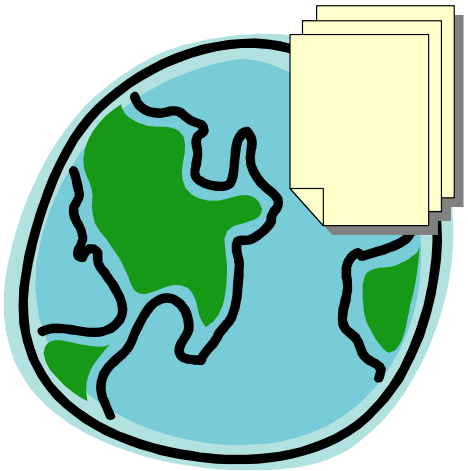
Zaawansowane przetwarzanie tekstu dla potrzeb bibliotek cyfrowych

Stanisław Osiński
stanislaw.osinski@man.poznan.pl

POZNAŃSKA FUNDACJA BIBLIOTEK NAUKOWYCH

Ilość dostępnych informacji rośnie bardzo szybko

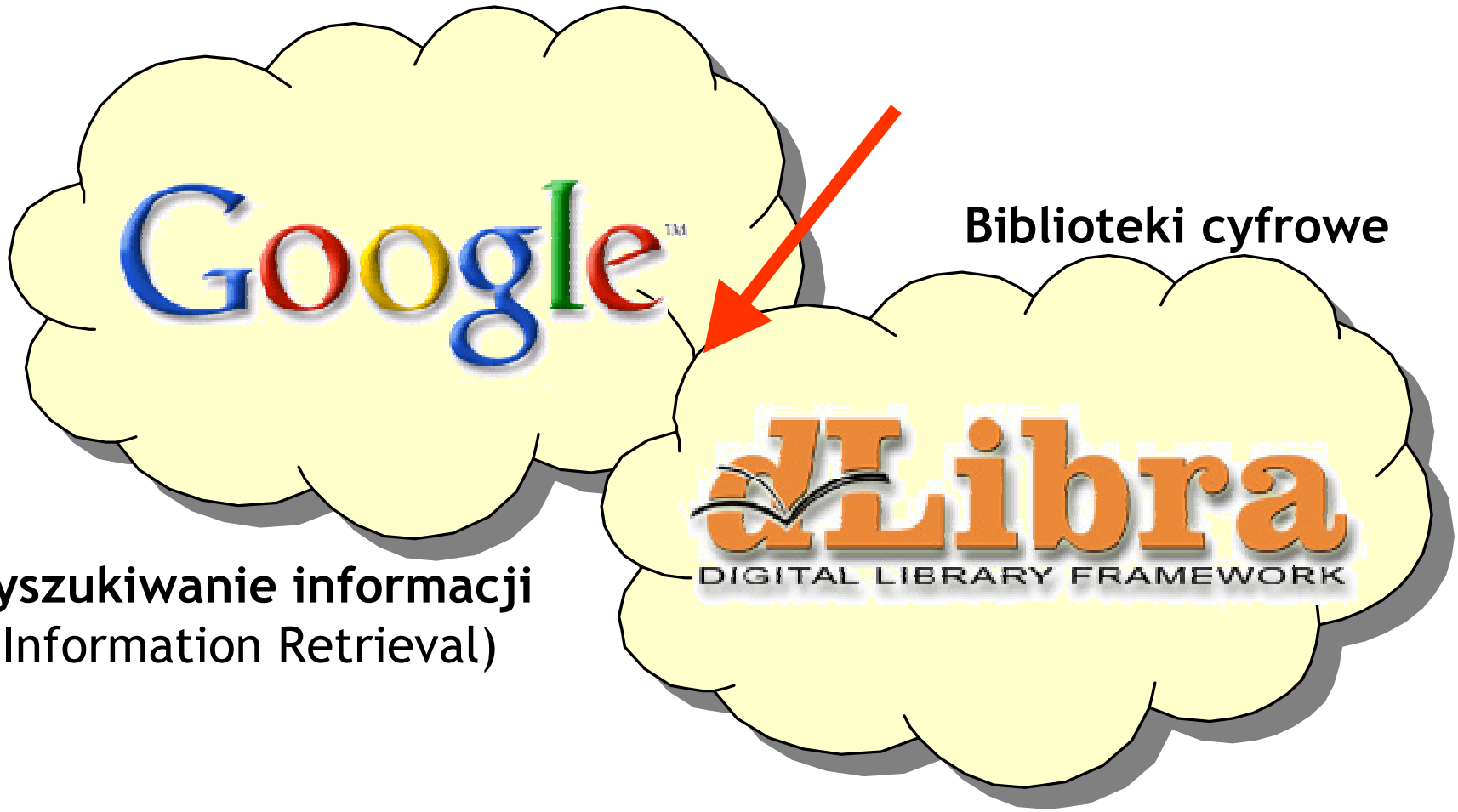


Problemy dostępu do informacji pojawiają się na wielu różnych polach

Google™

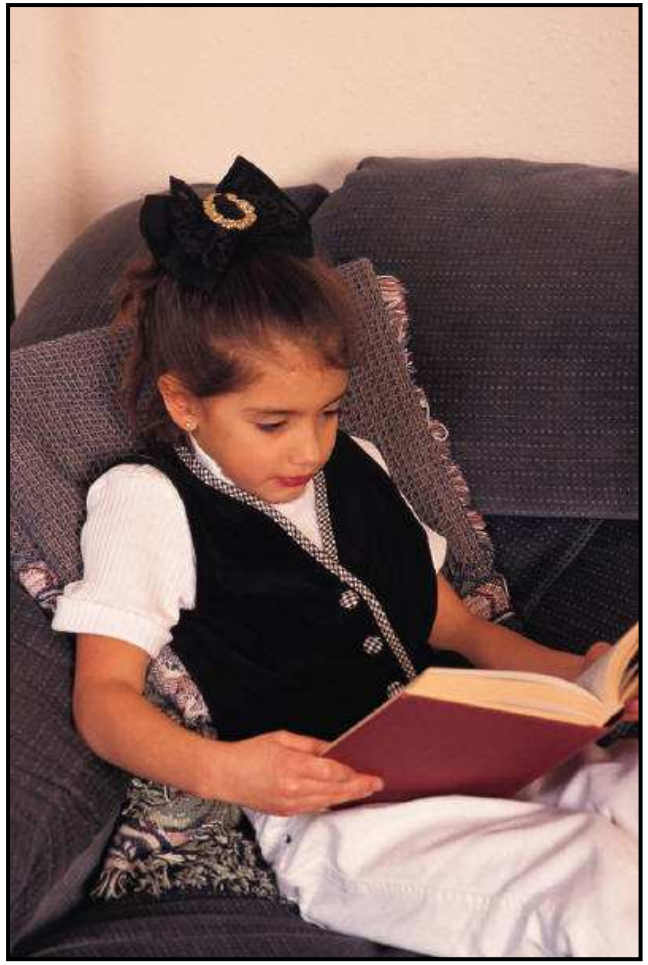
dLibra
DIGITAL LIBRARY FRAMEWORK

Czy wyszukiwanie informacji i biblioteki cyfrowe mają wspólne obszary?

