Marcin Wilkowski
Laboratorium Cyfrowe Humanistyki
Uniwersytet Warszawski
m@wilkowski.org

Wayback Machine – podstawy wykorzystania

Streszczenie: Autor analizuje zaprojektowane już w 1996 r., a udostępnione publicznie pięć lat później Wayback Machine, które jest internetowym archiwum zasobów World Wide Web. Celem artykułu jest przedstawienie podstawowych metod wykorzystywania Wayback Machine do pracy z archiwalnymi wersjami stron www zabezpieczanych w tej usłudze. Wersja beta archiwum została udostępniona w październiku 2016 r. Fundacja Internet Archive obchodziła wówczas 20-lecie działań na polu archiwizacji webu[1].

Słowa kluczowe: Wayback Machine, archiwizacja internetu, API, historia cyfrowa,

Kontekst

Zaprojektowane już w 1996 r., a udostępnione publicznie pięć lat później Wayback Machine jest internetowym archiwum zasobów World Wide Web: stron internetowych oraz związanych z nimi lub udostępnianych niezależnie obiektów, takich jak pliki graficzne, skrypty Java Script, pliki PDF czy TXT. Zasoby te pozyskiwane są metodą crawlingu (*web crawling*), co oznacza, że do archiwum mogą być gromadzone jedynie publicznie dostępne obiekty WWW, które dodatkowo nie są blokowane przez indeksowanie i zczytywanie za pomocą odpowiednich wpisów do robots.txt[2].

Zespół Internet Archive wypracował stosowną terminologię opisującą poszczególne rodzaje gromadzonych i przetwarzanych zasobów, przydatną podczas projektowania struktury archiwum Wayback Machine oraz do celów analitycznych (tab. 1).

Tab. 1. Pojęcia stosowane w dokumentacji Wayback Machine

webpage	Treść poprawnej odpowiedzi na zapytanie do serwera źródłowego, udostępniającego pozyskiwany do archiwum obiekt dostępny pod określonym adresem URL. Zatem w terminologii Wayback Machine *webpage* to nie tylko strony WWW w postaci dokumentów HTML (i związanych z nimi plików graficznych Java Script, CSS itp.), ale także pliki PDF oraz inne pliki tekstowe.
domain	Termin funkcjonujący w systemie DNS (Domain Name System), zbudowany z wykorzystaniem jednej ze standardowych końcówek (np. pl, org), np. *wikipedia.org.*
host	Pełna nazwa domenowa (*fully qualified domain name*, FQDN), np. *pl.wikipedia.org.*
website	Host publikujący *webpages*, do którego linkuje przynajmniej jedna strona z innej domeny[3]. Liczba *webpages* jest zawsze większa niż liczba *websites*. Np. *pl.wikipedia.org.*

Źródło: oprac. własne na podstawie GOEL, V. Defining Web pages, Web sites and Web captures. W: *Internet Archive Blogs* [online]. Internet Archive Blogs. [Dostęp 26.04.2017]. Dostępny w: https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/.

[1] Za ich symboliczny początek można uznać artykuł Brewstera Kahle: KAHLE, B. Preserving the Internet. *Scientific American* 1997, 276 (3), s. 82–83. ISSN 0036-8733.

[2] Polityka respektowania robots.txt przez Fundację Internet Archive zmienia się i przynajmniej w przypadku zasobów WWW publikowanych przez instytucje rządowe bywa ignorowana. Zob. ROSSI, A. Robots.txt Files and Archiving .gov and .mil Websites. W: *Internet Archive Blogs* [online]. Internet Archive Blog. [Dostęp 26.04.2017]. Dostępny w: https://blog.archive.org/2016/12/17/robots-txt-gov-mil-websites/.

[3] Jest to oczywiście konieczne, aby roboty crawlujące mogły trafić na domenę i ją zindeksować.

