Submitted on: 07.08.2015

From historical newspaper to e-newspaper. Challenge for libraries

Krista Kiisa
Digital Archive Department, National Library of Estonia, Tallinn, Estonia.
E-mail address: Krista.Kiisa@nlib.ee

Abstract:

More and more often we face the reality that news are moving from the static paper media to the web environment. For libraries it is quite a challenge to cope with the situation. In order to deal with the variability of data, different kind of media is used. Very often the key is in the operativity to build the relations between archival systems and digital objects, and to keep track of the decisions made for data acquisition, long term preservation and access. Usually it is not easy to use automated processes on a vast scale for that – what is needed is quite a considerable amount of human activity. The decisions we have to take are very different in origin.

In the National Library of Estonia there are 3 kind of different workflows used for newspapers. Preservation digitisation is used for old paper newspapers, fragile in origin. Here digitisation minimises the risks on media and formats, makes it possible to limit the access to the original.

Voluntary deposit of newspaper preprint files is used as a means of preserving the outlook and the content of the printed newspaper in electronic format. This one allows remarkable saving for libraries, avoiding the need for further digitisation, and makes it possible to build new library services on the article-level content.

Web archiving is a challenge that we cannot overlook any longer. What is the sufficient frequency and depth when archiving the newspaper publications from the web?
The paper tries to give a short overview of the current situation in Estonia in terms of processing the newspapers in different media.

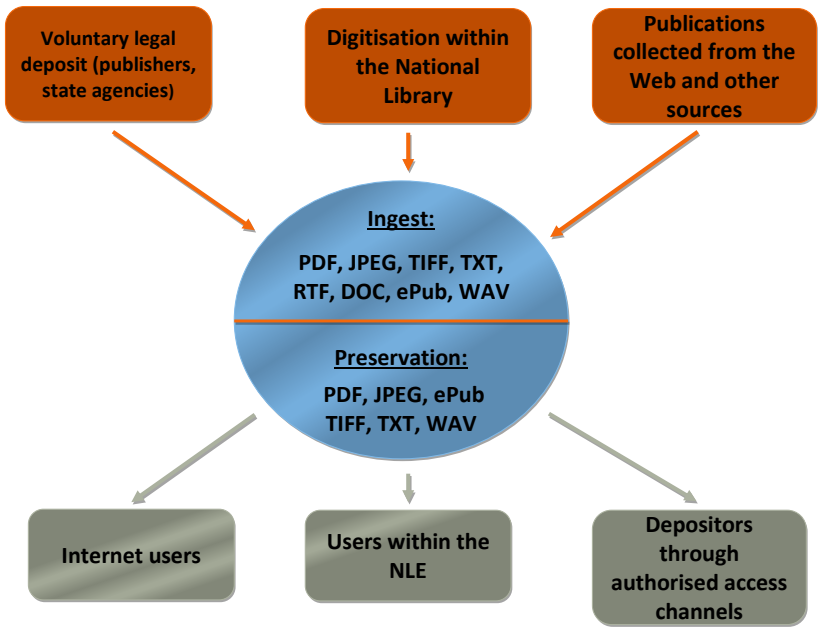
Keywords: newspapers, digitisation, web archiving, digital collections, publishing.

Introduction

A problem that libraries are facing today is well known for all of us: surplus of formats, mediums and even information itself. Paper has been used as the storage medium for text and images for centuries. Due to its inherent chemical instability, it is not getting better in years. It's especially relevant to the historical ones, but it is appropriate also for the currently printed publications. The situation is even more complicated with the digitally born issues and newspaper websites. For the citizens of Estonia, using Internet is as natural as any other of human rights. You can't find a place in Estonia where you don't have Internet connection. This means that essential part of news and information comes to us via news portals, RSS feed, blogs etc. We try to keep up hard on the heels of the publisher new trends, but it's not easy. To compile a complete and full collection of newspapers we need to apply different workflows and work simultaneously with newspapers in paper format and the ones born and published in digital form. Plus harvest the sites published only in e-format. My presentation will describe different ways we handle the newspapers in Estonia and describe the current solutions we have for access, trying to open some last year trends of publishers behaviour publishing and selling the news.

Practise of publishing the news is changing rapidly nowadays. Radical changes in media and news production take place so rapidly, sometimes so unexpectedly that it's very difficult for the libraries to fluently change their running workflows accordingly. In the library the job related to the newspapers has the most variable nature at the moment. More and more often we find ourselves asking the question: what is the actual situation we have to cope with when talking about handling the news and newspapers in the library?

