Lidia Derfert-Wolf
Biblioteka Główna Uniwersytetu Technologiczno-Przyrodniczego w Bydgoszczy

# Archiwizacja Internetu — wprowadzenie i przegląd wybranych inicjatyw

***Streszczenie:*** *Artykuł przedstawia początki i stan obecny archiwizacji Internetu na świecie. Opisano najważniejsze inicjatywy, organizacje międzynarodowe oraz przydatne serwisy, publikacje i wykazy linków. Zaprezentowano charakterystyczne cechy archiwów zasobów sieciowych oraz problemy wynikające z ich tworzenia i utrzymania. Przedstawiono cztery projekty zabezpieczania stron WWW, by wskazać praktyki, z jakimi mamy teraz do czynienia.*

***Słowa kluczowe:*** *archiwizacja Internetu, archiwizacja Web, The Library of Congress Web Archives, PANDORA Australia's Web Archive, Portuguese Web Archive, Hrvatski arhiv weba*

## Wprowadzenie

Internet — jako pierwsze z dotychczas wynalezionych mediów — łączy w sobie cechy zarówno nośnika informacji i danych, jak również środka komunikacji i masowego przekazu. Formy przekazu treści w sieci, m.in. strony WWW są przedmiotem badań i analiz, a wiele z nich jest przejawem twórczości artystycznej. Internet jest medium przekazu informacji, reklamy, publikowania i komunikowania się, również naukowego, prezentowania dorobku w różnych dziedzinach. Wszystko to doprowadziło do uznania stron internetowych za dziedzictwo kulturowe — światowe, narodowe czy regionalne. Oczywiście to dziedzictwo obejmuje tylko część zasobów sieciowych, z których pozostałe stanowią doskonały materiał dla przyszłych badaczy, jak również każdego zainteresowanego wszelkiego rodzaju informacją. Z drugiej strony, podkreśla się efemeryczność sieci — strony i witryny są ulotne, znikają bądź zmieniają swój wygląd. Marcin Wilkowski przywołuje badania pokazujące, że przeciętne życie strony WWW trwa ok. 50 dni[1]. Maria A. Jankowska podaje, że 44% witryn internetowych znika w ciągu jednego roku[2], a Daniel Gomes i in., że 80% stron jest aktualizowanych lub znika po roku[3]. Dzieje się tak z powodu świadomego usuwania witryn z różnych przyczyn, np. likwidacji instytucji, zakończenia projektu itp., jak również zmian technologicznych oraz rozwoju sieci, gdy jedne serwisy zastępowane są innymi. Tymczasem te usunięte witryny czy pierwotne wersje stron bywają bardzo przydatne do różnych celów. Dostrzeżono przy tym, że część serwisów zawiera informacje dostępne wyłącznie on-line, np. witryny wyborów, wydarzeń sportowych, konferencji naukowych i wielu, wielu innych.

[1] WILKOWSKI, M. Trzy argumenty przeciwko archiwizowaniu Internetu. W: *Historia i Media* [on-line]. 04.10.2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://historiaimedia.org/2011/10/04/trzy-argumenty-przeciwko-archiwizowaniu-internetu/.

[2] JANKOWSKA, M.A. Biblioteki akademickie — trendy dotyczące zasobów elektronicznych. W: GANIŃSKA, H.(red.). *Informacja dla nauki a świat zasobów cyfrowych* [on-line]. Poznań: Biblioteka Główna Politechniki Poznańskiej, 2008, s. 168 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.library.put.poznan.pl/konf_idn/art/4_3.pdf.

[3] GOMES, D., MIRANDA, J., COST, M. A Survey on Web Archiving Initiatives. *Lecture Notes in Computer Science* 2011, Vol. 6966, s. 408.