Znajomość znaczenia tych pojęć, wykorzystywanych w systemie Wayback Machine, pozwala poprawnie odczytywać nie tylko ogólne statystyki archiwum, ale także statystyki dotyczące poszczególnych zarchiwizowanych hostów, dostępne w nowej wersji usługi. Należy dodać, że zaproponowana przez zespół Internet Archive kategoryzacja archiwizowanych obiektów opiera się wyłącznie na ich cechach technicznych w ramach systemu DNS i struktury WWW. Alternatywne kategoryzacje uwzględniać mogą już inne cechy obiektów WWW i bardziej sprawdzać się przy opracowywaniu metod badań na historycznym webie[4].

Znaczniki czasu

Wszystkie obiekty pozyskiwane i archiwizowane w ramach Wayback Machine publikowane są z odpowiednim znacznikiem czasu (*timestamp,* tzw. *snapshot*). Jeden obiekt (*webpage*) może być dostępny dla danej daty w żadnym, jednym lub kilku snapshotach. Przykładowo, strona główna serwisu (*website*) Onet.pl z 1 marca 2001 r. posiada w Wayback Machine trzy kopie (*snapshots*) z różnych godzin dnia – 08:32:32, 20:02:22, 21:05:40[5], przy czym czas i data podawane są dla czasu uniwersalnego (GMT). Kolory poszczególnych znaczników informują odpowiednio o statusie pozyskanego zbioru (tab. 2).

Tab. 2. Informacje o statusie snapschotów w Wayback Machine

Kolor	Status odpowiedzi HTTP	Opis
niebieski	2xx	Snapshot dostępny pod linkiem został wygenerowany przez prawidłową odpowiedź serwera źródłowego.
zielony	3xx	Snapshot zawiera jedynie przekierowanie do innego snapshota (ew. może kierować do obiektu niedostępnego w archiwum).
czerwony	4xx	*Klient error* – snapshot niewygenerowany ze względu na błąd pobierania z serwera źródłowego; wina po stronie klienta (np. próba archiwizacji URLa, który nie istniał).
czerwony	5xx	*Server error* – snapshot niewygenerowany ze względu na błąd pobierania z serwera źródłowego; wina po stronie serwera.

Źródło: oprac. własne na podstawie GRAHAM, M. FAQs for some new features available in the Beta Wayback Machine. W: *Internet Archive Blogs* [online]. Internet Archive Blogs. [Dostęp 26.04.2017]. Dostępny w: https://blog.archive.org/2016/10/24/faqs-for-some-new-features-available-in-the-beta-wayback-machine/.

Poprawne przeglądanie obiektów archiwalnych umożliwiają jedynie snapshoty w kolorze niebieskim. Zachowanie przez Wayback Machine informacji o błędnych statusach snapshotów oraz ich prezentacja pozwalają na badanie braków w strukturze zarchiwizowanych witryn.

Znaczniki czasu definiujące snapshoty publikowane są także w nagłówkach odpowiedzi http, takich jak:

- X-Archive-Orig-Date: Thu, 01 Mar 2001 20:02:32 GMT
- Memento-Datetime: Thu, 01 Mar 2001 20:02:22 GMT[6]

[4] Zob. BRÜGGER, N. Web History and the Web as a Historical Source. *Zeithistorische Forschungen* [online]. 2012, 9 (2), s. 316–325. [Dostęp 26.04.2017]. Dostępny w: http://www.zeithistorische-forschungen.de/2-2012/id=4426.

[5] Dostępność kopii strony Onet.pl dla 2001 roku: *Wayback Machine Internet Archive* [online]. Wayback Machine. [Dostęp 26.04.2017]. Dostępny w: .https://web.archive.org/web/20010601000000*/http://onet.pl/

Warto dodać, że nagłówek odpowiedzi http Memento-Datetime umożliwia integrację zasobów udostępnianych przez Wayback Machine z aplikacjami wykorzystującymi protokół Memento[7]. Niestety, w Wayback Machine różne snapshoty mogą udostępniać różne nagłówki, co jest problemem szczególnie przy maszynowym zczytywaniu i analizowaniu zbiorów. Znaczniki czasu pełnią także fundamentalną rolę w metodach API serwisu Wayback Machine, umożliwiając automatyczne sprawdzanie dostępności obiektów o określonym URL w określonym czasie[8].

