Propozycja modelu udostępniania zbiorów historycznych polskiego Webu zarchiwizowanych w usłudze Wayback Machine

Marcin Wilkowski (LaCh UW)
@marcinwilkowski

wayback machine

- https://archive.org/web/
- **Wayback Machine ≠ Internet Archive**
- od 2001 r.
- ponad 273 miliardów stron* (webpages): html, txt, pdf / 361 milionów serwisów (websites)
- API (status i URL archiwum w JSON)
- od niedawna nowa wersja beta
- jedyne* zbiory polskiego historycznego Webu

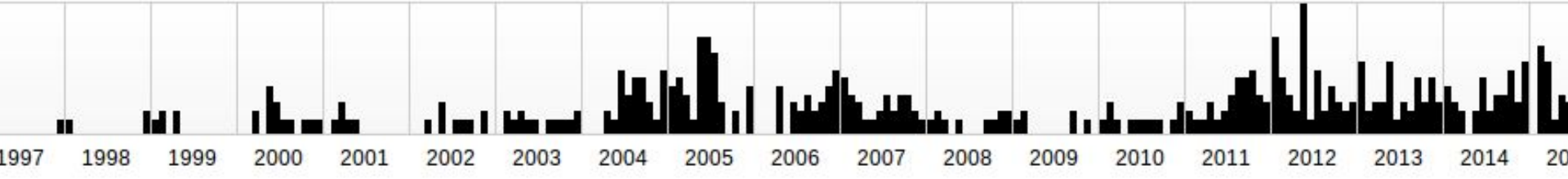
INTERNET ARCHIVE
WayBackMachine

http://prezydent.pl/ BROWSE HISTORY

http://prezydent.pl/

Saved **552 times** between December 12, 1997 and October 29, 2016.

PLEASE DONATE TODAY. Your generosity preserves knowledge for future generations. Thank you.



JAN							FEB							MAR							APR						
				1	2			1	2	3	4	5	6		**1**	2	**3**	**4**	5							1	**2**
4	5	6	**7**	8	9		7	8	9	10	11	12	13	**6**	7	8	**9**	**10**	11	**12**	3	4	5	**6**	7	8	9
11	**12**	13	**14**	15	16		14	15	16	**17**	18	19	20	13	**14**	15	16	17	18	**19**	10	11	**12**	**13**	14	15	**16**
18	19	20	21	22	23		21	22	23	24	25	26	27	20	21	22	**23**	24	25	26	**17**	**18**	**19**	20	**21**	22	**23**

ograniczenia

- brak katalogu (nie wiemy, czy coś zostało zarchiwizowane, dopóki nie mamy konkretnych URLi)
- brak wiedzy o jakości kopii (jaka część serwisu została zachowana)
- java script (dynamiczne renderowanie stron)
- niejasny status prawnoautorski zbiorów

Chat or hit the road @ChuckBaggett · 13.08.2013
What would it take to make the WayBack Machine internet archive searchable by word in the pages instead of by urls? Money?

Internet Archive
@internetarchive

Obserwowany

@ChuckBaggett Yes, money is exactly what it takes to make Wayback full text searchable. Feel free to donate! :-) archive.org/donate/

Zobacz tłumaczenie

PODANE DALEJ 3 POLUBIENIE 1

02:54 - 13.08.2013

3 1

free.polbox.pl

Zareklamuj sie u nas - jestesmy skuteczni !

Informacje Nasi uzytkownicy E-mail/WWW Co nowego? Pomoc

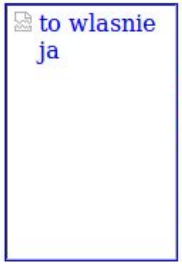
Polbox

- pierwszy darmowy hosting (1997-2008)
- free.polbox.pl, 2MB WWW, zakaz stron komercyjnych
- "polskie Geocities"
- ważne źródło historyczne dokumentujące początki polskiego "oddolnego" Webu
- zob: Marcin Jagodziński, *Polbox: historia pewnego falstartu*, http://netto.blox.pl/2008/04/polbox-historia-pewnego-falstartu.html
- inspiracja: Internet Archive: GeoCities Special Collection 2009 (1996-2009)

GOOSIE'S HOME PAGE - experimental launch of English version - dedicated to all my English-speaking pals from all over the world...

Strona Domowa
G O O S I E

Witam na stronie domowej Goosie !!


to wlasnie ja

Pewnie nurtuje Was mysl, kim moze byc owa tajemnicza Goosie. Jak mozecie podejrzec na zalaczonym obrazku jest kobieta - i to w dodatku blondynka. Jesli nie zniechecilo Was to do dalszej inwigilacji mej skromnej osoby, zapraszam do lektury Zyciorysu - tam mozecie znalezc wszystkie (no nie przesadzajmy...) pikantne wycinki z mej biografii...

