Review of the NCD 45 (2024), 34–54

Nikola Smolenski, **Ivana Gavrilović**
University Library "Svetozar Marković"

AN ANALYSIS OF THE USE OF THE SEARCHABLE DIGITIZED HISTORICAL NEWSPAPERS PROJECT

Abstract. This paper presents the results of the analysis of the Searchable Digitized Historical Newspapers project of the University Library "Svetozar Marković". The results of the analysis shall be used for improvements of its search functions and implementation of the same in similar future projects.

The first part of the analysis is a usual web analysis of page views, however its second and more important part is the analysis of search functions and queries.

An immediate application of the work is in the Serbian Literary Criticism (*Srpska književna kritika*) and The Epoch (*Epoha*) projects, that also include the search of digitized documents, using similar technologies.

Keywords: Europeana, digitalization, historical newspapers, web traffic analysis, search engine

1. Introduction

Searchable Digitized Historical Newspapers is a project of the University Library "Svetozar Marković" from Belgrade, that provides search and display of digitized historical newspapers and other documents. The website offers simple and advanced search interfaces, and an interface for browsing through the digitized works. It was located at the address http://www.unilib.rs/istorijske-novine/, but at the time of writing paper the site is under updating process and the URI of the site will be changed to https://pretraziva.rs/.

The previous website design is shown in the Figure 1:

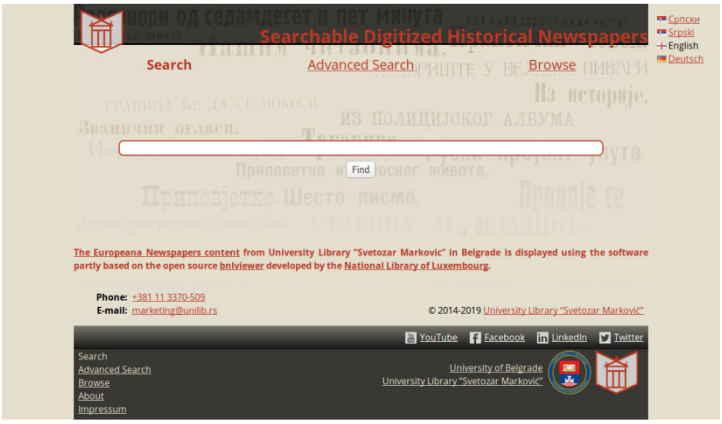

Figure 1. Old website design

Current website design is shown in the Figure 2:

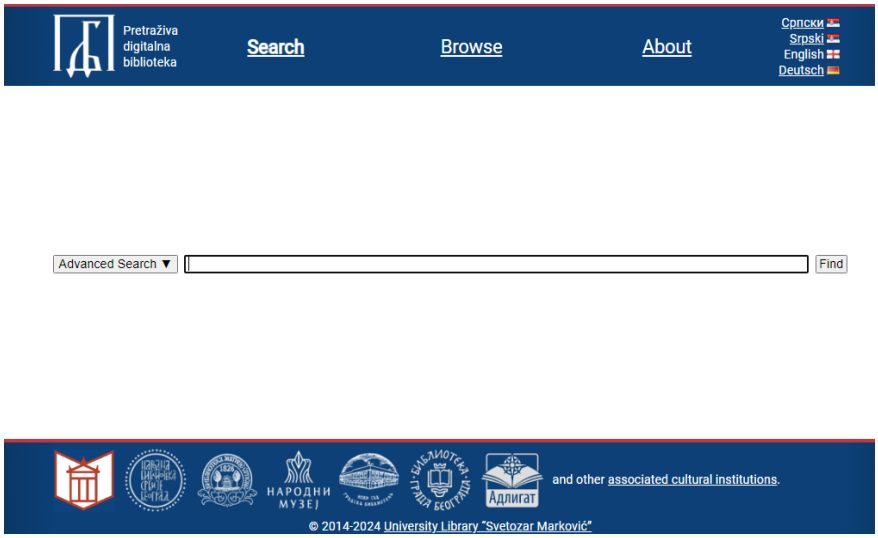

Figure 2. New website design

The website relies on the BnLViewer software by the National Library of Luxembourg [1] to display digitized documents, while the search backend uses Lucene/Solr [2] search engine with enhancements made by the library. The website infrastructure [3] is shown in the following diagram (Figure 3):

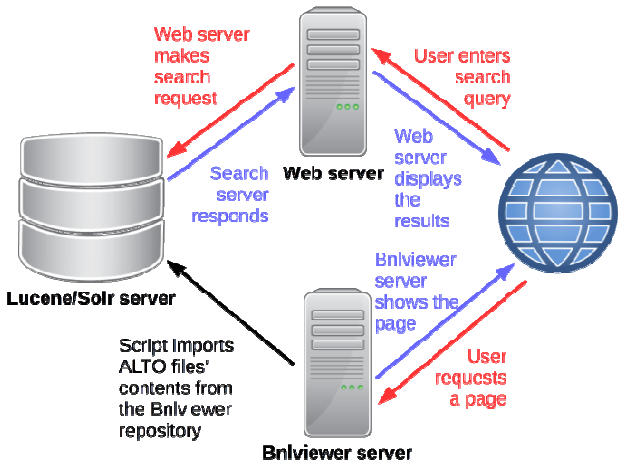

Figure 3. The website infrastructure

Collection descriptions and other texts of the website itself are published under the Creative Commons Attribution (CC-BY) license [4], to facilitate their reuse in other projects.

1.1. History. European Digital Library Network / EDLnet, is a project initiated in 2007. by the European Commission with the goal of digitizing European cultural and scientific heritage. The EDL project was incrementally upgrading and growing, and in 2008 was created the Europeana portal, which enables the access to a large number of digital objects: books, images, films, museum pieces [5].Over 2,000 European institutions are taking part in upgrading the Europeana digital library. The digital objects that are represented at the portal remain stored in the network of the central institution that is publishing them on the portal.

