Full-Text Search in the Resources of Polish Digital Libraries

Arkadiusz Pulikowski
ORCID 0000-0003-1807-8642
Institute of Culture Studies, Faculty of Humanities
University of Silesia in Katowice, Poland

Abstract

Purpose/Thesis: The article aims to analyze the conditions and possibilities of full-text search in Polish digital libraries (DL), taking into account the access to full-text search in individual DLs, the file formats and the software used, as well as the visibility of the DL resources in the Google search engine.
Approach/Methods: Forty of the largest Polish DLs, whose resources primarily comprise digitized traditional library resources, were selected for the study. The study examined the type of the software used, the availability and efficiency of the full-text search, and the extent to which the resources were indexed as PDF files in Google and Google Scholar. Finally, the study compared the results of full-text search in ten DLs with those obtained from Google.
Results and conclusions: All DLs selected for the study allow for full-text search. There are significant differences between specific interfaces. Each has advantages and disadvantages that require further development. The Google search engine is not currently a viable alternative to accessing content provided in DLs.
Originality/Value: The issue of full-text search in DLs is rarely addressed, even though users consider it one of the most important functions of DLs. The result of the study present a picture of DL's current capacity in this area.

Keywords

Content Visibility. Digital libraries. Full-text search. Google search engine. Information retrieval.

Received: 2 December 2022. Reviewed: 11 December 2022. Accepted: 20 January 2023.

1. Introduction

1st of October 2022 marked the twentieth anniversary of establishing the Greater Poland Digital Library (Pl. *Wielkopolska Biblioteka Cyfrowa*). It was the first digital library (DL) to employ the dLibra software, developed by the Poznan Supercomputing and Networking Center (Pl. *Poznańskie Centrum Superkomputerowo-Sieciowe* – PSNC). Its dissemination in the following years allowed for a dynamic

growth of similar projects all over the country (Kolasa, 2007). Over the last two decades, the number of institutions hosting or co-hosting DLs grew, as did the number of the objects they made accessible. The user base increased, too. A significant part of the users – c. 50% (Parkoła et al., 2016, 5) – are academics, who use the resources of DLs for their research. These are mostly historians, linguists, and scholars pursuing literary and culture studies. They were quick to appreciate the convenience of remote access ensured by the DLs, allowing them to use the resources at any time and place.

With time, the traditional DLs (based on library resources) hosted by the dLibra platform were joined by the digital resources of museums and multimedia archives, as well as scientific repositories. The repositories were created to share current scientific resources, unlike the digital libraries, museums and archives, which were established in order to protect and promote the premodern artefacts collected by the scientific and cultural institutions (Bednarek-Michalska, 2017, 47). Accordingly, the repositories' collections are dominated by objects which were 'born digital,' while the libraries, museums and archives focus on the digitized images of stored objects.

Digitized collections of various Polish institutions are scattered in more than one hundred and fifty DLs of various types (as per data collected in November 2022)[1]. The resources of these DLs might be browsed simultaneously thanks to the Federation of Digital Libraries aggregator (Pl. *Federacja Bibliotek Cyfrowych* – FBC, fbc.pionier.net.pl). Established in 2007, the aggregator allows the user to browse the resources hosted on the dLibra platform, or by the means of another type of software system supporting the OAI-PMH protocol (Open Archives Initiative Protoctol for Metadata Harvesting). Using the protocol, the database systematically harvests the metadata describing the objects from the DLs, and creates a single summary nationwide collection, which might be browsed on the FBC website (Lewandowska et al., 2007).

Over the last fifteen years, the FBC search engine has undergone many changes; the most apparent shift occurred in 2015, when the engine's interface was completely reconstructed, allowing for faceted filtering of the results. The current interface is clear and user-friendly, supporting efficient research. The only significant limitation of the database is the absence of the full-text search option. Because the system aggregates only the metadata, rather than the objects' content, full-text search is impossible.

[1] Digital libraries, digital museums, digital archives, scientific repositories. All these resources are often referred to as digital libraries, although it is not semantically correct. The practice is partly due to the name of the FBC platform, which was established at a time when the only digital resources were the resources of libraries, and partly due to the lack of a better alternative, which would encompass the digital resources of other types of institutions. This article will continue to use this simplified term, excluding the analysis which will consider only one type of digital libraries.

In order to search the content of the files stored in the scattered digital libraries, it is necessary to conduct separate searches on the websites of specific digital libraries. It is a time-consuming and challenging process. Given that the full-text search of files hosted by DLs is widely recognized as one of the websites' most useful functions (Parkola et al, 2016, 8), we should consider the reality of conducting such searches and potential alternatives.

The aim of the article is to analyze the conditions and possibilities of full-text search in Polish digital libraries (DL), taking into account the availability of full-text search in individual DL, the file formats and the software used, as well as the visibility of the DL resources in the Google search engine.

2. Selection of the DLs for the study

For the purposes of this study, it was necessary to make a selection out of over one hundred and fifty DLs registered in the FBC database. The chief criterion was the collection's emphasis on text files (a prerequisite for the full-text search). The FBC's list of sources is dominated by the traditional DLs which digitized their collections. These constitute over a half of the sources available via the FBC. DLs of this type were selected for analysis not only because of their uniform character, allowing for a comparative analysis, but primarily because of the issues with the full-text search that have been consistently obstructing access to the content of the files since the incipience of the DLs (Pulikowski, 2009).

Purely scientific resources, gathered in repositories and on the websites of scientific journals and publishers, are hosted on separate aggregators, such as Google Scholar, which allows its users to search not only the descriptions of the files, but also their contents. The functionality encompasses not only those publications which were created digitally, but also those that have been digitized, with the OCR layer added. The DLs created on the basis of traditional library resources contain files of various theme and origin, which by definition renders impossible the project of indexing them in their entirety on Google Scholar. It is viable only for the scientific sources (see chapter 5).

