News Media Section

Getting to Know Users of Digital Newspaper and Journal Library – What Can Statistics of Use Tell Us

Juha Rautiainen
The Centre for Preservation and Digitisation, The National Library of Finland, Mikkeli, Finland.

Abstract:

Organizations need information about the users and use of digital online collections for many reasons. Within the organization the information is increasing understanding of users' needs, thus helping allocation of resources. The external stakeholders, such as funders and copyright holders, are interested in the use of material also.

The Historical Newspaper and Journal library of The National Library of Finland currently has over 10 million pages of digitalized newspapers and journals. 33 percent of the collection is freely available online. An ongoing project, which examines the possibilities of extending the online availability of digitized material subject to copyright, is developing monitoring of the use of the library at the same time.

Based on the statistics of use, it is possible to determine the most used materials, for example, by title, and by era. The statics also show the variations of the popularity between years or months. When the statistics are connected to other information, one could try to make more far-reaching conclusions about why some material is more popular than some other.

Keywords: digital library, statistic of use, analyses, online.

1 INTRODUCTION

The Historical Newspaper and Journal library of The National Library of Finland currently has over 10 million pages of digitalized newspapers and journals. Papers published in Finland between 1771 and 1910 are freely available online. That is about 33 percent of the collection. Newspapers and journals published after 1910 are in limited use available only on selected locations. [1]

The internet service of the Historical Newspaper and Journal library (digi.kansalliskirjasto.fi) is developed in-house. It utilizes commonly used web and database server software and the system logic is based on Java technologies.

An ongoing project Aviisi examines the possibilities of extending the online availability of digitized material. For this purpose, we have opened the volumes of two newspapers Länsi-Savo and Maaseudun Tulevaisuus from the first published papers (1916 and 1917) to the end of 2013 for specified user groups. The pilot groups include Mikkeli area schools, museums and the editorial staff of the two newspapers. [2]

Aviisi project is funded by European Union European Regional Development Fund - Leverage from EU 2014 – 2020. The National Library of Finland co-operates with Kopiosto, Kaakon Viestintä, Viestilehdet and City of Mikkeli.

Monitoring and analyzing of the use of the collection is an important part of the project for a couple of reasons. First, a big part of the pilot material is subject to copyright. Naturally copyright holders and publishers would like to get some compensation for the use in the future and require prevention of unauthorized use. Both of these needs will benefit from improved statistics and monitoring. Analyses of the user statistics will provide data to be used when proposed pricing models are evaluated and understanding of normal use will help to detect unauthorized use.

Second, the National Library would benefit from better statistics. The statistics would enable better resource allocation, for example. Also, if it’s possible to find the reason why some material is more popular than some other, the information can be utilized when actions are made to promote the use of digitized newspapers and magazines.

The data about use of the collection will also supplement the understanding of user behavior based on other methods. For example, when questionnaires reveal what user say they do, statistics of use will show what they actually do. Another way around the analysis could potentially bring forth questions for questionnaires.

Third, from the project point of view, the statistics of use are important because we want to monitor the use of the pilot user groups. We need to see how much and what kind of material the groups are using and learn to understand their needs better. We should also be aware if some pilot group is not as active as project goals would require.

There are many ways to collect and analyze the information about the use and users of a web service. Some guides (e.g. [3]) prefer the use of publicly available ready-made analyzing tools, such as AWStats and Google Analytics. These tools either utilize the log files of the web server or collect the data by themself. Some of those tools are already used in the National Library. They do provide basic information about the users and the use of the site, such as how many users visited the site during a month or what is the geographical distribution of the users.

These analyzing tools have some benefits. They are, for example, relatively easy to use. Unfortunately, it seems that in the context of The Historical Newspaper and Journal library they don’t provide enough information about how the collection itself is used. We have also noticed that there is an inconsistency between a number of web page loads on the site and the

number of loaded pages of the papers. Therefore, we need to analyze the user statistics collected by the system itself.

2 THE DATA

The internet service of Historical Newspaper and Journal library collects and stores data about the loaded pages and volumes of the papers. The raw data of the page loads includes page indicator, timestamp and the IP address where the load request was issued. Based on that information it is possible to find out, for example, the title, publication day and the language of the title.