"Wszystko co naprawde lubie jest albo niemoralne, albo nielegalne, albo tuczace" **..... czyli moje zainteresowania.**

Temat jest to o tyle niewdzieczny, iz zawsze marzeniem moim bylo moc pochwalic sie niecodziennym hobby - np. hodowla pytonow badz gra na mandolinie...
No bo czy fakt, ze:

- uwielbiam towarzystwo moich przyjaciol

budowa podstaw archiwum

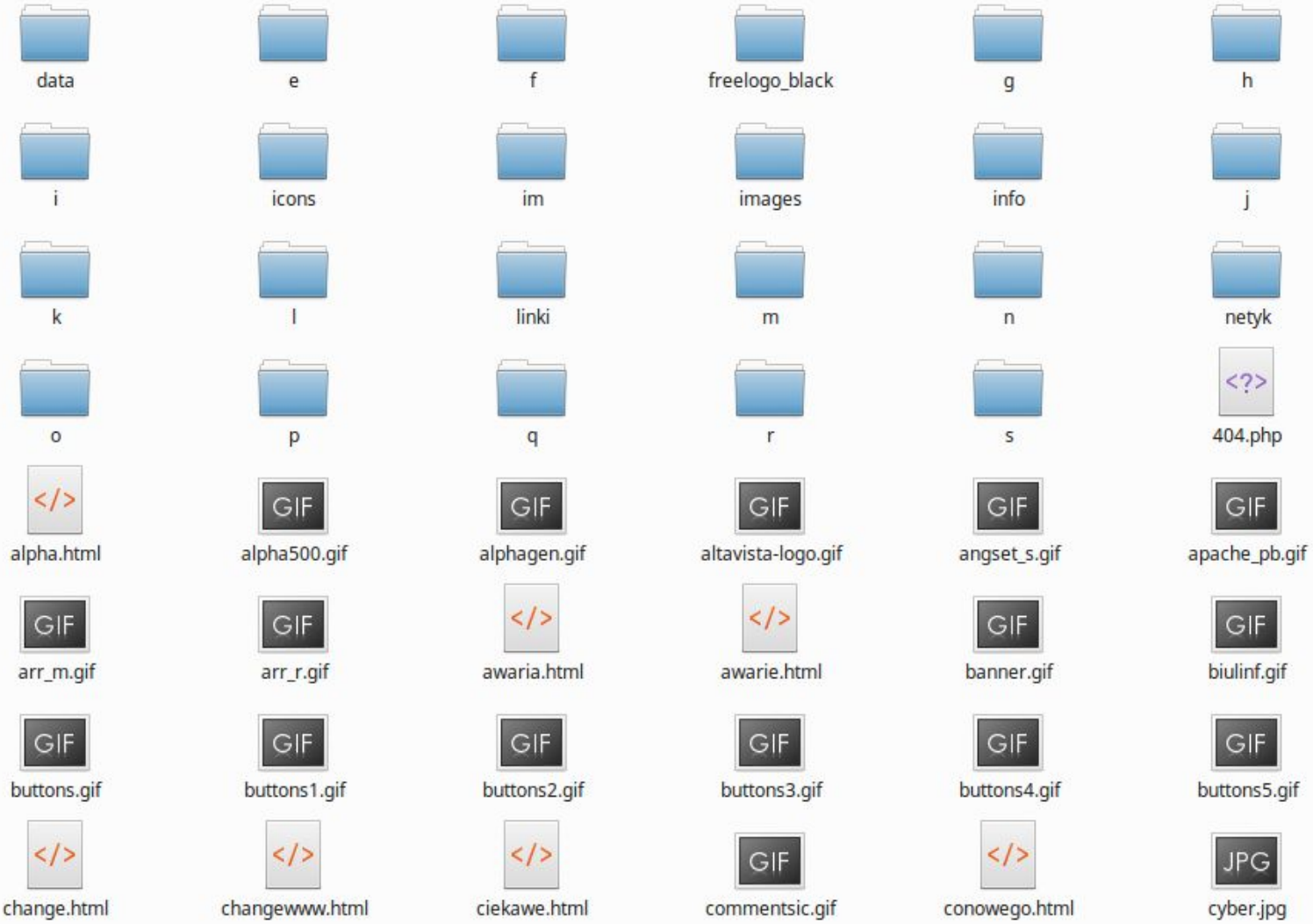
- pobranie z WM całej domeny free.polbox.pl
- indeksowanie
- stworzenie wyszukiwarki pełnotekstowej

- wykorzystanie schematu konstrukcji URL `free.polbox.pl/n/nickname`
- wykorzystanie `wayback_machine_downloader` (Ruby) https://github.com/hartator/wayback-machine-downloader

```
Usage: wayback_machine_downloader http://example.com

Download an entire website from the Wayback Machine.

Optional options:
    -d, --directory PATH             Directory to save the downloaded files into
                                     Default is ./websites/ plus the domain name
    -f, --from TIMESTAMP             Only files on or after timestamp supplied (ie. 20060716231334)
    -t, --to TIMESTAMP               Only files on or before timestamp supplied (ie. 20100916231334)
    -o, --only ONLY_FILTER           Restrict downloading to urls that match this filter
                                     (use // notation for the filter to be treated as a regex)
    -x, --exclude EXCLUDE_FILTER     Skip downloading of urls that match this filter
                                     (use // notation for the filter to be treated as a regex)
    -a, --all                        Expand downloading to error files (40x and 50x) and redirections (30x)
    -c, --concurrency NUMBER         Number of multiple files to dowload at a time
                                     Default is one file at a time (ie. 20)
    -p, --snapshot-pages NUMBER      Maximum snapshot pages to consider (Default is 100)
                                     Count an average of 150,000 snapshots per page
    -l, --list                       Only list file urls in a JSON format with the archived timestamps, won't
    -v, --version                    Display version
```


