Janusz S. Bień
Department of Formal Linguistics, University of Warsaw
jsbien@uw.edu.pl

Facilitating access to digitalized dictionaries in DjVu format

September 21, 2009
May 6, 2010

Abstract

One of the best formats for scanned documents is DjVu. An essential feature of the format is the hidden text layer, usually containing the results of Optical Character Recognition. Another important feature is the ability to store (and serve over Internet) the documents as a collection of individual pages.

From the very beginning the DjVu format has been used also for dictionaries, in particular there are several Polish dictionaries available in this format. So the question is how to search efficiently the text layer in such large multi-volume works. For this purpose we intend in particular to adapt Poliqarp (*Polyinterpretation Indexing Query and Retrieval Procesor*), a GPLed corpus query tool developed in the Institute of Computer Science of Polish Academy of Sciences. Some preliminary experiments are described in the talk.

In our „quick and dirty" approach we treat every page as a single document with the metadata consisting of the name of the document index and the name of the file with the page content. For every word, instead of grammatical tags, we provide its localization on the page in the form of the line number and its position in the line. All the data taken together allow to link the search results to the appropriate fragments of the original scans.

We mention also another approach to the problem, exemplified by djvu-xfgrep program.

Keywords: digitalization, DjVu, dictionaries, Poliqarp, djvu-xfgrep

1. Digitalization

In the proper sense digitalization consist in representing an object by means of bits or numbers, but it can be done in many different ways. A page of text can be treated as a picture and divided into small points classified as black or white, so the page can be represented by an array of binary digits. Such points and their representations are called *pixels*, i.e. picture elements (*pix* was used as an abbreviation for *picture*, cf. http://en.wikipedia.org/wiki/Pixel or *Webster's New World Dictionary of the American Language*). Of course classifying all point as just black or white is often not sufficient, so there are also grayscale and color pixels in use.

Texts represented as pixels are usually the results of scanning printed texts, so they are called simply scanned texts or just scans (although the traditional scanners are more and more often replaced by various devices using digital photography). As such texts can be also digitally born, we advocate here a more precise term *pixel texts*. Such texts are often subject to the process of Optical Character Recognition (OCR). If the process is unattended and the results are neither verified nor corrected by humans, it is called „dirty OCR". There is no convenient term for the textual results of OCR, so we propose here to call them *textel texts*, as they consist of characters, which can be considered the primary 'textels', i.e. text elements (the term *textel* is sometimes used as meaning texture element, but for this notion the term *texel* seems more popular). A typical digitally born electronic text is also a textel text.

Both pixel and textel texts can be supplemented by various additional data, extracted automatically from the text content or added by human intervention. The simplest and most useful forms of such information are outlines describing the document structure and facilitating navigation in the electronic text, annotations providing additional information about text fragments and hyperlinks e.g. to footnotes etc. These features can be provided using a commonly used standard or with a special purpose program. Although the term *digitalization* can be still used, the latter case is better described as computerization of a text.

2. DjVu technology

To quote [4], DjVu technology is

> an image compression technique, a document format, and a software platform for delivering documents images over the Internet

It was developed by Yann Le Cun, Léon Bottou, Patrick Haffner, and Paul G. Howard at AT&T Laboratories in 1996. A document in DjVu format may consist of several layers: the pixel layer (actually split into foreground, background and stencil) representing the page images, and the textel one (called hidden text layer) which can be used for searching, cut and paste operations etc. For several years this feature was unique. In 2001 similar possibility was introduced by Adobe to the Portable Document Format (PDF) specification, but DjVu still has several advantages over PDF. The feature of DjVu, absent in PDF and most relevant for the present paper, is the possibility to store individual pages in separate files (so called *unbundled* or *indirect* documents) and to serve them over the Internet in any order. This is an important reason for preferring DjVu over PDF also for digitally born documents.

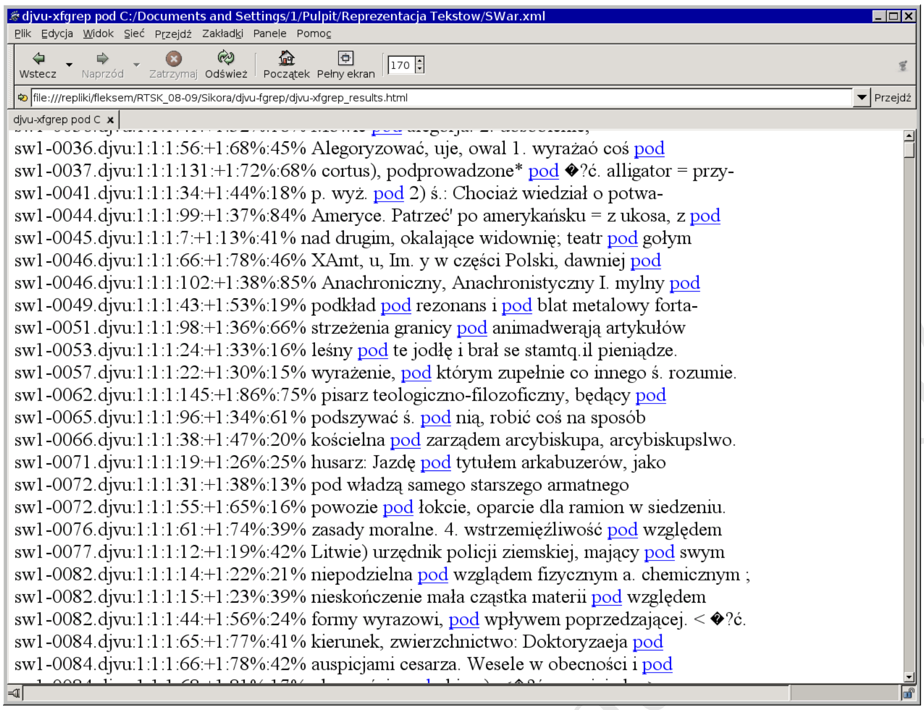
Several dictionaries available in the DjVu format have been mentioned already in [1] and [3]. One should note also *Jamieson's Etymological Dictionary of the Scottish Language Online* (`http://www.scotsdictionary.com/`) which is available both in DjVu and PDF format and is provided with sophisticated search facilities. There also interesting new acquisitions of Polish digital libraries, such as 19th century editions of Latin-Polish glosses by Bartłomiej from Bydgoszcz, first published in 1488. To locate all its volumes, visit the site of Federacja Bibliotek Cyfrowych (`http://fbc.pionier.net.pl`) and search for `Bartłomiej z Bydgoszczy`.

