Nr 9 (11) / 2018, s. 49–62

ARCHIWA – KANCELARIE – ZBIORY

Data przesłania artykułu: 31 I 2018 r.
Data przyjęcia artykułu do druku: 19 VI 2018 r.
DOI: http://dx.doi.org/10.12775/AKZ.2018.003

Bartłomiej Konopa
(Archiwum Państwowe w Bydgoszczy)

ARCHIWA INTERNETU JAKO NOWE BAZY ŹRÓDŁOWE

Słowa kluczowe
Internet Archive; PANDORA Archive; UK Web Archive; CyberCemetery; archiwa Internetu; archiwizacja Internetu

Keywords
Internet Archive; PANDORA Archive; UK Web Archive; CyberCemetery; Internet archives; Internet archiving

Streszczenie
W artykule podjęte zostały rozważania nad archiwami Internetu jako specyficzny twór wydzielony różnorodnych projektów archiwizacji zasobów sieciowych. Autor zaprezentował kilka przykładowych inicjatyw: Internet Archive, PANDORA Archive, UK Web Archive, CyberCemetery oraz '09 European Election Web Harvesting Project. Na ich podstawie oraz definicji spróbował on scharakteryzować to zjawisko, wymieniając elementy, które składają się na archiwum Internetu. Na koniec przedstawił dostępne kry-

Bartłomiej Konopa. Tytuł zawodowy magistra na kierunku Archiwistyka i Zarządzanie Dokumentacją na Uniwersytecie Mikołaja Kopernia w Toruniu. Obecnie pracowni Archiwum Państwowego w Bydgoszczy i uczestnik I roku studiów doktoranckich w zakresie historii na Wydziale Nauk Historycznych UMK w Toruniu. Jego zainteresowania badawcze obejmują przede wszystkim archiwizację Internetu, a także historię Torunia, źródłoznawstwo oraz teorię archiwistyki.
E-mail: bartlomiejkonopa@gmail.com
ORCID ID: 0000-0001-9843-5552

teria ich podziału oraz refleksje dotyczące ich możliwości oraz posiadanego przez nie potencjału dla badaczy.

Summary

Internet archives as new source bases

The article contemplates Internet archives as specific creations dedicated to various Web archiving projects. The Author presents several examples of such initiatives: the Internet Archive, the PANDORA Archive, the UK Web Archive, the CyberCemetery, and the '09 European Election Web Harvesting Project. On the ground of these examples, as well as the definition of Internet archive, the Author attempted to characterise the phenomenon by listing elements that comprise an Internet archive. Finally, he presented possible criteria for their classification, as well as some reflections on their capabilities and research potential.

Zawsze gdy zaczyna mówić się o archiwizacji Internetu zwraca się uwagę na rolę jego we współczesnym świecie. Tak jak teoretyk hipertekstu i badacz nowych mediów Jay David Bolter uznał komputer za technologię definiującą koniec XX w., a więc wynalazek o szczególnym znaczeniu dla ludzkości oraz istotnym wpływie na kulturę i naukę (wskazać tu można m.in. na porównywanie ludzkiego mózgu do komputera)[1], za takową dla początków XXI w. uznać możemy Sieć. Wśród licznych i wielkich odkryć i wynalazków, chociażby w dziedzinie fizyki lub medycyny, to właśnie ona w największym stopniu oddziałuje na przeciętnego człowieka, i to dzięki niej możemy mówić dziś o istnieniu społeczeństwa informacyjnego. W wyniku gwałtownego rozwoju technologii mobilnych, dostęp do Internetu możliwy jest nie tylko przez komputery stacjonarne, ale także przez laptopy i smartfony, które wielu ludzi nosi przy sobie w kieszeni. Posiadając możliwość połączenia się z nim dwadzieścia cztery godzinę na dobę, siedem w dni w tygodniu, ludzie przenoszą do niej coraz większą część swojej aktywności. Dziś w Sieci prowadzi się dyskusje polityczne, publikuje prace literackie i naukowe, pracuje, realizuje własne hobby i zainteresowania, robi zakupy, socjalizuje etc.[2] Internet zdaje się być gotowy przejąć niemal wszystkie możliwe aspekty życia człowieka, tak więc nie może dziwić przyrównywanie

[1] J. D. Bolter, *Człowiek Turinga. Kultura Zachodu w wieku komputera*, tłum. i wstęp T. Goban-Klas, Warszawa 1990, s. 35–40.