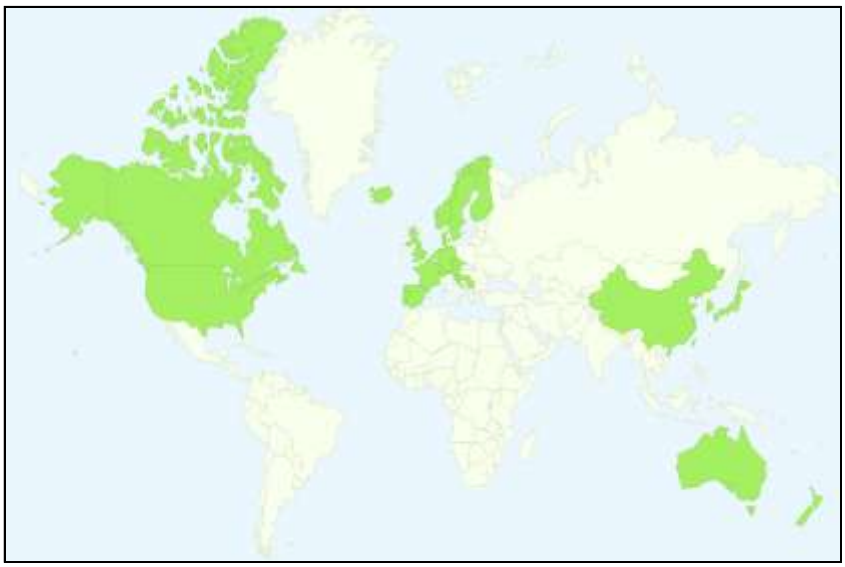
Powyższe zjawiska sprawiły, że *problem archiwizacji zasobów internetowych od wielu lat porusza wyobraźnię informatyków i badaczy. Do sfery marzeń należy jeszcze stworzenie takiego projektu, który zabezpieczałby całość dostępnych w Internecie treści, jednocześnie przechowując je w niezmienionej postaci i udostępniając internautom na całym świecie*[4]. Takie marzenie było i jest trudne do zrealizowania, głównie z uwagi na rozmiar zasobów sieciowych, który jest ogromny, ale nikt nie jest w stanie dokładnie go określić. Najczęściej cytowane są badania analizujące wielkości indeksów standardowych wyszukiwarek albo informacje ich twórców, ostatnio bardzo rzadkie. Google w swoim oficjalnym blogu poinformował w 2008 r., że rejestruje trylion (wg systemu w krajach anglojęzycznych $10^{12}$) unikalnych adresów URL, ale przyznaje też, że wielu z nich nie opłaca się indeksować[5]. Należy przy tym pamiętać, że wyszukiwarki nie zachowują kopii stron WWW, czego każdy internauta doświadcza, będąc kierowanym do nieistniejących zasobów. Powraca zatem potrzeba archiwizacji Internetu, nazywanej też archiwizacją Web i określanej jako proces polegający na poszukiwaniu, gromadzeniu i organizacji źródeł informacji w celu zabezpieczenia ich przed zniknięciem z WWW[6]. Definicja z Wikipedii ogranicza zakres archiwizacji do fragmentów WWW zachowywanych dla badaczy, historyków i społeczeństwa, co przy ogromie zasobów sieciowych wydaje się bardziej realne do spełnienia[7]. W niniejszym artykule terminy „archiwizacja Internetu", „archiwizacja WWW" i „archiwizacja zasobów sieciowych" są rozumiane jako synonimy i oznaczają zabezpieczanie zasobów World Wide Web.

## Początki, stan obecny, ważne projekty i raporty

Za pierwszą inicjatywę związaną z archiwizacją sieci powszechnie uznawana jest założona w 1996 r. Wayback Machine[8] należąca do amerykańskiej organizacji non-profit Internet Archive. Dziś istnieje na świecie kilkadziesiąt projektów o bardzo zróżnicowanym charakterze, pod względem zakresu archiwizowanych zasobów, ram organizacyjnych, aspektów prawnych i wielu innych. Cechy archiwów Internetu i problemy z nimi związane omówione zostaną w kolejnej części artykułu. W tym miejscu warto wspomnieć, że trud zabezpieczania witryn internetowych podejmują zarówno indywidualnie instytucje, jak również na szerszą skalę biblioteki narodowe, archiwa, uniwersytety czy organizacje międzynarodowe. Najpełniejsza lista inicjatyw zamieszczona w Wikipedii[9] zawiera 56 projektów z 29 krajów, w tym 19 z Europy, sześć z Azji i po dwa z Ameryki Północnej oraz Australii i Oceanii. W większości krajów realizowany jest jeden projekt, głównie przez biblioteki narodowe, ale np. w Stanach Zjednoczonych dziewięć, a w kilku krajach europejskich od dwóch do

[4] GMITEREK, G. *Archiwum Internetowe — czy możliwa jest archiwizacja zasobów sieci?* [on-line]. 25.08.2010 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.biblioteki.org/pl/wiadomosci/czytaj/786.
[5] We knew the web was big... W: *Official Google Blog* [on-line]. 25.07.2008 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.
[6] JANKOWSKA, M.A. dz. cyt., s. 168.
[7] Web archiving. W: *Wikipedia — the free encyclopedia* [on-line]. 08.02.2012 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://en.wikipedia.org/wiki/Web_archiving.
[8] Serwis Wayback Machine przedstawiła K. Gmerek w niniejszym numerze „Biuletynu EBIB".
[9] List of Web archiving initiatives. W: *Wikipedia — the free encyclopedia* [on-line]. 20.01.2012 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives.

czterech. Analiza listy inicjatyw według roku utworzenia archiwum pokazuje, że cztery projekty zainicjowano w 1996 r.: wspomniane już Wayback Machine Internet Archive, Kulturarw3 w Szwecji, Australia's Web Archive oraz Tasmanian Web Archive. W kolejnych latach powstawały pojedyncze archiwa, a od 2003 r. po kilka każdego roku.



*Rys. 1. Inicjatywy archiwizacji Internetu w czerwcu 2011 r.*
*Źródło: Wikipedia — the free encyclopedia [on-line]. 01.06.2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web:* http://en.wikipedia.org/wiki/File:Map_of_Web_archiving_initiatives_worldwide_in_June_2011.png.