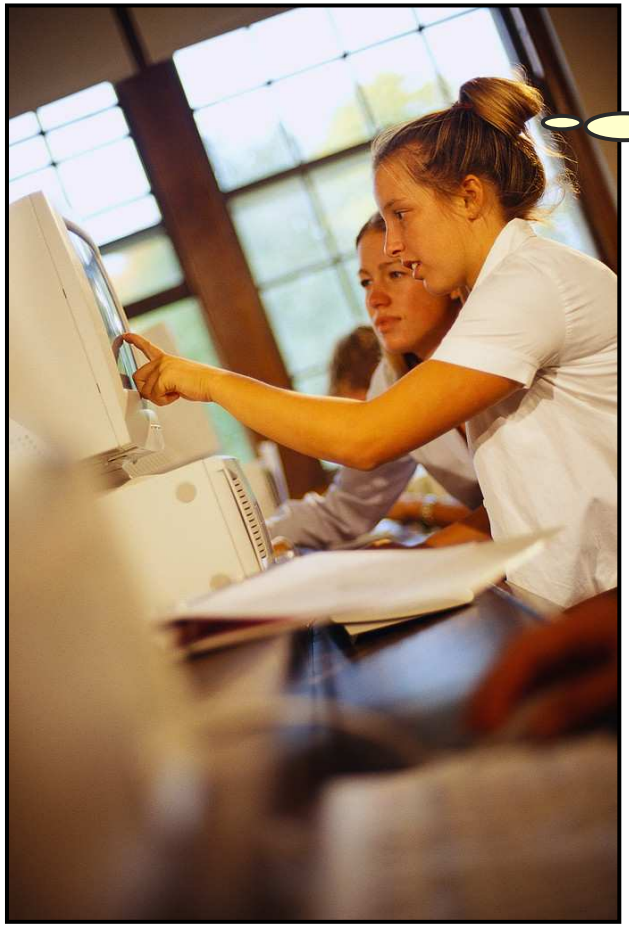


Wyszukiwanie informacji
(Information Retrieval)

Skorzystać mogą zarówno **czytelnicy** jak i **bibliotekarze**



Czytelnicy chcą szybko **dotrzeć do właściwej informacji**



Biografia Fryderyka Chopina

chopin biografia

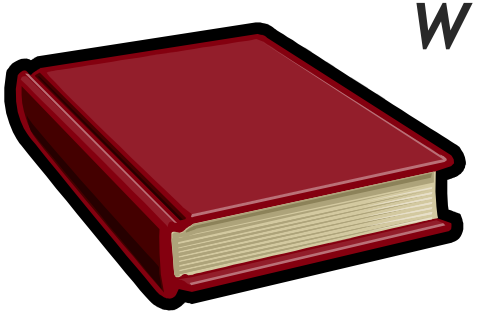
szopen biografia

„w *chołdzie* chopinowi”

biografie kompozytorów

?

Bardzo prosta wyszukiwarka
nie zawsze się sprawdza



W hołdzie Chopinowi:
rys biograficzny kompozytora

chopin biografia	0 publikacji
szopen biografia	0 publikacji
w *chołdzie* chopinowi	0 publikacji
biografie kompozytorów	196 publikacji

Lematyzacja usuwa **końcówki fleksyjne**

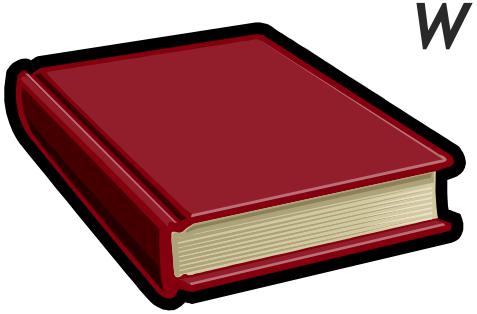
chopin
chopina
chopinem
chopinowi
post-chopinowski
chopinowskiemu

chopin

biograf

biografia
biografii
biograficzny
biograficznej
biografów
biografiami

Lematyzacja **usprawnia wyszukiwanie**



W hołdzie Chopinowi:
rys biograficzny kompozytora

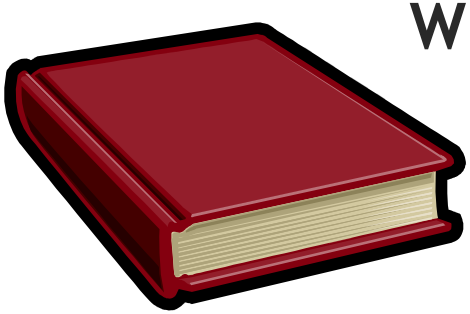
chopin biografia	**1 publikacja** ✓
szopen biografia	0 publikacji
w *chołdzie* chopinowi	0 publikacji
biografie kompozytorów	196 publikacji

