IFLA WLIC 2019 Athens

Submitted on: 21.06.2019

Opening Digitized Newspapers for Different User Groups - Successes and Challenges

Juha Rautiainen
Research Library, the National Library of Finland, Mikkeli, Finland.
E-mail address: juha.rautiainen@helsinki.fi

Abstract:

In recent years the National Library of Finland (NLF) has taken several initiatives to enable access to the digitized Finnish newspapers for wider use. Technical improvements in the presentation system (Digi) and agreements with Finnish copyright organizations have made it possible to provide access to copyrighted material in different ways.

Since 2016 NFL has opened over 4 million pages of digitized newspapers and journals from years 1911-1929 to open online use in the Digi system. This has doubled the digitized material available outside of the legal deposit libraries. The opening has benefited both the general public and researchers.

In a pilot project, digitized newspapers and journals from years 1930-2010 are opened for research use. During the one and half year's pilot period, authorized users are able to access the materials in restricted use through the Digi system from their own premises and with their own devices.

The NLF has promoted the use of newspapers as data in research by providing ready-made datasets available in the Digi system. The datasets contain all the digitized newspaper pages from 1771 to 1910 in the ALTO XML format and some other data collections.

The datasets are sufficient for many users, but customized packages have also been required. One of these cases is the Horizon 2020 funded EU project NewsEye in which the NLF is participating. The aim of the project is to develop new integrated tools and methods for effective exploration and exploitation of digital newspapers by means of new technologies. The NLF provides to the project a set of 0.5 million pages of digitized newspapers selected with the researchers in the project.

All in all, efforts to increase the use of digitized newspapers have been successful in many ways. However, a number of issues still needs to be considered in the future. The paper summarizes experiences so far.

Keywords: online use, newspaper, digital collection, digitization, copyright.

Introduction

In recent years the National Library of Finland (NLF) has taken several initiatives to enable wider access to the digitized Finnish newspapers, journals, and books. The focus has been to enable access to the material under copyright outside of the legal deposit libraries.

Based on the law, the NLF is in charge of recording and maintaining Finnish national publishing heritage and ensuring its accessibility[1]. In addition, NLF is in charge of the assignments defined in the act on the depositing and conservation of cultural materials[2] including conserving Finnish publishing products for the use of researchers and others (legal deposit operations), in cooperation with the operators of the publishing sector.

Making copyrighted digitized resources more available for research and teaching is a part of the strategy of the National Library of Finland for years 2016–2020[3], and according to its Openness policy[4], The NLF promotes open access to the cultural heritage and aims to ensure maximum access to resources in its agreements.

Since 2016 the NFL has opened over 4 million pages of digitized newspapers and journals from years 1911-1929 to open online use in the Digi system. Also, the NLF is piloting arrangements enabling the wider research use of the copyrighted material.

The paper summarizes experiences of the actions taken by NLF focusing on openings of the digitized newspapers. Chapter 1 describes the most essential cases the NLF has promoted wider access to the digitized newspapers. Chapter 2 describes the experiences and evaluates the results of the openings from different perspectives. Chapter 3 presents some conclusions and plans.

Digi System and the Digitized collections

All newspapers and periodicals digitized by the NLF as well as electronically deposited new newspapers are in the digi.kansalliskirjasto.fi (Digi) system.[5] Also, a selection of books and some other digitized materials are in the same system.

In spring 2018 Digi contains over 16 million pages of digitized materials, which can be searched with free text. Overall, the service has almost 8.5 million pages of newspapers, of which about 48 percent are in open online use.[6] The rest of the newspapers may be used on the customer workstations in six legal deposit libraries.

The NLF has a digitization plan guiding the digitization production. All the historical newspapers in the collection have been digitized up to year the 1935 and plan is to digitize the newspapers up to 1940 in the year 2019. Since 2016 all the new legal deposit newspapers have been digitized right after arrival. From the period between those years, only a few newspaper titles have been digitized.

1 PROMOTING THE WIDER ACCESS TO THE DIGITIZED NEWSPAPERS

This chapter describes four different cases where the NLF has provided digitized newspapers for wider use. In openings for public online use and Haka project great share of the material is

under copyright, whereas in export packages and NewsEye project most of the material is copyright free.