Dostępne w archiwum kopie mogą być wyświetlane w przeglądarce. Adres URL archiwalnej kopii zawiera także znacznik czasu, a w treść kopii wpisywany jest automatycznie niewielki kod HTML i JavaScript, wyświetlający nagłówek (banner) Wayback Machine oraz zliczający odsłony. Aby móc pracować z kopiami pozbawionymi dodatkowego, nieoryginalnego kodu, należy podmienić adres URL, dodając do znacznika czasu element id_:

- URL kopii strony głównej serwisu Onet z 20 lutego 1997 r. z bannerem nawigacyjnym Wayback Machine – https://web-beta.archive.org/web/20010106072200/http://onet.pl/,
- URL kopii tej samej strony głównej pozbawiona nieoryginalnego kodu – https://web-beta.archive.org/web/20010106072200id_/http://onet.pl/.

Nowa wersja Wayback Machine

Nowa wersja Wayback Machine udostępniona w październiku 2016 r. wciąż nie jest publikowana na stronie głównej usługi[9]. W porównaniu z dotychczasową wersją pojawiły się w niej dwie nowe funkcjonalności, ułatwiające pracę z zawierającym dziś ponad 284 mld serwisów (*webpages*) w archiwum.

Pierwszą nowością jest możliwość pełnotekstowego przeszukiwania, ograniczona niestety tylko do zawartości stron głównych (*homepage*) poszczególnych witryn (*websites*). Pozwala ona na odnajdywanie archiwalnych zasobów webu bez konieczności znajomości ich źródłowego adresu URL. To ważna opcja, w znaczący sposób ułatwiająca pracę z Wayback Machine i zwiększająca użyteczność tego archiwum w pracy badawczej. Wyszukiwanie pełnotekstowe wspiera także wykorzystywanie wieloznaczników. Przykładowo, aby odnaleźć w zbiorach Wayback Machine archiwalne wersje serwisów internetowych firm czy instytucji mieszczących się w Warszawie, można posłużyć się poniższym zapytaniem, które wygeneruje wynik z listą dostępnych hostów[10].

[6] Kopia strony głównej portalu Onet.pl z 1 marca 2001 roku: *Wayback Machine Internet Archive* [online]. Wayback Machine. [Dostęp 26.04.2017]. Dostępny w: https://web-beta.archive.org/web/20010301200222/http://onet.pl/.

[7] Zob. *About the Time Travel Service* [online]. [Dostęp 26.04.2017]. Dostępny w: http://timetravel.mementoweb.org/about/; WILKOWSKI, M. *Performatywne archiwa Webu i ich ograniczenia* [online]. [Dostęp 26.04.2017]. Dostępny w: http://wilkowski.org/notka/1274.

[8] Wayback Machine API. W: *Wayback Machine Internet Archive* [online]. W: Wayback Machine Internet Archive. [Dostęp 26.04.2017]. Dostępny w: https://archive.org/help/wayback_api.php.

[9] *Wayback Machine Internet Archive* [online]. Wayback Machine. [Dostęp 26.04.2017]. Dostępny w: https://web-beta.archive.org.

[10] Lista *websites* zawierających frazę „Warszawa, ul. *" na swojej stronie głównej: *Wayback Machine Internet Archive* [online]. Wayback Machine. [Dostęp 26.04.2017]. Dostępny w: https://web-beta.archive.org/web/*/Warszawa,%20ul.%20*.