Problemy archiwizacji Internetu i potrzeba współpracy doprowadziły do powstania wielu międzynarodowych organizacji i inicjatyw. Najważniejsze, posiadające najbogatszy dorobek i skupiające najwięcej członków jest powstałe w 2003 r. Międzynarodowe Konsorcjum Archiwizacji Internetu (International Internet Preservation Consortium, IIPC, http://netpreserve.org/about/index.php[10]). Konsorcjum koordynowane przez British Library liczy obecnie 42 członków, głównie biblioteki narodowe. Celem IIPC jest przede wszystkim wsparcie dla organizacji członkowskich w zakresie stosowania wspólnych narzędzi, technik i standardów umożliwiających tworzenie archiwów Internetu, jak również wspólne inicjatywy archiwizowania[11]. Na uwagę zasługują opublikowane w 2008 r. wyniki badań poszczególnych projektów członków IIPC, ich udziału w pracach konsorcjum[12] oraz raport na temat aktualnego stanu archiwizacji w 11 krajach członkowskich[13].

[10] Wszystkie odesłania do stron internetowych przedstawiają wersję aktualną w dn. 10.02.2012 r.
[11] Więcej na temat International Internet Preservation Consortium oraz udziale w nim Biblioteki Narodowej w artykule K. Ślaskiej i A. Wasilewskiej w niniejszym numerze „Biuletynu EBIB".
[12] GROTKE, A. *International Internet Preservation Consortium. 2008 Member Profile Survey Results* [on-line]. 2008 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.netpreserve.org/publications/IIPC_Survey_Report_Public_12152008.pdf.
[13] MAYR, M. *International Internet Preservation Consortium. Harvesting Practices Report* [on-line]. June 10, 2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://netpreserve.org/publications/IIPC_Harvesting_Practices_Report.pdf.

W 2004 r. powstała organizacja non profit European Archive (obecnie Internet Memory Foundation, http://internetmemory.org/), której jednym z pierwotnych celów było utworzenie archiwum europejskiego Internetu, we współpracy z Internet Archive. Obecnie fundacja uczestniczy w wielu międzynarodowych projektach dotyczących m.in. rozwoju nowej generacji narzędzi służących do archiwizacji, w tym treści tworzonych w serwisach społecznościowych. W ramach europejskiego projektu Liwa, Internet Memory Foundation przeprowadziła w 2010 r. badania inicjatyw i problemów archiwizacji Internetu w Europie[14]. W raporcie przedstawiono ciekawe dane statystyczne z 37 krajów obrazujące stan tej działalności, rodzaje zaangażowanych instytucji, kontekst prawny, politykę doboru zasobów, zarządzanie oraz formy dostępu do informacji.

Z innych ciekawych inicjatyw warto wspomnieć o Preserving Access to Digital Information (PADI, http://www.nla.gov.au/padi/index.html) — serwisie Biblioteki Narodowej Australii będącym swego rodzaju międzynarodową bramą do źródeł informacji dotyczących szeroko pojętego zabezpieczania zasobów cyfrowych, w tym archiwizacji Internetu. Na stronie Web archiving[15] przedstawiono modele procesu archiwizacji, opisy wiodących projektów z 17 krajów. Dostępny jest również bardzo bogaty wybór linków do stron dotyczących archiwizacji Internetu w poszczególnych krajach oraz tekstów artykułów związanych z tym tematem.

Do wspomnianych wyżej raportów i wyników badań należy dodać kilka innych publikacji dotyczących ogólnych problemów archiwizacji Internetu. Powszechnie polecana i cytowana jest książka *Web Archiving* pod redakcją Juliana Masanèsa, omawiająca metody archiwizacji zasobów sieciowych, procesy selekcji materiałów, techniki kopiowania witryn internetowych, problemy związane z archiwizacją tzw. sieci ukrytej oraz zagadnienia dostępu do zarchiwizowanych zasobów[16]. D. Gomes i in. przeprowadzili w 2010 r. badania 42 inicjatyw archiwizacji Internetu na całym świecie i dokonali ich przeglądu w artykule *A Survey on Web Archiving Initiatives*[17]. Autorzy poddali analizie statystycznej m.in. wielkości dotyczące archiwizowanych danych (liczba plików, zajmowane miejsca na dysku), formaty plików w archiwum, liczbę osób zaangażowanych w projektach. Metodologia zastosowana w tym badaniu pozwoliła ankietowanym zaprezentować nie tylko surowe dane, ale również własne opinie i spostrzeżenia na temat ich inicjatyw. D. Gomes i in. przedstawili wyniki swoich badań również w publicznie dostępnym filmie[18]. Na koniec warto wymienić publikacje Joint Information Systems Committee (JISC), opracowane w ramach projektu Preservation of Web Resources (PoWR), którego rezultaty przeznaczone są dla zarządzających serwisami WWW w szkolnictwie wyższym Wielkiej Brytanii, ale mogą być z powodzeniem wykorzystywane w innych krajach. Z materiałów tych na

[14] *Web Archiving in Europe. A survey provided by the Internet Memory Foundation* [on-line]. 2010 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://internetmemory.org/images/uploads/Web_Archiving_Survey.pdf.
[15] Web archiving. W: *PADI Preserving Access to Digital Information* [on-line]. [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/92.html.
[16] MASANÈS, J. (Ed.). *Web Archiving*. New York: Springer, 2006.
[17] GOMES, D., MIRANDA, J., COST, M. dz. cyt., s. 408-420.
[18] *A Survey on Web Archiving Initiatives* [on-line]. [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.youtube.com/watch?v=s4AgAeMj4Is.

szczególną uwagę zasługują podręczniki dotyczące archiwizacji na poziomie instytucjonalnym stron internetowych i zasobów zamieszczanych w sieci, w tym treści tworzonych w serwisach Web 2.0: *Preservation of Web Resources Handbook*[19] oraz *A Guide to Web Preservation. Practical advice for web and records managers based on best practices from the JISC-funded PoWR project*[20]. Pod kierunkiem JISC powstało również ważne opracowanie na temat archiwizacji zasobów WWW[21].