As of 10th September 2009, in Polish digital libraries (cf. e.g. [5]) there are 199 701 publications in DjVu format, which is 72% of the total (`http://fbc.pionier.net.pl/owoc/attr-stats`).

3. Dictionaries as texts

Every DjVu viewer allows for searching the hiddent text layer, but for unbundled documents it is unefficient as it defeats the purpose of splitting the document into separate pages: to access the hidden text, all the pages have to be loaded, and if the search is repeated, they are reloaded multiple times.

Several possible solutions has been outlined in my note [2] and tentatively implemented in Java by Piotr Sikora. Two programs, djvu-fgrep and djvu-xfgrep, are available at `http://code.assembla.com/djvu-fgrep/subversion/nodes` on GNU GPL license; for the present paper only the latter program is relevant (x in the name stands for *eXtended*). The programs are named after a very popular Unix utility fgrep, which is a simplified version of grep. The grep program performs the function of searching globally for regular expressions and printing (i.e. displaying) the matches; the first letters of words *global*, *regular expression*,


djvu-xfgrep pod C:/Documents and Settings/1/Pulpit/Reprezentacja Tekstow/SWar.xml
Plik Edycja Widok Sieć Przejdź Zakładki Panele Pomoc
Wstecz
Naprzód
Zatrzymaj
Odśwież
Początek
Pełny ekran
170
file:///repliki/fleksem/RTSK_08-09/Sikora/djvu-fgrep/djvu-xfgrep_results.html
Przejdź
djvu-xfgrep pod C

```
sw1-0036.djvu:1:1:1:56:+1:68%:45% Alegoryzować, uje, owal 1. wyrażaó coś pod
sw1-0037.djvu:1:1:1:131:+1:72%:68% cortus), podprowadzone* pod �?ć. alligator = przy-
sw1-0041.djvu:1:1:1:34:+1:44%:18% p. wyż. pod 2) ś.: Chociaż wiedział o potwa-
sw1-0044.djvu:1:1:1:99:+1:37%:84% Ameryce. Patrzeć' po amerykańsku = z ukosa, z pod
sw1-0045.djvu:1:1:1:7:+1:13%:41% nad drugim, okalające widownię; teatr pod gołym
sw1-0046.djvu:1:1:1:66:+1:78%:46% XAmt, u, Im. y w części Polski, dawniej pod
sw1-0046.djvu:1:1:1:102:+1:38%:85% Anachroniczny, Anachronistyczny I. mylny pod
sw1-0049.djvu:1:1:1:43:+1:53%:19% podkład pod rezonans i pod blat metalowy forta-
sw1-0051.djvu:1:1:1:98:+1:36%:66% strzeżenia granicy pod animadwerają artykułów
sw1-0053.djvu:1:1:1:24:+1:33%:16% leśny pod te jodłę i brał se stamtq.il pieniądze.
sw1-0057.djvu:1:1:1:22:+1:30%:15% wyrażenie, pod którym zupełnie co innego ś. rozumie.
sw1-0062.djvu:1:1:1:145:+1:86%:75% pisarz teologiczno-filozoficzny, będący pod
sw1-0065.djvu:1:1:1:96:+1:34%:61% podszywać ś. pod nią, robić coś na sposób
sw1-0066.djvu:1:1:1:38:+1:47%:20% kościelna pod zarządem arcybiskupa, arcybiskupslwo.
sw1-0071.djvu:1:1:1:19:+1:26%:25% husarz: Jazdę pod tytułem arkabuzerów, jako
sw1-0072.djvu:1:1:1:31:+1:38%:13% pod władzą samego starszego armatnego
sw1-0072.djvu:1:1:1:55:+1:65%:16% powozie pod łokcie, oparcie dla ramion w siedzeniu.
sw1-0076.djvu:1:1:1:61:+1:74%:39% zasady moralne. 4. wstrzemięźliwość pod względem
sw1-0077.djvu:1:1:1:12:+1:19%:42% Litwie) urzędnik policji ziemskiej, mający pod swym
sw1-0082.djvu:1:1:1:14:+1:22%:21% niepodzielna pod względem fizycznym a. chemicznym ;
sw1-0082.djvu:1:1:1:15:+1:23%:39% nieskończenie mała cząstka materii pod względem
sw1-0082.djvu:1:1:1:44:+1:56%:24% formy wyrazowi, pod wpływem poprzedzającej. < �?ć.
sw1-0084.djvu:1:1:1:65:+1:77%:41% kierunek, zwierzchnictwo: Doktoryzaeja pod
sw1-0084.djvu:1:1:1:66:+1:78%:42% auspicjami cesarza. Wesele w obecności i pod
```

