them in this writing but will discuss articles on the websites of online newspapers, whose harvest is challenging due to paywalls. In an online news article, constantly evolving events are updated in coverage. A printed article or a digital edition consists of one version of the news story at a given moment in time.

1.2 The Finnish Web Archive

In addition to publications and recordings, the National Library harvests and preserves publicly available Finnish online materials. The preserved online materials are a representative and diverse sample of materials available to the public in information networks at various times. Approximately 20 terabytes of online content are preserved annually, with a total collection of over 300 terabytes. The Finnish Web Archive has been collecting online content since 2006, and more material is harvested from websites (e.g., HTM and HTML websites) as well as social media platforms (e.g., Twitter, YouTube, Facebook and TikTok). The collection currently comprises more than 4.3 billion files.

There is also an extension of the Finland collection in its pilot phase, where a more detailed collection is conducted four times a year on the websites of certain key organizations.

Harvests are divided into three types. The annual Finnish domain harvest covers country code top-level domains .fi and .ax, as well as others which have been identified as Finnish language. This harvest does not select websites for harvesting based on their topic, theme or content. Approximately 600,000 websites are covered by the harvest.

Secondly, thematic web harvests supplement the annual harvests. They collect, as comprehensively as possible, online content relating to a specific issue or topical event. Content is harvested from both websites and social media platforms.

Topics for thematic web harvests include:

- Governmental and social events, phenomena and changes (e.g., elections, state visits, strikes, international conferences)
- National events (e.g., sports competitions, cultural events)
- Major events in global politics (e.g., conflicts and wars) and natural disasters

The content for thematic harvests is identified either in collaboration with partners or within the National Library. The harvests are conducted alongside museums, archives, research institutions, researchers, and enthusiasts. Recent thematic harvests included harvests about political strikes in Finland this spring, the 2024 presidential election, the 2024 presidential election, and a collaboration harvest with a group of researchers focusing on Karelian language and culture online.

Thirdly, there are harvests of continuously updating contents. These are divided into the newspaper harvest discussed here and the X harvest (formerly known as Twitter harvest) carried out from 2020 to 2023, which targeted approximately 3 500 organizations and individual accounts.

The URL addresses of the online materials can be browsed in the index of the Finnish Web Archive[1] which can be accessed anywhere online. The materials themselves are only available at the legal deposit workstations, which, in addition to the National Library of Finland, locate in five other legal deposit libraries across Finland, and in the Library of Parliament, and in the National Audiovisual Institute.

In Accordance with the Act on Collection and Preserving Cultural Materials, the collection plan for online materials[2] (covering the period 2021-2024), specifies the extent of harvesting and preserving content as well as the practices of releasing online resources. The collection plan also considers the needs of research and the archiving of cultural history as well as equal treatment of online publishers. The plan is reviewed at intervals of four years or less According to the Act, the National Library must present a plan on the scope of the collecting of online materials and on the related deposit practices, to be approved by the Ministry of Education and Culture. The plan may

[1] https://verkoarkisto.kansalliskirjasto.fi/?lang=en (accessed 23 April 2024)

[2] https://www.kansalliskirjasto.fi/en/legal-deposit-office/online-materials/collection-plan-online-materials-2021-2024 (accessed 17 April 2024)

be reviewed during its period of validity if the Finnish publishing industry or the technical or financial resources available to the National Library change significantly.

In addition to the Act on Collecting and Preserving Cultural Materials, collecting of cultural heritage is also guided by the guidelines of the United Nations Educational, Scientific and Cultural Organization (UNESCO) (Choy et al., 2016). Furthermore, international collaboration, for instance with the International Internet Preservation Consortium (IIPC) and other organizations collecting online materials, is increasingly important. The continuous development and internationalization of online publishing require cooperation in selecting materials to collect, as well as in developing and implementing harvesting and preservation technologies.

If collecting is not possible due to barriers like paywalls or registration requirements, the National Library may request permission for harvesting or for the depositing of materials. Harvesting these materials requires cooperation with online publishers or licensors.

In addition to the web content itself, related metadata is collected, and the long-term digital preservation of the material is ensured.

1.3 Newspapers in Finland

Finland is a country with a vast surface area, home to many strong local identities, including the Swedish-speaking regions and the northern Sami areas. As of 2023, there over 250 newspapers, including local newspapers and free of charge-newspapers (based on statistics collected by the National Library of Finland). The number of newspapers in Finland is high when compared to the population, and when comparing the number of residents to the total circulation of the press, Finland ranks third in the world.