In most cases only the IP address of the user is known, which limits the possibility to identify individual users. It is quite common that many people share the same public IP address. For example, all 4000 users in Mikkeli area schools use two public IP addresses. Also, an IP address assigned to a user might change over the time, so on a long time period, same IP address could be used by many different users. For this reason, during the analyses of the use, we must keep in mind that usually we are not able to say, whether certain page loads are done by the single user or a large group of users.

Even though in some cases more specific identification could be useful, The National Library has no interest to identify individual users, unless it is necessary. Tracking an individual user or an IP is done only when some problem or misuse is investigated.

The papers published before the end of 1910 are freely available online and papers published after 1910 are only in limited use. Therefore, the number of page loads of the recent volumes are remarkably lower. Especially page loads of the papers published after 1946 are so low that one must be very careful if any conclusions are made based on that data.

The raw data recorded by the system needed some processing before it was usable for reliable analyses. During the first attempts to analyze the data it came apparent that significant part of the page loads were made by Google bot. Its share of the total number of page loads varied between 18 and 70 percent per month.

Usually, we don’t consider bots to be actual users, so before further analyzes traffic from Google bot was removed from the data. Page loads from the internal network and IP addresses used by MSN bot were filtered out too. The data also included page loads from some other bots (e.g. Yandex and Baidu), but their share was so low that those could be ignored.

Most of the analyses have been done to data collected during the year 2015 and in the beginning of 2016. The system has stored usage data from the year 2009, but the use of data older than a couple of years would involve some error sources. First, the system was changed during 2014 and the new version of the system does some operations in a different way. That has had some effect on the recorded use.

The second error source is related to bots. The bot filtering was made afterwards based on the IP address and it is possible that the IP addresses change over time. It seems that there is no reliable way to identify the IPs used by bots afterwards and, therefore, the filtering might go wrong.

Thirdly, the number of digitized pages is increasing over time. On average about 1 million new pages have been added to the collection each year. Because the number of page loads is in some relation to number of pages available, that should be taken account if older data is used in the analysis.

3 ANALYSES OF THE DATA

Two approaches are used in the analyses. The first approach is to form a question or hypothesis first. Then the data is analyzed to get an answer to the question or to see if the data supports the hypothesis. The second approach is to study the data itself trying to find anomalies and their explanations. Following four examples present these approaches in practice.

Example 1 There have been efforts to increase the use of The Historical Newspaper and Journal library by sharing links to material on social media. Users can either share a link to a certain page of a paper or make a clip of interesting content and share link to the clip. The general idea is that interesting samples would lure new users to the library.

A couple of cases were analyzed to see if the social media shares had any effect on the use.

On January 3rd, 2015 at 22:55 local time a link to a clip of journal Pääskynen was shared on Facebook. The text on the clip was an old Finnish fairytale. The update got 32 likes, 1 comment, and 20 shares.

The clip was opened on the library system 41 times during January 2015. The page where the clip was taken from was loaded twice on January 4th and after that, there were any loads on that month. To compare a total number of page loads on Sunday, January 4th was 8802. On average there were 9564 page loads on Sundays in January 2015.

On February 7th, 2016 one page of magazine Huliwili was shared on Twitter. The tweet was retweeted 7 times and liked 5 times. The linked page of the magazine was loaded 6 times on that day and 4 times more by the end of February. Four out of these ten users browse through the whole issue.

At least in these cases, the result is clear. Social media updates got some attention, but they did not increase the use of the Digital Library. Interesting detail is that link to the page led users to read the other pages of the issue whereas most users who opened the clip did not read the original paper at all. Whether this applies only to these cases or is more common phenomenon remains to be studied.

Example 2 The questionnaires (e.g. [4]) and informal feedback indicate that the users of the Digital library would prefer to see a quite recent material, and it would be reasonable to assume that the statistics of use will support this too. At first glance this seems to be the case: When page loads are allocated by the year of publication, the number of page loads per volume is increasing from 1771 to 1910 as seen in Figure 1.