In addition to the uses described in the chapter, a few archives and museums have access to some digitized newspapers and journals. For example, the historical papers published by labour movement are available at the Library of the Labour Movement, the Labour Archives, the People's Archives and the Finnish Labour Museum via Digi system.[7] Also, a couple of newspapers have their digitized volumes in editorial use.

1.1 Public Online Use for General Public and Researchers

For several years only newspapers and journals published before 1911 were on public online use in the Digi and all the newer material was only available in legal deposit libraries. This changed in February 2017 when the NLF opened also the newspapers and journals published between 1911 and 1920 for public online use.[8]

In the beginning of 2018 the NLF opened digitized newspapers and journals published in years 1918-1929 for public online use for one year based on an agreement with copyright organization Kopiosto.[9] Next year the agreement was extended until the end of 2020[10]. The NLF pays to Kopiosto the compensation specified in the agreement.

Before February 2017 there were about 2 million pages of digitized newspapers in open online use. The openings increased the number of pages in open online use by 1 million pages each, so in January 2018 the total number of digitized newspaper pages in open online use was about 4 million pages.

In recent years the NLF has also digitized some limited selections of newer materials and opened them for public online use by separate agreements. These selections include about 27 000 pages of five newspapers digitized with Käkisalmi Foundation funding.

Before the wider opening of the digitized collection, in years 2015 and 2016, Aviisi-project was looking for ways to extend the availability of digitized newspapers under copyright for both education and research purposes. During the project users from participating organizations were able to use two digitized newspapers, Länsi-Savo and Maaseudun Tulevaisuus, in the Digi system. The participating organizations were schools, public libraries, archives, museums and research organizations mainly in the Mikkeli region.[11]

1.2 Restricted Online Use and Data Sets for Researchers

To promote the use of digitized Finnish newspaper and journal content in research the NLF started a pilot project with Kopiosto, Mikkeli University Consortium and the Ministry of Education and Culture[12]. In the project NLF opens digitized newspapers and journals from years 1930-2010 for authenticated researchers until the end of year 2019. The so-called Haka-project is carried out in two phases.

In the first phase of the project Haka-login was implemented into the Digi system in 2017[13]. By the Haka-login users are able to authenticate themselves to the system, which makes it possible to identify authorized users of a selected set of the material.

Haka is the identity federation of the Finnish universities, polytechnics and research institutions operated by CSC - IT Center for Science Ltd. It is the most used user authentication system on its field in Finland by 326 000 end users. In Haka user identities are provided by the users' home organization. Services, such as Digi, do get the user attributes from home organization and the users are able to access federation services using a single user account and password.[14]

The second phase of the project started in the summer 2018 and will end by the end of the year 2019. In this phase selected user groups are able to study copyrighted digitized newspapers and journals. The size of user groups vary between an individual research group and a whole university. All users from University of Jyväskylä are able to use the material, if they meet the criteria, i.e. are doing research or have supporting role in it. University of Helsinki participates with two groups, Faculty of Arts and Ruralia Institute. In universities of Oulu, Tampere and Turku the user groups are more precisely defined and smaller.[15]

The right to use the defined digitized material is based on the agreements between the NLF and Kopiosto and between the NLF and the home organizations of the user groups. The agreements allow users to use the material in the Digi system and do data mining. Use of the pilot material is free of charge to the participating organizations.

To fulfill requests from several separate researchers the NLF made available the export packages of the digitized newspaper and journal content in 2017. While the text content of each page is downloadable separately from the Digi system, there is a need for the export packages. If the data is available in larger sets, mass operations in the users' own environment are easier.

The packages contain the page-specific XML of the all digitized papers from the year 1771 to 1910 and they are available in the Digi system[16]. The packages are available for all users, but their use requires some expertise.

While the export packages are sufficient for many users, in some cases customized datasets are still required. For example, the NLF is providing tailored datasets for the NewsEye project.

The NLF participates in the NewsEye project, which uses digitized newspapers as the base material. The multi-disciplinary research project, funded by the European Union's Horizon 2020 research and innovation programme, involves national libraries, humanities, and social science research groups and computer science research groups. The project started at the beginning of May 2018 and will end after 3 years at the end of April 2021.[17]

The NewsEye project aims to develop a seamlessly integrated armory of tools and methods will be created that will improve users' capability to access, analyze and use the content in historical newspapers. The project will seek to improve existing tools on text recognition and article separation, multilingual and uncertainty-aware semantic text enrichment and dynamic test analysis for example.[18]

The total amount of digitized newspaper pages processed during the project is planned to be 1.5 million pages. NLF's share of the total is 0.5 million pages, dividing between Swedish- and Finnish-language newspapers. The material has been chosen so that its use is as free as possible.