[2] Zob. A. Rosa, *Human trace on the Internet – the issue of archiving the Web from the point of view of antropology-oriented archival science*, „Archiwa – Kancelarie – Zbiory", t. 6 (8), 2015, s. 193–205.

go do archiwum dokumentującego jego działalności. Przedstawiciele nauk humanistycznych i społecznych, którzy ową działalność badają i w tych badaniach korzystają z nowego rodzaju źródeł *born-digital*, wykazują coraz większe zainteresowanie Internetem. Sieć i jej zawartość stają się ważnym elementem dziedzictwa kulturowego całej naszej współczesnej cywilizacji[3].

Zawartość Internetu rozrasta się w niesłychanym tempie, co można zobrazować następującymi liczbami: według danych za rok 2014 w ciągu zaledwie 60 sekund zarejestrowano 70 nowych domen, powstało 571 nowych stron, na Facebooku zamieszczono 293 tys. postów, na Twitterze pojawiło się 433 tys. nowych statusów oraz wysłano ponad 140 mln e-maili[4]. Wraz z rozwojem technologicznym oraz wciąż poszerzającym się dostępem do Sieci, spodziewać się można, że przytoczone dane za rok 2017 będą zdecydowanie większe. W tym miejscu należy rozprawić się z mitem, który głosi, że „to co zostało zamieszczone w Internecie pozostanie w nim na zawsze". Najnowsze badania wykazują, że przeciętna strona internetowa działa się od 90 do 100 dni, a następnie ulega zmianom lub zostaje usunięta. To samo tyczy się zawartości mediów społecznościowych, gdzie co roku tracone jest około 10% zawartości związanej z różnymi wydarzeniami[5]. Ta niezwykła płynność i niestałość Internetu jest niebezpieczna, zwłaszcza w przypadku coraz popularniejszych w cytowaniach hiperłączy. Badania przeprowadzone w 2014 r. w Narodowym Laboratorium w Los Alamos i na Uniwersytecie w Edynburgu pod kierownictwem Martina Kleina wykazały, że blisko ¼ przeanalizowanych artykułów z 2012 r. zawiera nieaktualne linki. W przypadku starszych artykułów procent dotknięty tym zjawiskiem jest większy, chociażby w bazie PubMed Central dla tekstów z 1997 r. wyniósł aż 80%[6]. Doniosła rola Internetu w dzisiejszych czasach, bogactwo i różnorodność jego zasobów wraz z jego zmiennością i dynamicznym rozwojem prowadzą do

---

[3] A. Sobczak, *Internet jako globalne archiwum społeczne – rozważania na temat roli Internetu w dokumentowaniu dziejów ludzkości*, [w:] *Toruńskie konfrontacje archiwalne*, t. 4: *Nowa archiwistyka - archiwa i archiwistyka w ponowoczesnym kontekście kulturowym*, red. W. Chorążyczewski, W. Piasek, A. Rosa, Toruń 2014, s. 238–242.

[4] *Online in 60 seconds. A year later* [infografika], http://blog.qmee.com/wp-content/uploads/2014/07/infographic-resized1.jpg (dostęp 11.11.2017).

[5] Dane pochodzą z *Archiving&Preserving Web content* [prezentacja multimedialna], http://www.coppul.ca/sites/default/files/uploads/COPPUL%20and%20OCUL%20Archive-It%20Informational%20webinar%20slides.pdf (dostęp 11.11.2017).

[6] M. Klein, H. Van de Sompel, R. Sanderson, H. Shankar, L. Balakireva, K. Zhou, R. Tobin, *Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot*, PLoS ONE 9(12): e115253. https://doi.org/10.1371/journal.pone.0115253 (dostęp 24.12.2017).

dość oczywistych wniosków. Należy dążyć do jego zachowania, zarówno dla celów bieżących (np. cytowania lub dowody w sprawach sądowych), jak i dla obecnych i przyszłych uczonych, którzy mogą chcieć wykorzystywać archiwalne witryny internetowe w swoich badaniach naukowych nad współczesnym nam człowiekiem.

