EBIB Biuletyn EBIB, nr 6 (183)/2018,
Współpraca bibliotek z naukowcami
Badania, teorie, opinie

Marcin Wilkowski
Uniwersytet Warszawski
Centrum Kompetencji Cyfrowych
m.wilkowski@uw.edu.pl

Archiwa webu otwarte na współpracę z użytkownikami?

Streszczenie: Instytucje archiwizujące web oraz środowisko naukowe są równorzędnymi partnerami w rozwijaniu teorii archiwalnej webu oraz metod jego zachowywania i badania. Archiwa webu wciąż jednak nie otwierają się na wiedzotwórczy wkład użytkowników w ich tworzenie i rozwijanie. W artykule naszkicowano obraz współczesnej archiwistyki webu w kontekście współpracy między instytucjami i środowiskiem akademickim a jednostkami i społecznościami zainteresowanymi zabezpieczaniem wybranych historycznych zbiorów WWW i korzystaniem z ich archiwów.

Słowa kluczowe: archiwizacja internetu, archiwizacja webu, World Wide Web, WWW, folksonomia

Biblioteki akademickie i narodowe podejmują zróżnicowane inicjatywy związane z archiwizacją World Wide Web (WWW). Obok regularnego pozyskiwania i zabezpieczania domeny krajowej przez biblioteki narodowe, np. w Austrii, Wielkiej Brytanii, Portugalii, Nowej Zelandii czy USA, projekty archiwizacji wybranych zasobów webu prowadzone są przez biblioteki akademickie w ramach ich stałych zadań. Projekty takie są też realizowane przez pracowników naukowych w ramach grantów badawczych, przy wsparciu bibliotek[1]. Wsparcie ze strony bibliotek może polegać np. na udostępnianiu oprogramowania, komputerów i nośników niezbędnych do przeprowadzenia archiwizacji[2] oraz wyszukiwania w zgromadzonych zbiorach czy na udzielaniu dostępu do zbiorów archiwalnego webu, przechowywanych i zarządzanych przez instytucję[3]. Wsparciem skierowanym do środowiska akademickiego jest też działalność edukacyjna, polegająca na przekazywaniu podstawowej wiedzy i kompetencji związanych z archiwizacją webu i korzystaniem z archiwów tego typu[4] czy pomoc w planowaniu badań i pisaniu wniosków na projekty badawcze wykorzystujące te zasoby.

Instytucje zajmujące się archiwizacją webu oraz środowisko naukowe są równorzędnymi partnerami w rozwijaniu podstaw teoretycznych archiwizacji oraz metod zachowywania i badania webu. Pionierską i niezwykle istotną rolę w tym zakresie odgrywa też Internet Archive jako organizacja pozarządowa (fundacja). Wśród zaangażowanych w archiwizacje webu znajdziemy także podmioty działające na rynku komercyjnym – jednym z członków Międzynarodowego Konsorcjum Archiwizacji Internetu (International Internet Preservation Consortium, IIPC) jest brytyjskie MirrorWeb LTD[5], które ściśle współpracuje z UK Web Archive przy archiwizacji domeny rządowej Wielkiej Brytanii.

[1] Badania w ramach grantów nie muszą być związane z dziedziną archiwistyki webu, ale mogą to być także badania, dla których niezbędne jest skorzystanie z archiwalnych zbiorów WWW.
[2] *About Netarkivet* [online]. [Dostęp 12.12.2018]. Dostępny w: http://netarkivet.dk/in-english/.
[3] Np. w Wielkiej Brytanii dostęp do części archiwów domeny krajowej możliwy jest wyłącznie na miejscu w wybranych instytucjach. W Austrii ograniczenie to dotyczy wszystkich zarchiwizowanych zbiorów.
[4] Zob. np. zaproszenie na warsztat z archiwizacji stron WWW odbywający się w bibliotece akademickiej Uniwersytetu Pensylwanii https://twitter.com/boriarchivista/status/1067510414077042688.
[5] IIPC members [online]. [Dostęp 12.12.2018]. Dostepny w: http://netpreserve.org/about-us/members/.