"Europeana Newspapers" is a similar initiative by the European Commission CIP 2007-2013, which aims to aggregate digitized newspapers for The European Library and Europeana. One of the main goals of this project is upgrading solutions for specific questions about search of digital newspapers collections.

The University Library "Svetozar Marković" is participating in the Europeana Newspapers project since 2013, at first by digitizing around 400,000 pages of historical Serbian newspapers, by scanning them within the library, with the University of Innsbruck later making METS/ALTO [6],[7] files out of them. At that moment, the library did not have adequate technical solutions for searching the digitized content; it was only possible to view it, using the BnLViewer. The search of the newspapers was implemented as an own solution, in the form of the website.

Later, the entire collection was enhanced with the search interface. This interface enables the visitors to search the entire collection in several search modes: basic, advanced or by browsing by newspaper name and date. Additionally, the website has the search instructions with detailed explanations about the search, and is available in Serbian, Serbian Latin, English and German languages.

Eventually, the library has mastered preparing METS/ALTO documents on its own, so that the entire collection grew to over 500,000 pages of digitized material.

1.2. Definitions. This section defines the terms used throughout the work that might otherwise be ambiguous or undetermined.

An "issue" is a single digital document that can be viewed as an independent whole. This could be a single issue of a magazine, or a book.

A "collection" is a group of issues. These could be the issues of a periodical publication, or a group of thematically-related books.

The "entire collection" includes all the documents available on the Searchable Digitized Historical Newspapers.

A "page view" is a view of any page by a visitor. For example, if a visitor makes a search query, and browses through 10 pages of results, this is considered to be 10 page views.

A "search query" is a distinct query by a visitor, regardless of later browsing through the results. The example above would be considered as a single search query. In addition, if a visitor repeats the same query multiple times within a day, this is also considered to be a single search query, as detailed in the section 3.2. Search Queries.

2. Method

The site traffic has been analyzed by analyzing the traffic logs. Because of the server configuration, we have no access to the actual server logs, however the website software independently logs every page view in the Apache combined log format [8].

The period of the analysis is 2016-2018, inclusive. This period includes a total of 1,378,823 page views.

The logs were converted to SQL format and imported into a relational database. During this step, we have filtered out page views that were made by bots and web crawlers, using the Crawler-Detect[9] package. We have found that 1,251,253 (90.75%) page views were made by humans, while the remaining 127,570 (9.25%) were made by bots and web crawlers.

The analyses that could be performed by SQL queries alone were performed that way, while for more complex analyses, the query results were exported to CSV format and analyzed by specific programs, some of which are presented in section 5. Appendix.

3. Results

3.1. Page Views. Out of 1,251,253 human page views, 1,231,577 (98.43%) were valid, while the remaining 19,676 (1.57%) were attempts to view non-existent pages (404 errors [10]).

3.1.1. By Page. One of the features of this web site, that is somewhat unusual, is that the root page takes the visitor to the last page previously used by the visitor, if there is any; if not, it takes the visitor to the default, search page.

The numbers of page views regardless of their language are given in the following table:

Total	1,231,577	100.00%
Advanced search	624,442	50.70%
Search	416,381	33.81%
Browse	138,473	11.24%
Root page	50,346	4.09%
About	1,359	0.11%
Impressum	576	0.05%

It is interesting to note that the advanced search page is used more than the basic search page, which is the opposite of what might be expected. The proportion of the pages' usage remains similar when the number of search queries or the number of viewed results using them are compared. This is examined in further detail in section 3.2.2. By Advanced Search Abilities.

3.1.2. By Language. The website is localized to four interfaces: Serbian, Serbian Latin, English and German.

The English and German interfaces are not, strictly speaking, useful for the search, because the entire collection is practically entirely in the Serbian language,

therefore there are no search results in English or German, and so everyone who can understand the search results could also understand the Serbian interface. Similarly, the Serbian Latin interface is not fully useful for people who don't know Serbian Cyrillic, since a number of search results are still in Cyrillic. However, all of these visitors may still read the pages about the website and descriptions of the magazines on the Browse page.

The numbers of page views by languages are given in the following table:

Serbian	951,818	81.24%
Serbian Latin	219,814	18.76%
English	4,869	0.42%
German	4,730	0.40%

It is interesting to note that visitors using the Serbian interface and the Latin advanced search interface show no preference in the script used for their search queries; while the visitors using the Latin basic search interface are using Latin alphabet 4 times more than Cyrillic. Somewhat similarly, while most of the visitors who used the English interface still searched in the Serbian language, most of the visitors who used the German interface tried to search in the German language, probably to disappointing results.

Our conclusion is that the search interfaces in foreign languages are not necessary and might even be confusing for the visitors; while, of course, pages about the project are still necessary.

3.1.3. By Country. We have detected visitors' countries by finding the locations of their IP addresses using MaxMind GeoLite2 database [11]. While the database is valid for 2019, and therefore newer than the observed period, we have not identified any major changes in IP ranges that have happened during the period that could significantly affect the results.

We have then paired the number of page views with the number of Internet users by country [12] in order to estimate the relative number of visitors from every country. This database, on the other hand, contains data for 2016, but the changes in the numbers of Internet users should be small and proportional enough not to affect the results significantly.