Figure 1. Sample djvu-xfgrep output

print make up the name of the program. The letter *f* in fgrep means that this version of the program allows only to search for *fixed* strings.

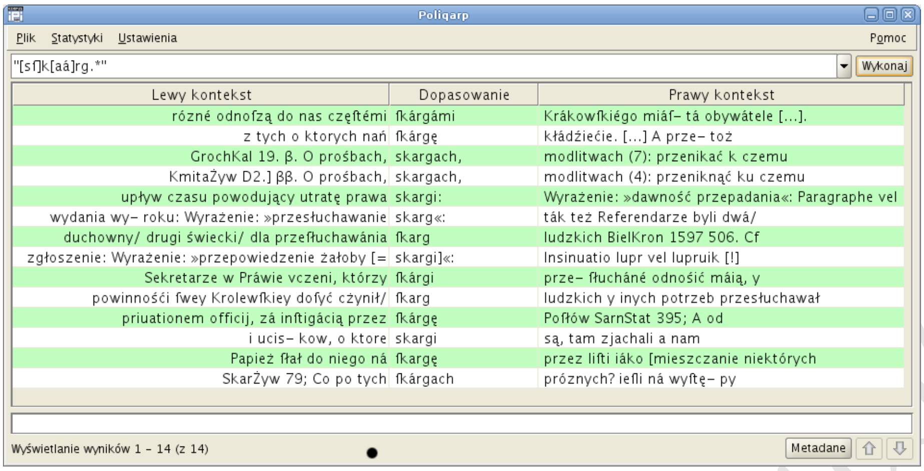
At present downloading the whole document to be searched is unavoidable, but it has to be done only once. Then the hidden text is extracted in XML format (with djvutoxml, a program from the free DjVuLibre library) and provided as the input to the djvu-xfgrep program together with the word to be searched. The results are in the form of a HTML file. Figure 1 shows the result of searching the word *pod* in the hiddent text layer (obtained by „dirty OCR") of the first volume of the so called Warsaw Dictionary (cf. `http://ebuw.uw.edu.pl/publication/255`).

Several columns on the left describe the localisation of the found word. First there is the name of the file containing the page in question, next the numbers of the column, the region, the paragraph and the line containing the word (when so detailed division of the page is inappropriate or not available, then all lines are assumed to belong to the same column, region and paragraph). After the plus sign there is the number of the word in the line and the approximate location on the pixel image of the page, expressed as percents of the horizontal and vertical size of the page. Last but not least, the context of the word is specified, and the word itself is a hyperlink to the image of the page with the word in question highlighted in a similar way as demonstrated on Figure 4.

Although the program is usable, implementing it was not a goal in itself, but just a milestone on the way to the converter described in the next section.

4. Dictionaries as corpora

Treating dictionaries as corpora has been suggested first, to the best of my knowledge, in [7]. Following this suggestion we propose to use corpus tools to investigate and search the text of dictionaries. For several reasons our preferred tool is Poliqarp (*Polyinterpretation Indexing Query and Retrieval Processor*), a GPLed corpus query tool developed in the Institute of Computer Science of Polish Academy of Sciences (cf. `http://korpus.pl/index.php?page=publications`).



"[sſ]k[aá]rg.*"

Lewy kontekst	Dopasowanie	Prawy kontekst
rózné odnoſzą do nas częſtémi	ſkárgámi	Krákowſkiégo miáſ– tá obywátele [...].
z tych o ktorych nań	ſkárgę	kłádźiećie. [...] A prze– toż
GrochKal 19. β. O prośbach,	skargach,	modlitwach (7): przenikać k czemu
KmitaŻyw D2.] ββ. O prośbach,	skargach,	modlitwach (4): przeniknąć ku czemu
upływ czasu powodujący utratę prawa	skargi:	Wyrażenie: »dawność przepadania«: Paragraphe vel
wydania wy– roku: Wyrażenie: »przesłuchawanie	skarg«:	ták też Referendarze byli dwá/
duchowny/ drugi świecki/ dla przeſłuchawánia	ſkarg	ludzkich BielKron 1597 506. Cf
zgłoszenie: Wyrażenie: »przepowiedzenie żałoby [=	skargi]«:	Insinuatio lupr vel lupruik [!]
Sekretarze w Práwie vczeni, którzy	ſkárgi	prze– ſłucháné odnośić máią, y
powinnośći ſwey Krolewſkiey doſyć czynił/	ſkarg	ludzkich y inych potrzeb przesłuchawał
priuationem officij, zá inſtigácią przez	ſkárgę	Poſłów SarnStat 395; A od
i ucis– kow, o ktore	skargi	są, tam zjachali a nam
Papież ſłał do niego ná	ſkargę	przez liſti iáko [mieszczanie niektórych
SkarŻyw 79; Co po tych	ſkárgach	próznych? ieſli ná wyſtę– py

Wyświetlanie wyników 1 – 14 (z 14)

Figure 2. Poliqarp — a query

Poliqarp is a client-server system. If a server is available on a local or global network, then there is no need to download the whole document (in particular a multivolume dictionary) for the purpose of searching. Another important feature of Poliqarp is sophisticated support of regular expressions, which can be used not only to circumvent the errors of dirty OCR, but also to accounts for different spellings in historical dictionaries.