Newspapers are important reflections of their time. Each significant publication has an online presence, from daily newspapers to those published less frequently. Altogether, they cover Finnish news reporting from global news to local happenings.

Newspapers in particular are widely read. According to the National Media Survey (News Media Finland, 2023), 96 percent of Finns aged 15 and over read newspapers. This equates to over 4.1 million people. A printed newspaper is read by every other Finn (52%), with just over a quarter (28%) of the youngest and three-quarters of the eldest reading them.

Digital newspapers alone are read by 1.9 million Finns. The same number (1.9 million) reads newspapers both in print and digital form. Only in print, newspapers are read by 344,000 Finns.

Newspapers' reach is very thorough: they reach 91–97 percent of Finns aged 15 and over, 91–99 percent of those working in various professional groups, and 94–98 percent of households with different income levels.

Digital newspaper content is read by 88 percent of Finns. Digital reading is most common among the 35–44 age group, where 95 percent engage in it. The least usage of digital newspaper channels is by those aged 65 or over, even though more than three-quarters of them (77%) read digital newspapers either alongside the paper version or solely in digital format. In the past year, digital

newspaper reading surpassed print reading among those over 65 (75%). Today, the digital use of news media is more common than print among all age groups.

In Finland, it's the young, affluent and well-educated who trust and engage with the news, who retrieve their news directly from news media websites and who more commonly pay for news. There is now no difference between men and women in terms of paying for news. However, in 2021, men (22%) paid for news slightly more often than women (19%) (Reunanen et al., 2022).

1.4 Magazines in Finland

Currently, around 2,500 magazines are published in Finland. The number has halved since the turn of the millennium, but relative to the population, Finland still publishes a considerable number of magazines. A magazine is defined as a publication that is available for subscription or is widely accessible and primarily contains editorial material. A magazine can be either a printed publication or an online publication.

According to the Finnish Magazine Media Association (2021), 63 percent of Finns read magazines at least weekly, and 94 percent at least monthly. Two-thirds of readers also read the entire magazine. The most-read categories include women's magazines, news, current affairs and economic magazines, science magazines, general-interest magazines, as well as free customer magazines.

56 percent of magazine readers also read magazine content digitally, which can be considered a high figure given the nature of the medium. Digitally, magazines are most read as articles on websites (40%), but also as digital editions and on mobile phone screens.

1.5 The difference between an online article and a printed one

The fact that clickbait headlines online differ from headlines in the print version is very well known, but currently there are also significant differences in the actual news content. An online article evolves as the news story develops, unlike a printed newspaper's interpretation which is set in stone. Secondly, many stories are no longer even printed.

Born digital articles are targeted at those who read printed newspapers less frequently**.** Nowadays, a printed newspaper can be regarded as a summary rather than a comprehensive result of all the news content that newspaper companies produce. In addition, there are readers' comments on articles, which in themselves are often of interest: the ability to comment on online news according to Kangaspunta (2020) is a meaningful resource for users of online media and is perceived, for example, as an important form of civic engagement.

Therefore, from the perspective of collecting and preserving cultural materials, it is necessary to collect online newspapers as they more fully show what media companies want to offer to different audiences. From a research perspective, they may not be as clearly defined material as digitized newspapers, and search functions in web archives may not match the precise search results produced by text recognition. Hence, the role of web archiving may be more to act as a general representation of how news is reported at a certain point in time.

1.6 Paywalls

Challenges in harvesting online newspapers, of course, include paywalls, which everyone has likely encountered at some point. Paywalls in online newspapers are a digital method to limit users' access to content without a subscription or payment. The core idea of paywalls is to provide a revenue stream for the newspaper instead of continuous distribution of free content. This is in response to the fact that traditional media outlets have lost advertising revenues to free distribution channels online, such as social media and search engines.

There are different types of paywalls. Globally common paywalls can be roughly categorized into the following (Arrese, 2015).

Hard paywall: This prevents all content from being viewed without a subscription. Only certain sections, often non-editorial, such as the front page or job listings, may be free. An example of a hard paywall is The Wall Street Journal.

Soft paywall: soft paywall allows users to read a certain number of articles for free over a given time (for example, ten articles per month) before requiring a subscription. The New York Times uses this type of system.

Hybrid/freemium model: In this case, some stories are freely accessible, but the most expensive premium content is behind a paywall. Generally, what is free to read is usually what is the cheapest for newspapers to produce: namely short stories, often obtained from news agencies, or articles aimed at young readers. The most expensive, and laborious articles, which have required a journalist on the field, are worth keeping behind a paywall.