The NLF is providing newspaper data for the project in three ways. Research groups can download the selected material from the Digi system page by page and download readymade

data packages. Also, the NLF is providing a tailored set of page images for the project to process.

2 MEASUREMENTS AND FEEDBACK

As can be seen in chapter 1, the ways how the NLF has promoted the wider access to the digitized material are quite different from each other. Therefore, it is not necessarily meaningful to measure them all with the same indicators. While the quantitative approach is a good baseline, qualitative indicators are also necessary to understand the whole.

During the Aviisi project, the NLF improved the monitoring and analyzing of the use of the digitized collection in the Digi system.[19] The most used metrics is the number of loaded page images of the papers, hereafter page views.

The NLF has collected feedback from users and information on the uses of the digitized material mainly by surveys. In addition, research users must give some background information when they login with Haka credentials the first time, and before downloading datasets users are asked to tell how they plan to use the data.

2.1 Public Online Use

The opening of the newspapers and journals published between 1911 and 1920 for public online use in February 2017 got quite much attention in the media[20]. The number of page views in February doubled from the previous year and remained higher throughout the year, leveling off slightly after the start. The digitized newspaper pages were viewed 751 014 times in February and 475 557 times in March. In 2017 the average number of page views per month was 342 146.

The opening of the newspapers and journals until 1929 at the beginning of January 2018 had less media attention, but page views of the digitized newspapers in open online use got up 1 120 098. That is 3.5 times more compared to January 2017 (218 423 page views).

Since the number of pages on open online use doubled due to the openings, the absolute number of page views could be misleading. Therefore also the ratio between page views and pages on open online use is interesting. In newspapers the ratio was 0.11 in January 2017, in February it rose to 0.25, and in January 2018 the ratio was 0.28.

In general, it seems that the more recent material is used more. The relative number of page views of the older material has stayed quite stable, while interest in new material in open online use is leveling off within one month of opening. In January 2018 the relative number of page views of the newspapers published before year the 1918 was 0.13, and for newspapers published 1918-1929 it peaked to 0.58. In January 2019 the relative number of page views for newspapers published before 1918 was 0.11, while for newspapers published between 1918 and 1929 it was 0.24.

In comparison, the newspapers opened for public online use by Käkisalmi Foundation funding have been popular. The selection contains only about 27 000 pages of newspapers, but the relative number of page views has stayed over 0.75 since the opening in the October 2018.

In their article[21] Pääkkönen and Lilja review the results of a survey done after the openings between 14th of June and end of June 2018. In the survey genealogy (36.1%) was the most popular reason to use the Digi system, private research (17.4%) and browsing (17%) followed. Scientific research (11.3%) was fourth most popular. The answers indicated that one user quite often had more than one reason to use the material.

The newspapers were the most used material, 90% of the participants told that they use them. The answers correspond to the statistics of use, the number of page views of newspapers are higher than other materials.

Based on the answers the papers published before 1917 are the most popular part of the material, which seems to be in conflict with statistics of use. On the other hand, in the open answers many respondents hoped the copyrighted newer material to open online use.

2.2 Use in Research

The Haka project is still ongoing, ending in December 2019, but some preliminary experiences can already be presented at this stage. The first users were able to access copyrighted newspapers and journals in the Digi system during summer 2018 and the material was available in all six organizations in September 2018.

The Haka federation gives general requirements and guidelines to the identity providers, but since user identities are provided by the home organizations, implementation of identity and access management may vary. One of the first observation was that defining a certain group of users was a time-consuming and difficult task if the group did not correspond to some existing group of users, such as a degree program.

Before the Digi system grants user access to the material, the user is required to answer a few questions considering the purpose of use. 292 users have answered the questions by the end of April 2019. 44 percent of the respondents are from the University of Helsinki and 38 percent from the University of Jyväskylä. The remaining 18 percent of respondents are from the University of Turku (8%), the University of Tampere (6 %) and the University of Oulu (4%).

In the University of Jyväskylä, 72 percent of users are from the Faculty of Humanities and Social Sciences. Humanities and social sciences also form the largest group of all the respondents, but that is a direct result of the composition of the other groups.