We will illustrate the latter aspects using the 32nd volume of the Dictionary of the 16th century Polish. Thanks to the late Prof. Franciszek Pepłowski, the former head of the dictionary team, there is a digitally-born version of the volume — to make a long story short, the PDF files used for printing the paper version have been converted to DjVu using Jakub Wilk's excellent pdf2djvu program (`http://jwilk.net/software/pdf2djvu.html`, cf. also [6]). The converter from DjVu to Poliqarp has been also specified first in the note [2], then implemented in Java by Piotr Sikora and made available under the terms of the GPL license at `http://code.assembla.com/djvu-fgrep/`. An additional utility has been written by Jakub Wilk, who is also the current developer of Poliqarp.

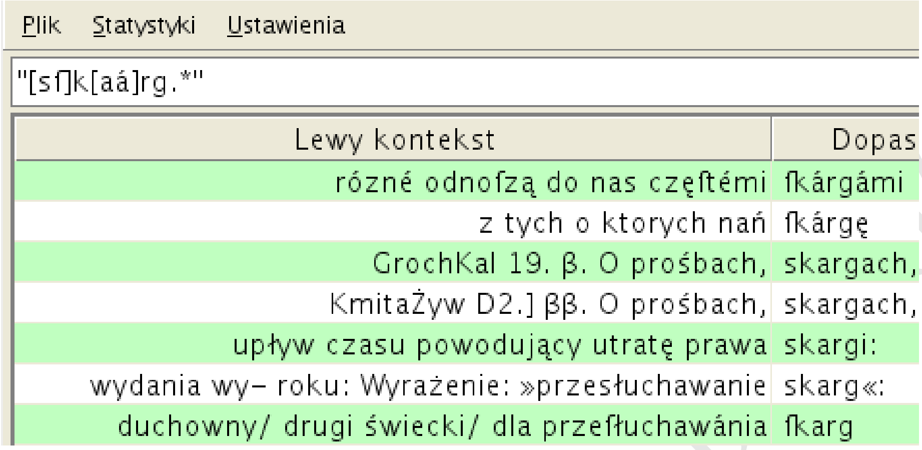
Our intention is to search for the word *skarga* (meaning *complaint*) in all inflexional forms, so we search for words starting with the appropriate string — `.*` (the final part of the regular expression) means any number of any characters. On the other hand we cannot use the string *skarga* directly, because it can be spelled also with long *s* and/or with accented *a*, so we have to use list of alternative characters such as `[aá]`. Figures 2 and 3 show the result of our query (Poliqarp interface can be switched to English, but for this example this doesn't seem appropriate).

Figure 4 demonstrates a standard feature of Poliqarp, namely the sub-window with the larger context of the match.

Figure 5 demonstrates another standard feature of Poliqarp, namely displaying the metadata of the document, but we use this in a non-standard way. Because we treat the dictionary as a corpus, the equivalent of a document is in our case just a page of the dictionary. In consequence the metadata consist of two fields: the reference to the dictionary as a whole (the index file of the unbundled DjVu document) and the reference to the component file containing the specific page.

Similarily Figure 6 demonstrates yet another standard feature of Poliqarp, namely displaying the tags of the matched word, and again we use this in a non-standard way. Although intended for tags such as part of speech etc., we use them simply to provide localisation of the word on the page, specyfing for this purpose the line number and the running number of the word in the line.

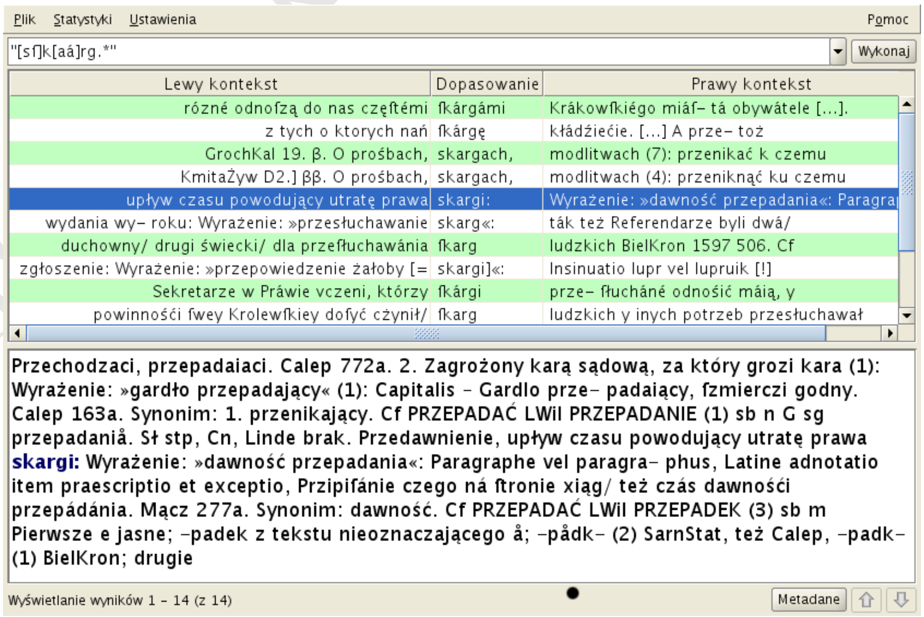
Hence we have shown that Poliqarp can store all the information needed to