Cookie paywall: This version requires acceptance of ads and third-party cookies, which is prohibited in many countries and is often considered unethical.

Finnish newspapers most commonly use the freemium model, where the most editorially labor-intensive articles and particularly expansive feature-type stories are typically made to be paid content. Smaller-circulation local newspapers may use a hard paywall, where all articles, apart from non-editorial notices and freely distributed stories, are behind a paywall.

Newspapers are established channels of information in their own region, so the articles that become more expensive through editorial work are worth putting behind a paywall. The exception would be agencies like the Finnish News Agency (STT), which produce articles for free.

Generally, paywalls are justified by the fact that they enable funding for the work of journalists and content producers. If credible news activity cannot secure its funding, the resulting news vacuum can benefit various propagandist fake media outlets; quite often they are financially backed by state actors.

Paywalls are practically a feature of all subscription newspapers; recently, even the online sites of Finnish tabloids have sections that require payment.

In contrast magazines, paywalls are less common: some use newspaper-like paywalls (like Finland's leading weekly magazine, Suomen Kuvalehti), while others only leave freely readable feature stories on their sites (often recipes or knitting patterns), with the main purpose of making readers interested in the magazine itself without frustrating them with barriers.

1.7 National Library's Online Newspaper Harvest

The National Library of Finland has been harvesting online news content since 2009. Initially, only the sites of daily newspapers were harvested, but in 2010 the harvest expanded to include other types of periodicals on a monthly rotation. In 2014, sites of weekly magazines were also added. Currently, the National Library archives 70 newspaper and magazine sites daily, about 330 newspapers and magazines weekly, and the same number monthly, making over 700 periodicals. The collection includes daily newspapers, magazines, local newspapers, free distribution newspapers, scientific journals, cultural magazines, membership bulletins, party publications, customer magazines, organization newsletters, as well as news media websites (such as those of the Finnish Broadcasting Company, Yle), which are structurally and content-wise similar to newspaper websites.

Over 10 years ago, when the project began, many online publications did not have paywalls, but they became more common over the 2010s as subscription numbers for newspapers declined sharply.

At the same time, problems with collecting online newspaper content became evident; merely collecting the headline, image, and the beginning of an article—or in the worst case, an empty front page—was no longer satisfactory. A process is needed to be developed to regain access to news content.

1.8 Collecting Content Behind Paywalls

The first attempts to harvest news content behind paywalls were made in the mid-2010s, but due to their complexity which involved stitching together warc files, the results were meager. A new attempt was started in 2023, employing the Browsertrix software, which is more suitable for these types of harvests, alongside the traditional Heritrix software.

Technically, these two software solutions differ significantly from one another. While Heritrix uses a specially developed browser, Browsertrix's collection is based on Chromium-based browsers. The first version of Heritrix was released in 2004, at which time the web was substantially different from what it is now, 20 years later. Browsertrix, on the other hand, was first released in 2021. Often, when attempting to harvest modern web pages with Heritrix, its age shows as it simply fails to harvest content behind complex JavaScript applications.

In practice, the harvest takes place in two ways. First, there is logging in via credentials, where user login information is fed directly to the harvesting tool, allowing the tool to do its job.

The second method involves the publisher opening an IP address connection to the website and harvesting articles through this direct link. Media groups use varying login procedures in their publishing systems, so the same approach does not work for all newspapers. Each publishing system is different, and the input of login credentials does not always work in the same manner, leading to the necessity for group-specific login procedures that need to be tested and maintained individually. The diversity of these methods requires considerable customization work for

harvesting tools. Therefore, the IP connection approach is preferred, as it offers the easiest access to news material. However, in some cases, even with an IP opening, login on the website itself may be required.

Given the myriad ways in which modern web page JavaScript can implement login functionality, it often needs to be carried out manually, page by page, simulating page use with JavaScript. Yet, this is not always successful, as modern sites may utilize complex login systems with numerous mechanisms designed to thwart automated logins, requiring sophisticated methods to bypass them. Even custom login procedures tailored to specific groups can break at any moment due to changes in their backend systems.

This, in turn, makes quality control challenging. There is a continuous need to monitor how well the harvest is performing. Quality control is a significant future challenge, and it should not be assumed that every article behind every newspaper paywall will be successfully collected.

Sometimes, content may be successfully harvested but not visible in the archive interface. In such cases, although the content is collected, it is not immediately accessible to researchers. There is hope, however, that challenges with the user interface can be overcome in the future.