The results, for all the countries with more than 100 page views, are given in the following table:

Name	Page views		Internet users (in millions)	Page views per 100,000 users
Total	1,251,253	100.00%	3,385	37
Serbia	999,402	79.87%	4.7	21,264
Bosnia and Herzegovina	92,897	7.42%	1.9	4,889
Croatia	37,206	2.97%	3.1	1,200
Montenegro	25,490	2.04%	0.44	5,793
Macedonia	13,993	1.12%	1.5	933

Austria	11,803	0.94%	7.3	162
United States	8,289	0.66%	250	3
Germany	6,696	0.54%	73	9
Netherlands	5,407	0.43%	15	36
China	4,999	0.40%	750	1
Japan	4,717	0.38%	120	4
Greece	3,635	0.29%	7.7	47
Slovenia	3,487	0.28%	1.6	218
Sweden	3,206	0.26%	8.8	36
Italy	3,111	0.25%	36	9
France	2,668	0.21%	55	5
Canada	2,524	0.20%	33	8
Russia	2,229	0.18%	110	2
Australia	2,064	0.16%	21	10
Switzerland	1,644	0.13%	7.5	22
Hungary	1,477	0.12%	7.7	19
Albania	1,261	0.10%	1.9	66
Norway	1,087	0.09%	5.1	21
United Kingdom	1,068	0.09%	62	2
Ukraine	500	0.04%	22	2
Czechia	489	0.04%	8.1	6
South Africa	476	0.04%	30	2
Bulgaria	427	0.03%	4.3	10
Ireland	384	0.03%	4	10
Poland	374	0.03%	28	1
Spain	334	0.03%	37	1
Romania	311	0.02%	12	3
Slovakia	253	0.02%	4.4	6
Iceland	192	0.02%	0.33	58
Malta	119	0.01%	0.33	36
Denmark	107	0.01%	5.5	2
Other	999	0.08%	1,644.8	0
Unknown	5,928	0.47%		

As could be expected, the large majority of the visitors are from Serbia, followed by other countries where the Serbian language is spoken, followed by countries with significant presence of Serbian diaspora.[1]

We believe that the number of visitors from other countries could be significantly increased by including collections from these countries or relevant for the countries, by adding informational pages about the project in the languages of the countries, and by informing the potential audiences in the countries about the project.

3.1.4. By Device. We have counted the number of kinds of devices used to access the website, by analyzing the User-Agent strings [13] of the visitors using the WhichBrowser [14] package. The results are in the following table:

Total	1,251,253	100.000%
Desktop	1,114,728	89.089%
Mobile phone	79,373	6.343%
Tablet	56,952	4.552%
Media player	10	0.001%
Television	9	0.001%
Unknown	181	0.014%

Given that almost 11% of the visits come from mobile phones or tablets, we conclude that the design of our website and similar projects needs to support mobile devices and take into account their smaller screens.

However, this is not presently necessary on our website, given that the display of digitized documents itself does not support mobile devices.

3.1.5. Browse Page. The Browse page gives visitors an overview of all the digitized collections, as well as the ability to select the desired collection and its issue by date. Most collections are accompanied by a short text about them.

Out of the total 138,473 views of the page, 39,995 (28.88%) were views of the overview page, while the remaining 98,478 (71.12%) were views of an individual collection.

Despite the existence of the powerful search feature, since the browse page has 11.24% of all page views, we conclude that this kind of page is still needed by the visitors.

1 After an additional review, we have concluded that page views from China and Japan are probably undetected bots or proxies, or IP range changes; we do not believe that there is a significant number of actual views from China or Japan.

3.2. Search Queries. In the observed period, there have been 396,046 search queries in total. We have, however, noticed that the visitors will often enter the same query multiple times, often in rapid succession. This may occur because a visitor accidentally reloads the page, returns to a previous query after searching for another one, repeats the same query with different advanced parameters, or for various other reasons.

In the analysis of the search queries, we wanted to analyze distinct queries, which led us to the question which queries are distinct. Obviously, queries made by distinct visitors are distinct, however we have no direct way to find out which visitors are distinct other than observing the IP address used for a page view. We have considered that each IP address of a page view belongs to a distinct visitor,[2] and counted the number of repeated queries from that address that occurred within a certain time range from each other.[3]

The results are given in the following table:

Total	396,046	100%
Within 1 second	2,411	0.61%
Within 1 minute	55,056	13.90%
Within 1 hour	89,735	22.66%
Within 1 day	98,090	24.77%
Within 1 week	102,149	25.79%

Given that the number of non-distinct queries that occur within 1 hour, day and week is similar, each of these choices would yield similar results. We have decided to choose 1 day, as it is possible that a visitor might perform distinct research on two consecutive days, and so will legitimately enter the same query twice, but distinctly.
The final result is that, in the observed period, there were 300,367 distinct search queries.

3.2.1. By Script. The Serbian language uses two alphabets: Cyrillic and Latin. In addition, Serbian Internet users often use "bald" Latin alphabet, using letters without diacritics [15]. The entire collection contains documents written in both alphabets, and the website offers the ability to search all documents using either alphabet, including without diacritics [16].

Offering the ability to search without diacritics makes the search queries less precise, since there are many cases of word pairs that match, except for diacritics. Thus it is very important to find out how much is this feature needed by the visitors.

We have counted the number of search queries by the alphabet used[4], and the results are given in the following table:

---

2 IP addresses often change through time, however, in a time period as short as one of these, each IP address should belong to a distinct device. It is also possible that the same IP address is shared by multiple real people, or that one person uses multiple IP addresses, however, these cases should be rare, and cancel each other.

3 "From each other" means that, for example, five queries occurring over three minutes are still counted as one query, if either two of the queries are within a minute from each other.

4 The exact method of alphabet identification could be seen in the source code provided in section 5.2. Query Script Identification.