"[sſ]k[aá]rg.*"

Lewy kontekst	Dopas
rózné odnoſzą do nas częſtémi	ſkárgámi
z tych o ktorych nań	ſkárgę
GrochKal 19. β. O prośbach,	skargach,
KmitaŻyw D2.] ββ. O prośbach,	skargach,
upływ czasu powodujący utratę prawa	skargi:
wydania wy– roku: Wyrażenie: »przesłuchawanie	skarg«:
duchowny/ drugi świecki/ dla przeſłuchawánia	ſkarg

Figure 3. Poliqarp — query details



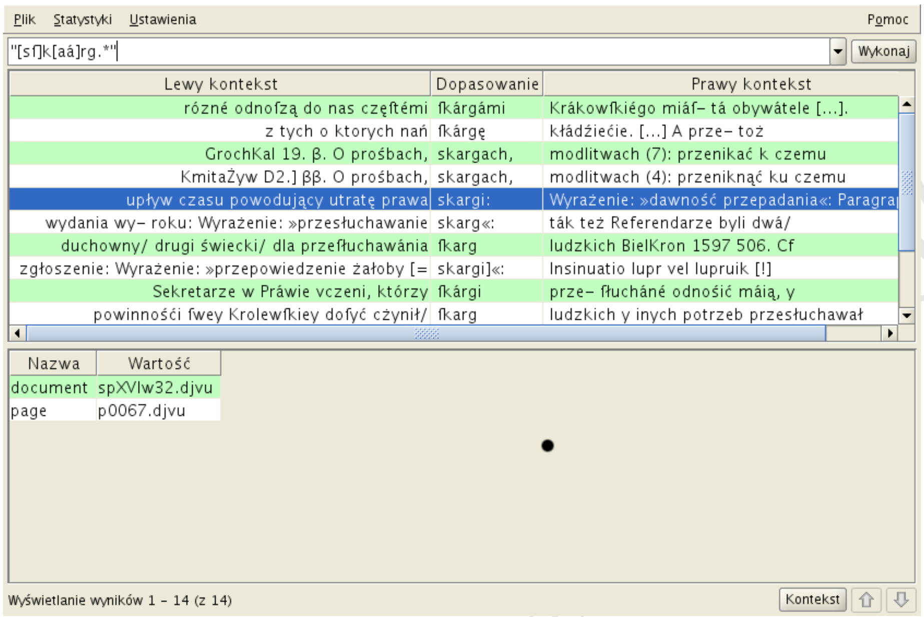
"[sſ]k[aá]rg.*"

Lewy kontekst	Dopasowanie	Prawy kontekst
rózné odnoſzą do nas częſtémi	ſkárgámi	Krákowſkiégo miáſ– tá obywátele [...].
z tych o ktorych nań	ſkárgę	kłádźiećie. [...] A prze– toż
GrochKal 19. β. O prośbach,	skargach,	modlitwach (7): przenikać k czemu
KmitaŻyw D2.] ββ. O prośbach,	skargach,	modlitwach (4): przeniknąć ku czemu
upływ czasu powodujący utratę prawa	skargi:	Wyrażenie: »dawność przepadania«: Paragra
wydania wy– roku: Wyrażenie: »przesłuchawanie	skarg«:	ták też Referendarze byli dwá/
duchowny/ drugi świecki/ dla przeſłuchawánia	ſkarg	ludzkich BielKron 1597 506. Cf
zgłoszenie: Wyrażenie: »przepowiedzenie żałoby [=	skargi]«:	Insinuatio lupr vel lupruik [!]
Sekretarze w Práwie vczeni, którzy	ſkárgi	prze– ſłucháné odnośić máią, y
powinnośći ſwey Krolewſkiey doſyć cżynił/	ſkarg	ludzkich y inych potrzeb przesłuchawał

Przechodzaci, przepadaiaci. Calep 772a. 2. Zagrożony karą sądową, za który grozi kara (1): Wyrażenie: »gardło przepadający« (1): Capitalis – Gardlo prze– padaiący, ſzmierczi godny. Calep 163a. Synonim: 1. przenikający. Cf PRZEPADAĆ LWil PRZEPADANIE (1) sb n G sg przepadanią̊. Sł stp, Cn, Linde brak. Przedawnienie, upływ czasu powodujący utratę prawa **skargi:** Wyrażenie: »dawność przepadania«: Paragraphe vel paragra– phus, Latine adnotatio item praescriptio et exceptio, Przipiſánie czego ná ſtronie xiąg/ też czás dawnośći przepádánia. Mącz 277a. Synonim: dawność. Cf PRZEPADAĆ LWil PRZEPADEK (3) sb m Pierwsze e jasne; –padek z tekstu nieoznaczającego å; –pådk– (2) SarnStat, też Calep, –padk– (1) BielKron; drugie