Taking into account these circumstances, the DLs selected from the FBC register for the purposes of this study do not include repositories, platforms of scientific journals and publishers, as well as digital museums and multimedia archives focusing on non-textual objects[2]. The analysis was conducted on the CSV file downloaded from the FBC website. A further selection was necessary to meet

[2] With the exception of the Multimedia Library of the NN Theatre, which contains a rich collection of text publications. The isolated "Library" collection comprising books, journals, articles, studies, etc., amounts to over 26 000 objects.

the time constraints, and to yield reliable data. To that end, the DLs were sorted according to the number of available objects in a descending order. It was assumed that analyzing the largest DLs will be the best solution, as such an analysis will cover the largest summary number of objects. The preliminary cut-off mark was set at 10 000 objects.

The FBC aggregator, collecting data concerning DLs nationwide, seemed like the perfect tool for selecting the DLs. It served its purpose as a vast register, but its data were revealed to be far from accurate. The disparities between the number of objects noted by the FBC and the websites of specific DLs came up to thousands (occasionally – ten thousands), both over and below. The only solution was to collect the data concerning the size of the collections from the DLs' websites over the course of a single day, to ensure that the selection would be valid. The analysis was conducted on November 12th, 2022. All DLs available in the FBC register were examined as otherwise it would not have been possible to determine which DLs meet the accepted criteria regarding the size of the collection. During the process of data collection, it turned out that some of the links in the FBC are expired, or inaccessible. It was doubtful that the register was complete; therefore, the author supplemented his data with those in an extensive DL register developed by Barbara Morawiec, hosted on the Lustro Biblioteki (En. *The Library Mirror*) portal (Morawiec, 2021). As a result, the final list encompassed 69 DLs collecting at least 1000 objects, among which 40 DLs collected as many as 10 000 objects, thus meeting the criteria for analysis. The decision to set the cut-off mark at 10 000 proved sound, as the next collection was significantly smaller (7400 objects, Krośniewska DL).

Definite majority of the 40 largest Polish DLs is hosted on the dLibra platform. Only three DLs use their own software (Polona, CRISPA, Polonijna DL). There are three versions of dLibra: 4.x – used by two DLs; 5.x – five DLs; 6.x – thirty-three DLs. The newest version, 6x, is predominant. It was introduced in 2016, and significantly improved on the previous versions (Parkoła et al., 2016, 10–20). Despite its high price, not only the largest DLs in the country chose to purchase the software. When confirming the real number of objects in individual DL, the author also identified the software used in the remaining 29 DLs with collections numbering more than 1000 and less than 10 000 objects. Ten DLs use dLibra 6.x; twelve – 5.x; three – 4.x. Four DLs use alternative software[3]. The information concerning the number of objects and the software used in the DLs selected for further analysis is presented in Table 1.

[3] Wolne Lektury (Free Schooltexts) – own software; DL of the Siedlce University of Natural Sciences and Humanities – Dspace; Digital Statistical Library – Aleph; DL of Sieradz Land – Sowa. The last two cases use integrated modules of library systems adjusted for the needs of DLs.

Tab. 1. The number of objects and the software used in the DLs selected for further analysis

No.	DL name	Number of objects	Software
1.	Polona	3720884	own
2.	Jagiellonian Digital Library | Jagiellońska Biblioteka Cyfrowa	854476	6.3.14
3.	Silesian Digital Library | Śląska Biblioteka Cyfrowa	528369	6.2.11
4.	CRISPA University of Warsaw Digital Library | CRISPA Biblioteka Cyfrowa Uniwersytetu Warszawskiego	464576	own
5.	Greater Poland Digital Library | Wielkopolska Biblioteka Cyfrowa	396783	6.3.13
6.	Kujawsko-Pomorska Digital Library | Kujawsko-PomorskaBiblioteka Cyfrowa	252073	6.3.16
7.	Małopolska Digital Library | Małopolska Biblioteka Cyfrowa	130535	5.8.5
8.	NN Theatre Multimedia Library | Biblioteka Multimedialna Teatru NN	127861	6.2.14
9.	Digital Library of University of Wrocław | Biblioteka Cyfrowa Uniwersytetu Wrocławskiego	119472	6.3.15
10.	Regional Materials of Łódź Land | Regionalia Ziemi Łódzkiej	113486	6.2.9
11.	Pomeranian Digital Library | Pomorska Biblioteka Cyfrowa	106691	6.3.16
12.	Lower Silesian Digital Library |Dolnośląska Biblioteka Cyfrowa	92701	6.3.15
13.	Digital Library of the University of Lodz | Biblioteka Cyfrowa Uniwersytetu Łódzkiego	90702	6.2.14
14.	Mazovian Digital Library | Mazowiecka Biblioteka Cyfrowa	85727	5.8.4
15.	Baltic Digital Library | Bałtycka Biblioteka Cyfrowa	72439	6.2.14
16.	Digital Library of the Silesian University of Technology | Biblioteka Cyfrowa Politechniki Śląskiej	70712	6.3.16
17.	Elbląg Digital Library | Elbląska Biblioteka Cyfrowa	67071	6.3.15
18.	Digital Library KUL | Biblioteka Cyfrowa KUL	66292	6.2.8
19.	Podlaska Digital Library | Podlaska Biblioteka Cyfrowa	58423	6.3.13
20.	KARTA Center Digital Library | Biblioteka Cyfrowa Ośrodka KARTA	54542	6.3.15
21.	Świętokrzyska Digital Library | Świętokrzyska Biblioteka Cyfrowa	49816	6.0.2
22.	The West Pomeranian Digital Library 'Pomerania' | Zachodniopomorska Biblioteka Cyfrowa	48421	6.2.12
23.	Lower Silesian Digital Library | Cyfrowy Dolny Śląsk	47200	6.2.14
24.	Digital Library of Zielona Góra | Zielonogórska Biblioteka Cyfrowa	47016	6.2.12
25.	UMCS Digital Library | Biblioteka Cyfrowa UMCS	44991	6.2.13
26.	Radom Digital Library | Radomska Biblioteka Cyfrowa	44659	6.2.13