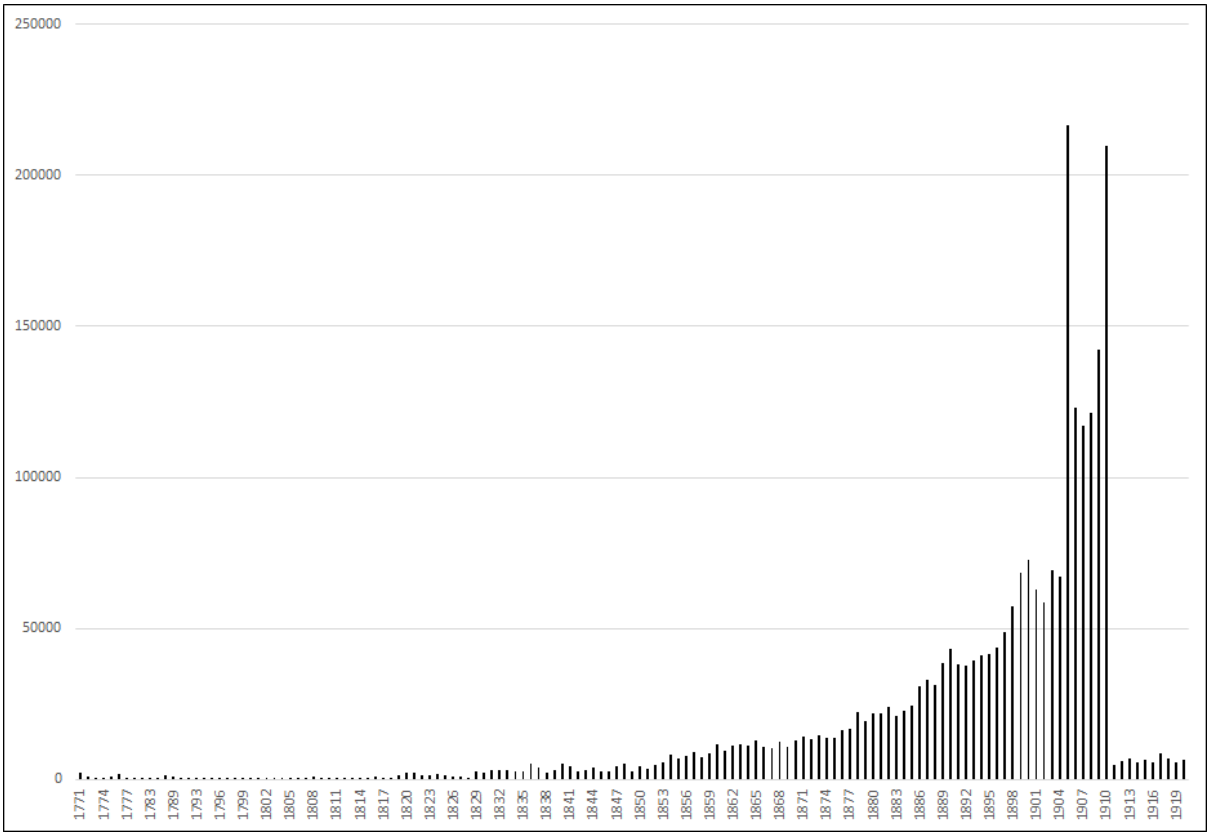

Figure 1: Number of page loads allocated by the year of publication on year 2015.