Chociaż wydawać by się mogło, że praktykowanie archiwistycznego, historycznego podejścia do zasobów webu jest domeną środowisk profesjonalnych, można wskazać przynajmniej trzy fakty mocno przeczące temu przypuszczeniu. Po pierwsze, jedne z największych archiwalnych zbiorów webu, gromadzone na serwerach Internet Archive, pochodzą – ostatnio w coraz większym stopniu[6] – od zwykłych użytkowników internetu, którzy w łatwy sposób mogą zabezpieczać wybrane strony WWW za pomocą usługi Save Page Now, wklejając link do formularza dostępnego na stronie Wayback Machine lub korzystając z otwartego API. Po drugie, od 2015 r. realizowany jest program *Documenting The Now*[7] w celu demokratyzacji dostępu do narzędzi i zbiorów archiwalnego webu przy zachowywaniu profesjonalnych standardów archiwizacji. Inicjatywa ta, rozwijana przez trzy amerykańskie uczelnie, skierowana jest nie tylko do badaczy, ale też do szerokiego grona osób pracujących w mediach czy pozarządowych, obywatelskich organizacjach strażniczych[8]. Po trzecie, jak wykazał to Savvas Zannettou i współautorzy, archiwizowanie webu oraz jego archiwa są narzędziami ułatwiającymi codzienną komunikację w ramach wielkich społeczności internetowych takich jak Reddit[9] czy 4chan[10], których użytkownicy korzystają z Wayback Machine czy Archive.it, aby omijać filtry blokujące publikowanie linków do określonych domen czy przytaczać treści z witryn usuniętych z powodów prawnych[11].

Archiwistyka webu jako dyscyplina badawcza oraz praktyka naukowa i bibliotekarska czy archiwalna wykształciła własne profesjonalne środowisko. Z drugiej strony zabezpieczanie wybranych stron i witryn staje się istotne dla zwykłych użytkowników internetu, środowiska dziennikarskiego czy organizacji pozarządowych, które korzystają z wypracowanych przez środowisko naukowe oraz instytucje metod i narzędzi. W innych dziedzinach wiedzy między tymi dwiema przestrzeniami (profesjonalną i obywatelską czy społeczną) rozwijają się wspólne inicjatywy o efektach użytecznych dla obu stron. Metoda wspierania badań i edukacji przy pomocy projektów nauki obywatelskiej nie jest już niczym nowym w naukach biologicznych czy astronomii, a crowdsourcing czy folksonomie (społecznościowe budowanie taksonomii) stały się użytecznymi narzędziami w rozwiązywaniu niektórych zadań badawczych. Czy archiwistyka webu – jako dziedzina wiedzy i praktyka badań – może korzystać ze wsparcia spoza środowiska profesjonalnego, poza wspomnianym już wyżej mniej lub bardziej świadomym rozbudowywaniem zasobów? Możliwość takiego wsparcia byłaby widoczna w projekcie archiwizacyjnym opartym o model folksonomii, w ramach którego odpowiedzialność za całość lub część opisu zbiorów czy relacji między nimi oddano by społeczności użytkowników.

---

[6] SUMMERS, E. (@edsu) [online]. [Dostęp 12.12.2018]. Dostępny w: https://twitter.com/edsu/status/1058788752749981698.

[7] *Documenting The Now* [online]. [Dostęp 12.12.2018]. Dostepny w: https://www.docnow.io/.

[8] Sieci obywatelskie podejmujące działania kontrolujące władzę (tzw. watchdogs).

[9] Reddit – serwis internetowy przedstawiający linki do różnorodnych informacji, które ukazały się w internecie. Serwis jest głównie anglojęzyczny, chociaż interfejs serwisu jest przetłumaczony na wiele języków (w tym polski). Za: Reddit [online]. Wikipedia : wolna encyklopedia, 28.10.2018. [Dostęp 12.12.2018]. Dostępny w: https://pl.wikipedia.org/w/index.php?title=Specjalna:Cytuj&page=Reddit&id=54863908

[10] 4chan – amerykański imageboard, klon japońskiego Futaba Channel. Uruchomiony w październiku 2003 r. przez nastolatka Christophera Poole'a, ps. moot. Pierwotnie o tematyce mangi oraz anime. Użytkownicy z reguły piszą oraz zamieszczają obrazy anonimowo. Za: 4chan [online]. Wikipedia : wolna encyklopedia, 28.10.2018. [Dostęp 12.12.2018]. Dostępny w: https://pl.wikipedia.org/wiki/4chan.

[11] ZANNETTOU, S., BLACKBURN, J. DE CRISTOFARO, E., SIRIVIANOS, M., STRINGHINI, G. *Understanding Web Archiving Services and Their (Mis) Use on Social Media* [preprint]. W: arXiv.org [online]. [Dostęp 12.12.2018]. Dostępny w: https://arxiv.org/abs/1801.10396.