Słownik synonimów znajdzie **podobne słowa i pojęcia**

szopen
- chopin
- szopen

fryderyk szopen
- Fryderyk Chopin
- Fryderyk Szopen
- Frédéric Chopin

Słownik synonimów również **usprawnia wyszukiwanie**



W hołdzie Chopinowi:
rys biograficzny kompozytora

chopin biografia	**1 publikacja** ✓
szopen biografia	**1 publikacja** ✓
w *chołdzie* chopinowi	0 publikacji
biografie kompozytorów	196 publikacji

Słownik ortograficzny **poprawi błędy**

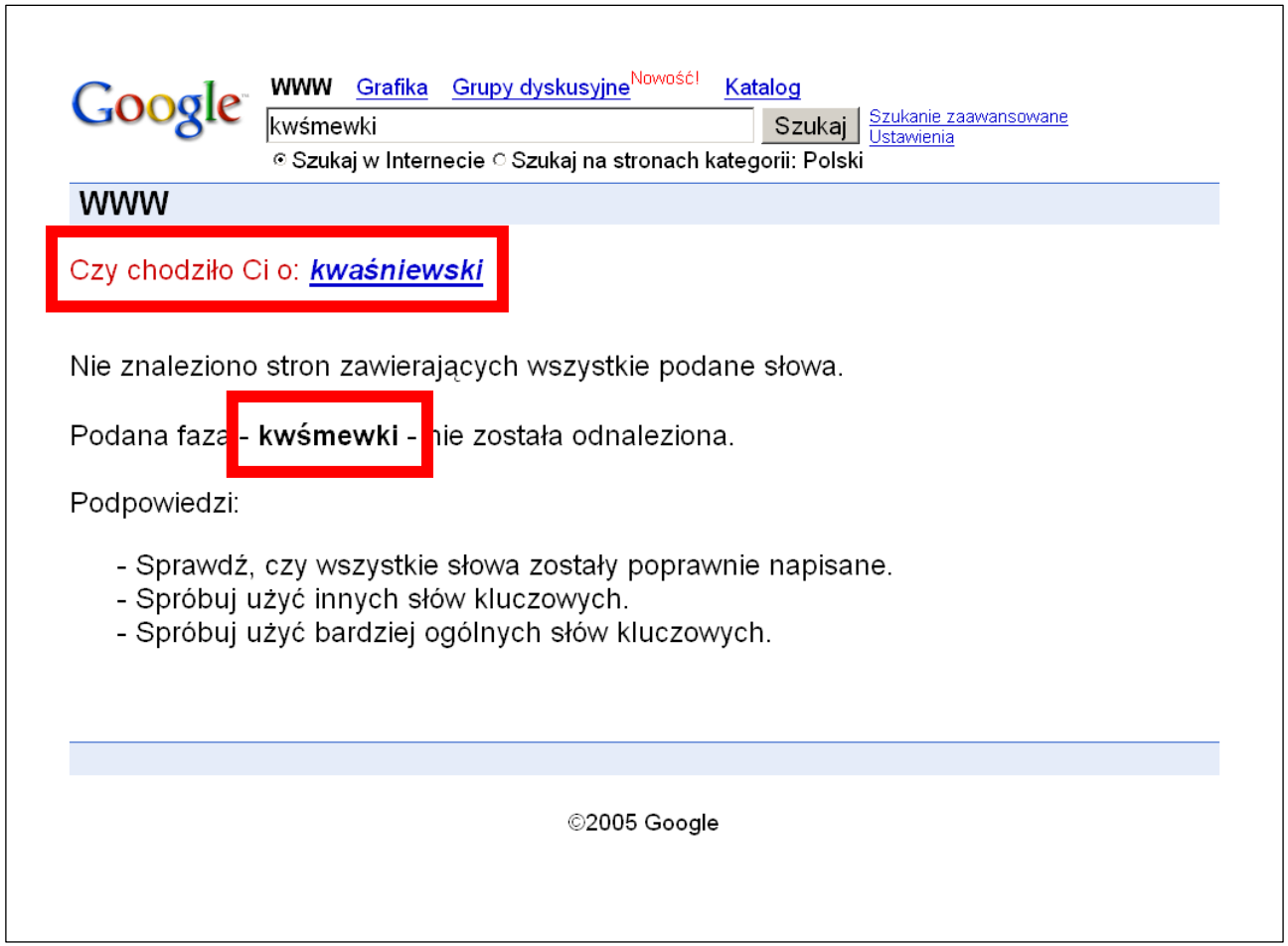
chłodzie (87%)
hałdzie (72%)
hołdzie (63%)