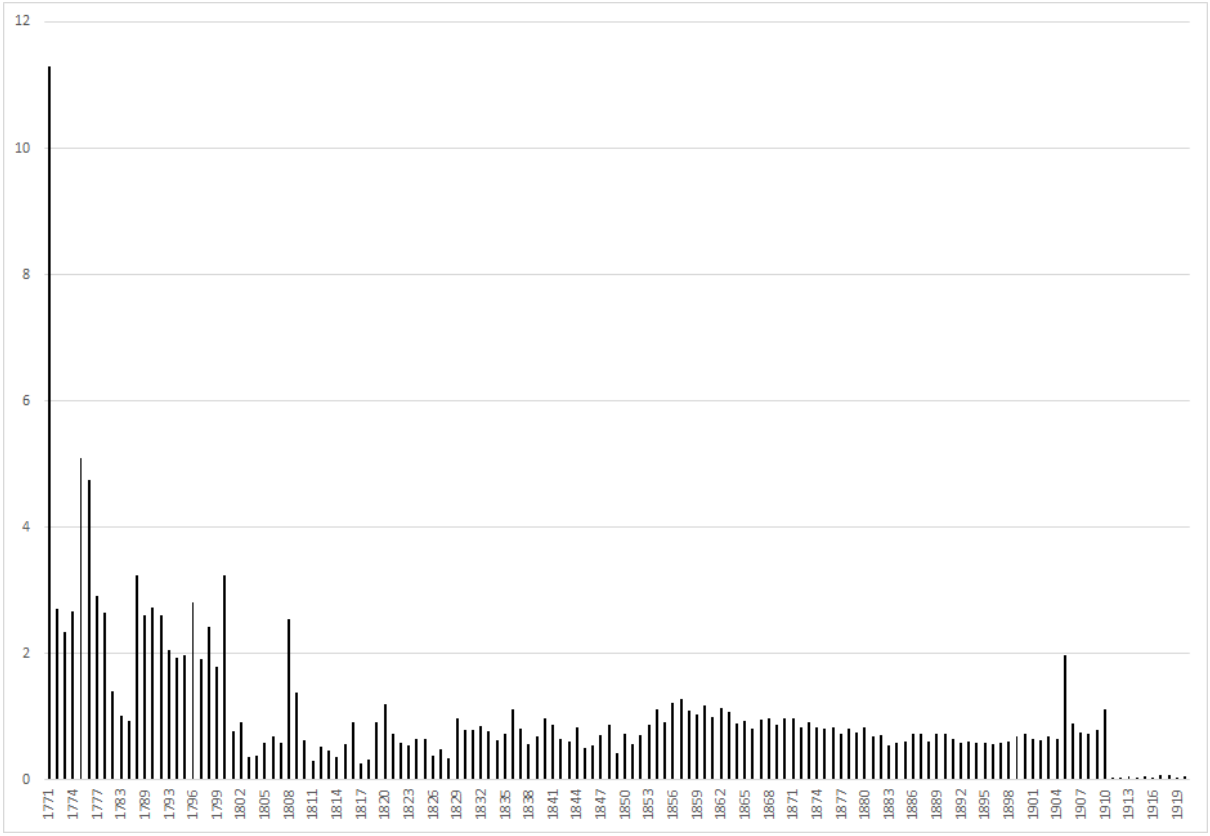

Figure 2: The ratio between page loads and pages available.

However, we know that at the same time period number of pages of published papers is also increasing. When this is taken into account, we see (Figure 2) that while every page from the year 1771 has been loaded 11,3 times, pages of the papers published in the year 1910 have been loaded only 1,1 times on an average.

So, it seems that the papers from 1771 to 1800 are actually relatively popular. The ratio between page loads and pages available is dropping after the year 1800. Starting from 1850 the ratio rises a little and stays quite stable until the year 1910. The collection is in limited use starting from the year 1911, which explains the drop from thereafter.

The number of pages available has effects on number of page loads, but this correlation is not straightforward. Further analyses are needed to find out what are the explaining reasons behind the variation. First guesses are the language of the publication (most of the users are Finnish speakers and the oldest papers are in Swedish) and the typeface used.

Example 3 When page loads are allocated by the year of publication we can see anomalies on the data by looking at the ratio between page loads and pages available. For example, we see in the Figure 2 that the ratio in the year 1905 is remarkably higher than on years before and after.

A plausible explanation for the popularity of these papers is the Newspapers of the day column on the front page of the Newspapers section because it presents newspapers published on the date 110 years ago. The data supports this theory: each day there is a remarkable number of page loads on the newspapers published on that same day year 1905. If same happens to papers published in 1906 in the year 2016, the hypothesis will be confirmed.

Example 4 Daily monitoring is one special case of data based analysis. We have created some alarms based on daily analysis and it has proven to be very valuable tool supplementing other protection mechanisms. For example, daily monitoring has revealed a misconfiguration on access rights and a test server that Google was showing on search results. In both of these cases, the firewall wasn’t able to detect any anomalous traffic.