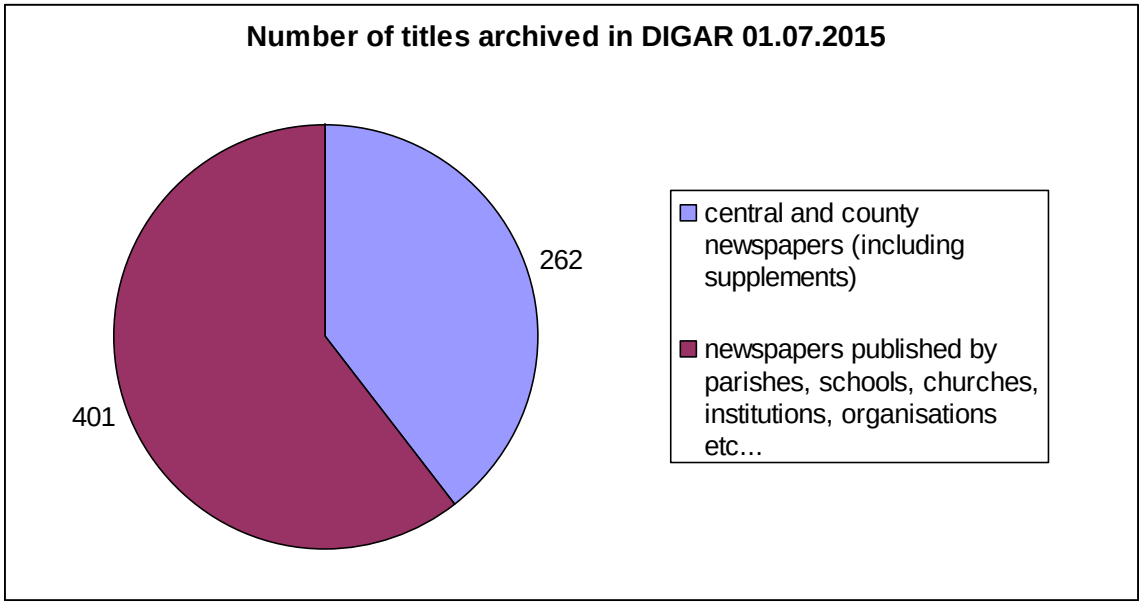
Collecting and archiving the newspapers



Picture 1. Newspapers workflow

There are different channels and ways how the content acquisition and ingest of newspapers content takes place. The National Library's first priority in working with newspapers is to process and archive

the electronic versions of currently published materials. On daily basis NLE acquires and archives the pre-print versions of currently published newspapers in print. That's quite a considerable amount of data, mostly in pdf format, that are sent to the library's server every morning. We still don't have the Legal Deposit Law supporting us for that activity, so in negotiations with publishers we still have to stress the good will and a solemn purpose doing that, future preservation of cultural heritage.
Our workflow is organised in a way that the publisher of the newspaper is the one who sends the pre-print files to our FTP-server. As we don't have a law forcing them to do that, they do that voluntarily. Means the first call is usually still done by the library. We have a long history of negotiations with publishers behind, to explain the value of digital archiving.
In the moment we got electonic print files from 65 publishers. Alltogether it makes 663 titles of newspapers in digital form. 262 of them are either daily or weekly published central and local county newspapers. The number is quite big as supplements are counted in the database as separate items. The amount of so-called small scale newspapers published by parishes, schools, churches, institutions, organisations, different societies and companies is 401 at the moment. This is big in number, but actually that is quite few in data volume as in most cases these titles are published rarely and the issues are small in pages.



Picture 2. Statistics for newspapers archived in the digital archive of National Library of Estonia

The supply of the electronic content is not always as regular and technically proper as we expect, but basically we can say that we have almost all central and regional newspapers in our ftp-server early in the morning before the working day starts. Besides the purpose of archiving the issue, this is also an opportunity for the library to save some money. No need to purchase several additional paper copies for library users visiting the physical library or for employees cataloguing and processing the articles. All that can be done using the electronic print files, ready to use in our servers latest at 9.00 am
Today we are in the situation where also the Estonian Newspaper Association (EALL) supports us. Part of the agreement is that publishers can use NLE's server as an intermediate station, from where media monitoring companies download the pre-print files from the library's ftp server for their commercial use. The most difficult aspect here is access.
Access to the publications, archived in NLE's digital archive is organised following the access restrictions and embargo time which are defined by the publisher. From one point you cannot expect the readers to be interested in your collection unless it is comprehensive and current. The other way round - more concurrent, fascinating and innovative user interface you have for access, more strict

access restrictions are placed by the publishers to their content. About the last year tendencies in this particular field will be opened under the access chapter.