Wyświetlanie wyników 1 – 14 (z 14)

Figure 4. Poliqarp — larger context



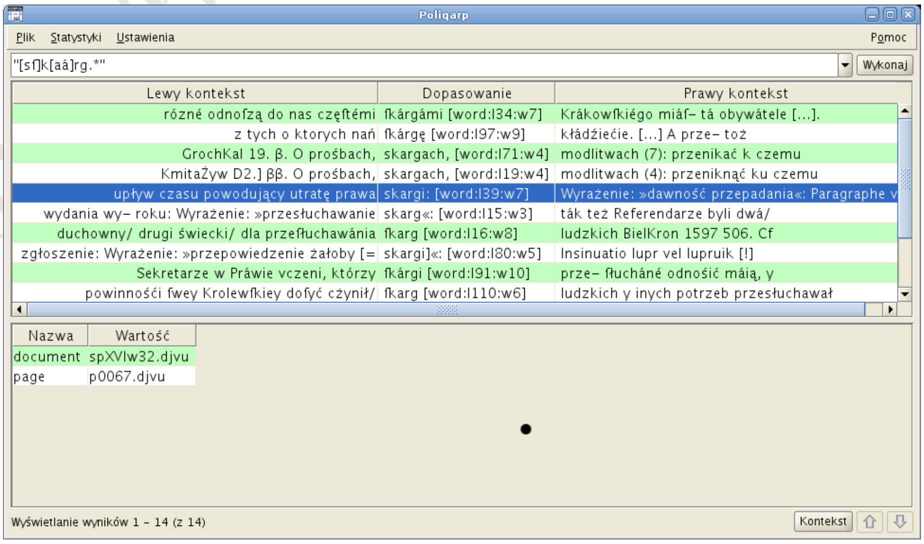
"[sſ]k[aá]rg.*"

Lewy kontekst	Dopasowanie	Prawy kontekst
rózné odnoſzą do nas częſtémi	ſkárgámi	Krákowſkiégo miáſ– tá obywátele [...].
z tych o ktorych nań	ſkárgę	kłádźiećie. [...] A prze– toż
GrochKal 19. β. O prośbach,	skargach,	modlitwach (7): przenikać k czemu
KmitaŻyw D2.] ββ. O prośbach,	skargach,	modlitwach (4): przeniknąć ku czemu
upływ czasu powodujący utratę prawa	skargi:	Wyrażenie: »dawność przepadania«: Paragra
wydania wy– roku: Wyrażenie: »przesłuchawanie	skarg«:	ták też Referendarze byli dwá/
duchowny/ drugi świecki/ dla przeſłuchawánia	ſkarg	ludzkich BielKron 1597 506. Cf
zgłoszenie: Wyrażenie: »przepowiedzenie żałoby [=	skargi]«:	Insinuatio lupr vel lupruik [!]
Sekretarze w Práwie vczeni, którzy	ſkárgi	prze– ſłucháné odnośić máią, y
powinnośći ſwey Krolewſkiey doſyć czynił/	ſkarg	ludzkich y inych potrzeb przesłuchawał

Nazwa	Wartość
document	spXVIw32.djvu
page	p0067.djvu

Wyświetlanie wyników 1 – 14 (z 14)

Figure 5. Poliqarp — selected match localisation



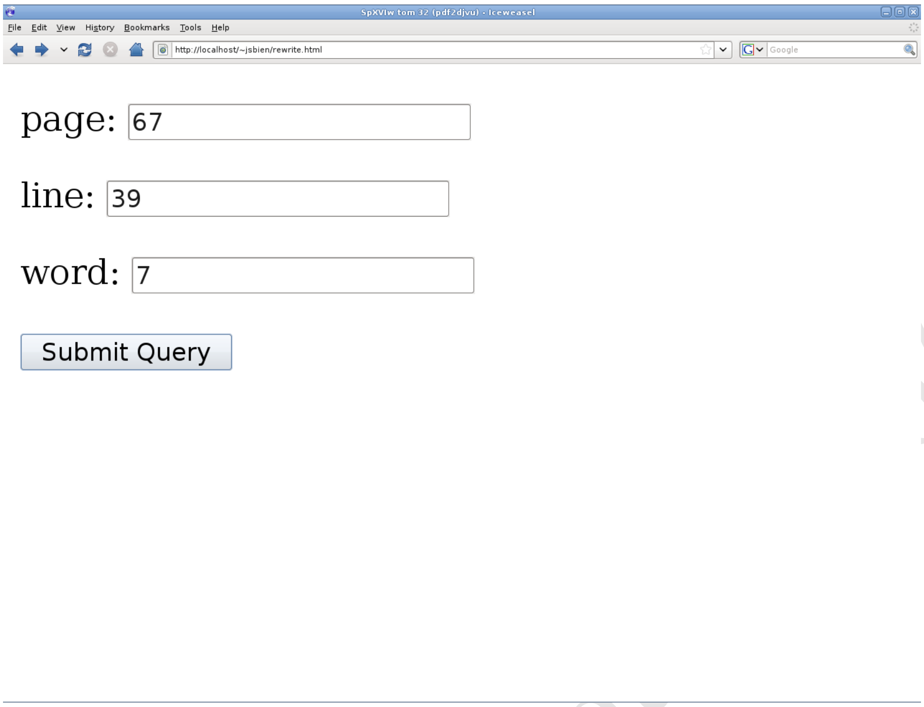
"[sſ]k[aá]rg.*"