Over 90 percent of respondents told that they needed the material for a scientific article and over 40 percent needed the material for masters or doctoral thesis or for a scientific monograph (respondents were able to select more than one option).

So far, the total number of page views under the license for Haka users has varied between 9 500 and 29 000 from September 2018 to April 2019. Share of digitized newspaper pages of this has varied from 2000 to over 7000 page views per month.

When asked how long the respondents expect they need the material, over 40 percent selected the option 12 months or longer and about 33 percent selected 3-6 months or less than 3 months. Other feedback from participating organizations has revealed that pilot period is too short to show all the benefits that access to the material may give: many studies take over a year and a

half before the results are ready for publication. The participants have a common view that there should be data from a longer period before conclusions are drawn.

All digitized newspapers and journals have been available on the customer workstations in six legal deposit libraries long before the latest initiatives to enable wider access to the digitized collections online. The figures in legal deposit libraries are about at the same level as in Haka project. The statistics show that use of the material hasn't been affected by the openings for the public and researchers. However, the number of page views is quite low and varies considerably between months. Therefore making further conclusion based on it is difficult.

Use of the export packages is hard to measure. After the package is downloaded, the user may use the data as much - or as little - he or she likes. Answers to the question about intended use can give some clues.

Often users have told that they are just testing or playing around with the data, but the packages have also been used in hackathons, research projects or as a source of other datasets, for example. This indicates that while there has been just a few hundred downloads since the publishing of the packages, users have benefited much from them.

The newspaper titles to process in the NewsEye project were selected by the libraries in co-operation with digital humanities researchers working for the project. During the selection process, there were several aspects to consider. For example, it was found that some parts of the newspapers may be under copyright and it was necessary to define the planned use of the papers in more detail before the final selections were made.

Making customized packages inevitably requires human work. Still, in some cases customization is justified.

3 CONCLUSIONS AND PLANS

The new material in public online use has been very popular and there is a demand to open more recent newspapers and journals for the public. The material in open online use serves many user groups, such as genealogists and researchers.

Making more digitized resources available for different user groups is a natural goal for the NLF. However, a couple of reasons are slowing down the opening of the newer material. The lack of digitized material from decades after 1940s is one and the increasing share of material under copyright in newer publications is another. It is debatable how much it would benefit to open the years where only a few titles have been digitized when compensation for opening up newer material is also higher.

In any case, NLF hopes that it is able to open the digitized newspapers and journals published before 1940s to public online use in the near future. The negotiations on this are underway at the time of writing. It is also possible that some selected newspapers and journals will be released to open online use with a similar model used with the Käkisalmi foundation newspapers and journals.

The Haka project will end by the end of the year 2019. At the moment NLF is negotiating with copyright organization Kopiosto about a three-year follow-up project called Tutkain. In that

project, the aim is to grant access to copyrighted material for a wider group of researchers and find a permanent solution by which research use will be possible in the future.

In the Tutkain project, the plan is to give access to the material for all users in participating organizations if they accept the terms of use. While Haka provides the technical capability to define user groups with great precision, it may not be necessary or appropriate based on experiences of the Haka project. Defining and setting up user groups causes extra work and is an error-prone process. At the same time, the number of users and page views have remained at a moderate level also in the University of Jyväskylä where the access to the material is not limited based on the user information.

The export packages are less used, but they serve an important user group. Users have asked packages of the newer material, and the plan is to make them publicly available when copyrights and licenses make it possible.

Because the ready-made packages are not suitable for all, the NLF should have the ability to tailor the data packages for specific purposes, such as the development of tools in NewsEye project. However, the needs do vary to such extent that all request can't be answered. Also, copyright or license agreements may limit the data sharing.

The NLF has, and probably will have, multiple agreements on the use of the copyrighted digitalized material with copyright organizations. Since many of the agreements have been pilot contracts by nature, some inconsistencies have occurred. It should be in the interest of all parties that the terms of use are clear, so the aim is to harmonize the different agreements as much as possible in the future.

All in all the National Library of Finland has been able to promote wider the use of digitized newspapers and other materials during the last few years. Good experiences urge to continue despite some difficulties.

Acknowledgments
NewsEye has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 770299.

Aviisi project was funded by the EU commission through its European Regional Development Fund and the program Leverage from the EU 2014–2020.