Total	300,367	100.00%	Example query
Serbian Cyrillic	149,164	49.66%	Радовић
Other Cyrillic	180	0.06%	Бакичь
Mixed Cyrillic and Latin	2,985	0.99%	pero врањешевић
Latin with diacritics	48,439	16.13%	Radović
Latin without diacritics	15,955	5.31%	Mikijelj
Latin with possible diacritics	82,467	27.46%	Klisic
Other Latin	301	0.10%	Jüterbog
Other	20	0.01%	Νίψον
No letters	856	0.28%	1914

However, just from this it is not possible to know whether the queries with possible diacritics are using "bald" Latin alphabet or are composed from words that actually consist only of letters without diacritics. To find this out, we have counted a random sample of 1024 such queries. The results are given in the following table:

Total	1,024	100%	82,467
Bald	381	37%	30,513
Not bald	615	60%	49,480
Uncertain	28	3%	2,474

We can conclude with 99% confidence that, out of 82,467 queries with possible diacritics, 37±4% are in fact "bald" Latin, which is 10±1% of all search queries. This result shows that the ability to search without diacritics is necessary for our visitors.

Thus, our recommendation is that similar projects should have this feature, or if they decide not to, should warn its visitors about it, and offer a virtual keyboard or another means for entering queries with diacritics.

3.2.2. By Advanced Search Abilities. The advanced search interface offers the abilities: to choose the number of the results on a single page; to search within only a single collection; to sort the results by score (relevance) or date; and to search within a range of dates. The layout of the interface is shown in the Figure 4:

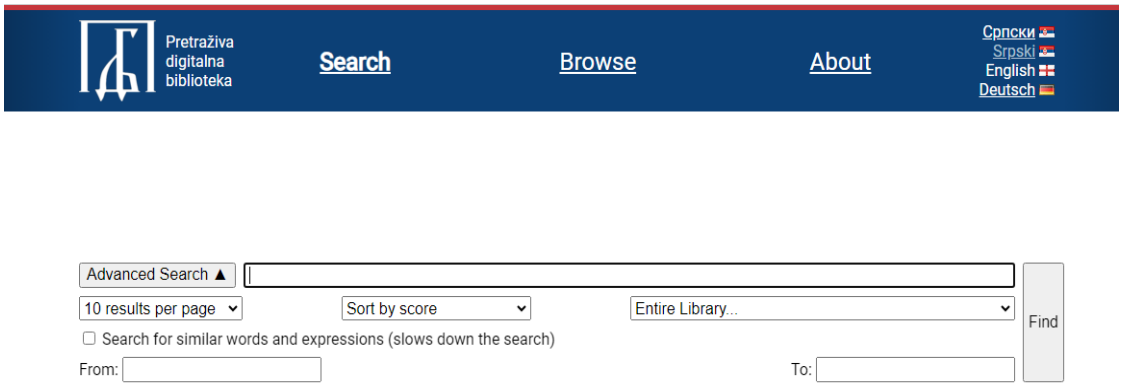

Figure 4. Advanced Search Abilities interface

In the analysis of the use of the advanced search interface, we have again wanted to analyze distinct queries, however, in this case, the use of a same query with a distinct advanced search ability, again if not repeated within a day, is regarded as a distinct query.

In the observed period, there were 250,113 queries using the advanced interface, out of which 213,303 were distinct. It should be noted, however, that 88,332 (41.41%) distinct advanced queries do not use any advanced abilities. This can be contrasted with the fact that the advanced search interface is used more than the basic interface, as remarked in section 3.1.1. By Page. Given that in the observed period there were 87,064 distinct queries using the basic search interface, when added to the above, the conclusion is that 175,396 (58.39%) of all distinct search queries do not use advanced search abilities, and thus that the quality of the results of the basic search is of particular importance.

The number of queries by the number of results sought to be displayed is given in the following table:

Total	213,303	100.000%
10	180,671	84.702%
20	4,860	2.278%
50	3,268	1.532%
100	24,495	11.484%
Other	9	0.004%

We conclude that the choice of the number of displayed results is a necessary feature for our visitors, albeit the choices of 10 (default) or 100 results seem to be the only ones needed.

The number of queries which were performed within (or not) a collection is given in the following table:

Total	213,303	100.00%
All collections	164,552	77.14%
A distinct collection	48,751	22.86%

We conclude that searching within a collection is also a feature necessary for our visitors.

The number of queries by the use of sorting is given in the following table:

Total	213,303	100.00%
Score	131,875	61.83%
Date ascending	57,946	27.17%
Date descending	23,482	11.01%

We conclude that the choice of sorting is a necessary feature for our visitors. It is interesting to note that searching for the earliest results (sorting by date ascending) is almost three times as common as for the latest, meaning that our visitors prefer older content to newer. This is further confirmed in the subsection.

The queries that use a range of dates are analyzed in detail in subsection 3.2.2.1. By Date Range.

During the writing of this work, after the observed period, we have introduced an additional advanced search ability, searching for similar words and phrases. This feature searches for words within Levenshtein distance [17] of 1 for 3-4 letter words and 2 for longer words; and phrases that differ by up to two words. Initial results, for May 2019, show that this feature is used in around 8.5% of queries, thus it is likely also necessary for our visitors.