data
e
f
freelogo_black
g
h
i
icons
im
images
info
j
k
l
linki
m
n
netyk
o
p
q
r
s
404.php
alpha.html
alpha500.gif
alphagen.gif
altavista-logo.gif
angset_s.gif
apache_pb.gif
arr_m.gif
arr_r.gif
awaria.html
awarie.html
banner.gif
biulinf.gif
buttons.gif
buttons1.gif
buttons2.gif
buttons3.gif
buttons4.gif
buttons5.gif
change.html
changewww.html
ciekawe.html
commentsic.gif
conowego.html
cyber.jpg

indeksowanie i udostępnianie

- wygenerowanie listy adresów wszystkich plików htm w katalogach głównych (np. n/nick) - około 6 tys. adresów (na 2GB danych)
- prosta pętla php i `file_get_contents` (scrappowanie htmla)
- treść pliku htm do bazy danych
- wyszukiwarka znajduje frazę dostępną w kodzie html konkretnej strony
- wersja testowa wyszukiwarki http://wilkowski.org/plbxs1

```
6397
6398 for ($i = -1; $i < 6382; ++$i) {
6399         $adres = $lista[$i];
6400         print '<a href='.$adres.'>&clubs;</a> ';
6401         $tresc = file_get_contents($adres);
6402         $database->insert("link2", array(
6403         "url" => $adres,
6404         "content" => $tresc
6405         ));
6406         sleep(60);
6407         }
6408
```

--- brak tytułu ---

Warszawa Sadyba Inowroc

http://free.polbox.pl/0/02924n21

Sprawdź status strony w Internet Archive

Teatr Studio

Teatr Studio body{background-color:white; font-family: Verdana, Arial,sans-serif; font-size:10pt; color:#424242} A{color:#424242; text-decoration:none} A:hover{color:blue} A:visited:{color:808080} #divMain{position:absolute; left:50; width:300; top:50; font-family: Verdana, Arial,sans-serif; font-size:10pt; color:#424242; font-weight:bold} /*## This script is made by bratta

http://free.polbox.pl/a/abc254

Sprawdź status strony w Internet Archive

ABC druk

ABC druk 01-646 Warszawa ul. Jelinka 52 tel/fax: 833 84 14 i 832 16 39 e mail: abcdruk@free.polbox.pl DRUKARNIA OFFSETOWA druk do formatu C 3 (33 X 46 cm) sk

http://free.polbox.pl/a/abcdruk

Sprawdź status strony w Internet Archive

błędy i ograniczenia

- zła jakość indeksowania (tylko strony główne index.htm, niedobre czyszczenie z css/js)
- duplikaty
- tylko konta A-S
- brak wyszukiwania po dacie/okresie udostępnienia
- prawa autorskie :/

prawa autorskie

- Wayback Machine (Internet Archive) formalnie udostępnia kopie
- ja udostępniam linki do konkretnych URL
- za pomocą Wayback Machine API automatyczne sprawdzanie statusu i generowanie linku do źródła w WM
- dodatkowy problem: dane osobowe
- dodatkowy problem: *prawo do zapomnienia*

Kod odpowiedzi http: 200

Zarchiwizowano: 07.12.1998

URL: http://web.archive.org/web/19981207014716/http://free.polbox.pl:80/a/ako1/

SIGN IN

Search

INTERNET ARCHIVE
WayBackMachine
BETA

Explore more than 273 billion web pages saved over time

Enter a URL or words related to a site's home page

Tools

Wayback Machine Availability API
Build your own tools.

WordPress Broken Link Checker

Subscription Service

Archive-It enables you to capture, manage and search collections of digital content without any technical expertise or hosting facilities. Visit Archive-It to build and browse the collections.

Save Page Now

https://

SAVE PAGE

Capture a web page as it appears now for use as a trusted citation in the future.

Feedback

0

propozycja modelu udostępniania polskich zbiorów Webu

- research głównych domen (z DMOZ, historycznych katalogów portali itp.)
- pozyskiwanie kopii 1:1 z Wayback Machine
- porządna indeksacja (treść + metadane)
- wyszukiwarka pełnotekstowa + wyszukiwanie zaawansowane
- lepsza integracja z API Wayback Machine
- wygenerowanie katalogu
- kopie wciąż **TYLKO** po stronie Wayback Machine (prawa autorskie)

dziękuję

Przeszukuj strony domowe Polboxa z lat 1997-2008
http://wilkowski.org/notka/1314

Przeszukiwanie pełnotekstowe w Wayback Machine (Internet Archive)
http://wilkowski.org/notka/1349

Web traffic analytics as a historical source
http://wilkowski.org/notka/1329