Lewy kontekst	Dopasowanie	Prawy kontekst
rózné odnoſzą do nas częſtémi	ſkárgámi [word:l34:w7]	Krákowſkiégo miáſ– tá obywátele [...].
z tych o ktorych nań	ſkárgę [word:l97:w9]	kłádźiećie. [...] A prze– toż
GrochKal 19. β. O prośbach,	skargach, [word:l71:w4]	modlitwach (7): przenikać k czemu
KmitaŻyw D2.] ββ. O prośbach,	skargach, [word:l19:w4]	modlitwach (4): przeniknąć ku czemu
upływ czasu powodujący utratę prawa	skargi: [word:l39:w7]	Wyrażenie: »dawność przepadania«: Paragraphe v
wydania wy– roku: Wyrażenie: »przesłuchawanie	skarg«: [word:l15:w3]	ták też Referendarze byli dwá/
duchowny/ drugi świecki/ dla przeſłuchawánia	ſkarg [word:l16:w8]	ludzkich BielKron 1597 506. Cf
zgłoszenie: Wyrażenie: »przepowiedzenie żałoby [=	skargi]«: [word:l80:w5]	Insinuatio lupr vel lupruik [!]
Sekretarze w Práwie vczeni, którzy	ſkárgi [word:l91:w10]	prze– ſłucháné odnośić máią, y
powinnośći ſwey Krolewſkiey doſyć czynił/	ſkarg [word:l110:w6]	ludzkich y inych potrzeb przesłuchawał

Nazwa	Wartość
document	spXVIw32.djvu
page	p0067.djvu

Wyświetlanie wyników 1 – 14 (z 14)

Figure 6. Poliqarp — selected match in-page localisation


SpXVIw tom 32 (pdf2djvu) - Iceweasel
File Edit View History Bookmarks Tools Help
http://localhost/~jsbien/rewrite.html
Google

page: 67

line: 39

word: 7

Submit Query

Figure 7. Locating the match

locate the word in question in the pixel version of the dictionary. To make this process less cumbersome, Jakub Wilk prepared a simple tool presented on Figure 7.

After filling the form the user is redirected to the appropriate page with the relevant fragment highlighted, as demonstrated on Figure 8. Usually it will be convenient to zoom into the interesting area, as illustrated on Figure 9.

5. Concluding remarks

We have demonstrated that Poliqarp, supplemented by programs by Piotr Sikora and Jakub Wilk, can be used „as is" to facilitate access to large documents in DjVu format. Our plans are to make the process much more user friendly. It will be one of the goals of the *Digitalization tools for philological research* project supported by Grant N N519 384036 of the Ministry of Science and Higher Education. The project lasts from 13 May 2009 to 12 November 2011, more information will be in due time available at `http://wbs.klf.uw.edu.pl`[1].

References

[1] Bień, Janusz S., 2006. Kilka przykładów dygitalizacji słowników *Poradnik Językowy* z. 8 (637), s. 55-63. `http://ebuw.uw.edu.pl/publication/250`.

[2] Bień, Janusz S., 2008. Narzędzia do analizy tekstowej warstwy dokumentów DjVu. Unpublished note. `http://bc.klf.uw.edu.pl/105/`.

[3] Janusz S. Bień. 2009 Digitalizing dictionaries of Polish. [w:] *Methods of Lexical Analysis: Theoretical assumption and practical applications*. Białystok, s. 37-45. `http://bc.klf.uw.edu.pl/71/`

[4] Yann Le Cun, Léon Bottou, Andrei Erofeev, Patrick Haffner, and Bill W. Riemers. "DjVu document browsing with on-demand loading and rendering

[1] As of May 6, 2010, the Poliqarp based search engine is available at `http://poliqarp.wbl.klf.uw.edu.pl/`.