Zainteresowani szczegółami zagadnień omawianych w artykule skorzystać mogą ponadto z następujących serwisów:

- obszernej bibliografii publikacji na temat archiwizacji Internetu pogrupowanej według różnych projektów *Web Archiving — Bibliography*[22];
- strony Wikipedii *List of Web Archiving Initiatives*, powstałej jako uzupełnienie wspomnianej pracy D. Gomesa i in.[23]; w założeniu aktualizowanej na bieżąco przez autorów i społeczność internatów (http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives);
- wykazu i opisów archiwów Internetu tworzonych w krajach członkowskich IIPC (http://netpreserve.org/about/archiveList.php);
- listy dyskusyjnej *web-archive@cru.fr* dotyczącej archiwizacji Internetu, spraw związanych z egzemplarzem obowiązkowym oraz zabezpieczaniem danych (https://listes.cru.fr/sympa/info/web-archive);
- promocyjnych filmów w YouTube, np. Web Archiving (http://www.youtube.com/watch?v=uqHs6pA2DBo&feature=related);
- polskiego blogu Historia i Media (http://historiaimedia.org/).

**Problemy archiwizacji Internetu i cechy charakterystyczne kolekcji zasobów**

Do najbardziej istotnych zagadnień różnicujących archiwa zasobów sieciowych, a jednocześnie pozwalających je analizować, porównywać, ulepszać i rozwijać należą:

- wielkość kolekcji (liczba obiektów, przestrzeń dyskowa),
- metody gromadzenia zasobów, zakres archiwizowanych stron i typy przechowywanych obiektów,
- kwestie prawne,
- wykorzystywane narzędzia informatyczne, częstotliwość pozyskiwania danych, format przechowywania zasobów,
- dostęp do archiwów,
- zarządzanie, personel, koszty.

[19] *The Preservation of Web Resources Handbook* [on-line]. 2008 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.jisc.ac.uk/media/documents/programmes/preservation/powrhandbookv1.pdf.
[20] FARRELL, S. (Ed.). *A Guide to Web Preservation. Practical advice for web and records managers based on best practices from the JISC-funded PoWR project*. [on-line]. 2010 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://jiscpowr.jiscinvolve.org/wp/files/2010/06/Guide-2010-final.pdf.
[21] DAY, M. *Collecting and preserving the World Wide Web. Version 1.0–25* [on-line]. University of Bath, February 2003, s. 24-25 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf.
[22] *Web Archiving — Bibliography* [on-line]. April, 2004 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.ifs.tuwien.ac.at/~aola/links/WebArchiving.html.
[23] GOMES, D., MIRANDA, J., COST, M. dz. cyt.

Ile zasobów Internetu zostało już zarchiwizowanych? Scott G. Ainsworth i in. dowiedli w badaniach przeprowadzonych na przełomie 2010 i 2011 r., że 35-90% stron WWW pochodzących sprzed 2008 r. ma co najmniej jedną kopię archiwalną, 17–49% — od dwóch do pięciu, 1–8% — od sześciu do dziesięciu, a 8–63% — minimum dziesięć. Jednocześnie tylko 14,6–31,3% stron jest archiwizowanych częściej niż raz w miesiącu. Badania prowadzono w zasobach Internet Archive Wayback Machine, pamięciach podręcznych trzech wyszukiwarek (Google, Bing, Yahoo!) oraz w Diigo, Archive-It, UK National Archives i WebCite. Wyniki pokazały, że najwięcej kopii witryn znajduje się w Internet Archive Wayback Machine[24]. Porównywanie archiwów Internetu pod względem liczby skopiowanych zasobów sieciowych i zajmowanej powierzchni dyskowej nie zawsze będzie obiektywne ze względu na ich zróżnicowany charakter. Niektóre zawierają wyłącznie kopie statycznych stron WWW w języku HTML, które nie dorównują wielkością zasobom multimedialnym. Liczba obiektów zależy od polityki gromadzenia i też trudno ją uczynić kryterium porównań. W wykazie archiwów w Wikipedii, w tabeli *Archived data* podano ich wielkość w milionach zarchiwizowanych zasobów. W zestawieniu widać, że niektóre przekraczają miliard, inne sięgają kilkunastu miliardów, a największe — Internet Archive Wayback Machine — liczy 150 miliardów stron[25]. Należy zdawać sobie sprawę, że obiekt w archiwum Internetu nie zawsze oznacza stronę WWW. Jeśli na przykład na jednej stronie osadzono trzy obrazy, łączna liczba obiektów wyniesie cztery[26].