References

[1] Yliopistolaki (558/2009). Available online: https://www.finlex.fi/fi/laki/ajantasa/2009/20090558
[2] Laki kulttuuriaineistojen tallettamisesta ja säilyttämisestä (28.12.2007/1433). Available online: https://www.finlex.fi/fi/laki/ajantasa/2007/20071433
[3] The National Library of Finland, National Library's Strategy for 2016–2020 : Digital services and openness for change – National Library services open to all. Available online: http://www.doria.fi/handle/10024/131074
[4] The National Library of Finland, Open National Library: Openness Policy for the National Library of Finland. Available online:

https://www.kiwi.fi/display/avoinkk/Avoin+Kansalliskirjasto%3A+politiikka+ja+toimenpideohjelma, referred 31.5.2019.
[5] The National Library of Finland, Digi. Available online: https://www.kansalliskirjasto.fi/en/services/digitising-and-preservation-services/digi, referred 31.5.2019.
[6] The National Library of Finland, Statistics - Digital Collections. Available online: https://digi.kansalliskirjasto.fi/stats, referred 31.5.2019.
[7] Kopiosto. Historialliset työväenlehdet digitaalisina käyttöön. Available online: https://www.epressi.com/tiedotteet/kulttuuri-ja-taide/historialliset-tyovaenlehdet-digitaalisina-kayttoon.html, referred 31.5.2019.
[8] Pääkkönen T, Kervinen J (2017). Kansalliskirjasto avasi digitoidut sanoma- ja aikakauslehdet yleisön saataville vuoteen 1920 asti – alkutunnelmia avauksesta. Tietolinja, 2017(1). Permanent address: http://urn.fi/URN:NBN:fi-fe201702151606
[9] Pääkkönen T (2018). Vuosien 1918-1929 lehtiaineistot käyttöön vuodeksi 2018. Tietolinja, 2018(1). Permanent address: http://urn.fi/URN:NBN:fi-fe201802133368
[10] Available online: https://www.kansalliskirjasto.fi/fi/uutiset/kansalliskirjaston-digitoimat-lehdet-vuoteen-1929-asti-ovat-avoimessa-verkkokaytossa-vuoden
[11] Pääkkönen, T., 2016. Increasing availability, data privacy and copyrights of digital content via a pilot project of the National Library of Finland. LIBER Quarterly, 26(3), pp.163–180. DOI: http://doi.org/10.18352/lq.10169
[12] Promoting the use of newspapers and journals in research and education. Available online: https://www.kansalliskirjasto.fi/en/projects/promoting-the-use-of-newspapers-and-journals-in-research-and-education
[13] Rautiainen J (2017). Digitoitujen sanoma- ja aikakauslehtien käyttö yliopistoissa ja korkeakouluissa helpottuu. Tietolinja, 2017(2). Permanent address: http://urn.fi/URN:NBN:fi-fe201709258723
[14] CSC - IT Center for Science Ltd. Luottamusverkosto. Available online: https://wiki.eduuni.fi/display/CSCHAKA/Luottamusverkosto, referred 31.5.2019.
[15] Hakkarainen J-P, Pääkkönen T, Rautiainen J (2018). Omalta koneelta käyttämään tekijänoikeuden alaisia aineistoja. Tietolinja, 2018(2). Permanent address: http://urn.fi/URN:NBN:fi-fe2018092236394
[16] Pääkkönen T (2016). Exporting Finnish Digitized Historical Newspaper Contents for Offline Use. D-Lib Magazine, July/August 2016, Volume 22, Number 7/8. Available online: http://www.dlib.org/dlib/july16/paakkonen/07paakkonen.html
[17] European Commission. NewsEye: A Digital Investigator for Historical Newspapers. Available online: https://cordis.europa.eu/project/rcn/216024/factsheet/en
[18] ibid.
[19] The National Library of Finland (2017). Aviisi loppuraportti.
[20] Pääkkönen T, Kervinen J (2017). Kansalliskirjasto avasi digitoidut sanoma- ja aikakauslehdet yleisön saataville vuoteen 1920 asti – alkutunnelmia avauksesta. Tietolinja, 2017(1). Permanent address: http://urn.fi/URN:NBN:fi-fe201702151606
[21] Pääkkönen T, Lilja J (2018). Hieno palvelu, mutta sisältöä lisää – Kansalliskirjasto kyseli Digi-palvelun käyttökokemuksia. Tietolinja, 2018(2). Permanent address: http://urn.fi/URN:NBN:fi-fe2018092336401