Digitisation for preservation

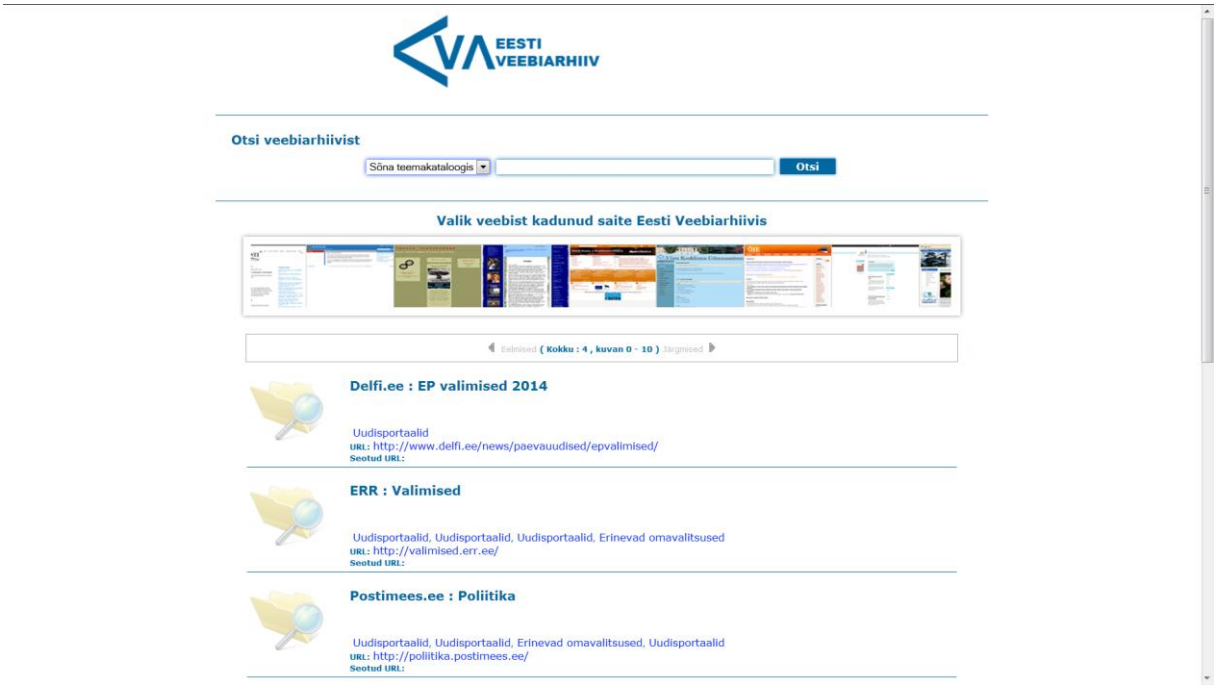
This concerns mostly the historical newspapers which are fragile in origin. But we do digitise also the missing pages and single issues of current daily newspapers the publishers have forgotten to deposit us as pre-print files.
Historical newspaper collection is the material we'll digitise from our own collection. In the moment we can say that we have managed to digitise most of the titles in Estonian language, published before 1944. Currently there are more than 400 titles, 1.5 million pages digitised from originals. Big portion of these are unfortunately still accessable as images only, because processes like OCR, manual text correction and page layout recognition of historical newspapers will apparently be continued for very many years from now on. Crowdsourcing for correcting the automated OCR text is already available for public, but the amount of pages, corrected this way is not a remarkable one yet. It is important to put the material online even if it's only on image or metadata level. Because more and more often we see that the user comes to the physical library to read the newspapers only when it's the only option and he won't have the information otherwise, like using libraries e-services or digitisation services etc.