~~chołdzie~~
chłodzie

~~chołdzie~~
hołdzie

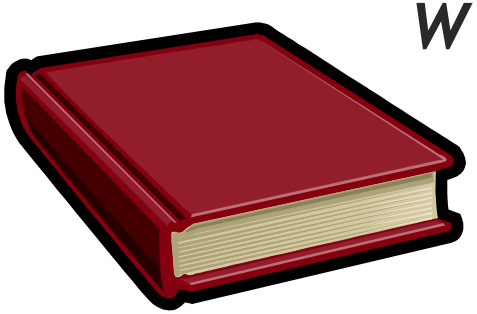
hołdzie (99%)
chłodzie (12%)
hałdzie (2%)

Słownik ortograficzny **poprawi błędy**



http://www.google.com

Korekta błędów ortograficznych **usprawnia wyszukiwanie**



W hołdzie Chopinowi:
rys biograficzny kompozytora

chopin biografia	1 publikacja	✓
szopen biografia	1 publikacja	✓
w *chołdzie* chopinowi	1 publikacja	✓
biografie kompozytorów	196 publikacji	

Duże ilości wyników można automatycznie grupować

biografie kompozytorów
(196 dokumentów)

Muzyka klasyczna
(12 dokumentów)

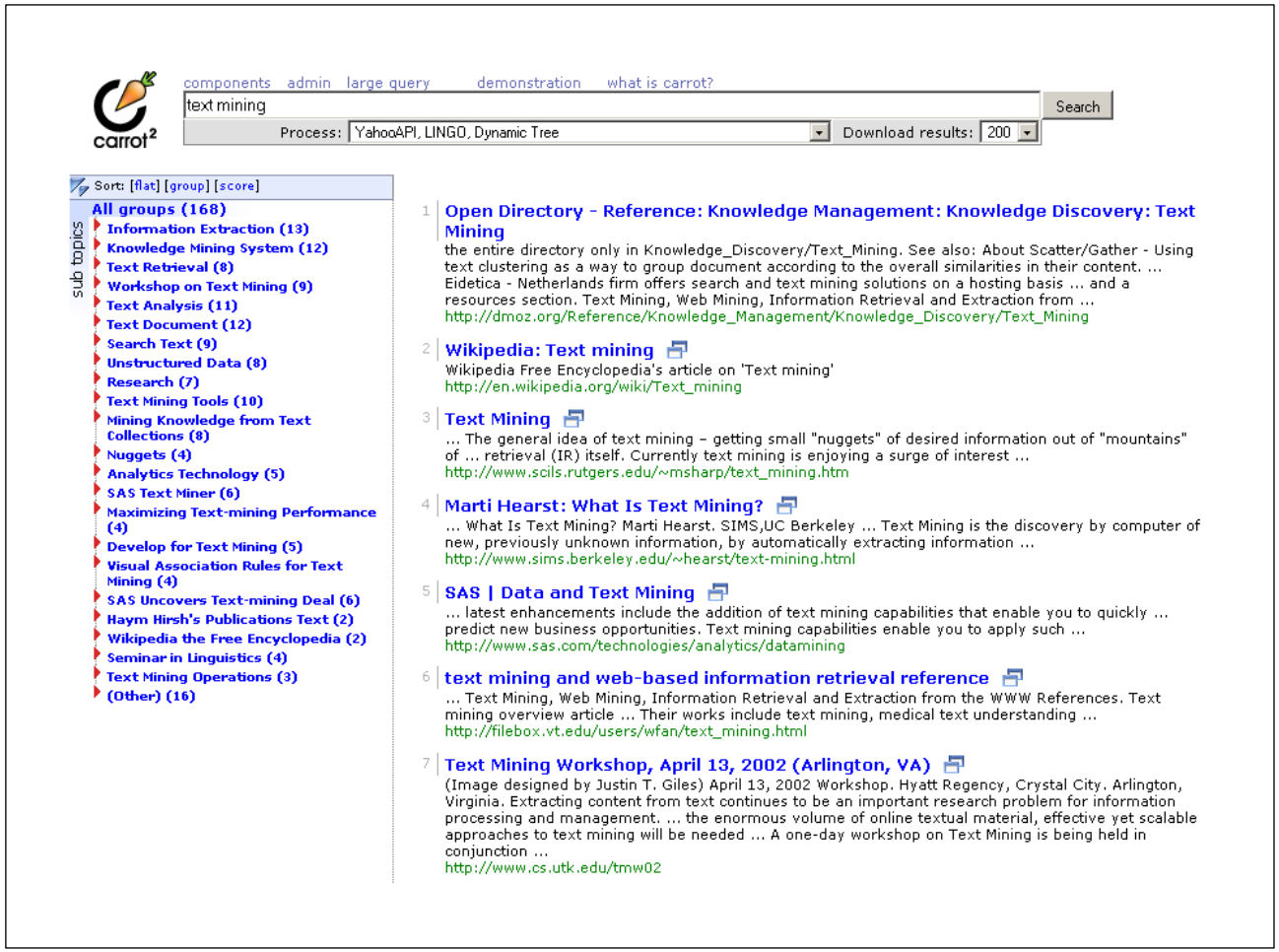
Jan Sebastian Bach
(5 dokumentów)