We wspomnianym wykazie archiwów w Wikipedii podano również zajmowaną przestrzeń dyskową w terabajtach. Z tabeli wynika, że są to wielkości bardzo skromne, ale również przekraczające 100 TB (np. Netarkivet.dk, BnF Web Legal Deposit, Library of Congress Web Archives), aż do 5,5 petabajtów (Internet Archive Wayback Machine). D. Gomes i in. wykazali, że archiwa Internetu przez nich przebadane zachowały od 1996 r. łącznie 181 978 mln zasobów (6,6 petabajtów).

Archiwa internetowe są również zróżnicowane pod względem zakresu i zasięgu gromadzenia danych. Istnieją niewielkie kolekcje instytucjonalne, tematyczne czy okazjonalne, kolekcje w danym języku, zasoby treści związanych z danym krajem czy najbardziej obszerne — zasoby z domeny określonego kraju czy regionu. Na przykład Web archive of Čačak ma na celu zachowanie stron związanych z tym miastem Serbii, podczas gdy zamiarem Internet Archive jest archiwizacja globalnej sieci[27]. Istnieje kilka podejść archiwizacji stron internetowych, w zależności od polityki instytucji odpowiedzialnej oraz ograniczeń prawnych. Pierwsze podejście to selektywna archiwizacja ściśle określonych zasobów (witryn), druga — zautomatyzowane gromadzenie przy pomocy robotów internetowych (crawlers). Selekcja może się sprowadzać do określonych typów zasobów np. witryn rządowych (Government of Canada Web Archive), wydarzeń (np. wybory parlamentarne, klęski żywiołowe, wydarzenia sportowe), tematów (brytyjskie UK Web Archive), zasobów

[24] AINSWORTH, S.G. i in. How much of the web is archived? W: *JCDL '11 Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*. [on-line]. 2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.cs.odu.edu/~mweigle/papers/ainsworth-jcdl11.pdf.
[25] List of Web archiving initiatives. dz. cyt.
[26] GOMES, D., MIRANDA, J., COST, M. dz. cyt., s. 413.
[27] Tamże, s. 411.

danego kraju o dużej wartości kulturowej i naukowej (australijskie archiwum PANDORA) albo witryn, których właściciele wyrazili zgodę na archiwizację. Drugie podejście to najczęściej gromadzenie wszystkich zasobów z domeny danego kraju (np. fińskie Finnish Web Archive, duńskie Netarkivet.dk, czeskie WebArchiv, francuskie BnF Web Legal Deposit). Takie rozwiązania mogą stosować tylko kraje, które mają odpowiednie uregulowania prawne. D. Gomes i in. analizując 42 inicjatywy archiwizacji Internetu, wyodrębnili 11 (26%), w których dokonuje się pozyskiwania zasobów w najszerszym zakresie, czyli w ramach całej domeny danego kraju. Wykazali też, że 80% archiwów przechowuje dane dotyczące wyłącznie kraju, regionu lub instytucji, utrzymujących dane archiwum[28]. Oczywiście gromadzenie zasobów domen narodowych Francji (.fr) czy Szwecji (.se) jest stosunkowo łatwe. Nie da się w sposób automatyczny zgromadzić witryn USA[29].

W kolekcjach archiwów Internetu dominują publicznie dostępne strony WWW i zasoby udostępniane przez protokół http, z reguły są to strony statyczne. Ponadto gromadzone są elementy witryn, np. obrazy, wideo, pliki PDF. Problemem — podobnie jak w indeksowaniu sieci przez wyszukiwarki — pozostaje archiwizacja tzw. ukrytej sieci, np. stron powstałych w technologii Flash czy tworzonych dynamicznie w trakcie ingerencji użytkownika (strony baz danych). Natomiast wielkim wyzwaniem jest archiwizacja zasobów serwisów społecznościowych, które z jednej strony stanowią problem technologiczny ze względu na ich dynamikę, a z drugiej — problem natury prawnej, ponieważ twórcy serwisów zapewniają swoim użytkownikom ochronę prywatności. Jednak Biblioteka Kongresu w 2010 r. rozpoczęła prace nad systemem archiwizacji wiadomości Twittera publikowanych od 2006 r., uznając je za część dziedzictwa kulturowego, które należy chronić dla przyszłych pokoleń[30].

Największy wpływ na zakres i zasięg gromadzonych materiałów w archiwach Internetu mają uregulowania prawne. Mają one też istotne znaczenie przy udostępnianiu archiwizowanych zasobów. Zdecydowana większość zasobów sieciowych podlega prawu autorskiemu i jest tym samym chroniona przed kopiowaniem. W niektórych krajach problem rozwiązano w przepisach dotyczących egzemplarza obowiązkowego, do którego włączono strony WWW, dzięki czemu mogą być archiwizowane bez naruszania praw autorskich[31]. Z badań Internet Memory Foundation przeprowadzonych w 2010 r. wynika, że 60,4% instytucji z 37 krajów Europy może powoływać się na przepisy prawne związane z archiwizacją Internetu i nie musi prosić właścicieli stron o zgodę na kopiowanie[32]. Z kolei w badaniach 17 członków International Internet Preservation Consortium z 2010 r. wykazano, że 12 krajów (71% badanych) ma uregulowane prawnie sprawy archiwizacji Internetu: Austria, Dania, Hiszpania, Francja, Izrael, Japonia, Korea,

[28] Tamże, s. 413.
[29] GROTKE, A. Web Archiving at the Library of Congress. *Computers in Libraries* [on-line]. December 2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.infotoday.com/cilmag/dec11/Grotke.shtml.
[30] WILKOWSKI, M. Archiwum Twittera w Bibliotece Kongresu. W: *Historia i Media* [on-line]. 14.06.2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://historiaimedia.org/2011/06/14/archiwum-twittera-w-bibliotece-kongresu/.
[31] Szerzej na ten temat można przeczytać w pozostałych artykułach niniejszego „Biuletynu EBIB".
[32] *Web Archiving in Europe*. dz. cyt.