Podobne konkluzje wypracowano już ponad 20 lat temu, gdyż w roku 1996 zaczęły powstawać pierwsze projekty archiwizacji Internetu, w tym największe obecnie tego rodzaju przedsięwzięcie w postaci Internet Archive, które pierwszą witrynę zarchiwizowało w maju wspomnianego roku. Według danych zawartych w Wikipedii aktualnie funkcjonuje około 80 takich inicjatyw[7], dołączyć do nich można również chociażby blisko 5 tys. różnych kolekcji zgromadzonych w ramach należącej do Internet Archive komercyjnej usługi Archive-It, która pozwala gromadzić zasoby sieciowe instytucjom nieposiadającym odpowiedniego ku temu zaplecza[8], inne tego rodzaju usługi oraz mniejsze, przez co trudniej dostrzegalne, projekty. Wraz z rozwojem zainteresowania tą problematyką i uświadomieniem sobie potrzeby zabezpieczania zawartości Sieci, zaczęły powstawać organizacje, które pomagają w konsolidacji działań podejmowanych przez instytucje rozsiane po całym globie. Najważniejszą tego rodzaju inicjatywą było założenie w 2003 r. International Internet Preservation Consortium (IIPC), które obecnie zrzesza 54 członków, głównie biblioteki narodowe, ale także archiwa i organizacje non-profit. Jego zadaniem jest wspieranie zainteresowanych instytucji w archiwizacji zasobów internetowych poprzez dostarczanie metod i narzędzi, przygotowanie konferencji i różnych wydarzeń, które umożliwiłyby wymianę wiedzy i doświadczeń oraz inne aktywności[9].

Z technicznego punktu widzenia archiwizacja Webu, w swojej najbardziej popularnej odmianie, polega na zautomatyzowanym gromadzeniu zrzutów stron internetowych przy użyciu przystosowanych do tego robotów przeszukujących Sieć (crawlerów) i przechowywanie ich w przestrzeni dyskowej, a następnie udostępnianie za pomocą odpowiednio przystosowanych narzędzi[10]. W tym

---

[7] *List of Web archiving initiatives*, https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives (dostęp 11.11.2017).

[8] *Archive-It Blog – About Us*, https://archive.org/about/ (dostęp 11.11.2017).

[9] K. Ślaska, A. Wasilewska, *Archiwizacja Internetu — sytuacja w polskim prawie z punktu widzenia bibliotekarzy*, Biuletyn EBIB, nr 1 (128)/2012, s. 2, http://www.ebib.pl/images/stories/numery/128/128_slaska.pdf (dostęp 11.11.2017).

[10] L. Derfert-Wolf, *Archiwizacja Internetu – wprowadzenie i przegląd wybranych inicjatyw*, Biuletyn EBIB, nr 1 (128)/2012, s. 8–9, http://www.ebib.pl/images/stories/numery/128/128_derfert.pdf (dostęp 11.11.2017).

momencie można zadać pytanie czym są archiwa Internetu, które powstają ramach tego procesu? Czy są to instytucje lub projekty powołane w tym celu, a może miejsca przechowywania albo portale internetowe służące upublicznianiu zrzutów, a może tylko określone zarchiwizowane już zasoby? Aby móc na nie odpowiedzieć konieczne będzie wskazanie lub sformułowanie odpowiedniej ich definicji. W dalszej kolejności można zastanawiać się nad systematyką archiwów Webu i jej kryteriami, nad charakterem ich zasobów oraz możliwym wykorzystaniem m.in. dla potrzeb naukowych. Przed próbą sformułowania odpowiedzi na zadane przed chwilą pytania, należy scharakteryzować kilka przykładowych projektów archiwizacji zasobów Webu.

Jako pierwszy przykład, najbardziej oczywisty, posłuży wspomniane już wcześniej Internet Archive[11]. Instytucja ta jest organizacją non-profit założoną przez Brewster'a Kahle w 1996 r. w San Francisco, która określa się mianem cyfrowej biblioteki (a niekiedy także archiwum) z misją uniwersalnego dostępu do całej wytworzonej przez ludzi wiedzy. Oprócz swojej podstawowej działalności jaką jest gromadzenie archiwalnych wersji witryn internetowych, w swoich zbiorach posiada ona także zdigitalizowane książki, nagrania dźwiękowe, wideo, obrazy oraz oprogramowanie komputerowe. Archiwizacją Webu zajmuje się od momentu swojego powstania, jednak zebrany w ten sposób zasób został udostępniony dopiero w 2001 r., gdy uruchomiono platformę Wayback Machine. Dano w ten sposób wszystkim użytkownikom World Wide Web wgląd do największej do tej pory kolekcji zrzutów witryn internetowych, która według danych z 2016 r. posiadała ich 361 mln[12]. Swoim zasięgiem Internet Archive stara się objąć możliwie największą część publicznie dostępnych zasobów sieciowych, do czego wykorzystuje crawlera Heritrix, a także opcję „Save Page Now", która pozwala użytkownikom samodzielnie wskazać witrynę do archiwizacji poprzez podanie jej adresu URL. Przy doborze materiałów nie są wykorzystywanie żadne dodatkowe kryteria, a także nie pyta się właścicieli witryn o zgodę na wykonanie zrzutu. Możliwe jest jednak uchronienie się przed crawlerem poprzez

[11] *Internet Archive: Digital Library of Free Books, Movies, Music & Wayback Machine*, https://archive.org/ (dostęp 11.11.2017); o Internet Archive m.in. K. Gmerek, *Archiwa internetowe po obu stronach Atlantyku – Internet Archive, Wayback Machine oraz UK Web Archive*, Biuletyn EBIB, nr 1 (128)/2012, s. 1–6, http://www.ebib.pl/images/stories/numery/128/128_gmerek.pdf (dostęp 11.11.2017).