Archiving the Web

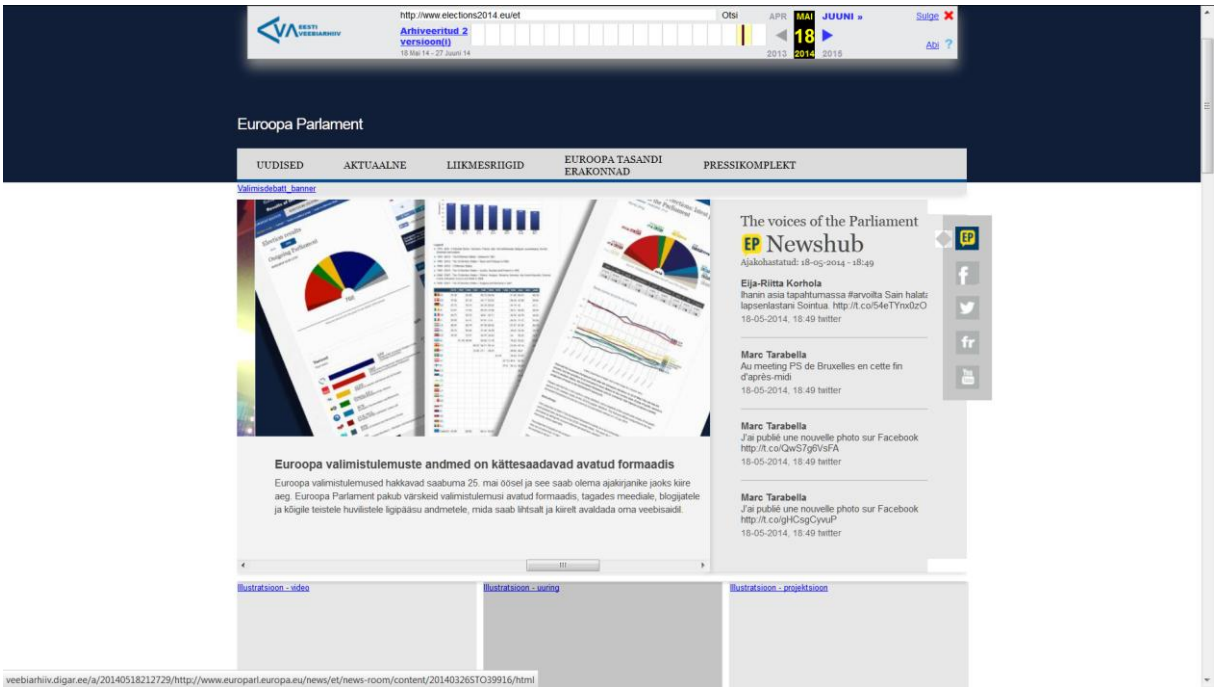
As the content published on web and the content on paper are usually 2 totally different things, there is a need to harvest dynamic data also from newspapers websites. Web harvesting is another process being actively developed during the very last years. Here we do have a Legal Deposit Law, supporting us to do that, but we don't have enough resources and skills to extensively harvest dynamic data from newspapers' websites as the sites, multi-platform in first place, are getting more and more complicated in nature every other day. New formats overtake the position faster than we can even identify their existence. This leaves no place for automation when quality is the priority.

Up to now we haven't archived online media regularly, but used mainly selective approach for that. Estonian Web Archive follows the selection principles worked out jointly with other Estonian memory institution specialists within the Working Group of Web Archiving Experts. The general criteria for selecting and archiving online materials are: their publication, identifiability, exhaustiveness, long-term and permanent value and place of publication. Beside that everybody can send hints and recommendations for archiving a valuable website via special form on the archive's webpage. Following the selective archiving approach our Web Archive is quite small in number at the moment. Basic growth is gained harvesting the certain subject based information. For example biggest news portals have been archived before and after some important event and a special theme collection is compiled from that data. We have archived information published in newspapers and news portals about the parliament elections, Olympic Games, national cultural events like song festival etc.

In 2014 during the 2 weeks time of the Olympic Games we daily archived all special editions issued online by newspapers and news portals. It was quite an enlightening activity for us speaking of the capacity and extent of the news. Quite soon we realised that for the purpose of disk space saving, every other time the already downloaded and archived photo or video was appeared on the next sub-feature, the de-duplication function has to be used to eliminate the photo. Actively used are the thematic news collection archived during the local, Parliament and European Union elections. For example more than 200 web sites were archived in connection with the local government council elections in cooperation with the central libraries of the counties.



Picture 3. Thematic collection. Elections of European Parliament, archived in Web Archive.