3.2.2.1. By Date Range. One of the features of the advanced search interface is the ability to enter a date range for the search. The dates for the date range can be selected in a calendar interface, but it is also possible to enter them manually. The interface is shown in the Figure 5:

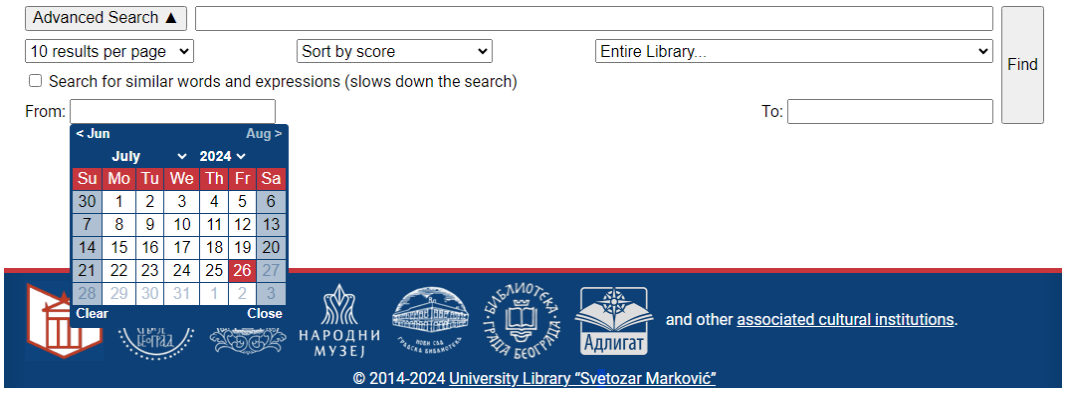

Figure 5. Calendar interface

Out of 213,303 distinct advanced queries, 26,185 (12.28%) have a date range specified. We conclude that the search within a range of dates is a necessary feature for our visitors.

A query can have its date range specified as closed between two dates, open from a starting date to present, or open from the beginning of time to an ending date. The count of queries by the type of range is given in the following table:

Total	26,185	100.00%
Closed range	20,522	78.37%
Start date to present	2,941	11.23%
Past to end date	2,722	10.40%

Out of all the queries, 222 (0.008%) have one or both of the dates manually entered malformed in various ways, so that they are not recognizable by the website. Since this is a very small number, we conclude that manual entering of date ranges in this fashion is acceptable, however, similar projects should be liberal with the date formats that are accepted.

The count of queries with closed date range, by range, is given in the following table:

Total	20,522	100.00%
1 day	716	3.49%
1 day - 1 week	430	2.10%
1 week - 1 month	593	2.89%
1 month - 1 year	3,014	14.69%
1 year - 10 years	4,970	24.22%
10 years - 100 years	9,308	45.36%
More than 100 years	1,380	6.72%
Invalid	111	0.54%

Given that 8.52% of queries does require a range smaller than one month, our conclusion is that the ability to select the exact date in the search interface is necessary for our visitors. It is interesting to note that, while it could be assumed that visitors searching for information on a certain historical event would restrict the date range close to that event, most queries have a very wide range.

Another analysis of queries with date ranges shows what years are the most interesting to our visitors. The number of queries with date range that includes a certain year are paired with the number of pages in the entire collection from the year (this includes the pages that are added after the observed period, which should not affect the results significantly). The results are shown in the following graph ():

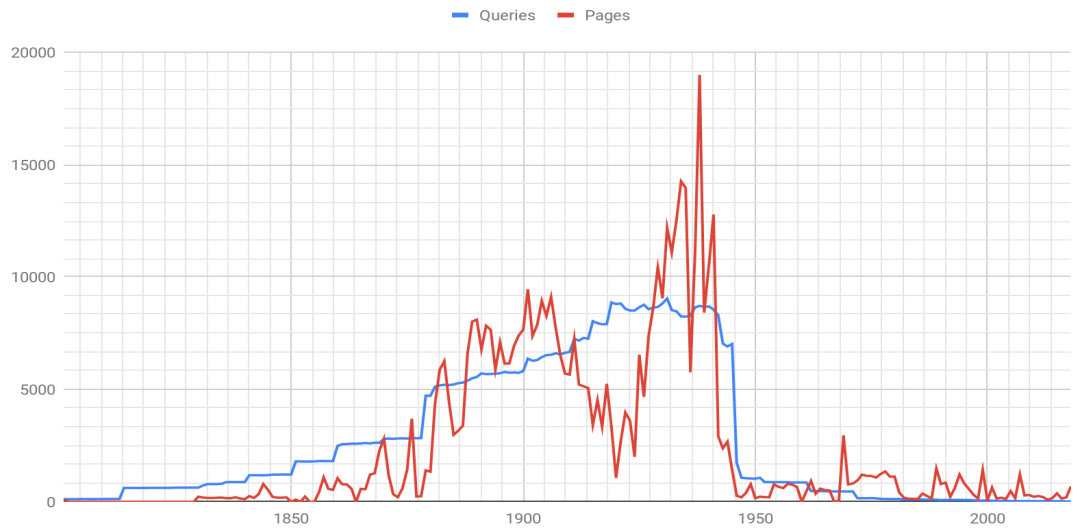

Figure 6. Analysis of queries with date ranges

This shows that the visitors are searching mostly for the years where the entire collection contains the most pages. However, among them, it shows that the First World War and the immediate pre-war and post-war periods are relatively the most interesting for our visitors, since the number of pages in these periods is smaller, yet the number of searches is not; similarly, earlier results are relatively more interesting than later ones, which correlates with the fact that visitors prefer sorting by ascending date.

It appears that in the searches there is no preference for specific events. Apparent peaks on the graph do not correspond to actual historical events, but are located at every 5th year, showing simply that visitors prefer "clean" years to select a date range.

During the writing of this work, a visitor has complained by email that the interface for entering date range in the advanced search interface is confusing. We do not see why this interface is confusing.

3.2.3. By Advanced Query Abilities. In both the basic and the advanced search interface, the search will by default find all the pages which contain all the sought words, however there are three advanced search capabilities of the search query [18]:

- it is possible to search for an expression that has multiple words in order, by putting them between quotation marks;
- it is possible to use the (localized) OR operator to find pages that contain any of the searched words;
- It is possible to exclude pages that contain a word by prefixing it with the minus sign.