Requests for cooperation have been directed at the management of the major newspaper groups, thereby ensuring the highest possible level of attention for the request and delegating the technical process to their experts. This has also given simultaneous access to all newspapers owned by the media group. All newspaper groups have, sooner or later, responded to our requests, and as a result, we are currently (as of May 2024) collaborating with 14 newspaper conglomerates, which gives us access to approximately 150 online newspapers.

Ownership of newspapers in Finland, as in many other countries, is quite concentrated. According to a recent government report (Lehtisaari et al., 2024), the trend towards concentration in the newspaper industry continues, unlike in the radio and TV sectors where there has been a move towards increasing competition. However, there are still some independent newspapers, and due to the amount of work required, it may not necessarily be feasible to initiate cooperation with them. We only archive the front pages of these papers, but we do not seek access behind their paywalls.

1.9 ELK

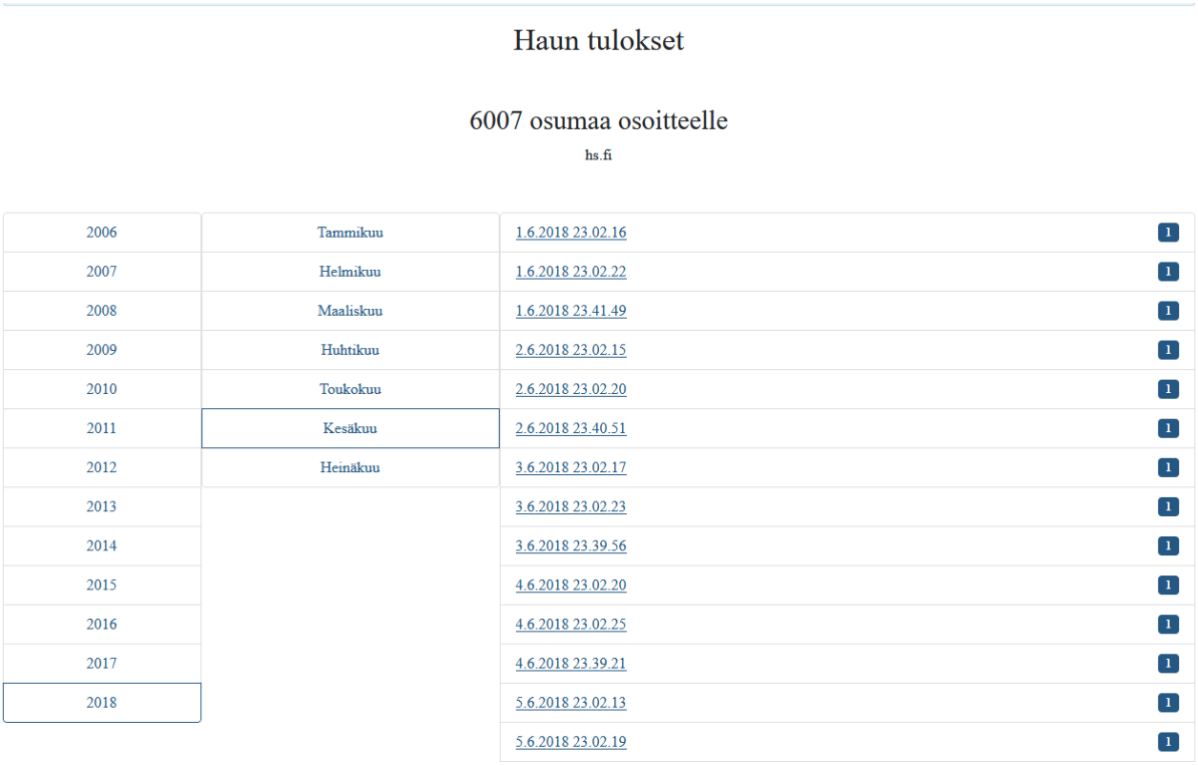
The Legal Deposit Office of the National Library of Finland has developed its own in-house tool, named Elk (a Finnish acronym for 'Elonleikkuukone', meaning 'Harvesting Machine'), for adding and configuring content to be harvested in the web archive. Within this system, new seeds (most often web URLs) can be added, and their harvesting depth can be defined, ranging from level 0 (the starting page only) to level three (the starting page + three links forward). The frequency of seed collection can be set to recur daily, weekly, or monthly if desired. Comments can also be added to a seed to facilitate identification of the harvesting target. The seeds that have been removed from a harvest are kept in the Elk to retain information about previous harvesting, but it can be hidden from the user interface. The future development of Elk aims to allow the employee who added the seed to also perform the harvesting and check the quality of the harvest through a preview feature.