No.	DL name	Number of objects	Software
27.	Digital Library of Provincial Public Library in Lublin | Biblioteka Cyfrowa Wojewódzkiej Biblioteki Publicznej w Lublinie	28394	5.8.3
28.	Podkarpacka Digital Library | Podkarpacka Biblioteka Cyfrowa	24131	6.1.3
29.	Lodz Regional Digital Library | Łódzka Regionalna Biblioteka Cyfrowa	20914	6.2.11
30.	Polonia Digital Library | Polonijna Biblioteka Cyfrowa	20795	own
31.	Opole Digital Library | Opolska Biblioteka Cyfrowa	19754	6.2.9
32.	Bialska Digital Library | Bialska Biblioteka Cyfrowa	17996	6.2.11
33.	Digital Library in Leszno | Leszczyńska Biblioteka Cyfrowa	17902	6.3.13
34.	Warmia-Mazury Digital Library | Warmińsko-Mazurska Biblioteka Cyfrowa	13727	6.2.11
35.	Digital Library of Lublin University of Technology | Biblioteka Cyfrowa Politechniki Lubelskiej	13473	6.0.2
36.	Wejherowo Digital Library | Wejherowska Biblioteka Cyfrowa	13440	4.0
37.	Military Digital Library | Wojskowa Biblioteka Cyfrowa	13365	5.8.4
38.	Digital Library of Warsaw University of Technology | Biblioteka Cyfrowa Politechniki Warszawskiej	12587	6.3.16
39.	Chelm Digital Library | Chełmska Biblioteka Cyfrowa	11865	4.0
40.	Digital Library of Inowroclaw | Inowrocławska Biblioteka Cyfrowa	10114	5.8.2
	Total	7994375	

3. The conditions for full-text search in DLs

Full-text search requires a text. This obvious statement becomes more meaningful when we consider the files created during digitization, which constitute a major part of the traditional DLs' collections, analyzed in this paper. Only the files processed by OCR (Optical Character Recognition) might be subject to a full-text search. The recognized text is usually inscribed as an invisible layer 'beneath' the image of a given page, where the recognized text matches the text visible on the image. This allows for highlighting the searched text and copying it, as it is the case with the documents created digitally.

It is difficult to determine what part of the DL collections is constituted by files with the OCR layer. This information is not a part of the available metadata. Considering that OCR has been employed in large DLs since they were established, we might assume that it features in the majority of their files. This might not be the case for the DLs established by smaller institutions, which might not have the software required for creating the OCR layer. The largest Polish DL – Polona,

which employs its own software for managing its DL, additionally marks the OCR files, which allows the user to narrow the scope of query in advanced search ("Search only objects with a text layer available"), and filtering the results ("text layer available"). Thanks to the marking, it is easy to determine that two thirds of texts available online in Polona have the OCR layer.

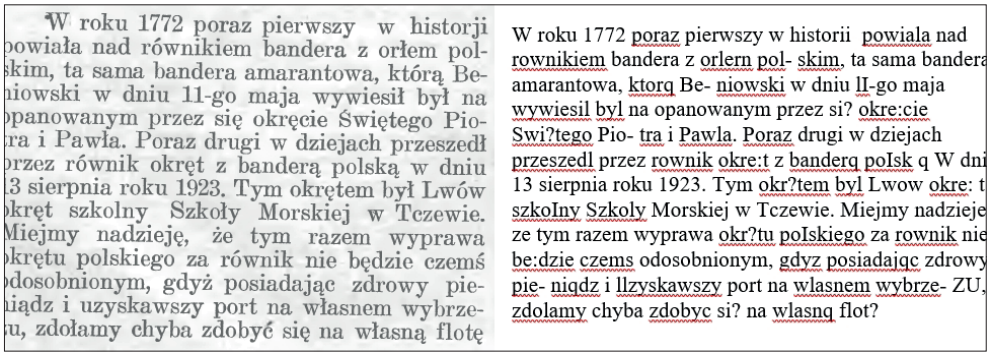
Scans/images with a hidden OCR layer are saved as PDF or DjVu files. The PDF is considered to be a standard file for saving information; it is widely recognized and supported by a variety of software and systems. The opposite is the case for DjVu which, despite its many advantages, remains an obscure file format and requires dedicated software. Many users encountered it for the first time in a DL. For many it was a cause for frustration, as they could not open the file and were obliged to install additional extensions for their internet browsers. Furthermore, it was soon impossible to use the extensions, as the browser security updates disabled them. To counteract these problems, in 2015 dLibra introduced a programming library for its version 5, which allowed the users to open the DjVu files without installing any further extensions (Parkoła et al., 2016, 14). The library is also available in the version 6 of dLibra. But the users of the older version 4 still have to face the challenge of opening the DjVu files.

The DjVu file format has not been updated in a long time, and presently is recommended only for very large documents. It has been widely used during the initial stages of digitization in Polish DLs, when its compression capacity was significantly higher than that of a PDF. Even ten years ago, DjVu files constituted almost 80% of all sources in Polish DLs (Szafrański, 2013, 8). It is difficult to determine what part of all sources they constitute now, but according to the author's calculations, there are at least 1.5 million DjVu files, which would suggest that they will remain in use at Polish DLs[4].

The full-text search depends not only on the presence of the OCR layer, but also on its efficacy. The accuracy of text recognition in scans/images of documents depends on many factors: the quality of scans/images (resolution, colour depth, dynamic range of the scanner/camera), the condition of the document (stains, discoloration, deformation), the type of paper and ink used (print contrast, bleed-through from the other side of the page), the text (font, language) and the OCR software used (Kotyńska, 2013, 8–9).