Picture 4. Archived website

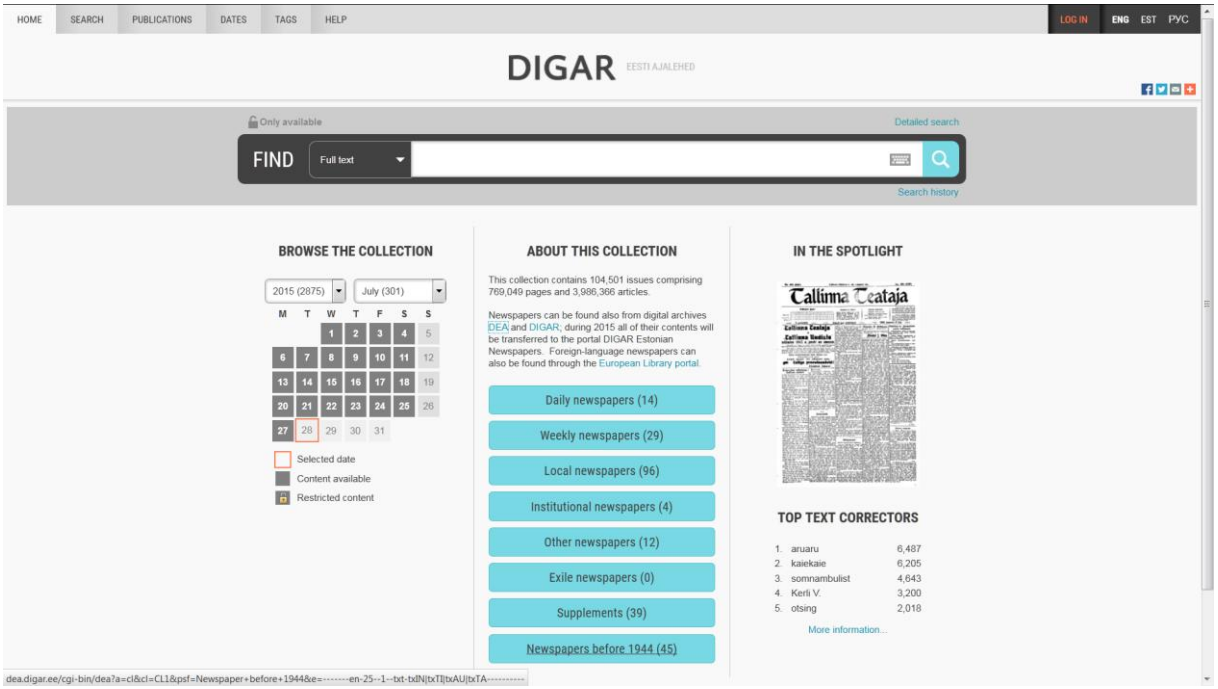
Future activity - newspapers online

Since it is impossible to foresee what will be important for future researchers and users, it is necessary to harvest a snapshot of entire Estonian national Web regularly. The first harvest of web-based periodicals, based on their frequency of publishing, starts in November 2015.

At first 4 most important newsportals (postimees.ee, err.ee, delfi.ee, ohtuleht.ee) will be harvested on the editions first page level once a day. Deeper harvesting (articles and news items with related comments) will be following according to their peculiarity with the frequency at least once a month. There is a list of 50 online newspapers, which do have the analog in print, but as their online version has considerably longer articles and much more photos and videos in it, harvesting and archiving of such publications will also take place according to their importance and frequency of publishing from once a day up to once a month.
Next to newspapers there are also 65 online journals in the list of harvesting. 17 of them are published online only. These will be harvested at least once a month.

Access

In the National Library of Estonia the main purpose is to assemble all newspapers and news items to the one and only place. For that reason a portal called DIGAR (DIGital ARchive) Estonian Newspapers (dea.digar.ee) was launched in October 2014. The aim of DIGAR's collection of Estonian newspapers is to provide a single point of access, unite web portal for all digitally created pre-print files, historical digitised newspapers and to those harvested from the web, presumed that they are published abroad in Estonian or here in Estonia.