Norwegia, Nowa Zelandia, Szwecja, Słowenia. W raporcie podano linki do odpowiednich aktów prawnych i zaleceń w dziesięciu z wymienionych krajów[33].

Z technicznego punktu widzenia pozyskiwanie zasobów internetowych (*harvesting*) jest zautomatyzowanym procesem gromadzenia zbiorów i metadanych, które są następnie indeksowane i składowane w archiwum cyfrowym według ściśle określonych parametrów[34]. Służą do tego specjalne oprogramowania bazujące na pracy robotów (*web crawlers*) „przeczesujących" wybrane obszary sieci zgodnie z zadanymi wymaganiami danego archiwum[35]. W projektach archiwizacji Internetu wykorzystuje się najczęściej narzędzia rozwijane przez IIPC, np. robota Heritix z otwartym kodem źródłowym. Robot przegląda zawartość strony WWW, kopiuje jej fragmenty (zgodnie z ustawionymi parametrami) oraz wyszukuje hiperłącza do wszystkich podstron w serwisie oraz innych stron WWW[36]. Narzędzia można rozbudowywać albo korzystać z innych programów wspomagających takie procesy jak deduplikacja (pomijanie stron, których zawartość nie zmieniła się), organizowanie kopiowanych danych, nadawanie uprawnień, ustalanie i kontrola częstotliwości pobierania danych, kontrola jakości, opis danych, zarządzanie danymi, filtrowanie i selekcja najistotniejszych informacji (*web curation*). W niektórych projektach korzysta się z zewnętrznych usług pozyskujących zasoby sieciowe „na żądanie", np. Archive-it[37]. Kolejna grupa programów to narzędzia służące do przeszukiwania (w tym pełnotekstowego) i nawigacji w serwisach udostępniających zarchiwizowane dane użytkownikom (np. NutchWAX)[38]. Istotną cechą odróżniającą archiwa jest częstotliwość pozyskiwania danych, która zależy od polityki danej instytucji. Harvesting dokonywany jest systematycznie w określonych odstępach czasu (np. kilka razy w miesiącu) albo okazjonalnie. Gromadzenie stron z całej domeny danego kraju jest procesem czasochłonnym i kosztownym, więc odbywa się rzadziej. Według danych opublikowanych w raporcie IIPC częstotliwość w takich wypadkach kształtuje się od jednego do trzech razy w roku (Hiszpania, Szwecja, Dania, Norwegia, Francja) po operacje przeprowadzane co kilka lat (Nowa Zelandia, Austria)[39]. Archiwizowane zasoby WWW są zapisywane i przechowywane w specjalnie do tego celu opracowanych formatach, z których standardem jest obecnie format WARC opublikowany w 2009 r. przez ISO[40,41] i opracowany na podstawie zdefiniowanego przez Internet Archive formatu ARC. Z przebadanych przez D. Gomesa i in. inicjatyw archiwizacji sieci, 54% korzysta ze wspomnianych formatów[42]. Standardowy format

[33] MAYR, M. dz. cyt., s. 42-43.
[34] KWIATKOWSKA-ŻÁK, K., ŽÁK, P. Webarchiv — czeski projekt archiwizacji publikacji internetowych. *Biuletyn EBIB* [on-line]. 2009, nr 7 (107) [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.ebib.info/2009/107/a.php?kwiatkowska_zak.
[35] Web archiving. dz. cyt.
[36] KWIATKOWSKA-ŻÁK, K., ŽÁK, P. dz. cyt.; Zob. artykuł F. Kłębczyka i M. Jędralskiej w niniejszym numerze „Biuletynu EBIB".
[37] Web archiving. dz. cyt.
[38] Tamże oraz w: List of Web archiving initiatives. dz. cyt.; na stronie International Internet Preservation Consortium (http://www.netpreserve.org/software/downloads.php) wymieniono narzędzia rekomendowane przez IIPC oraz wykorzystywane powszechnie przez członków konsorcjum.
[39] MAYR, M. dz. cyt., s. 20.
[40] ISO 28500: 2009 Information and documentation — WARC file format.
[41] *WARC, Web ARChive file format* [on-line]. July 13, 2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml.
[42] GOMES, D., MIRANDA, J., COST, M. dz. cyt., s. 415.

przechowywania archiwizowanych treści ułatwia współpracę w tworzeniu wyszukiwarek i innych narzędzi do przetwarzania danych.