The software is a key factor here, as inappropriate selection or inappropriate use might result in errors depicted in Figure 1. Unfortunately, there are many documents in Polish DLs with weak OCR, which means that some part of the text remains unsearchable. The problem affects primarily the files in the DjVu format, created during the digitizing boom in the first decade of the 21st century.

[4] Calculations were based on advanced search for 40 DLs studied according to format.



W roku 1772 poraz pierwszy w historji powiała nad równikiem bandera z orłem polskim, ta sama bandera amarantowa, którą Beniowski w dniu 11-go maja wywiesił był na opanowanym przez się okręcie Świętego Piotra i Pawła. Poraz drugi w dziejach przeszedł przez równik okręt z banderą polską w dniu 13 sierpnia roku 1923. Tym okrętem był Lwów okręt szkolny Szkoły Morskiej w Tczewie. Miejmy nadzieję, że tym razem wyprawa okrętu polskiego za równik nie będzie czemś odosobnionym, gdyż posiadając zdrowy piędz i uzyskawszy port na własnem wybrzeżu, zdołamy chyba zdobyć się na własną flotę

W roku 1772 poraz pierwszy w historii powiala nad rownikiem bandera z orlern pol- skim, ta sama bandera amarantowa, ktorq Be- niowski w dniu lI-go maja wywiesil byl na opanowanym przez si? okre:cie Swi?tego Pio- tra i Pawla. Poraz drugi w dziejach przeszedl przez rownik okre:t z banderq poIsk q W dni 13 sierpnia roku 1923. Tym okr?tem byl Lwow okre: t szkoIny Szkoly Morskiej w Tczewie. Miejmy nadzieje ze tym razem wyprawa okr?tu poIskiego za rownik nie be:dzie czems odosobnionym, gdyz posiadajqc zdrowy pie- niqdz i llzyskawszy port na wlasnem wybrze- ZU, zdolamy chyba zdobyc si? na wlasnq flot?

Fig. 1. Problems with text recognition in digitized documents.
Source: own transcription based on: https://kpbc.umk.pl/dlibra/publication/87307/edition/92270/content

As it was already mentioned, OCR is not available for all digitized files. While in some cases it is a consequence of lacking OCR software, in other cases text recognition simply does not make sense, particularly for manuscripts, incunabula, maps, and other documents for which OCR accuracy does not reach higher than 50% (Szala, 2015, 267). We might also find files in the PDF or the DjVu format used for storing non-textual objects, primarily photographs. The DjVu file will always have a graphic layer, while the PDF file might not. The PDF is also the format for storing files created digitally. These texts may always be indexed for the purposes of full-text search. Files created digitally might be accessed in DLs in other formats, such as TXT, HTML, EPUB, or MOBI. These might be equally easily searched. Increasingly often, the user finds a single publication in several formats, which is possible thanks to the flexibility of modern DL platforms.

4. The possibilities for full-text search in DLs

As it is impossible to conduct a full-text search of scattered DL resources from a single access point, the users must rely on search engines of individual DLs. It is most likely that they primarily use the largest Polish DL – Polona, the digital library of the National Library in Poland. The regional DLs, and the DLs of the institutions where the users are employed, might be equally useful. Nevertheless, if a query concerning the content of the files is to be exhaustive, then it will be necessary to repeat it in all DLs. Fortunately, majority of DLs is hosted on the dLibra platform, which means that the users will interact with only two interfaces – version 6.x, and the earlier 5.x and 4.x.

All analyzed DLs, presented in Table 1, allow for full-text search. The functionality was tested on single terms and phrases. Regardless of the software used, there were

no serious issues with searching the text of the files. The analysis was conducted between 11th and 14th of November 2022. The remaining part of this chapter discusses the author's experiences using individual searching interfaces.

On the Polona website, the option of full-text search is marked on the right side of the search bar as a box described "Search the content". After selecting this option and entering the query, detailed results load (it usually takes a moment) – see Figure 2. The users might additionally apply filters listed on the left side, and select another method of sorting the results. Under the description of each file, which takes up a large part of the screen, the user will find the fragment of the sentence where the searched phrase occurs, highlighted in blue. The user might see further matches clicking "See more" on the right side. When the query is a full phrase, rather than a single term, one should apply quotation marks; nevertheless, the suggested matches will include sentences where only one element of the phrase will appear. It is particularly bothersome where prepositions are concerned, for example in the phrase "w pieskowej skale" (in Pieskowa Skala Castle), where "w" (in) will be considered as a significant term and generate dozens, if not hundreds of results. What is convenient is the transposition of the information concerning found content to the full text view. The advanced search allows for generating more complex queries, which use Boolean operators. The Polona website does not have a mobile version and requires manual scaling. Since November 2022, mobile users may use the mPolona app, optimized for phones. Unfortunately, for the time being full-text search in the app yields an error page instead of a list of records (state in December 2022).