[12] V. Goel, *Defining Web pages, Web sites and Web captures*, Internet Archive Blogs, https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/ (dostęp 24.12.2017).

odpowiednie skonfigurowanie pliku robots.txt lub poprzez wystosowanie prośby do Internet Archive o usunięcie zgromadzonych materiałów.

Oprócz Wayback Machine Internet Archive oferuje komercyjną usługę Archive-It[13], uruchomioną w 2006 r., która adresowana jest głównie do organizacji i instytucji zainteresowanych stworzeniem własnej kolekcji archiwalnych zasobów internetowych. Za pomocą dostarczanych przez nią narzędzi możliwe jest zbudowanie swojego małego „archiwum", którego zasób zbierany jest zgodnie z obranym kluczem, a następnie przechowywany na serwerach Internet Archive. Jego dysponent decyduje o tematyce, zasięgu terytorialnym, języku, rodzaju i formacie gromadzonych materiałów, a także o dostępie do nich. Archive-It sugerowane jest jako rozwiązanie dla instytucji, które nie są w stanie w samodzielny sposób przeprowadzić archiwizacji zasobów sieciowych, a chciałyby uzupełnić swoje zasoby o tego rodzaju materiały. Przykładami takich kolekcji prowadzonych za pomocą tej usługi są dokumenty stanu Maryland, blogi mormońskie czy też witryny internetowe należące do Instytutu Smithsona. Swoje kolekcje, poświęcone wydarzeniom o znaczeniu międzynarodowym, tworzy samo Internet Archive.

Innym pionierskim projektem, również uruchomionym w 1996 r., jest PANDORA Archive[14], czyli archiwum „australijskiego" Internetu prowadzone przez tamtejszą bibliotekę narodową wraz z bibliotekami stanowymi i innymi instytucjami dziedzictwa. Profil jego działalności jest inny niż Internet Archive. Przede wszystkim ogranicza się ono wyłącznie do zasobów dotyczących Australii i Australijczyków, a ponadto gromadzi je w sposób selektywny i o tym co zostanie zarchiwizowane decydują pracownicy instytucji uczestniczących w projekcie. Inną istotną różnicą jest również konieczność uzyskania zgody na gromadzenie i późniejsze udostępnianie od właściciela danej witryny, ponieważ ten rodzaj materiałów w Australii nie jest objęty przepisami o egzemplarzu obowiązkowym. Kwestia ta w znaczny sposób utrudnia i spowalnia powiększanie zasobu, a niekiedy ogranicza także dostęp do niego w wyznaczonych w czytelniach stanowiskach. PANDORA Archive korzysta również z własnego oprogramowania – obecnie jest to PANDAS w wersji 3. opublikowanej w 2007 r. Najnowsze dane wykazują, iż zgromadzono ponad 51 tysięcy unikalnych witryn internetowych,

---

[13] *Archive-It – Web Archiving Services for Libraries and Archives*, https://archive-it.org/ (dostęp 12.11.2017); Archiving & Preserving Web content.

[14] *Pandora Archive – Preserving and Accessing Networked Documentary Resources of Australia*, http://pandora.nla.gov.au/ (dostęp 12.11.2017); *PANDORA Fact Sheet* [broszura], http://pandora.nla.gov.au/factsheet.doc (dostęp 12.11.2017); o PANDORA Archive m.in. L. Derfert-Wolf, *Archiwizacja Internetu – wprowadzenie*, s. 11–12.

z czego przeszło połowę stanowią zasoby o proweniencji rządowej. Nadmienić można również fakt, iż w roku 2004 PANDORA została wpisana na australijską listę Pamięci Świata.