Folksonomia była jedną z obietnic Web 2.0 – koncepcji reprezentującej optymizm związany ze zmianą modeli działania instytucji wiedzy. Zmiana ta miała być wywołana głównie środkami technicznymi, odpowiadać za nią miały również zmiana paradygmatu witryny internetowej (przejście od płaskiej i statycznej strony do platformy) oraz uspołecznienie się webu, czyli powstawanie dynamicznych, trwałych i inkluzywnych wspólnot online oraz niemal powszechny dostęp do internetu. Folksonomia w bibliotece oznaczać miała istnienie wokół tej instytucji społeczności gotowej nieustannie rozbudowywać indeksy zasobów, czy to przez włączanie do nich własnych opisów, tagów i uwag, czy też rozwijanie osobnych indeksów i taksonomii uzupełniających potencjał standardowych katalogów. W pozytywnej opowieści na ten temat nie było jednak miejsca ani na brak środków i narzędzi, którymi biblioteki miałyby rozwijać takie społeczności, ani tym bardziej na krytyczne uwagi dotyczące wad i ograniczeń społecznościowej taksonomii jako metody zarządzania wiedzą. Problem braku słowników, nadmiarowych pojęć, niejednoznaczności czy wpływu liderów społeczności[12] to przecież tylko niektóre z nich.

Folksonomia jest jednak wartościową metodą opracowywania specyficznych rodzajów zasobów. Jak piszą Honk Yu i współautorzy *szczególnie w przypadku zasobów multimedialnych, takich jak muzyka, zdjęcia czy filmy,* [społecznościowe] *tagowanie zasobów jest jedynym możliwym sposobem organizacji danych multimedialnych i umożliwienia ich przeszukiwania*[13]. Wydawać by się mogło, że podobnie będzie w przypadku archiwalnych zbiorów webu i biblioteki je gromadzące będą ochoczo korzystać ze wsparcia użytkowników opisujących poszczególne witryny i strony. Mogliby oni wyodrębniać z dużych zbiorów poszczególne elementy (kolekcje, obiekty itp.), opisywać ich zawartość za pomocą słów kluczowych czy uzupełniać i poprawiać opisy dodane pierwotnie przez instytucję. Trudno jednak wskazać skuteczny projekt tego typu. Dlaczego?

Po pierwsze, w odróżnieniu od multimediów, strony WWW mogą być natywnie przeszukiwalne, tzn. łatwo można wyszukiwać w ich treści, bez pomocy innych programów. Poza niektórymi wyjątkami (np. strony budowane we Flashu czy dynamicznie ładujące zawartość przez JavaScript) treść, struktura strony (obecne w niej tagi HTML definiujące sekcje, np. tytuł) i metadane mogą być automatycznie przeszukiwane na tej samej zasadzie, co treść innych dokumentów tekstowych. Oczywiście, wyszukiwanie pełnotekstowe, jako metoda gromadzenia wiedzy na bazie zasobu, różni się radykalnie od przeglądania za pomocą kategorii, tagów itp., pozwala jednak na automatyczne generowanie taksonomii – odpowiednie kwerendy umożliwiają wyodrębnić odpowiednio sprofilowane metadane ze zbioru czy nawet automatycznie kategoryzować treść[14]. Jednocześnie prawidłowo i zgodnie ze standardem wykonane kopie stron i witryn, przechowywane w formacie WARC, pozwalają badaczom na szerokie kwerendy za pomocą wybranego języka programowania (np. Pythona i biblioteki PyWARC[15]).

[12] HU, J., WANG, B., LIU, Y., LI, D. Y. Personalized tag recommendation using social influence. *Journal of Computer Science and Technology* 2012, 27 (3), 527–528.
[13] YU, H., ZHOU, B., DENG, M., HU, F. Tag recommendation method in folksonomy based on user tagging status. *Journal of Intelligent Information Systems* 2018, 480.
[14] Możliwe jest także profilowanie samych archiwów Webu pod kątem udostępnianej przez nich zawartości. Zob. ALSUM, A., WEIGLE, M. C., NELSON, M. L., VAN DE SOMPEL, H. Profiling web archive coverage for top-level domain and content language, *International Journal on Digital Libraries* 2018, 14 (3–4), 149–166, doi: 10.1007/s00799-014-0118-y.
[15] PyWarc [oprogramowanie] [online]. [Dostęp 12.12.2018]. Dostępny w: https://github.com/cllu/PyWARC.