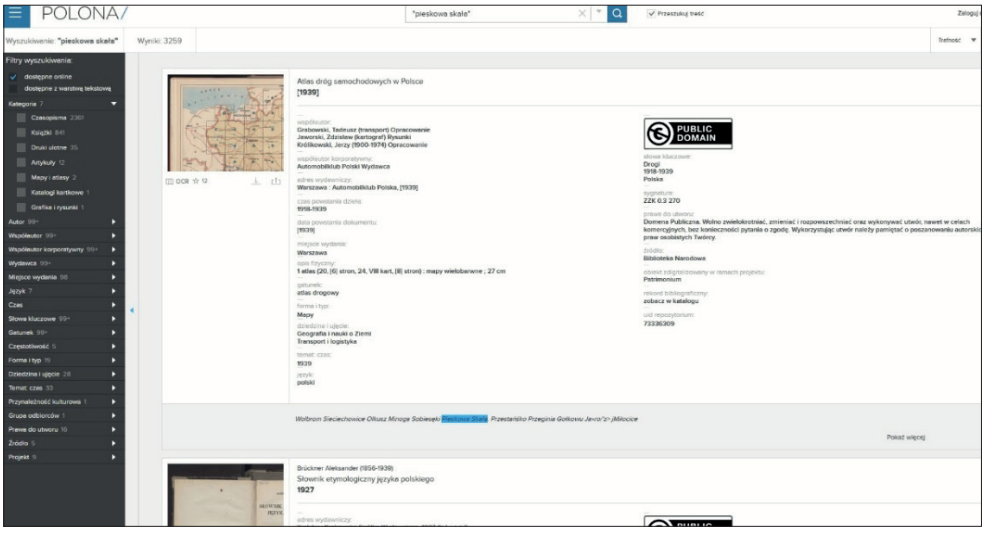


Fig 2. Search results in Polona.
Source: polona.pl

The 6.x dLibra, predominant in the largest Polish DLs (33 out of 40) provides the full-text search option in the advanced search. The users must select the "Search content" field. After entering the phrase, there appears (after a short while) a list of results, identical to the results yielded by searches of other types, allowing for dynamic filtering, sorting, etc. However, the results do not display the fragments of texts where the searched phrase appear. The user has to open the full text, and search the phrase again. The search is effective both for single terms and for whole phrases. There is no issue with searching the content of DjVu files. Searching further occurrences of a given term, we move to further pages of the file. The search engine automatically searches different grammatical forms of the searched terms and phrases. This should be borne in mind when searching the content of the files – it will be more effective to type in the root of the word. Constructing the query, the user may apply many Boolean operators expanding the possibilities of full-text search. The possible operators are suggested after clicking the question mark next to the search bar. The mobile version of the dLibra 6.x is well-matched to the size of the device and does not hinder the search.

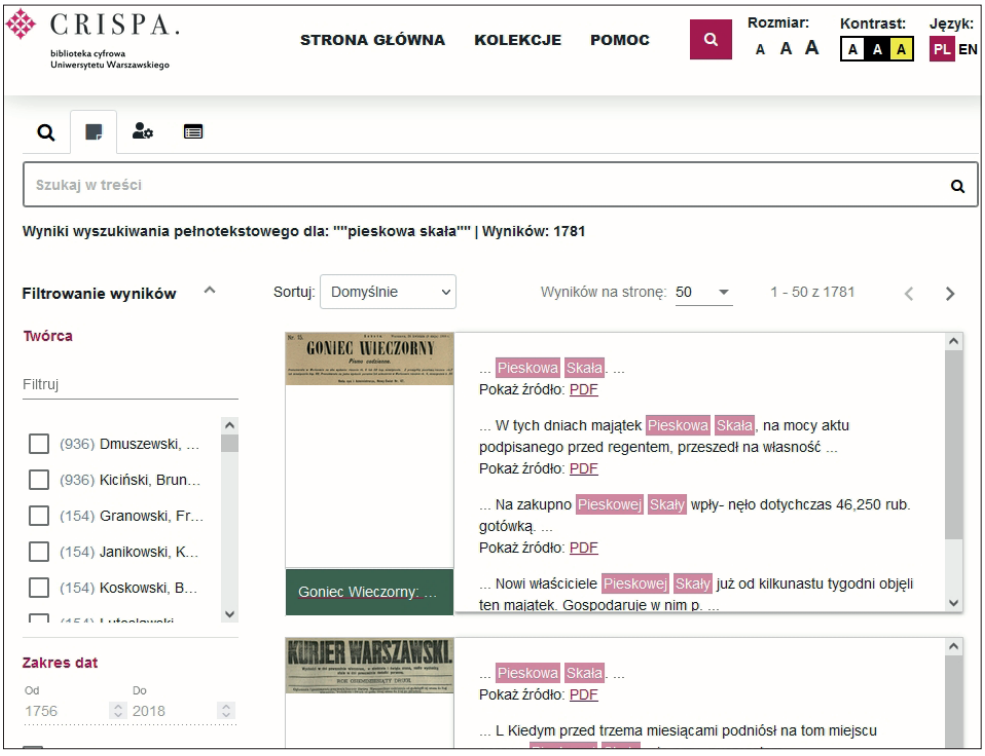


Fig 3. CRISPA search results.
Source: crispa.uw.edu.pl

The full-text search option in the former versions of dLibra is not much different. A slightly outdated interface allows for conducting analogous searches (without the option to filter the results). However, there might be problems with searching the content of DjVu files; in the fourth version, these files might fail to open. Curiously enough, the results in the 4.x version include an accuracy coefficient which has not been included in the later versions. Unlike the sixth version of dLibra, the earlier versions are not optimized for mobile use.

When we consider the largest DLs in the country, we should note the DL of the University of Warsaw, CRISPA, which uses its own software. The full-text search is conducted in a separate page, selected above the search bar. It takes time to generate the results, but they are well worth the wait. The results (Fig. 3), next to the cover of each document, show a nested text box, containing the fragments of the text with the search terms highlighted. This form of presenting the results makes it much easier to select files for further analysis. Unfortunately, the searched fragments are not transposed onto the full text view, as in Polona. It is necessary to search the text of each opened file. CRISPA takes into account the grammatical variations of each term, allows the user to use quotation marks to search by phrases, but unfortunately does not recognize other Boolean operators. The mobile version is scaled and fully operational.