We have counted how often are these abilities used, and the results are in the following table:

Total	300,367	100.000%
Multiple-word expression	32,923	10.961%
OR operator	12	0.004%
Word exclusion	105	0.035%
None	267,327	89.000%

As could be seen, visitors often search for multiple-word expressions, while other advanced abilities are practically never used. While it is possible that some visitors are simply not aware of the advanced query abilities, it is unlikely that they are aware of only one of them, but not all of them, thus we conclude that OR operator and word exclusion are simply not needed by the visitors.

However, this is not a recommendation to exclude these features in similar projects: they should be included, if they are easy to implement.

3.2.4. By Topic. In order to find out which topics are searched the most by our visitors, we have analyzed 1034 randomly selected search queries, assigning to each one its probable topic or topics. Each query with a reliably-assigned topic gets the score of 1, but if a topic could not have been reliably assigned, we have split the score between the probable topics (for example, if there are two probable topics, each one gets the score of 0.5). We have then summed all the scores. The results (99% confidence, ±4% error margin) are given in the following table:

		Example queries
Person	47%	stefan decanski velbuzd; ćatović ilijaz; broćović
Location	16%	opštine čestin; miljakovac; strazilovo
Concept	13%	romanu u nastavcima; antialkoholicno; presbiro
Microlocation or organization	12%	"Zrinjskog 10"; Šećerana Bač; hotel srpski kralj
Group, event or other	13%	muftije; Veče klavirskih kompozicija; а самим тим

The identified topics additionally confirm the conclusions in section 3.2.5. By Word Class.

3.2.5. By Strategy. We have noticed that visitors use multiple strategies when searching for people. We have thus assigned the strategy to every query about a person from the previous dataset, 498 queries in total. The results (99% confidence, ±5.77% error margin) are given in the following table:

		Example queries
Full name	49.20%	ćatović ilijaz; "наталија леко"; радош и зорка недић
Name	27.71%	broćović; путник; tonkovica
Name and activity	13.05%	СЛИКАР НИКОЛАЈ МАЈЕНДОРФ; Kraljica Marija
Name and placename	8.43%	Sreten Bogicevic Kragujevac; Баба Злата врање
Other	1.61%	1884 mitropolit

The words used in these search strategies additionally confirm the conclusions in the next section.

3.2.6. By Word Class. Using the same random sample as in section 3.2.4. By Topic, we analyzed the words in visitor queries in order to find out what word classesare used for search. For word classes where applicable, we also analyzed the words's grammatical number, whether it is in a case other than the nominative, and whether it is written correctly. The results (99% confidence, ±4% error margin) are shown in the following table:

Word class	Total	Plural	Case	Error	% of total	%plural	%case	%error
Total	2,176	105	328	41	100.00%	4.83%	15.07%	1.88%
Proper noun	1,167	9	194	25	53.63%	0.77%	16.62%	2.14%
Common noun	524	81	94	10	24.08%	15.46%	17.94%	1.91%
Adjective	216	14	36	5	9.93%	6.48%	16.67%	2.31%
Unchangeable	152	N/A	N/A	1	6.99%	N/A	N/A	0.66%
Number	73	N/A	2	0	3.35%	N/A	2.74%	0.00%
Verb	36	1	N/A	0	1.65%	2.78%	N/A	0.00%
Adverb	4	N/A	N/A	0	0.18%	N/A	N/A	0.00%
Pronoun	4	0	2	0	0.18%	0.00%	50.00%	0.00%
Unknown	5	N/A	N/A	N/A	0.23%	N/A	N/A	N/A

More than half of the words used in the queries are proper nouns, followed by common nouns and adjectives, while verbs are practically unused. We conclude that in order to improve search results, it is important to focus on improving the searching for nouns and adjectives, while verbs are of secondary importance. In agreement with the previous two sections, especially important are proper nouns, and among them especially human names.

Another conclusion is that visitors can not be relied upon to enter their queries in nominative case, which could potentially be used to make the search more precise.

The typing error rate of around 2% suggests that our visitors are experienced typists, albeit it is at the higher end of the reported typing error rates [19].

3.2.7. By Depth. We have examined how many search results for their searches the visitors wanted to see. Note that this is the number of results sought, not necessarily displayed (for example, queries with no results are also included). Here we have not tried to identify distinct queries, given that it is not possible to discern whether multiple views of the same query at the same depth are distinct, or repeated pagination through the same query, thus we have resorted to counting the page views. We have also not paid attention to the advanced search option for selecting the number of displayed results, which we believe does not affect the search depth significantly. The results are shown in the following graph (Figure 7):