Podobny charakter posiada UK Web Archive[15] prowadzone przez British Library od 2004 r. Do wprowadzonych w 2013 r. zmian w prawie o egzemplarzu obowiązkowym, również gromadziło witryny w sposób selektywny, także przy wykorzystaniu oprogramowania PANDAS. Dopiero wspomniane regulacje pozwoliły na gromadzenie zasobów sieciowych bez zgody ich właścicieli, co skutkowało zastosowaniem technologii udostępnianych przez Internet Archive i zmianę profilu działalności. Od tego momentu zaczęto archiwizować automatycznie całą brytyjską domenę internetową, natomiast aby uzupełnić lukę sprzed tego momentu pozyskano zrzuty witryn internetowych również od Internet Archive. Wpłynęło to jednak na możliwość udostępniania tak zgromadzonego zasobu, ponieważ bez uzyskania odpowiedniej zgody (nie dotyczy to zbiorów pochodzących od instytucji publicznych) jest on udostępniany wyłącznie w pracowni British Library oraz sześciu innych bibliotekach partnerskich. Według najnowszych statystyk zgromadzono w nim ponad 15 tysięcy unikalnych witryn internetowych.

Innym przykładem może być CyberCemetery[16], prowadzone przez bibliotekę Uniwersytetu Północnego Teksasu przy wsparciu National Archives and Records Administration oraz U.S. Government Printing Office. Jego zadaniem jest przechowywanie i udostępnianie witryn internetowych należących do nieistniejących już agencji i komisji rządowych, których kolekcja liczy około 90 tytułów. Współpraca z wymienionymi instytucjami rządowymi pozwala na nieograniczony dostęp do zgromadzonych materiałów. Zbliżonym projektem był prowadzony w ramach IIPC '09 European Election Web Harvesting Project, w którym uczestniczyła m.in. Biblioteka Narodowa. W ramach tego przedsięwzięcia wyselekcjonowano witryny komitetów wyborczych oraz kandydatów, a następnie nawiązano z nimi kontakt, aby uzyskać zgodę na archiwizację. Niestety ze względu na słaby odzew udało zgromadzić się niewielką liczbę zrzutów[17].

---

[15] *UK Web Archive*, https://www.webarchive.org.uk/ukwa/ (dostęp 12.11.2017); *UK Web Archiving Consortium*, https://en.wikipedia.org/wiki/UK_Web_Archiving_Consortium (dostęp 12.11.2017); o UK Web Archive m.in. K. Gmerek, *Archiwa internetowe po obu stronach Atlantyku*, s. 6–10.

[16] *UNT Libraries: CyberCemetery Home*, https://govinfo.library.unt.edu/ (dostęp 12.11.2017).

[17] K. Ślaska, A. Wasilewska, *Archiwizacja Internetu – sytuacja w polskim prawie*, s. 2–3.

W tym miejscu należy przyjrzeć się w jaki sposób określane jest archiwum Webu. Jedna z jego definicji, na jaką udało się natrafić, zamieszczona w raporcie *Information and documentation — Statistics and Quality Indicators for Web Archiving* z 2012 r., określa je jako „całość materiałów pobranych z Sieci w czasie, obejmująca jedną lub więcej kolekcji”[18]. Inna, zawarta w materiałach promocyjnych Archive-It, składa się z dwóch części, z których pierwsza rozumie je jako zarchiwizowany „zbiór adresów URL pogrupowanych według tematu, wydarzenia, obszaru lub adresów internetowych”[19].

Takie jego określenie wydaje się jednak być zbyt szerokie i objąć może, oprócz właściwych archiwów Webu, także nieduże rozmiarowo kolekcje zasobów sieciowych, gromadzonych przez określony, relatywnie krótki czas i z dość precyzyjnym kryterium doboru. Zbiory takie zazwyczaj stanowią uzupełnienie większych całości i charakteryzują się dość małą samodzielnością. Nie można ograniczać definicji także do samych kolekcji witryn, ponieważ innymi ważnymi elementami archiwum Internetu są infrastruktura, zwłaszcza sprzęt komputerowy, na którym są zgromadzone dane przechowywane, a także – bardziej istotne – wykorzystywane przy całym procesie oprogramowanie. Jego wpływ na sposób pozyskiwania zasobów sieciowych jest znaczący, nowsze wersje pozwalają na pobieranie dokładniejszych zrzutów zawierających bardziej zaawansowane elementy. Zastosowane oprogramowanie odpowiada również za późniejsze udostępnienie zarchiwizowanych materiałów i możliwości ich wykorzystania. Zatem, aby móc mówić o pełnoprawnym archiwum Internetu, powinno ono stanowić osobną instytucję, tak jak Internet Archive, lub komórkę w ramach większej struktury. Inicjatywa taka powinna przeprowadzać archiwizację zasobów sieciowych o szerokim zakresie i, co wydaje się najważniejsze, zaplanowaną jako działanie długoterminowe, z odpowiednio przygotowaną infrastrukturą oraz oprogramowaniem, a także własnym zarchiwizowanym zasobem sieciowym.