Po drugie, w przypadku analizy zbiorów archiwalnego webu w celach badawczych, oczekiwane przez użytkowników taksonomie mogą różnić się między sobą w zależności od profilu badania. Osoba badająca takie zbiory pod kątem struktury dokumentów HTML czy sieci wzajemnych połączeń między nimi może potrzebować taksonomii zupełnie innej od tej, którą należałoby wypracować do badania analizującego semantyczną treść zbiorów.

Po trzecie, zbiory archiwów webu gromadzone przez profesjonalne instytucje, organizowane są w bardzo zróżnicowany sposób, niekiedy nawet zupełnie ignorujący poziom metadanych i taksonomii. Przykładowo, zasoby Internet Archive prezentowane są bez jakiejkolwiek wyraźnej struktury metadanych[16], zbiory gromadzone w usłudze Archive-It opisywane są metadanymi Dublin Core bez dodatkowych tagów, ale przynajmniej z częściowo zamkniętymi i standardowo wykorzystywanymi słownikami[17]. W UK Web Archives poszczególne witryny opisywane są łącznie w ramach kolekcji i bez metadanych[18], podczas gdy archiwum webu w Bibliotece Kongresu prezentuje metadane archiwaliów w schemacie Metadata Object Description Schema (MODS)[19]. Żadne z tych archiwów nie zezwala na dodawanie przez użytkowników własnych opisów i tagów.

Warto zbadać, czy dzieje się to z powodu ograniczeń wynikających ze stosowanego oprogramowania do pozyskiwania i udostępniania zbiorów, skali danych czy przeświadczenia o tym, że to użytkownicy sami powinni wypracować sobie własne zasady opisu interesujących ich zasobów. Jednak pobieżna analiza narzędzi przeszukiwania i przeglądania zbiorów dostępnych wśród kilku projektów to zbyt mało na wyciągnięcie ogólnych wniosków. Przeprowadzone w roku 2011 badania ankietowe instytucji archiwistyki webu wykazały, że 89% archiwów pozwala na przeszukiwanie zbiorów za pomocą adresów URL, a 79% umożliwia przeszukiwanie w polach metadanych w wybranym schemacie[20]. Także ArchiveWeb[21], aplikacja udostępniona użytkownikom usługi Archive-It, umożliwia szerokie społecznościowe opracowywanie zbiorów webu, chociaż jej zasięg jest ograniczony wyłącznie do osób z zespołów bibliotecznych i archiwalnych, opłacających abonament w Archive-It. Trudno mówić tu o folksonomiach w skali znanej z funkcjonowania Wikipedii czy serwisu Flickr. Z punktu widzenia otwartości na twórczy wkład użytkowników w opis archiwalnych zbiorów webu wyróżnia się zdecydowanie inicjatywa Archive Team, nieformalna grupa wsparcia dla działań archiwalnych, pierwotnie zorganizowana wokół fundacji Internet Archive. Społeczność ta rozwija własną wiki, na której dokumentuje stan zachowania zagrożonych usunięciem/zniknięciem zasobów webu, głównie dużych, popularnych witryn[22] i prowadzi także społecznościową archiwizację tych zasobów w oparciu o narzędzia Internet Archive (Wayback Machine). Być może to najbardziej wyraźny przykład społecznościowego, ale również jakościowego zaangażowania w archiwistykę webu.

[16] https://web.archive.org/web/20131021165347/http://www.imdb.com/ – do dyspozycji jest timestamp itp.
[17] Np. dla pola format https://www.archive-it.org/collections/4399.
[18] Zob. https://www.webarchive.org.uk/en/ukwa/collection/44.
[19] Zob. https://www.loc.gov/item/lcwa00085389.
[20] GOMES, D., MIRANDA, J., COSTA, M. A survey on web archiving initiatives. W: *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries: Research and Advanced Technology for Digital Libraries (TPDL '11)*, 2011, s. 1045–1050.
[21] FERNANDO, Z. T., MARENZI, I., NEJDL, W. ArchiveWeb: collaboratively extending and exploring web archive collections – How would you like to work with your collections? *International Journal on Digital Libraries* 2017, 19 (1), 39–55. doi:10.1007/s00799-016-0206-2.
[22] *Archive Team wiki* [online]. [Dostęp 12.12.2018]. Dostępny w: https://www.archiveteam.org/index.php?title=Main_Page.