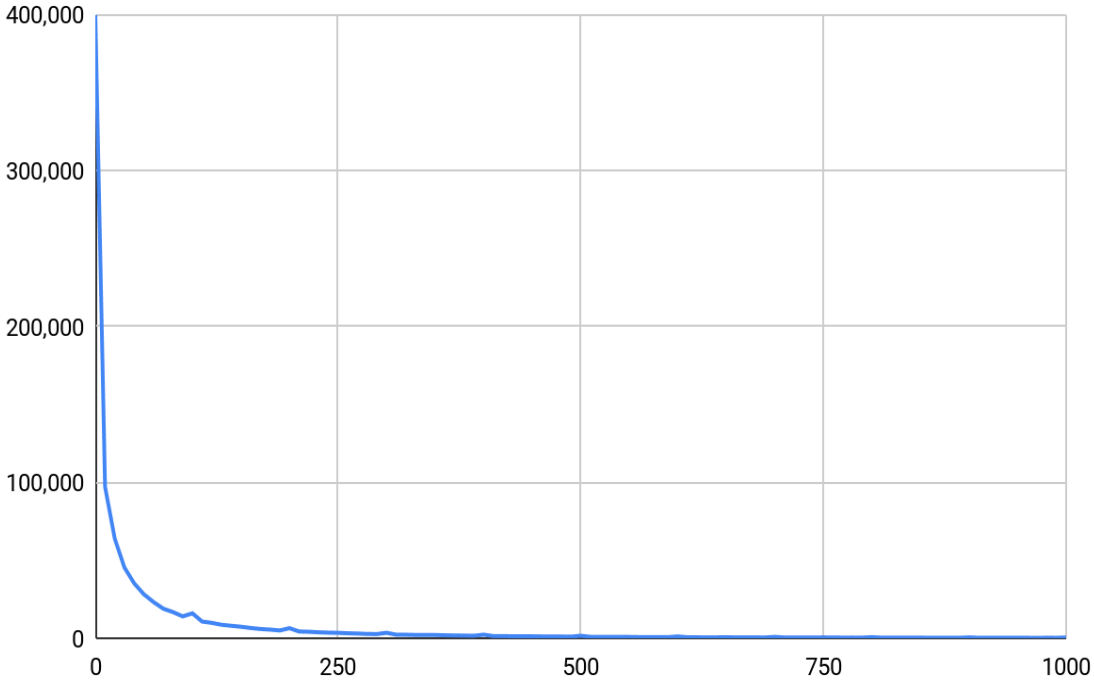

Figure 7. The number of search results the visitors wanted to see

The results show a typical long tail distribution, with most searches ending at the very first results, while very few needing a lot of results. This could be more precisely seen in the following table:

Results sought	Page views	Percent of searches
Total	942,332	100.00%
0	399,493	42.39%
10-90	342,719	36.37%
100-990	182,329	19.35%
1,000 and more	17,791	1.89%

Given that for more than half (52.70%) of the search queries the visitors are looking only at the first or second page of results, we find that it is very important that these pages have relevant results.

Another conclusion of ours is that, while there is no reason to artificially limit the number of results that can be displayed, if displaying large numbers of results is difficult to implement or would significantly slow down the system, the maximal number of displayed results can be set to 1,000, since that would leave less than 2% of desired page views unanswered.

3.2.8. Notes. During the making of this work, we had various observations that can not be statistically analysed, but still might offer interesting insights.

We have noticed that, in a couple of queries, the visitors were using word exclusion accidentally, while in fact intending to write a hyphenated word, for example:

• Спомен -дом Витешког Краља Александра

This happens rarely enough that this feature should not be excluded because of it.
In a number of queries, the visitors were using + operator or * operator, which do not have any effect, apparently believing that it will work similar to some other search engines.
Some of the visitors have serial searches for the same term in different grammatical forms, which would benefit from the OR operator, however it is still unused.
We have found examples of queries entered using the wrong keyboard layout, for example:

• milan ga;evi'
• marko vrawe[evi'
• arbitrayni sud

Although it would in principle be possible for the search engine to recognize and correct this, it happens very rarely.

Several visitors have mistakenly tried to use the search interface to get to the website of the library or to a URL.

4. Conclusions

Here is the summary of our conclusions:
• Search interfaces in foreign languages might be harmful.
• It is possible to increase visits from foreign countries by including relevant content and by promotion.
• Mobile devices should be supported.
• “Bald” Latin alphabet needs to be supported for the Serbian language.
• The present advanced search abilities (listed in section 3.2.2. By Advanced Search Abilities) are all needed, except 20 and 50 displayed results per page.
• Date range interface could be improved, although it is unclear how exactly.
• The OR operator and word exclusion in search queries are not necessary.
• Any search improvements should focus on, in order: human names, other proper nouns, nouns and adjectives.
• If necessary, number of displayable results can be limited to 1,000.

5. Appendix

Following calls to publish the source code used in scientific research [20], we have included here relevant source code used in making this work.

5.1. Conversion of Apache Log to SQL. The following program converts Apache combined log format to SQL, at the same time removing crawlers and filtering dates.

```
<?php

require_once( "vendor/autoload.php" );
use Jaybizzle\CrawlerDetect\CrawlerDetect;
$crawlerDetect=new CrawlerDetect;
```

```
$logFile=fopen( $argv[1], "r" );
while( ( $line=fgets( $logFile ) ) !==false ) {
	$line=mb_convert_encoding( $line, 'UTF-8', 'UTF-8' );

	preg_match( "/ [[]([^]]+)[]] /", $line, $m );
	$time=date( "Y-m-d H:i:s", strtotime( $m[1] ) );
	if( $time<"2016"||$time>"2019" ) {
		continue;
	}

	preg_match( "/\"([^\"]*)\"$/", $line, $m );
	$userAgent=$m[1];
	if( $crawlerDetect->isCrawler( $userAgent ) ) {
		continue;
	}

	preg_match( "/^([^ ]+) /", $line, $m );
	$ip=$m[1];

	preg_match( "/ \"GET ([^\"]+)\" /", $line, $m );
	if( empty( $m ) ) {
		continue;
	}

	$url=$m[1];
	$components=parse_url( $url );
	$path=rtrim( preg_replace( "#^/istorijske-novine/#", "", urldecode( $components['path'] ) ), "/" );
	$query=$queryParams= [];
	if( @$query=$components['query'] ) {
		$query=explode( "&", $query );
		foreach( $queryas$k=>$v ) {
			$v=explode( "=", $v );
			$queryParams[urldecode($v[0])] =@urldecode( $v[1] );
		}
	}

	$line=rtrim( $line );

	echo
"INSERT INTO analysis( ip, time, path, search, newspaper, startrow, dateFrom, dateTo,
collection, sort, results, raw ) VALUES(
	". prep( $ip ) .",
	". prep( $time ) .",
	". prep( $path ) .",
	". prep( @$queryParams['search'] ) .",
	". prep( @$queryParams['newspaper'] ) .",
	". prep( @$queryParams['startrow'] ) .",
	". prep( @$queryParams['dateFrom'] ) .",
	". prep( @$queryParams['dateTo'] ) .",
	". prep( @$queryParams['collection'] ) .",
```

```
        ". prep( @$queryParams['sort'] ) .",
        ". prep( @$queryParams['results'] ) .",
        ". prep( $line ) ."
);
";
        //      echo "$ip\t$time\t$path\t$line\n";
}

functionprep( $str ) {
        if( $str===null ) {
                return"NULL";
        } else {
                return"'".mysql_escape_string( $str ) ."'";
        }
}
```