By continuous daily monitoring, we also noticed that some of the pilot groups did not use the service at all and were able to react to the situation. For example, in the beginning of the pilot period, we send emails about the project to school administration and the headmasters, but only a few page loads were registered from the school network. As a reaction, we sent posters and some other material to the schools. The use rose almost immediately, so we might assume that most of the potential users did not know about our project.

We have got interesting data about the use of some workstations located on the premises of the National Library. In these cases, we can identify individual workstations but not individual users. For example, some days over 1000 pages have been loaded on a single workstation. When divided by opening hours of the library, it means that the user (or users) spend about 30 seconds per page on average.

4 GENERAL DISCUSSION

Data from statistics of use is useful on itself for certain purposes. It is quite an obvious source, if we need to know what material was used and how often it was used, for example. By analyzing the user stats we can also verify information from other sources, such as questionnaires, and check does our hypothesis hold.

The analysis could raise some new questions too. For example, as mentioned before, statistics of use reveal that occasionally user (or users) might browse through over 1000 pages on a day using 30 seconds per page on average. It would be interesting to know, why the users do that. We know from feedback that some of them read papers by volumes, but time spent on a page rather indicates that in these cases they are searching for something. If that is the case, then why they are not using the text search?

Also, based on the analysis we can see that simple the Newspapers of the day -column on the site generates more page loads than social media updates. We may assume that the social media updates do reach a different audience because to see the Newspapers of the day one must be on the site already. If this is the case, social media updates increase the conspicuousness of the collection. However, without data supporting this assumption it is just an educated guess. To this question, we might actually find an answer by analyzing the page load data a little bit further.

The quality of the data can be improved with some quite simple actions. For example, we have created a daily IP resolving process, which can reliably identify bots. This will make analysis both faster and more accurate in the future. For some purposes it could be possible to generate a normalized data set from a longer time period.

Overall, even simple analyses can reveal much about the users and use of the system. With data from the longer time period and more advanced analyzing techniques it probably is possible to get even more information out of the data. For example, it could be possible to find out what kind of material certain user group is interested in at a certain time of the year. This information can be utilized when social media campaigns and curated sections of the site are planned to promote the use of digitized newspapers and magazines.

Analyses of the statistics of use measures just the use of the current collection and actions of the current users, which is a remarkable constraint. We can make some predictions based on the data, but we can’t guarantee that some new material will be popular among the current users or bring new users to the service. That kind of prediction would be better when supported by information from other sources, such as questionnaires. Other information sources are needed also when we want to know how the users utilize the papers.

5 SUMMARY

The monitoring of the use and analyses of statistics of use of the Historical Newspaper and Journal library are developed in an ongoing project. Previously some ready-made analyzing tools have been used, but they provide more information about the use of the site than about the use of the actual collection.

Two basic approaches are used in analyses. On data-based approach aim is to find anomalies and their explanations. Another approach is to start from a question or hypothesis and see what the data can reveal on that matter.

The first experiments have already shown some results and given assurance that it is useful to analyze the statistics of use collected by the system. The analyses of the user data can provide information that is not available by other methods, such as anomalies in a number of pages loads. Although the user data is limited to current use, it could verify or prove wrong the assumptions made based on the other information sources and provide information for planning processes. In turn, sometimes the analysis of the data might disclose questions that can be best answered with questionnaires or some other data collecting methods.

References

[1] The National Library of Finland (2016). *Information about the service: The digital collections of the National Library*. Referenced 1.4.2016 http://digi.kansalliskirjasto.fi/info

[2] Pääkkönen, Tuula (2015). *Mikä Aviisi? Miksi? Miten?* Referenced 1.4.2016. http://blogs.helsinki.fi/digiaviisi/mika-aviisi-miksi-miten/

[3] Marek, Kate (2011). Using Web Analytics in the Library: A Library Technology Report. American Library Association.

[4] Hölttä, Tiina (2016). *Digitoitujen kulttuuriperintöaineistojen tutkimuskäyttö ja tutkijat*. University of Tampere, School of Information Sciences.