Fryderyk Chopin
(10 dokumentów)

John Lennon
(3 dokumenty)

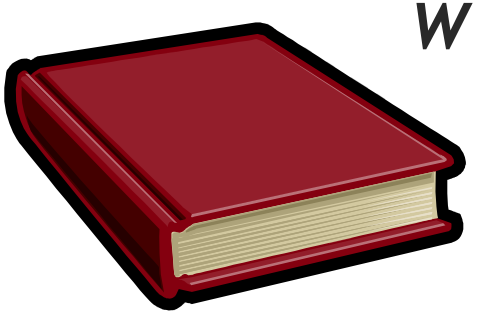
...

Duże ilości wyników można automatycznie grupować



http://carrot.cs.put.poznan.pl/carrot2-remote-controller/

Grupowanie wyników może przynieść dodatkowe usprawnienia



W hołdzie Chopinowi:
rys biograficzny kompozytora

chopin biografia	1 publikacja	✓
szopen biografia	1 publikacja	✓
w *chołdzie* chopinowi	1 publikacja	✓
biografie kompozytorów	12 kategorii	✓

Bibliotekarze chcą **łatwo udostępniać nowe zasoby**

Słowa kluczowe

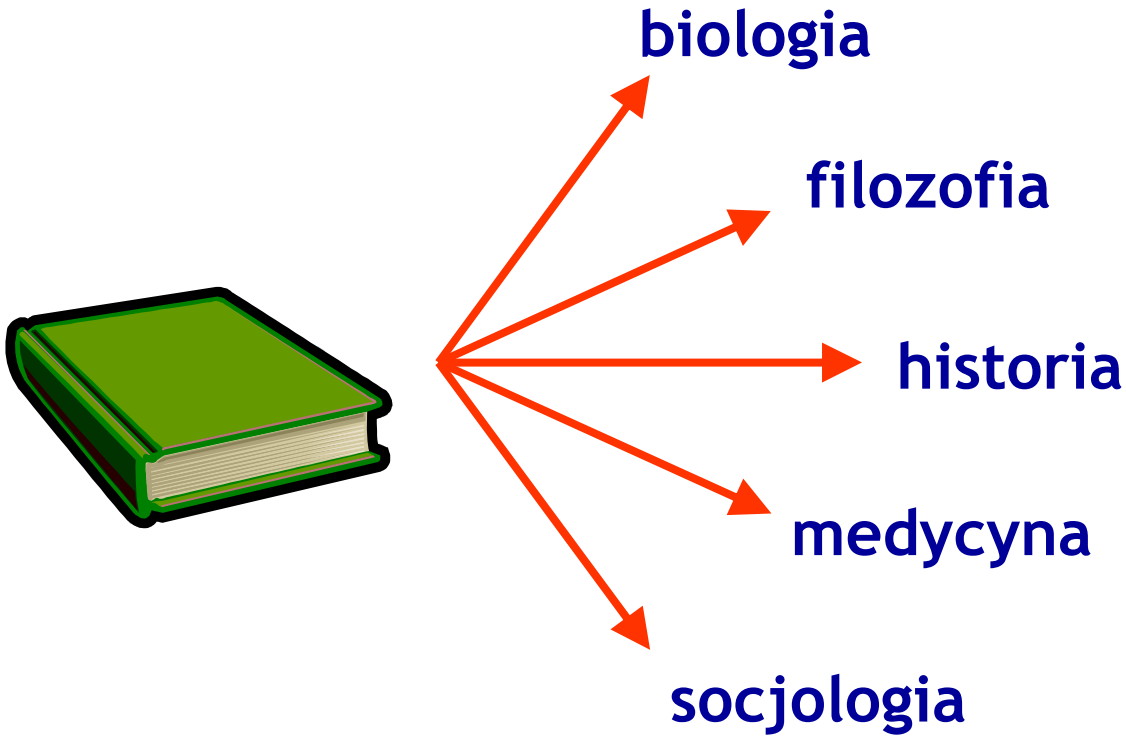
Opisy, streszczenia

Definicje, wiedza

Bardzo duża liczba nowych publikacji może wymagać częściowej automatyzacji

dLibra
DIGITAL LIBRARY FRAMEWORK

Algorytm klasyfikujący przypisuje dokument do kategorii



Algorytm klasyfikujący wybierze odpowiednie miejsce w katalogu

biologia