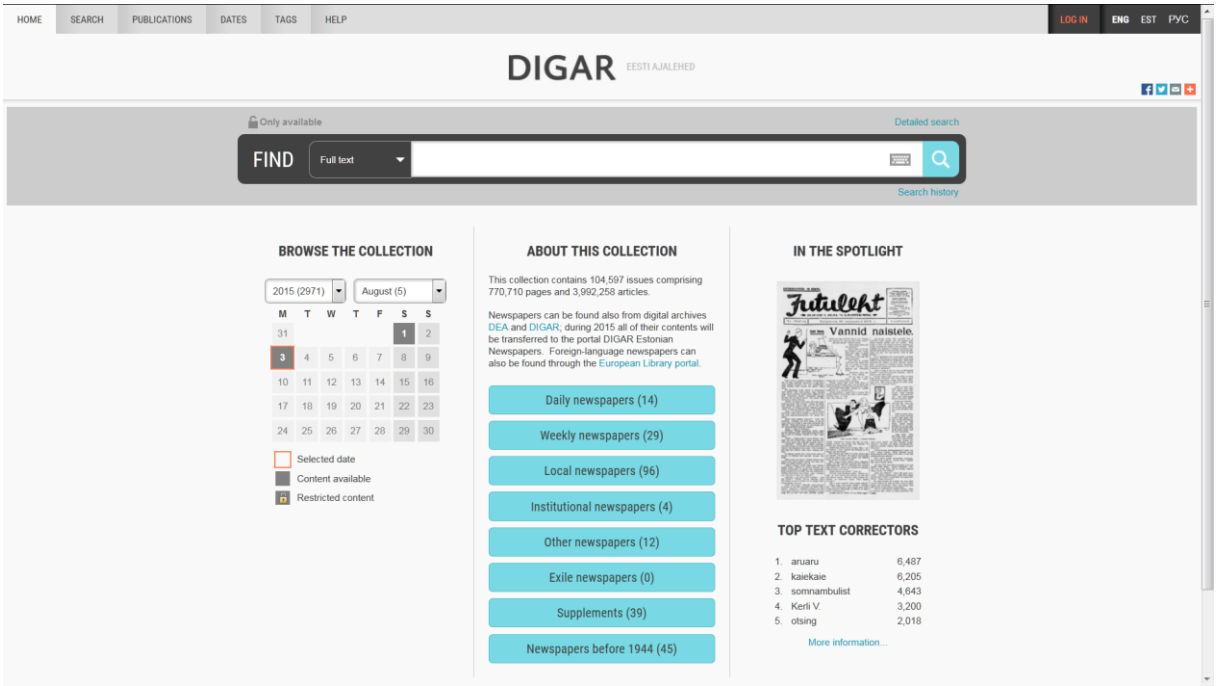


Picture 5. Newspapers web portal. DIGAR Estonian Newspapers. http://dea.digar.ee

For processing the content (searchable fulltext on the article level), we use the CCS | Content Conversion Specialists software docWORKS. The new user interface is developed in collaboration with the New Zealand company DL Consulting Ltd. This portal went online in October 2014. Unfortunately there are still three different databases online that users must search if they want to find and read newspapers from different periods. For older historical newspapers there is an 12 years old image database (dea.nlib.ee) with 1.4 million pages from 1821-1944. This will ocr-ed and transferred to the new newspaper portal during the years of 2015-2016. National Library's digital archive DIGAR has the collection of periodicals, containing digitally created newspapers and digital copies of older papers that are published in Estonia. That was used for giving access during the period we didn't have the new system. The aim is to transfer the newspapers from core digital archive interface to the newspapers portal latest in 2016.

Picture 6. User interface of the digital archive of NLE, available at www.digar.ee

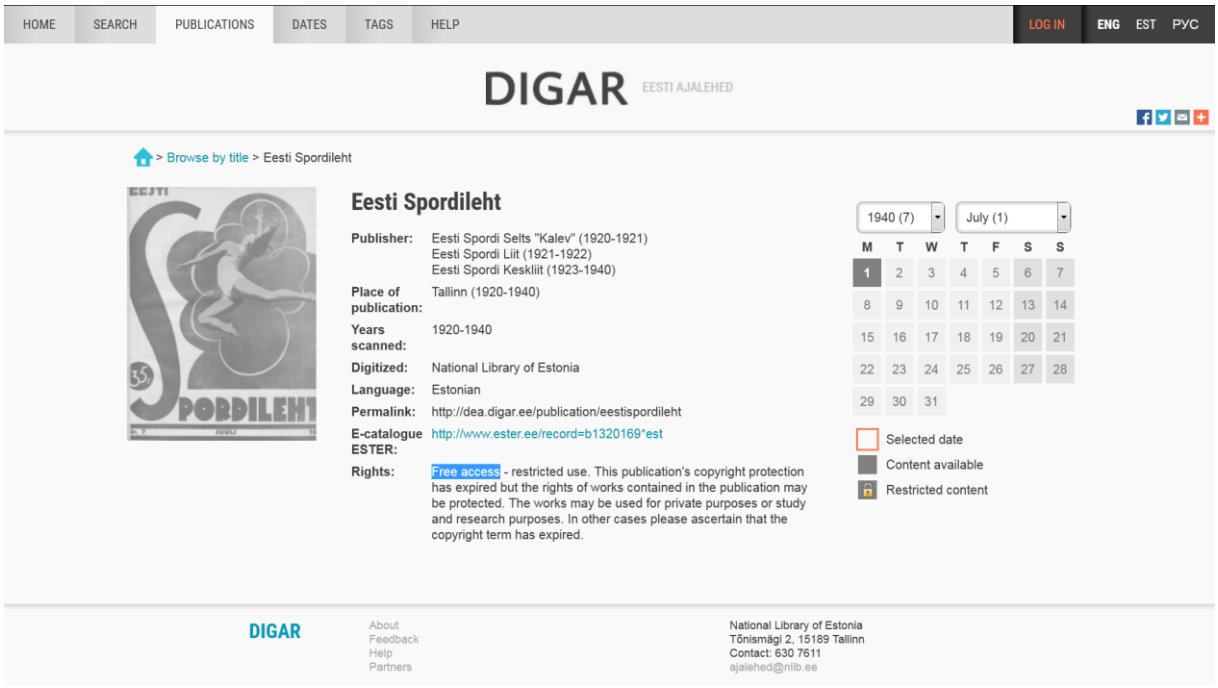
The basic effort is pointed to the new newspapers portal accessible at dea.digar.ee, where at the moment the users are provided with access to all newspapers published in 2014-2015 and a selection of older newspapers, we have managed to transfer from old databases.



Picture 7. Newspaper portal dea.digar.ee

The new database is supplied with current newspapers, at latest in the evening of the item's publishing date. The older newspapers (published 1821–2013) will be added here one by one from previous databases in compliance with the conversion plan.