Nie należy jednak rezygnować z budowania mniejszych kolekcji, ponieważ mogą one stanowić doskonałe uzupełnienie gromadzonych w sposób masowy zasobów, zwłaszcza ze względu na ściślej określony zakres i znacznie większą, w porównaniu z archiwami opartymi o automatyczny wybór, precyzją w doborze materiałów. Poszczególne mniejsze tego rodzaju zbiory mogą składać się na

---

[18] *Information and documentation – Statistics and Quality Indicators for Web Archiving*, s. 2, http://netpreserve.org/resources/IIPC_project-SO_TR_14873__E__2012-10-02_DRAFT.pdf (dostęp 18.12.2017).

[19] *Archiving & Preserving Web content*.

całość zasobu jednego archiwum. Na takiej zasadzie gromadzą zasoby sieciowe PANDORA Archive, w ramach którego poszczególne instytucje uczestniczące selekcjonują witryny i tworzą z nich kolekcje według własnych kryteriów. Podobne rozwiązanie może znaleźć zastosowanie w Niemczech, gdzie biblioteki naukowe poszczególnych krajów związkowych mają przypisane konkretne dziedziny nauki[20].

Druga część wspomnianej wcześniej definicji, pochodzącej z materiałów Archive-It, jest postulatem, który postawić można przed dużymi archiwami Webu, jak i małymi kolekcjami. Zakłada on, iż powinny one zawierać tak dużo zasobów pochodzących z oryginalnego źródła, aby jego archiwalna wersja w jak największej części naśladowała wrażenie korzystania z niego w momencie archiwizacji[21].

Możliwe jest zaproponowanie systematyki archiwów Internetu wykorzystując różnorakie kryteria podziału. Pierwszym jakie można zastosować jest pochodzenie inicjatywy, gdzie można wskazać projekty oddolne, których autorami są instytucje pozarządowe (organizacje non-profit, korporacje etc.), oraz organizacje o charakterze państwowym (biblioteki narodowe, archiwa, publiczne uniwersytety). Kolejnym może być zakres gromadzenia zasobów, gdzie wymienić można archiwa zainteresowane całym publicznie dostępnym wycinkiem sieci (Internet Archive), konkretną domeną (np. krajową jak w przypadku UK Web Archive) lub dobierające materiały przy wykorzystaniu jakiegoś klucza (tematyka, język etc.). Inną osią podziału jest sposób gromadzenia witryn – automatyczny lub selektywny. Zaznaczyć trzeba jednak, że archiwa działające w sposób automatyczny posiadają możliwość wskazania użytkownikom adresu URL, który powinien ich zdaniem zostać zarchiwizowany, a obie wymienione strategie mogą być stosowane równocześnie. Ostatnim kryterium podziału może być dostęp do zasobów, który według raportu IIPC może być otwarty, ograniczony (np. tylko do terminali w pracowni) lub całkowicie zamknięty. Dostęp do poszczególnych części zbiorów w obrębie jednego archiwum może się różnić, tak jak ma to miejsce np. w UK Web Archive lub PANDORA Archive[22].

Rodzącym najwięcej problemów elementem procesu archiwizacji Internetu są gromadzone w jej trakcie zasoby. Zaznaczyć w tym momencie wypada, iż okre-

---

[20] A. Kugler, T. Beinert, A. Schoeger, *Archiwizacja internetu jako usługa naukowa*, Biuletyn EBIB, nr 2 (172)/2017 s. 2–5, http://open.ebib.pl/ojs/index.php/ebib/article/download/524/676 (dostęp 12.11.2017).

[21] *Archiving & Preserving Web content*.

[22] *Information and documentation*, s. 12.