5. Full-text search in DLs by Google

As conducting a full-text search in DLs is very time-consuming, we might consider the alternative offered by the Google search engine. Thanks to the modifiers *site:* and *filetype:*, it is easy to determine to what extent Google indexes the DL resources. The *site:* modifier limits the search to a given domain; the *filetype:* modifier limits the results to a selected format, here, the PDF. Selecting a single format is necessary to avoid duplicated results – sites describing the files, and the files themselves. It is impossible to conduct an analogous search for the DjVu files as Google does not recognize the format.

The original plan was to compare the results of a Google search with the number of the PDF files in selected DLs. However, it turned out that it is impossible to determine the data credibly for such a high number of DLs. Furthermore, the author was forced to conclude that the results of the Google search might be considered only as estimations, as the numbers change every few days. Despite these limitations, it was possible to establish an approximate number of PDF files that Google "sees" in the DLs. It was also possible to confirm the presence of these documents in Google Scholar (GS).

The search was conducted on November 16th, 2022, and repeated after a week in order to average the results. The search term was as follows: "site:the DL-domain filetype:pdf".

For the search conducted in GS, the term only included site: as GS groups duplicate results together. The results are presented in Table 2.

Tab. 2. The presence of Polish DL resources in Google and Google Scholar

No.	DL name	Google PDF 1	Google PDF 2	Google PDF avg.	GS 1	GS 2
1.	Polona	8	8	8	0	0
2.	Jagiellonian Digital Library | Jagiellońska Biblioteka Cyfrowa	17300	16400	16850	749	751
3.	Silesian Digital Library | Śląska Biblioteka Cyfrowa	16200	16300	16250	0	0
4.	CRISPA University of Warsaw Digital Library | CRISPA Biblioteka Cyfrowa Uniwersytetu Warszawskiego	18	17	18	0	0
5.	Greater Poland Digital Library | Wielkopolska Biblioteka Cyfrowa	10700	11800	11250	2290	2290
6.	Kujawsko-Pomorska Digital Library | Kujawsko-Pomorska Biblioteka Cyfrowa	17100	18100	17600	391	390
7.	Małopolska Digital Library | Małopolska Biblioteka Cyfrowa	4870	4010	4440	0	0
8.	NN Theatre Multimedia Library | Biblioteka Multimedialna Teatru NN	7110	6770	6940	185	186
9.	Digital Library of University of Wrocław | Biblioteka Cyfrowa Uniwersytetu Wrocławskiego	15600	14900	15250	584	593
10.	Regional Materials of Łódź Land | Regionalia Ziemi Łódzkiej	4850	4360	4605	17	17
11.	Pomeranian Digital Library | Pomorska Biblioteka Cyfrowa	7920	7910	7915	434	434
12.	Lower Silesian Digital Library | Dolnośląska Biblioteka Cyfrowa	11900	12800	12350	3270	3310
13.	Digital Library of the University of Lodz | Biblioteka Cyfrowa Uniwersytetu Łódzkiego	32	47	40	14	14
14.	Mazovian Digital Library | Mazowiecka Biblioteka Cyfrowa	9250	9750	9500	38	35
15.	Baltic Digital Library | Bałtycka Biblioteka Cyfrowa	8330	7830	8080	35	35
16.	Digital Library of the Silesian University of Technology | Biblioteka Cyfrowa Politechniki Śląskiej	19400	19300	19350	3530	3520

No.	DL name	Google PDF 1	Google PDF 2	Google PDF avg.	GS 1	GS 2
17.	Elbląg Digital Library | Elbląska Biblioteka Cyfrowa	1240	1320	1280	60	60
18.	Digital Library KUL | Biblioteka Cyfrowa KUL	8350	8610	8480	441	441
19.	Podlaska Digital Library | Podlaska Biblioteka Cyfrowa	8540	9080	8810	539	537
20.	KARTA Center Digital Library | Biblioteka Cyfrowa Ośrodka KARTA	23	24	24	0	0
21.	Świętokrzyska Digital Library | Świętokrzyska Biblioteka Cyfrowa	3560	3900	3730	4	4
22.	The West Pomeranian Digital Library 'Pomerania' | Zachodniopomorska Biblioteka Cyfrowa	865	718	792	123	123
23.	Lower Silesian Digital Library | Cyfrowy Dolny Śląsk	3800	3900	3850	33	33
24.	Digital Library of Zielona Góra | Zielonogórska Biblioteka Cyfrowa	4590	4580	4585	33	2820
25.	UMCS Digital Library | Biblioteka Cyfrowa UMCS	5110	5080	5095	1290	1290
26.	Radom Digital Library | Radomska Biblioteka Cyfrowa	774	750	762	15	15
27.	Digital Library of Provincial Public Library in Lublin | Biblioteka Cyfrowa Wojewódzkiej Biblioteki Publicznej w Lublinie	768	877	823	0	0
28.	Podkarpacka Digital Library | Podkarpacka Biblioteka Cyfrowa	688	817	753	9	9
29.	Lodz Regional Digital Library | Łódzka Regionalna Biblioteka Cyfrowa	5120	4940	5030	546	546
30.	Polonia Digital Library | Polonijna Biblioteka Cyfrowa	4140	3980	4060	8	8
31.	Opole Digital Library | Opolska Biblioteka Cyfrowa	1610	1690	1650	47	47
32.	Bialska Digital Library | Bialska Biblioteka Cyfrowa	792	834	813	12	12
33.	Digital Library in Leszno | Leszczyńska Biblioteka Cyfrowa	25	25	25	0	0
34.	Warmia-Mazury Digital Library | Warmińsko-Mazurska Biblioteka Cyfrowa	2360	2300	2330	407	407