Warszawa, ul. *

Wyszukiwarka nie przeszukuje kodu źródłowego strony głównej, a jedynie jej widoczną w przeglądarce postać. We frazach wyszukiwania nie można także wykorzystywać wyrażeń regularnych.

Drugą nowością w wersji beta Wayback Machine jest automatyczne generowanie podsumowań dla poszczególnych archiwizowanych hostów. Strony podsumowań[11] zawierają informacje o charakterystyce zawartości archiwalnych kopii (statystyki liczby plików graficznych, plików HTML, kodów JS itp.), liczbie wykonanych archiwizacji danego hosta i liczbie pozyskanych w ich ramach adresów URL. Dane prezentowane są w ujęciu rocznym.

Podsumowanie

Wayback Machine to największe dostępne dziś archiwizacji sieci. Ponieważ w odróżnieniu od wielu innych projektów tego typu archiwizuje web w wymiarze globalnym (nie ograniczając się wyłącznie do określonych domen krajowych) i posiada zbiory już z drugiej połowy lat 90. XX wieku, jego znaczenie w pracy badawczej jest nie do przecenienia. Aby poprawnie i skutecznie pozyskiwać i analizować zgromadzone tam zasoby, należy poznać podstawy organizacji archiwum, wykorzystywaną w nim terminologię oraz ograniczenia wynikające z przyjętej metody archiwizacji webu (*web crawlingu*).

Dla nowo projektowanych archiwów Webu Wayback Machine jest przykładem nowoczesnego archiwum cyfrowego, otwartego na użytkownika i przykładem na ponowne wykorzystanie publikowanych tam zbiorów, które dodatkowo udostępniane są także poza interfejsem graficznym. Możliwość skorzystania z Wayback Machine przez interfejs programistyczny (API) zdecydowanie ułatwia pracę z obszernymi zbiorami tego archiwum.

Bibliografia:

1. *About the Time Travel Service* [online]. [Dostęp 26.04.2017]. Dostępny w: http://timetravel.mementoweb.org/about/.
2. *Wayback Machine Internet Archive* [online]. Wayback Machine. [Dostęp 26.04.2017]. Dostępny w: https://web-beta.archive.org.
3. BRÜGGER, N. Web History and the Web as a Historical Source. *Zeithistorische Forschungen* [online]. 2012, 9 (2), s. 316–325. [Dostęp 26.04.2017]. Dostępny w: http://www.zeithistorische-forschungen.de/2-2012/id=4426.
4. GOEL, V. Defining Web pages, Web sites and Web captures. W: *Internet Archive Blogs* [online]. Internet Archive Blogs. [Dostęp 26.04.2017]. Dostępny w: https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/.
5. GRAHAM, M. FAQs for some new features available in the Beta Wayback Machine. W: *Internet Archive Blogs* [online]. Internet Archive Blogs. [Dostęp 26.04.2017]. Dostępny w: https://blog.archive.org/2016/10/24/faqs-for-some-new-features-available-in-the-beta-wayback-machine/.
6. KAHLE, B. Preserving the Internet. *Scientific American* 1997, 276(3), s. 82–83. ISSN 0036-8733.

[11] Np. dla hostu prezydent.pl: https://web-beta.archive.org/details/prezydent.pl.

7. ROSSI, A. Robots.txt Files and Archiving .gov and .mil Websites. W: *Internet Archive Blogs* [online]. Internet Archive Blog. [Dostęp 26.04.2017]. Dostępny w: https://blog.archive.org/2016/12/17/robots-txt-gov-mil-websites/.
8. Wayback Machine API. W: *Wayback Machine Internet Archive* [online]. W: Wayback Machine Internet Archive. [Dostęp 26.04.2017]. Dostępny w: https://archive.org/help/wayback_api.php.
9. WILKOWSKI, M. *Performatywne archiwa Webu i ich ograniczenia* [online]. [Dostęp 26.04.2017]. Dostępny w: http://wilkowski.org/notka/1274.