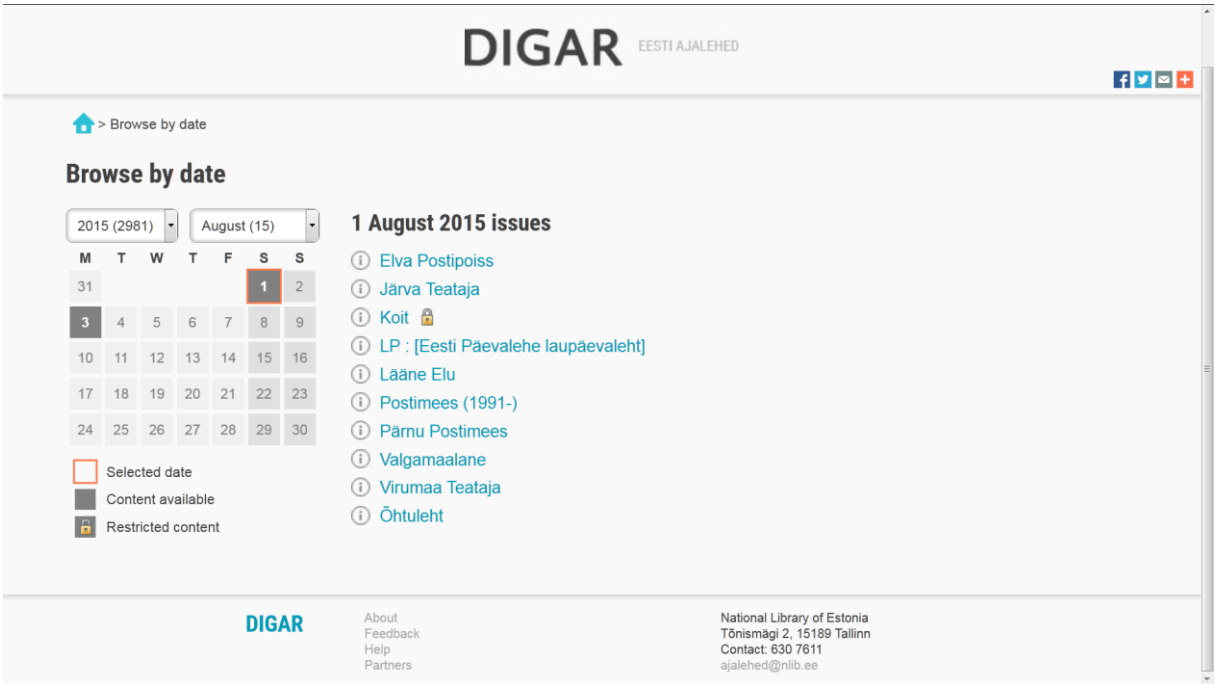
The system is developed keeping in mind the interest of the publishers. The newest numbers are usually not accessible for remote users at the day of the publishing. But the full texts are always accessible in the premises of National Library of Estonia and in the University of Tartu Library. Regardless the restrictions applied by the publishers to the articles fulltext, you can always make enquiries and view bibliographic data of all articles (authors, headlines, information about publications). The information about access restrictions, given according to the publishers prescriptions are separately shown for every item.



Pictrue 8. Newspaper with free access in digital archive

Situation in news market and how does it influence the given Access Restrictions

Daily newspapers have a clear trend to move from paper to the online website in Estonia. Content published on the publication website is much more thorough and covers the issue more in detail than the same article in print edition. Also the articles in the online edition are complemented several times during the day. It's quite a common behaviour especially for the hot topics of the day.
In the moment National Library of Estonia has succeeded to obtain most of the daily and weekly newspapers in the electronic pre-print version. It means we can electronically preserve it in the exact way it was printed. But we haven't started archiving the web sites online.
There is a very good Internet access all over Estonia. 3G coverage is excellent. You can freely read newspaper in the forest while picking berries or in a small village when visiting your grandmother.
While editions' online availability is growing and the user-statistics of online visits are rising, free of charge access of newspapers online is little by little cut down by the publishers. It's quite common that you can read free of charge only the first column of the article and if you are interested to go on, you must buy either a day ticket or become a permanent user of the online edition. That's a trend all big central newspapers have taken over one by one during the very last year.
Also the new and innovative newspaper portal of the National Library has given its contribution to that. The user doesn't have to move from one newspaper website to another, but can easily now come to the library portal, browse the date she wants, and gets all newspapers, published on that day from one access point quickly and easily.



Picture 9. All newspapers in the archive published on same date accessible from one page.

Though the system is developed keeping in mind the interest of the publishers, (it is possible for them to apply embargo for access), publishers anyway have started to restrict the Internet access to their publications in the library's digital archive.

County, parish and small scale publications freely give access to their publications via library's digital archive. This attitude hasn't changed. But there is an obvious trend that central daily and weekly publications who so far had applied only some weeks or some months embargo time for free access, now have changed the agreement and allow free access to their archived electronic pre-print files only and only in the premises of the physical library.