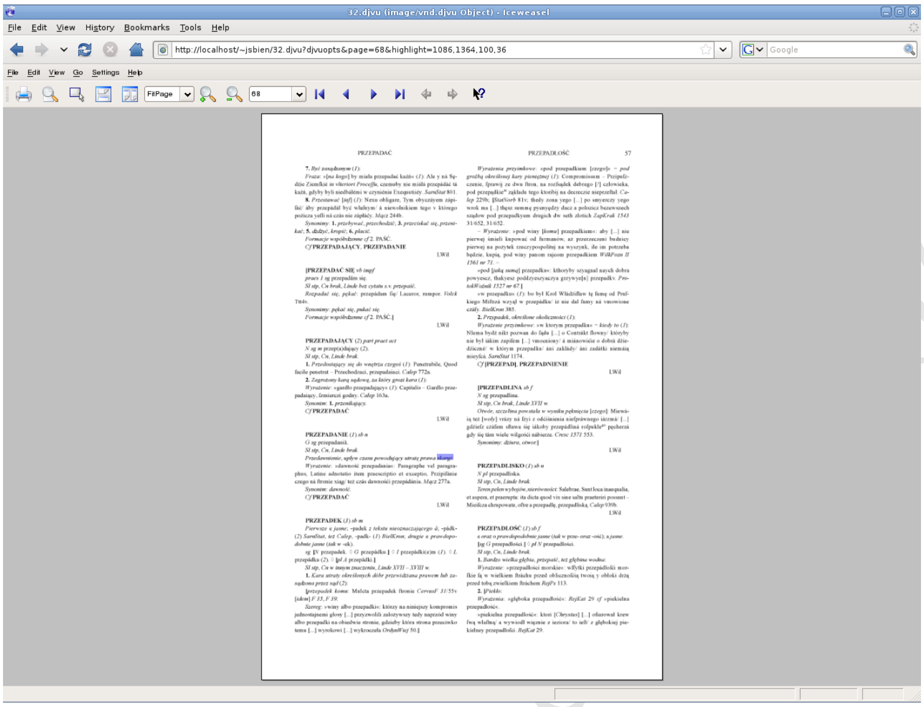
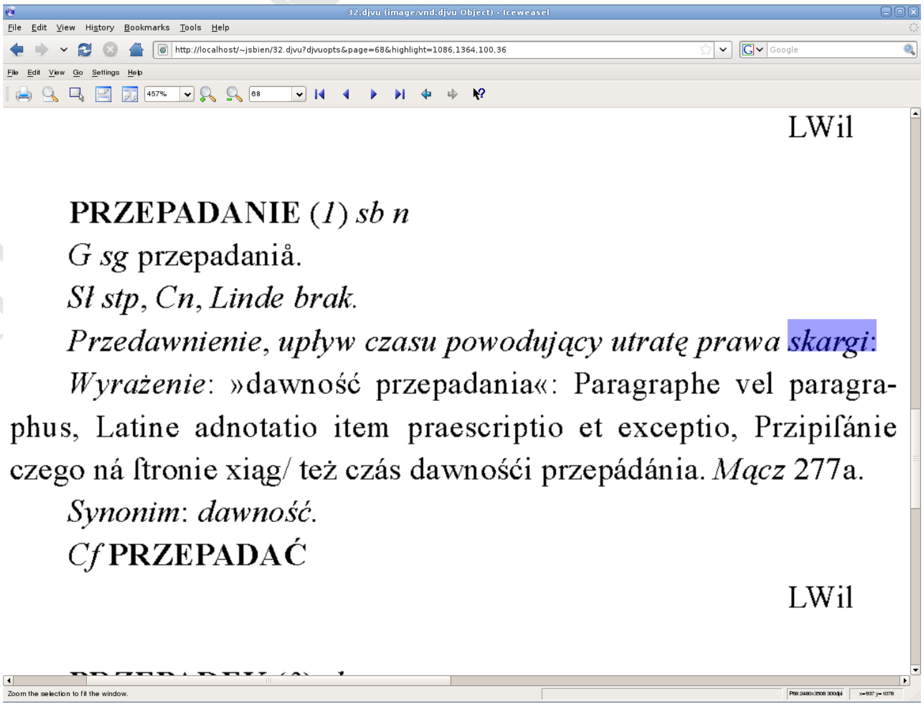

32.djvu (image/vnd.djvu Object) - Iceweasel
File Edit View History Bookmarks Tools Help
http://localhost/~jsbien/32.djvu?djvuopts&page=68&highlight=1086,1364,100,36
Google
File Edit View Go Settings Help
FitPage
68

Figure 8. The located match


32.djvu (image/vnd.djvu Object) - Iceweasel
File Edit View History Bookmarks Tools Help
http://localhost/~jsbien/32.djvu?djvuopts&page=68&highlight=1086,1364,100,36
Google
File Edit View Go Settings Help
457%
68
Zoom the selection to fit the window.

LWil

PRZEPADANIE (*1*) *sb n*

G sg przepadaniå.

Sł stp, *Cn*, *Linde brak*.

Przedawnienie, *upływ czasu powodujący utratę prawa skargi*:

Wyrażenie: »dawność przepadania«: Paragraphe vel paragraphus, Latine adnotatio item praescriptio et exceptio, Przipiſánie czego ná ſtronie xiąg/ też czás dawnośći przepádánia. *Mącz* 277a.

Synonim: *dawność*.

Cf **PRZEPADAĆ**

LWil

Figure 9. Details of the match

of image components” in *Internet Imaging*, San Jose, January 2001. http://leon.bottou.org/papers/lecun-2001.

[5] Anna Wałek. 2009. Bibilioteki cyfrowe na platformie dLibra. Wydawnictwo SBM, Warszawa.

[6] Jakub Wilk. Wilk, J., 2008. Rozbudowa pakietu oprogramowania DjVuLibre. http://jw209508.hopto.org/papers/thesis/.

[7] Piotr Żmigrodzki. Żmigrodzki, P., 2005. Słownik jako korpus tekstów — korpus tekstów jako słownik. Perspektywy polskiej leksykografii naukowej. *Poradnik Językowy* nr 6, s. 3-14.