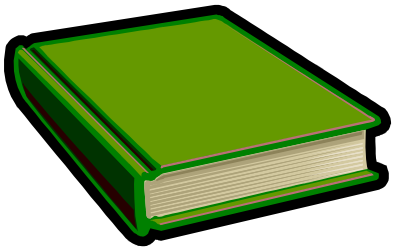
filozofia

historia

medycyna

???

Algorytm streszczający wybiera ważne fragmenty tekstu

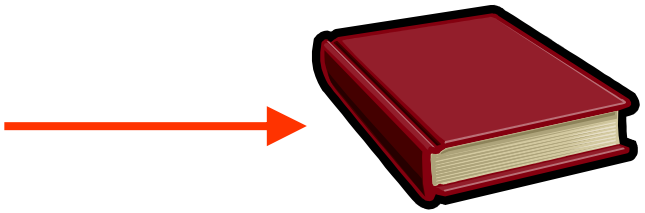


„Najwybitniejszy polski kompozytor, a także wybitny pianista”

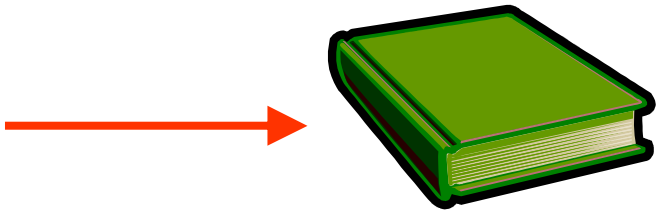
„W 1838 roku Chopin poznaje starszą od niego o 6 lat i dominującą nad nim George Sand”

„Umarł w 1849 roku, w otoczeniu najbliższych osób”

Algorytm streszczający
zasugeruje opis publikacji



„Najwybitniejszy polski kompozytor, a także wybitny pianista”



„Istnieją kontrowersje, kto był głównym autorem zwycięstwa” (???)

Istnieją algorytmy wyszukujące **nazwiska, daty, miejsca** etc.

Fryderyk Chopin
George Sand
Jarosław Iwaszkiewicz

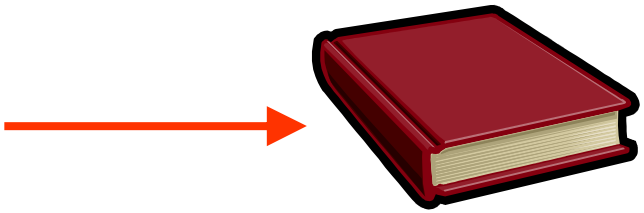
1 marca 1810 roku
17.10.1749
Po roku 1723

(w) Żelazowej Woli
(do) Paryża
Szkole Głównej Muzyki (???)