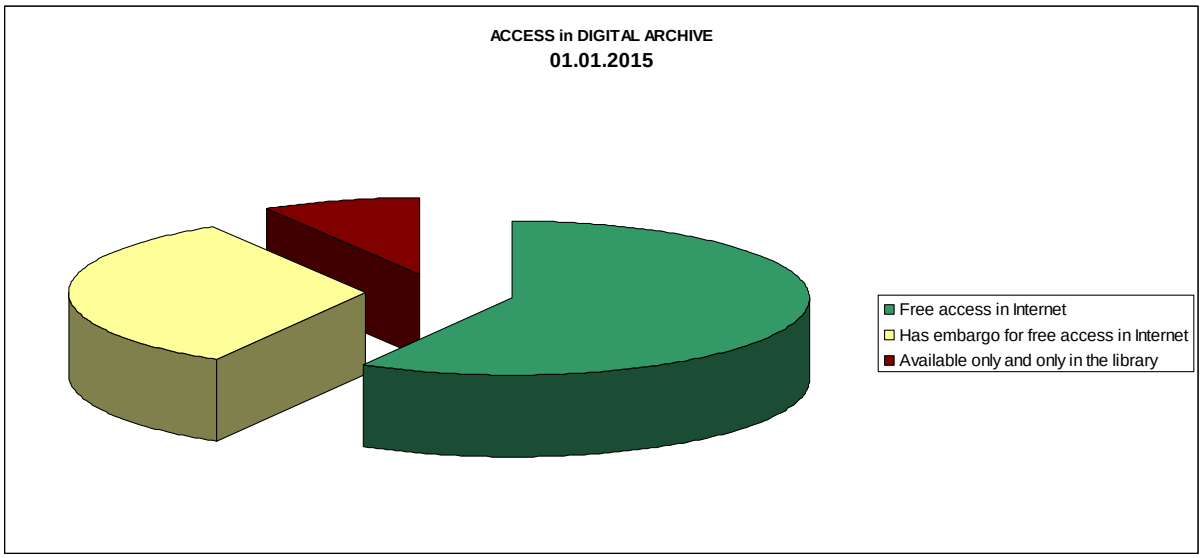


Figure 10. In January 2015 embargo used more than restricted access by publishers

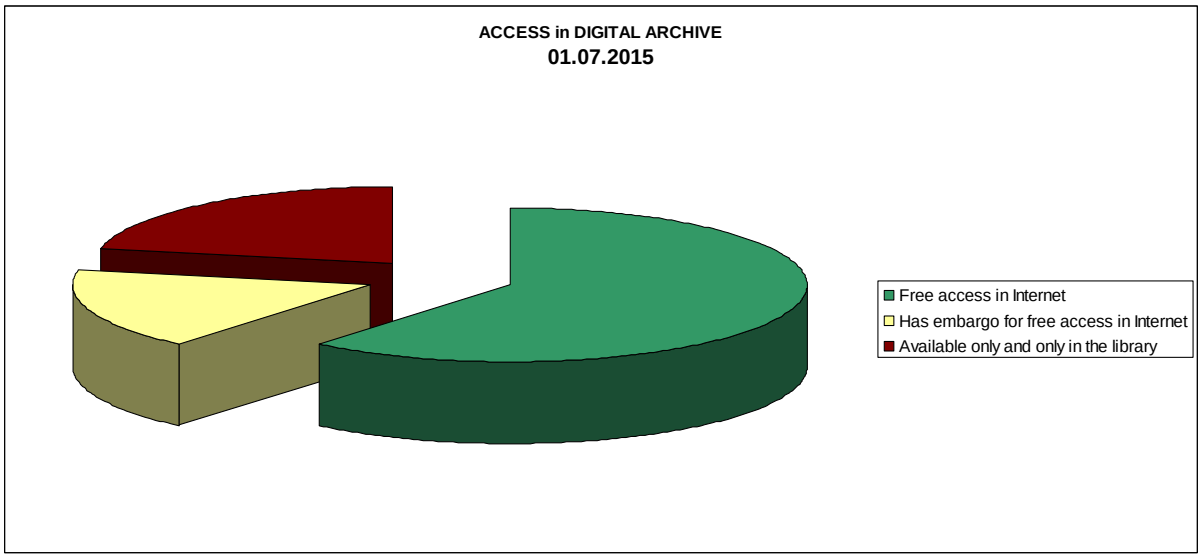


Figure 11. In July 2015 restricted access used more than embargo by publishers

According to the law the electronic fulltexts are always accessible in the National Library in North Estonia and in the University of Tartu Library in South Estonia. Regardless the restrictions to Internet access, you can always make enquiries and view bibliographic data of all articles (authors, headlines, summaries, information about publications).

Summary

All in all it is quite an ambitious task for the librarian. To leave behind the wellknown conservative, old-fashioned attitude of collecting and archiving and enter to the world of open market and start negotiating the publishers about offering the services as equal partners. Our advantage is that we do possess an enormous amount of data. It's up to us now, how good we are in negotiating and what kind of compromises we have to settle on, to offer the users library services, built on that valuable amount of data.
Presumption is that we have to fit into the trends and understand that big changes are unavoidable, they are happening with us, do we want it or not.