In the case of newspapers, the URLs to be harvested (seeds) are added to Elk and categorized into daily, weekly, and monthly collections. Within these collections, they are further divided into those

without paywalls and those with paywalls. Paywalled collections are subdivided into those harvested behind the paywall using Browsertrix software, and those for which an agreement to bypass the paywall hasn't been made, thus only the headlines and images are harvested using Heritrix (usually including a few lines of text from the beginning of the page as well).

Once stored in Elk, the harvested seeds pass through indexing and subsequently become available at the legal deposit workstations. Often, a newspaper's front page is harvested several times throughout the day since the harvesting program may return to the front page via other links on the site, and if the front page has been updated, the program will harvest it again because it is technically a different page.

These continuous newspaper harvests, like other ongoing harvests and thematic harvests, have also been cataloged in the National Bibliography to improve discoverability.



Index of the Helsingin Sanomat newspaper collection in the web archive. The menu displays daily captures from June 2018, which the user can access at the legal deposit workstation.

1.10 Finally

The paywall project was initiated with newspapers. In the future, it may be expanded to include magazines; however, as mentioned before, paywalls are less common in magazines.

Even now, it can be said that significant results have been achieved. Collecting outcomes are never perfect, and with the changing login protocols there must be alertness to constantly improve the harvesting process. The project is not one that can simply be started and then left to run on its own. The quality must be continuously monitored, and this requires resources. Of course, there is also

the risk that the volume of quality control and labor intensity required for many periodicals may make the project too demanding over time.

However, the quantities of news content successfully harvested from behind paywalls demonstrate that this project has been worthwhile. We estimate that, on average, we have a well-functioning connection to about one hundred newspaper sites and media platforms.

In Finland, cooperation with the publishing industry on collecting online newspapers has so far been smooth, with all parties aware of the importance of preserving and providing contemporary news content for future generations.

References

Arrese Á (2015) From Gratis to Paywalls. *Journalism Studies* 17(8): 1051–1067.

Choy SCC, Crofts N, Fisher R, Choh NL, Nickle S, Oury C and Slaska K (2016) The UNESCO/PERSIST guidelines for the selection of digital heritage for long-term preservation. Available at: https://unesdoc.unesco.org/ark:/48223/pf0000244280 (accessed 19 April 2024).

Finnish Magazine Media Association (2021) Magazine moment 2021. Available at: https://www.aikakausmedia.fi/media/3699/magazine-moment_finnish-magazine-media-association_2021.pdf (accessed 24 April 2024).

Kangaspunta V (2020): *Online news comments as a form of public participation*. PhD Thesis, University of Tampere, Finland.

Lehtisaari K, Grönlund M, Hellman H, Ranti T and Suikkanen R (2024) *Media concentration and diversity of media content in Finland*. Helsinki: Prime Minister's Office.

National Library of Finland (2021a) Collection plan for online materials. Available at: https://www.kansalliskirjasto.fi/en/legal-deposit-office/online-materials/collection-plan-online-materials-2021-2024 (accessed 17 April 2024).

National Library of Finland (2021b) Duties and Strategy. Available at: https://www.kansalliskirjasto.fi/en/about-us/duties-and-strategy (accessed 23 April 2024).

News Media Finland (2023) Sanomalehdet tavoittavat miltei kaikki suomalaiset. Available at: https://www.uutismediat.fi/ajankohtaista/sanomalehdet-tavoittavat-miltei-kaikki-suomalaiset/ (accessed 24 April 2024).

National Library of Finland (2024a) Online materials. Available at: https://www.kansalliskirjasto.fi/en/legal-deposit-office/online-materials (accessed 24 April 2024).

National Library of Finland (2024b) Publication statistics. Available at: https://www.kansalliskirjasto.fi/en/legal-deposit-office/publication-statistics (accessed 23 April 2024).

Reunanen E, Alanne N, Helske H, Lappalainen E, Niemi MK, Petterson M and Seuri V (2022) *Uutismedia verkossa 2022: Reuters-instituutin Digital News Report Suomen maaraportti*. Tampere: University of Tampere.

Reunanen E, Alanne N, Huovinen T, Järvinen U, Nevalainen R, Puolimatka R and Vehkasalo V (2023) *Uutismedia verkossa 2023: Reuters-instituutin Digital News Report Suomen maaraportti*. Tampere: University of Tampere.