5.2. Query Script Identification. The following program identifies query script and advanced query abilities used of queries supplied in a CSV file.

```
<?php

$regexps= [
        "All Cyrillic"=>"/^[\\p{Cyrillic}\\P{L}]+$/ui",
        "Serbian Cyrillic"=>"/^[абвгдђежзијклљмнњопрстћуфхцчџш\\P{L}]+$/ui",
        "Mixed Cyrillic and Latin"=>"/^(?=.*[a-zšđžčć])(?=.*[\\p{Cyrillic}].*)/ui",
        "All Latin"=>"/^[\\p{Latin}\\P{L}]+$/ui",
        "Latin with diacritics"=>"/^(?=.*[čćšđž])[\\p{Latin}\\P{L}]+$/ui",
        "Latin without diacritics"=>"/^[abefghijklmnoprtuv\\P{L}]+$/ui",
        "Latin with possible diacritics"=>"/^(?=.*[cdsz])[a-z\\P{L}]+$/ui",
        "Word exclusion"=>"/(^-|\\s-)[^\\s]/",
        "Multiple-word expression"=>"/\"/",
        "OR operator"=>"/\\s(OR|ИЛИ|ILI|ODER)\\s/",
];

$res= [];

$f=fopen( $argv[1], "r" );
while( $line=fgetcsv( $f ) ) {
        $query=$line[3];

        if( preg_match( "/^\\P{L}+$/", $query ) || trim( $query ) ==="" ) {
                @$res["No letters"]++;
                continue;
        }

        foreach( $regexpsas$name=>$regexp ) {
                if( preg_match( $regexp, $query ) ) {
                        @$res[$name]++;
                }
        }
}
```

```
print_r( $res );
```

References

[1] The National Library of Luxembourg, "bnlviewer," SourceForge, Feb. 16, 2016. [Online]. Available: https://sourceforge.net/projects/bnlviewer/
[2] The Apache Software Foundation, "Apache Solr," [Online]. Available: http://lucene.apache.org/solr/
[3] N. Smolenski, M. Kostic, A. M. Sofronijevic, "Intrapreneurship and Enterprise 2.0 as Grounds for Developing In-House Digital Tools for Handling METS/ALTO Files at the University Library Belgrade," in Developing In-House Digital Tools in Library Spaces, IGI Global, 2018, 86–111.
[4] Creative Commons, "About The Licenses," [Online]. Available: https://creativecommons.org/licenses/
[5] J. Kamps et al., Eds., Research and Advanced Technology for Digital Libraries: 21th International Conference on Theory and Practice of Digital Libraries, TPDL 2017, Thessaloniki, Greece, September 18-21, 2017 : Proceedings. Cham: Springer International Publishing, 2017.
[6] The Library of Congress, "Metadata Encoding and Transmission Standard (METS) Official Web Site," Jan. 17, 2019. [Online]. Available: http://www.loc.gov/standards/mets/
[7] The Library of Congress, "ALTO: Technical Metadata for Layout and Text Objects," May 26, 2016. [Online]. Available: https://www.loc.gov/standards/alto/
[8] The Apache Software Foundation, "Combined Log Format," 2019. [Online]. Available: https://httpd.apache.org/docs/2.4/logs.html#combined
[9] JayBizzle, "JayBizzle/Crawler-Detect," GitHub. [Online]. Available: https://github.com/JayBizzle/Crawler-Detect
[10] R. Fielding, "10.4.5 404 Not Found," [Online]. Available: https://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html#sec10.4.5
[11] MaxMind, Inc., "GeoLite2 Free Downloadable Databases," Mar. 4, 2019. [Online]. Available: https://dev.maxmind.com/geoip/geoip2/geolite2/
[12] International Telecommunication Union, "Internet users by region and country, 2010-2016," 2019. [Online]. Available: https://www.itu.int/en/ITU-D/Statistics/Pages/stat/Treemap.aspx
[13] "User agent," Wikipedia, Mar. 15, 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=User_agent&oldid=887947355#Use_in_HTTP
[14] WhichBrowser, "WhichBrowser/Parser-PHP," GitHub, Oct. 2, 2018. [Online]. Available: https://github.com/WhichBrowser/Parser-PHP
[15] D. Ivković, "Pragmatics meets ideology," Journal of Language and Politics, 12(3), Jan. 2013, 335–357
[16] N. Smolenski, "SerbianLanguageSupport - Solr Wiki," Nov. 2, 2015. [Online]. Available: https://wiki.apache.org/solr/SerbianLanguageSupport
[17] "Levenshtein distance," Wikipedia, Apr. 25, 2019. [Online]. Available: https://en.wikipedia.org/w/index.php?title=Levenshtein_distance&oldid=894091390
[18] University Library "Svetozar Marković", "Searchable Digitized Historical Newspapers," [Online]. Available: http://www.unilib.rs/istorijske-novine/about#search-guide (accessed jul. 13, 2019)

[19] J. T. Grudin, "Error Patterns in Novice and Skilled Transcription Typing," in Cognitive Aspects of Skilled Typewriting, W. E. Cooper, Ed. New York, NY: Springer New York, 1983, 121–143
[20] N. Barnes, "Publish your computer code: It is good enough," Nature, 467(7317), 2010, 753–753

smolenski@unilib.rs
gavrilovic@unilib.rs