ślenie „archiwizacja Internetu" jest swego rodzaju skrótem myślowym, ponieważ Internet jest to sieć połączeń pomiędzy komputerami, natomiast archiwizuje się udostępniane za pomocą usługi World Wide Web niezwykle różnorodne zasoby. Już sama ta różnorodność generuje problemy natury technicznej, ponieważ nie wszystko poddaje się archiwizacji (problematyczne są chociażby technologia Flash oraz media społecznościowe) lub nie wszystko opłaca się archiwizować, chociażby ze względu na duży rozmiar (np. zamieszczane na stronach pliki wideo)[23]. Zauważyć również trzeba fakt, że zrzuty wykonywane za pomocą crawlerów są kopiami, a ponadto ze względu, że wykonywane są od strony użytkownika, a nie serwera, nie są odwzorowaniami w stu procentach dokładnymi[24]. Utrudnienie stanowią prawa autorskie, sytuacja witryn i innych materiałów zamieszczanych w Sieci nie jest jasno zdefiniowana, a część elementów w obrębie jednej witryny może im podlegać, zaś inna część nie. Częstym rozwiązaniem jest obejmowanie ich przepisami o egzemplarzu obowiązkowym, co pozwala na ich pozyskiwanie, lecz nie zawsze pozwala na pełne udostępnienie. Opieranie się o zgody właścicieli praw do danej witryny również nie rozwiązuje wszystkich problemów, a dodatkowo może znacznie spowolnić proces archiwizacji. Problemami do rozwiązania na przyszłość jest określenie czym, z punktu widzenia nauki, dokładnie są zasoby sieciowe (obecnie można je umieścić w obrębie dość ogólnego pojęcia wirtualnych artefaktów – nowego rodzaju źródeł, które pojawiły wraz z technologiami komputerowymi), a także możliwości ich selekcji i wartościowania, które to narzędzia dostarcza archiwistyka.

Pomimo tych trudności archiwa Internetu kryją w sobie niezbadany jeszcze potencjał i mogą okazać się bardzo pożyteczne. Na pierwszym miejscu wskazać można ich użyteczność bieżącą, stanowiąc przysłowiową „deskę ratunku" w przypadku natrafienia na słynny komunikat o błędzie 404, czyli niemożliwości znalezienia strony internetowej na serwerze, lub zmian, w wyniku których część informacji na niej zawarta zniknęła. Archiwalne wersje witryn mogą mieć zastosowanie zarówno przy codziennym korzystaniu z Sieci, jak i jako dowód w sprawie sądowej. Następnie, dzięki rozwijaniu narzędzi służących ich udostępnianiu, chociażby o indeksowanie pełnotekstowe, oraz gromadzone w trakcie archiwizacji metadane, mogą one być wykorzystywane przez przedstawicieli wielu nauk, zarówno ścisłych, jak i humanistycznych. Wskazuje się na możliwość ich wykorzystania przy badaniach z zastosowaniem metod *big data*, *text mining*,

---

[23] Takie obostrzenie wymienia m.in. UK Web Archive: *Technical information*, *UK Web Archive*, http://www.webarchive.org.uk/ukwa/info/technical (dostęp 18.12.2017).

[24] *Information and documentation*, s. 8–9.

analizę trendów. Archiwalne witryny internetowe mogą również posłużyć naukowcom różnych dziedzin zajmujących się historią, socjologią, informatologią, a także rozwojem Internetu i nowoczesnych technologii. Informacje zawarte na archiwalnych witrynach mogą również stanowić uzupełnienie innych źródeł, dostarczając bardziej społecznego kontekstu dla analizowanych zjawisk.

Archiwizacja Internetu jest zjawiskiem dość nowym, rozwijanym zaledwie od drugiej połowy lat 90. XX w., przed którym odkrywają się coraz to nowe możliwości. Dość szybko zauważono wartości jakie posiadają zasoby Sieci wynikające przede wszystkim z ich rosnącej popularności, jednak w pierwszej kolejności podkreślano zastosowanie archiwów Webu przy problemach wynikających z dostępu do oryginalnych zasobów. Obecnie coraz większe jest grono uczonych zainteresowanych wykorzystaniem zawartości Internetu w swojej pracy m.in. w charakterze źródeł historycznych, a wśród nich liczni są przedstawiciele nauk humanistycznych i społecznych. Wraz z rozwojem tego rodzaju badań swoją przydatność mogą wykazać różnorodne archiwa Sieci. Właściwym zatem zdaje się dążenie do powstawania nowych tego rodzaju projektów. Ta inicjatywa powinna objąć powoływanie nowych archiwów Internetu w rozumieniu zaproponowanej wcześniej definicji, a więc posiadających szeroki zakres gromadzenia (np. domena krajowa) i zaplanowanych jako działania długofalowe o znaczącym stopniu samodzielności oraz posiadających odpowiednią ku temu infrastrukturę. Obok nich powinny wyłaniać się również małe kolekcje tematyczne, tak aby objąć zasięgiem archiwizacji jak największą część zasobów sieciowych. Należy także wciąż rozwijać technologie, które pozwolą na dokładniejsze i pełniejsze przeprowadzanie tego procesu. Ważne również będzie popularyzowanie tych działań i zasobów wśród użytkowników Internetu oraz jego badaczy, tak aby włożony w nie wysiłek nie został zmarnowany i wykorzystany został cały potencjał, który w sobie kryją.

Bibliografia

Archive-It. Web Archiving Services for Libraries and Archives. Dostęp 12.11.2017. https://archive-it.org/.