No.	DL name	Google PDF 1	Google PDF 2	Google PDF avg.	GS 1	GS 2
35.	Digital Library of Lublin University of Technology | Biblioteka Cyfrowa Politechniki Lubelskiej	10500	10900	10700	752	752
36.	Wejherowo Digital Library | Wejherowska Biblioteka Cyfrowa	0	0	0	0	0
37.	Military Digital Library | Wojskowa Biblioteka Cyfrowa	6030	6010	6020	0	0
38.	Digital Library of Warsaw University of Technology | Biblioteka Cyfrowa Politechniki Warszawskiej	10500	10500	10500	351	351
39.	Chelm Digital Library | Chełmska Biblioteka Cyfrowa	82	98	90	9	9
40.	Digital Library of Inowroclaw | Inowrocławska Biblioteka Cyfrowa	634	678	656	1	1

The column "Google PDF 1" contains the results of the first search; "Google PDF 2" – the search repeated after a week. The average values are presented in the third column. The disparities between the results in the first and second column reach thousands, both over and below. Nevertheless, it is clear which DLs are less visible in Google. It is certainly surprising that both Polona and CRISPA are practically invisible. The DLs of the technical universities are very well represented – despite the relatively small size of their collections, they have over 10 thousand documents filed in Google; the library of the Silesian University of Technology has the highest result of all DLs.

The results of a search conducted in Google Scholar are much more precise. The disparities between the results presented in columns "GS 1" and "GS 2" are minimal. The visibility of DL resources in GS is more varied. More DLs are less visible. It should not come as a surprise as GS prioritizes repositories over DLs. Nevertheless, some DL resources do appear in GS. The libraries with the highest number of files in GS were the Greater Poland Digital Library, the Lower Silesian Digital Library, the Digital Library of the Silesian University of Technology and the UMCS Digital Library.

The data in Table 2 concerning the visibility of the DL resources in Google shows that they are somewhat visible. But this is not enough to consider Google as a tool for browsing the DL resources, particularly where full-text search is concerned. A point of reference is required. According to the author's calculations, the number of PDF files in all DLs studied is not lower than a million.[5] For reference, the

[5] As in the case of the DjVu files, the calculations were based on the advanced search conducted for 40 DLs considered in the article, using the format field.

average number of results of a Google search (presented in Table 2) comes up to c. 230 thousand total. This would suggest that Google indexes no more than 23% of all DL resources. These are obviously mere estimations, but they indicate the scale of the disparity between the number of available files and the results of the Google search.

To achieve a more precise assessment of Google's capacity to enable a full-text search of the DL resources, additional analysis was conducted. It involved a direct comparison of the results yielded by searching the phrase "pieskowa skała" in DLs (content search) and in Google (limiting the results to the DLs' domains). The analysis focused on the ten largest DLs, not counting Polona and CRISPA, whose resources are unindexed. The results of the comparison are presented in Table 3.

Tab. 3. The comparison of the results of full-text search conducted in DLs and in Google

No.	DL name	DL	Google	% Google
1.	Jagiellonian Digital Library | Jagiellońska Biblioteka Cyfrowa	1074	22	2
2.	Silesian Digital Library | Śląska Biblioteka Cyfrowa	905	64	7
3.	Greater Poland Digital Library | Wielkopolska Biblioteka Cyfrowa	373	12	3
4.	Kujawsko-Pomorska Digital Library | Kujawsko-Pomorska Biblioteka Cyfrowa	189	16	8
5.	Małopolska Digital Library | Małopolska Biblioteka Cyfrowa	291	27	9
6.	NN Theatre Multimedia Library | Biblioteka Multimedialna Teatru NN	13	5	38
7.	Digital Library of University of Wrocław | Biblioteka Cyfrowa Uniwersytetu Wrocławskiego	26	2	8
8.	Regional Materials of Łódź Land | Regionalia Ziemi Łódzkiej	143	5	3
9.	Pomeranian Digital Library | Pomorska Biblioteka Cyfrowa	85	8	9
10.	Lower Silesian Digital Library | Dolnośląska Biblioteka Cyfrowa	27	5	19
	Total	3126	166	5

There is a vast difference between the results yielded by DLs and by Google. It is partially explained by the fact that Google does not index the DjVu files (Lewandowski, 2014). Google retrieved only 5% of the documents found in the DL resources. To confirm these results, other queries were tested. The disparity is so

large that it rules out Google as a viable tool for accessing the DL resources. At least for the time being, the tools offered by particular DLs are the best method of accessing their resources.

6. Conclusions

More than anything, this analysis of the conditions and realities of full-text search in Polish DLs shows how much remains to be done in order to improve the users' access to the content of the files collected by the DLs. The most serious issue is the lack of a single point of access which would allow for conducting a full-text search of files scattered in various libraries using a unified index. Google cannot fill this gap – Google Custom Search might have served this purpose if Google were capable of indexing a larger part of the DL resources. The situation might be improved if the software used by DLs is optimized for the visibility of the files in Google. The disparities in visibility of resources collected by particular DLs are apparent when we compare Polona and CRISPA with the digital libraries of Polish technical universities. Equally crucial for indexing the DL files in Google will be replacing the DjVu files with PDF files, or uploading the PDF files alongside DjVus. The quality of OCR layer in the documents created digitally will be important, too, as it will allow for accurate indexing by Google and by the DL software. Many older files, particularly those in the DjVu format, should have their OCR layer corrected, or added, so that the users may access their content. We should not neglect that convenience is another important factor in browsing the DL resources. It depends on the efficacy of the system, and the operability of the software employed. We might observe significant changes for the better in this regard, but there is room for improvement.