Kolejną ważną kwestią związaną z archiwizacją Internetu jest dostępność archiwalnych treści dla użytkowników. Z tym zagadnieniem wiążą się ograniczenia w dostępie, opis danych oraz poziomy i sposoby dostępu (wyszukiwanie i/lub przeglądanie). Jeśli chodzi o dostęp w ogóle, jest on uzależniony od uregulowań prawnych w danym kraju i bywa całkowicie otwarty dla wszystkich użytkowników Internetu, otwarty tylko lokalnie w bibliotekach/archiwach utrzymujących archiwum (niekiedy wyłącznie dla określonych grup użytkowników, np. badaczy)[43,44,45]. Z wyników badań D. Gomesa i in. wynika, że 21 inicjatyw (50%) zapewnia pełny dostęp on-line do zarchiwizowanych treści, niektóre umożliwiają dostęp do wybranych treści, a 16 inicjatyw (38%) w jakiś sposób ogranicza dostęp[46]. Kolejną sprawą jest rodzaj udostępnianych danych i sposób ich katalogowania. W wykazie archiwów w Wikipedii, w tabeli *Access methods* można prześledzić wszystkie inicjatywy pod względem prezentowania historii URL, opisu metadanowego oraz wyszukiwania pełnotekstowego[47]. Wyniki uzyskane w badaniach D. Gomesa i in. wskazują, że 89% archiwów WWW udostępnia historię danego adresu URL, 79% umożliwia wyszukiwanie metadanych i 67% zapewnia wyszukiwanie pełnotekstowe w zarchiwizowanych treściach[48]. Z kolei badania IIPC z 2008 r. wykazały, że 46,4% badanych archiwów wykorzystuje format Dublin Core do katalogowania zasobów[49]. Raport Internet Memory Foundation ujawnia sposoby dostępu do danych w tych archiwach, które udostępniają je publicznie: 67,5% archiwów umożliwia przeglądanie według URL, 70% wyszukiwanie według słów kluczowych, a 65% przeglądanie kolekcji tematycznych[50]. Na rys. 2. przedstawiono formularz wyszukiwania zaawansowanego w archiwum rządowych stron WWW Kanady. W menu po lewej stronie widać inne opcje dostępu: wyszukiwanie proste, przeglądanie listy URLi, przeglądanie listy instytucji.

[43] Zob. artykuły K. Gmerek i L. Nalewajskiej w niniejszym "Biuletynie EBIB".
[44] MAYR, M. dz. cyt., s. 22.
[45] *Web Archiving in Europe*. dz. cyt.
[46] GOMES, D., MIRANDA, J., COST, M. dz. cyt., s. 415-416.
[47] List of Web archiving initiatives. dz. cyt.
[48] GOMES, D., MIRANDA, J., COST, M. dz. cyt., s. 415.
[49] GROTKE, A. *International Internet Preservation...* dz. cyt., s. 13.
[50] *Web Archiving in Europe*. dz. cyt.


*Rys. 2. Formularz wyszukiwania zaawansowanego w Government of Canada Web Archive. Źródło: Government of Canada Web Archive [on-line]. July 13, 2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.collectionscanada.gc.ca/webarchives/index-e.html.*

Nie bez znaczenia dla funkcjonowania i rozwoju archiwów Internetu są sprawy związane z zarządzaniem, personelem i kosztami. Jeśli projekt jest niewielki lub nie dysponuje wykwalifikowanymi pracownikami, korzysta z outsourcingu. Cenna jest oczywiście współpraca w ramach organizacji międzynarodowych, dająca możliwość wymiany doświadczeń oraz korzystania ze wspólnie wypracowanych narzędzi. W badaniach przeprowadzonych przez D. Gomesa i in. wykazano, że w 42 analizowanych inicjatywach zaangażowanych jest łącznie 277 osób (112 osób w pełnym wymiarze godzin i 166 w niepełnym wymiarze czasu pracy). Zespoły pracowników są zazwyczaj niewielkie (średnio ok. 2,5 pełnych etatów) i składają się w większości z bibliotekarzy i informatyków[51]. Raport IIPC z 2011 r. zawiera ciekawe dane na temat zarządzania projektami, w tym promocji, badań użytkowników itp.[52]. Ze względu na to, że stosunkowo niedawno rozpoczęto proces archiwizacji WWW, nie ma bliższych danych na temat kosztów różnych przedsięwzięć. Z badań inicjatyw europejskich prowadzonych przez Internet Memory Foundation wynika, że 52,7% projektów nie dysponuje specjalnym budżetem, 5,5% posiada budżet mniejszy niż 10 tys. euro, a 16,4% ma do dyspozycji ponad 200 tys. euro[53].

## Wybrane projekty

*List of Web archiving initiatives* w Wikipedii zawiera w miarę aktualny wykaz inicjatyw archiwizacji zasobów sieciowych na świecie. W niniejszym numerze „Biuletynu EBIB" omówiono bliżej projekty w krajach skandynawskich[54], Wayback Machine Internet Archive[55], UK Web Archive[56] oraz polski serwis Archiwum Internetu[57]. Wcześniej, w

[51] GOMES, D., MIRANDA, J., COST, M. dz. cyt., s. 412.
[52] MAYR, M. dz. cyt., s. 5-20.
[53] *Web Archiving in Europe*. dz. cyt.
[54] Zob. artykuł L. Nalewajskiej w niniejszym numerze „Biuletynu EBIB"
[55] Zob. artykuł K. Gmerek w niniejszym numerze „Biuletynu EBIB".
[56] Tamże.