Archive-It. „About." Dostęp 11.11.2017. https://blog.archive.org/about/.

Archive-It. „Archiving & Preserving Web content." Dostęp 11.11.2017. http://www.coppul.ca/sites/default/files/uploads/COPPUL%20and%20OCUL%20Archive-It%20Informational%20webinar%20slides.pdf.

Bolter, Jay David. *Człowiek Turinga: kultura Zachodu w wieku komputera*. Przekł. i wstęp Tomasz Goban-Klas. Warszawa: Państwowy Instytut Wydawniczy, 1990.

Derfert-Wolf, Lidia. „Archiwizacja Internetu – wprowadzenie i przegląd wybranych inicjatyw." *Biuletyn EBIB* 128 (2012). http://www.ebib.pl/images/stories/numery/128/128_derfert.pdf

Gmerek, Katarzyna. „Archiwa internetowe po obu stronach Atlantyku — Internet Archive, Wayback Machine oraz UK Web Archive." *Biuletyn EBIB* 128 (2012). http://www.ebib.pl/images/stories/numery/128/128_gmerek.pdf.

Goel, Vinay. „Defining Web pages, Web sites and Web captures." Opublikowany 23.10.2016. https://blog.archive.org/2016/10/23/defining-web-pages-web-sites-and-web-captures/.

„Information and documentation – Statistics and Quality Indicators for Web Archiving." Dostęp 18.12.2017. http://netpreserve.org/resources/IIPC_project-SO_TR_14873__E__2012-10-02_DRAFT.pdf.

Internet Archive. „Digital Library of Free Books, Movies, Music & Wayback Machine". Dostęp 11.11.2017. https://archive.org/.

Klein, Martin, Herbert Van de Sompel, Robert Sanderson, Harihar Shankar, Lyudmila Balakireva, Ke Zhou, Richard Tobin. „Scholarly Context Not Found: One in Five Articles Suffers from Reference Rot." PLoS ONE 9(12): e115253. https://doi.org/10.1371/journal.pone.0115253.

Kugler, Anna, Tobias Beinert, Astrid Schoeger. „Archiwizacja internetu jako usługa naukowa." *Biuletyn EBIB* 172 (2017). http://open.ebib.pl/ojs/index.php/ebib/article/download/524/676.

Pandora Archive – Preserving and Accessing Networked Documentary Resources of Australia. Dostęp 12.11.2017. http://pandora.nla.gov.au/.

Pandora Archive. „PANDORA Fact Sheet." Dostęp 12.11.2017. http://pandora.nla.gov.au/factsheet.doc.

Qmee. „Online in 60 seconds. A year later." Dostęp 11.11.2017. http://blog.qmee.com/wp-content/uploads/2014/07/infographic-resized1.jpg.

Rosa, Agnieszka. "Human trace on the Internet – the issue of archiving the Web from the point of view of antropology-oriented archival science." *Archiwa – Kancelarie – Zbiory* 6 (8) (2015): 193–205. http://dx.doi.org/10.12775/AKZ.2015.006.

Sobczak, Anna. „Internet jako globalne archiwum społeczne – rozważania na temat roli Internetu w dokumentowaniu dziejów ludzkości." *Toruńskie konfrontacje archiwalne*, t. 4: *Nowa archiwistyka – archiwa i archiwistyka w ponowoczesnym kontekście kulturowym*, red. Waldemar Chorążyczewski, Wojciech Piasek, Agnieszka Rosa, 237–247. Toruń: Wydawnictwo Naukowe Uniwersytetu Mikołaja Kopernika, 2014.

Ślaska, Katarzyna, Anna Wasilewska. „Archiwizacja Internetu – sytuacja w polskim prawie z punktu widzenia bibliotekarzy." *Biuletyn EBIB* 128 (2012). http://www.ebib.pl/images/stories/numery/128/128_slaska.pdf.

UK Web Archive. Dostęp 12.11.2017. https://www.webarchive.org.uk/ukwa/.

UK Web Archive. „Technical information, UK Web Archive." Dostęp 18.12.2017. http://www.webarchive.org.uk/ukwa/info/technical.

University of North Texas Libraries. „CyberCemetery." Dostęp 12.11.2017. https://govinfo.library.unt.edu/.

Wikipedia. „List of Web archiving initiatives.” Dostęp 11.11.2017. https://en.wikipedia.org/wiki/List_of_Web_archiving_initiatives.

Wikipedia. „UK Web Archiving Consortium.” Dostęp 12.11.2017. https://en.wikipedia.org/wiki/UK_Web_Archiving_Consortium.