References

Bednarek-Michalska, B. (2017). Polish Digital Libraries and Repositories. Origins, Operation and Usage. *Przegląd Biblioteczny*, 85 (spec. iss.), 46–69. https://doi.org/10.36702/pb.854

Kolasa, W. M. (2007). dLibra Digital Library Framework – platforma do budowy bibliotek cyfrowych. In: J. Woźniak-Kasperek & J. Franke (eds.), *Biblioteki cyfrowe : projekty, realizacje, technologie* (67–88). Warszawa: Wydaw. SBP. http://eprints.rclis.org/16562

Kotyńska, E. (2013). *Korekta OCR – problemy i rozwiązania* [online]. Biblioteka cyfrowa dziś a wyzwania jutra. Międzynarodowa konferencja naukowa. Biblioteka Jagiellońska, Kraków, 24–25 stycznia 2013 [19.01.2023], https://jbc.bj.uj.edu.pl/Content/218090

Lewandowska, A., Mazurek, C., Werla, M. (2007). *Federacja Bibliotek Cyfrowych w sieci PIONIER – Dostęp do otwartych bibliotek cyfrowych i repozytoriów* [online]. IV Ogólnopolska Konferencja EBIB Internet w bibliotekach. Open Access. Toruń, 7–8 grudnia

2007 [19.01.2023], http://www.ebib.pl/publikacje/matkonf/mat18/lewandowska_mazurek_werla.php
Lewandowski, T. (2014). *Google Scholar a repozytoria i biblioteki cyfrowe w Polsce* [online]. *Otwarta nauka.pl* [19.01.2023], https://otwartanauka.pl/analysis/case-studies/google-scholar-a-repozytoria-i-biblioteki-cyfrowe-w-polsce
Morawiec, B., M. (2021). *Biblioteki cyfrowe w Polsce* [online]. Lustro Biblioteki [20.12.2022], https://lustrobiblioteki.pl/biblioteki-cyfrowe-polsce
Parkoła, T., Bohdanowicz, K., Werla, M. (2016). Realizacja potrzeb użytkowników bibliotek cyfrowych na przykładzie systemu dLibra 6. *Biuletyn EBIB* [online], 8 (170), http://ebibojs.pl/index.php/ebib/article/view/168
Pulikowski, A. (2009). *Wyszukiwanie pełnotekstowe w zasobach bibliotek cyfrowych* [online]. X Forum Informacji Naukowej i Technicznej. Zakopane, 22–25 września 2009 [19.01.2023], https://sbc.org.pl/dlibra/publication/15686/edition/13873/content
Szala, M. (2015). Cyfrowe oblicza zbiorów. In: G. Piotrowicz (ed.) *Wykorzystanie nowoczesnych technologii i mediów cyfrowych w Bibliotece Uniwersyteckiej we Wrocławiu. Stan na rok 2015* (259–276). Wrocław: Biblioteka Uniwersytecka we Wrocławiu. https://www.bibliotekacyfrowa.pl/publication/75789
Szafrański, L. (2013). *Dokumenty cyfrowe w JBC – próba charakterystyki* [online]. Biblioteka cyfrowa dziś a wyzwania jutra. Międzynarodowa konferencja naukowa. Biblioteka Jagiellońska, Kraków, 24–25 stycznia 2013 [19.01.2023], https://www.jbc.bj.uj.edu.pl/dlibra/publication/229579/edition/218080

Wyszukiwanie pełnotekstowe w zasobach polskich bibliotek cyfrowych

Abstrakt

Cel/teza: Celem artykułu jest analiza uwarunkowań i możliwości wyszukiwania pełnotekstowego w zasobach polskich bibliotek cyfrowych (BC), uwzględniająca dostępność wyszukiwania w treści dokumentów w poszczególnych BC, wykorzystywane formaty plików i oprogramowanie, a także widoczność zasobów BC w wyszukiwarce Google.
Koncepcja/Metody badań: Do badań wytypowano 40 największych polskich BC, których zasoby stanowią głównie digitalizowane zbiory tradycyjnych bibliotek. Sprawdzono rodzaj wykorzystywanego oprogramowania, dostępność i możliwości wyszukiwania pełnotekstowego oraz stopień indeksowania zasobów w formacie PDF w wyszukiwarce Google i Google Scholar. Na koniec porównano wyniki wyszukiwania pełnotekstowego w dziesięciu BC z uzyskanymi w Google.
Wyniki/Wnioski: Wszystkie wybrane do badań BC posiadają możliwość wyszukiwania pełnotekstowego. Wyodrębnione interfejsy wyszukiwawcze różnią się znacznie między sobą. Każdy ma zalety i wady wymagające dalszych prac rozwojowych. Wyszukiwarka Google nie nadaje się obecnie do wykorzystania jako alternatywny sposób dostępu do treści udostępnianych w polskich BC.

Oryginalność/Wartość poznawcza: Problematyka wyszukiwania pełnotekstowego w BC jest rzadko podejmowana, mimo iż użytkownicy uznają wyszukiwanie w treści dokumentów za jedną z najbardziej przydatnych funkcji BC. Wyniki przeprowadzonych badań tworzą obraz aktualnych możliwości BC w tym zakresie.

Słowa kluczowe

Biblioteki cyfrowe. Widoczność zasobów. Wyszukiwanie informacji. Wyszukiwanie pełnotekstowe. Wyszukiwarka Google.

ARKADIUSZ PULIKOWSKI, Associate Professor at the Institute of Culture Studies at the Faculty of Humanities of University of Silesia in Katowice. Research interests: information search; information behavior; infometrics; digitization of information. Major recent publications: Modelowanie procesu wyszukiwania informacji naukowej. Strategie i interakcje *(2018),* Searching for LIS scholarly publications: a comparison of search results from Google, Google Scholar, EDS, and LISA *(Journal of Academic Librarianship, 2021, co-author: A.Matysek),* The Relation Between the Structure of Abstracts in LIS and Anthropology Journals and Their Rank *(Zagadnienia Informacji Naukowej, 2020).*

Contact to the Author
arkadiusz.pulikowski@us.edu.pl
Institute of Culture Studies,
Faculty of Humanities,
University of Silesia in Katowice
ul. Bankowa 11
40-007 Katowice, Poland