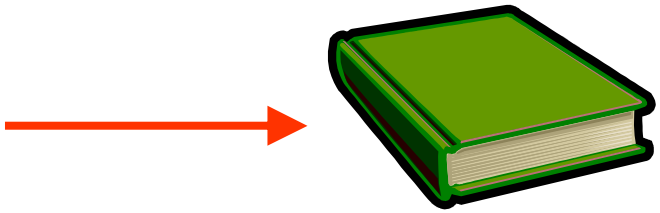
ul. Wiejska 1, 00-000 Warszawa

134 zł 12 gr
12,00 zł
6 EUR

Prowadzone są prace nad automatyczną **ekstrakcją metadanych**



Autor: Jarosław Iwaszkiewicz
Słowa kluczowe: romantyzm, fyderyk chopin, biografia



Autor: Józef Piłsudski (???)
Słowa kluczowe: marszałek, biografia, uczęszczał do gimnazjum (???)

Możliwe jest automatyczne **wyszukiwanie definicji pojęć**

„Tlen jest niemetalem z szóstej grupy głównej”

„Poznań to jedno z najstarszych i największych polskich miast, położone nad rzeką Wartą”

„Poznań jest ważnym ośrodkiem przemysłu spożywczego” (???)

Możliwe jest automatyczne wyszukiwanie definicji pojęć

Sign in

Google Web Images Groups News Froogle Local more »

define: digital library Search Advanced Search Preferences

Web

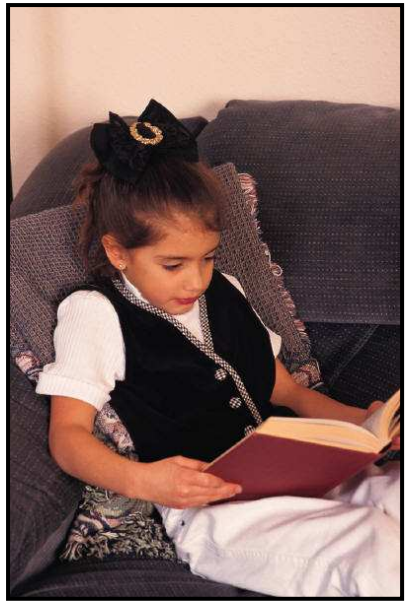
Related phrases: perseus digital library digital library federation list of digital library projects alexandria digital library digital library services digital library systems acm digital library greenstone digital library california digital library informedia digital library

Definitions of **digital library** on the Web:

- Collection of texts, images, etc., encoded so as to be stored, retrieved, and read by computer. Digital printing
  www.sir.arizona.edu/resources/glossary.html
- The meaning of the phrase digital library varies tremendously, but one simple definition is the use of computers to store library materials appearing in electronic (digital) format.
  www.law.harvard.edu/library/collections/digital/guidelines_glossary.php
- Digital libraries can include reference material or resources accessible through the World Wide Web. Digitized portions of a library's collection or original material produced for the web can also be included in a digital library.
  www.collectionscanada.ca/vrc-rvc/s34-151-e.html
- an integrated set of services for capturing, cataloging, storing, searching, protecting, and retrieving information
  www.wtec.org/loyola/digilibs/d_01.htm

http://www.google.com

Skorzystać mogą zarówno czytelnicy jak i bibliotekarze



lematyzacja

synonimy

ortografia

grupowanie wyników

klasyfikacja

streszczenia

ekstrakcja informacji

ekstrakcja metadanych

ekstrakcja definicji

Dziękuję za uwagę

Zaawansowane przetwarzanie tekstu dla potrzeb bibliotek cyfrowych

Stanisław Osiński
stanislaw.osinski@man.poznan.pl

POZNAŃSKA FUNDACJA
BIBLIOTEK NAUKOWYCH