Analiza polityk bibliotek archiwizujących WWW wobec opisu tych zbiorów wymagałaby uwzględnienia również procesu przyjmowania od użytkowników zgłoszeń kolejnych adresów do archiwizacji. Być może takie działanie, a nie folksonomia, jest bardziej użytecznym i oczekiwanym sposobem otwierania się ich na twórczy wkład odbiorców. Należałoby przy tym zbadać, jakich dokładnie danych i jakiej wiedzy oczekują instytucje archiwizujące web od użytkowników zgłaszających strony do zabezpieczenia. Np. Internet Archive żąda tylko jednej informacji – URL strony, którą należy zachować. Czy inne otwarte archiwa webu wymagają od osób zgłaszających linki jakichkolwiek rozbudowanych informacji[23]?

Zadawanie pytań o potencjał społecznościowego wsparcia archiwizacji webu ma sens nie tylko dla rozwoju jego podstaw teoretycznych czy badań społecznych nad partycypacją w cyfrowym dziedzictwie. Ponieważ archiwistyka webu to dziedzina nieustannie się zmieniająca (co poniekąd wymuszane jest ciągłymi zmianami technicznymi WWW, ale też zmianą w tym, jak definiuje się i określa granice dziedzictwa cyfrowego), wiedza o możliwościach wykorzystania społeczności przy gromadzeniu i opracowywaniu zbiorów webu może stać się kluczowa w planowaniu strategii archiwizacyjnych bibliotek i archiwów. I choć przy digitalizacji i opracowywaniu klasycznych zbiorów dziedzictwa zaznaczył się już mocno wkład społeczności i inicjatyw oddolnych (archiwa społeczne, program GLAM-Wiki itp.), archiwistyka webu to wciąż zadanie zdominowane przez środowisko profesjonalne.

Artykuł powstał w ramach projektu: „Upowszechnianie wiedzy o archiwizacji Webu i metodach korzystania z historycznych zasobów WWW w instytucjach publicznych i sektorze NGO” – zadanie finansowane w ramach umowy 868/P-DUN/2018 ze środków Ministra Nauki i Szkolnictwa Wyższego przeznaczonych na działalność upowszechniającą naukę.

Bibliografia

1. *About Netarkivet* [online]. [Dostęp 12.12.2018]. Dostępny w: http://netarkivet.dk/in-english/.
2. ALSUM, A., WEIGLE, M. C., NELSON, M. L., VAN DE SOMPEL, H. Profiling web archive coverage for top-level domain and content language, *International Journal on Digital Libraries* 2018, 14 (3–4), 149–166, doi: 10.1007/s00799-014-0118-y.
3. *Documenting The Now* [online]. [Dostęp 12.12.2018]. Dostepny w: https://www.docnow.io/.
4. FERNANDO, Z. T., MARENZI, I., NEJDL, W. ArchiveWeb: collaboratively extending and exploring web archive collections – How would you like to work with your collections? *International Journal on Digital Libraries* 2017, 19 (1), 39–55. doi:10.1007/s00799-016-0206-2.
5. GOMES, D., MIRANDA, J., COSTA, M. A survey on web archiving initiatives. W: *Proceedings of the 15th International Conference on Theory and Practice of Digital Libraries: Research and Advanced Technology for Digital Libraries (TPDL ’11)*, 2011, s. 1045–1050.
6. HU, J., WANG, B., LIU, Y., LI, D. Y. Personalized tag recommendation using social influence. *Journal of Computer Science and Technology* 2012, 27 (3), 527–528.
7. YU, H., ZHOU, B., DENG, M., HU, F. Tag recommendation method in folksonomy based on user tagging status. *Journal of Intelligent Information Systems* 2018, 480.
8. ZANNETTOU, S., BLACKBURN, J. DE CRISTOFARO, E., SIRIVIANOS, M., STRINGHINI, G. *Understanding Web Archiving Services and Their (Mis) Use on Social Media* [preprint]. W: arXiv.org [online]. [Dostęp 12.12.2018]. Dostępny w: https://arxiv.org/abs/1801.10396.

WILKOWSKI, M. Archiwa webu otwarte na współpracę z użytkownikami? *Biuletyn EBIB* [online] 2018, nr 6 (183), Współpraca bibliotek z naukowcami. [Dostęp 18.12.2018]. ISSN 1507-7187. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/view/674.

[23] Np. w procesie zgłaszania strony do archiwizacji w UK Web Archive osoba zgłaszająca musi podać swoje podstawowe dane (imię i nazwisko, adres email) oraz wybrany adres URL. Podanie informacji kontekstualizujących archiwizowaną stronę nie jest obowiązkowe. Zob. https://www.webarchive.org.uk/en/ukwa/info/nominate.