2009 r., Katarzyna Kwiatkowska-Žák i Petr Žák opisali bardzo szczegółowo czeski projekt Webarchiv[58], natomiast w tym artykule zostaną krótko omówione następujące projekty reprezentujące zasoby różnych krajów i tworzone różnymi metodami:

- The Library of Congress Web Archives (LCWA),
- PANDORA Australia's Web Archive,
- Portuguese Web Archive (PWA),
- Hrvatski arhiv weba (HAW).

Projekt Library of Congress Web Archives (pierwotnie Minerva, Mapping the Internet the Electronic Resources Virtual Archive, http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html) został zainaugurowany przez Bibliotekę Kongresu w 2000 r., początkowo jako program pilotażowy polegający na archiwizacji 35 wybranych stron internetowych. Równolegle codziennie kopiowano około 200 stron związanych z wyborami prezydenckimi w 2000 r. Obecnie LCWA jest zbiorem ok. 40 kolekcji tematycznych i wydarzeń, o których wyborze decydują specjaliści dziedzinowi Biblioteki Kongresu. Są to np. wybory, wojna w Iraku, wydarzenia 11 września, blogi z zakresu prawa (*Legal Blawgs*). Niektóre kolekcje zawierają materiały spoza USA i w innych językach niż angielski, np. Papal Transition 2005 Web Archive. Archiwa internetowe LCWA zawierają również kolekcje rekomendowane i utrzymywane przez pracowników oddziałów zbiorów specjalnych, np. strony internetowe fotografów, rysowników, oraz firm architektonicznych. W LCWA zarchiwizowano ok. 5 mln obiektów liczących łącznie ok. 250 terabajtów (przyrost ok. 4–5 terabajtów w miesiącu). W Stanach Zjednoczonych nie ma na razie przepisów prawnych sprzyjających pełnej archiwizacji zasobów sieciowych i dostępu do nich. Pracownicy Biblioteki Kongresu zwracają się do właścicieli stron WWW o zgodę na ich kopiowanie i udostępnianie. W razie nieuzyskania zgody, dostęp do stron jest ograniczony tylko do korzystania w celach badawczych albo możliwy wyłącznie na miejscu w bibliotece. Selektywne pozyskiwanie danych do LCWA odbywa się za pośrednictwem Internet Archive przy pomocy robota Heritrix. Harvesting dokonywany jest kilka razy w tygodniu, co miesiąc lub 1-2 razy w roku, w zależności od miejsca i wydarzenia. W miarę możliwości z jednej witryny archiwizuje się treść w języku HTML, obrazy, Flash, pliki PDF, audio i wideo. Zawartość archiwum przechowywana jest w formacie WARC. Do udostępniania danych używa się wersji open source Wayback Machine, a sposoby dostępu to: przeglądanie według dziedzin, tytułów oraz wyszukiwanie według tytułu, dziedziny, nazwiska, języka oraz daty archiwizacji[59].

[57] Zob. artykuł F. Kłębczyka i M. Jędralskiej w niniejszym numerze „Biuletynu EBIB".
[58] KWIATKOWSKA-ŽÁK, K., ŽÁK, P. dz. cyt.
[59] GROTKE, A. Web Archiving at the... dz. cyt.


*Rys. 3. Opis zasobu sieciowego zarchiwizowanego w Library of Congress Web Archives. Źródło: Library of Congress Web Archives. Papal Transition 2005 Web Archive [on-line]. [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://lcweb2.loc.gov/diglib/lcwa/mrva0010.0169/default.html.*

PANDORA Australia's Web Archive (http://pandora.nla.gov.au/) jest archiwum przydatnych dla badaczy treści publikowanych on-line, dotyczących Australii i z terenu tego kraju. Kolekcję zaczęła tworzyć Biblioteka Narodowa Australii w 1996 r. Obecnie rozwija ją we współpracy z 11 innymi bibliotekami i instytucjami[60]. PANDORA jest tzw. archiwum selektywnym. Oznacza to, że nie gromadzi się wszystkich australijskich publikacji sieciowych i stron WWW. Poza tym materiały sieciowe nie są w Australii objęte przepisami o egzemplarzu obowiązkowym, więc biblioteki muszą pytać wydawców i właścicieli zasobów o zgodę na ich kopiowanie i udostępnianie. Każdy z partnerów PANDORY ma własną politykę doboru materiałów[61]. Dla przykładu Biblioteka Narodowa Australii gromadzi zasoby ważne dla dziedzictwa narodowego, biblioteki stanowe, państwowe — zasoby znaczące dla danego regionu, Australian War Memorial archiwizuje witryny związane z historią wojskowości, a Australian Institute of Aboriginal and Torres Strait Islander Studies — witryny dotyczące ludności tubylczej. W zakresie typów gromadzonych zasobów są to głównie witryny rządowe, czasopisma i konferencje naukowe, blogi, strony instytucji i organizacji, wybrane strony firm, gazet oraz witryny dotyczące ważnych wydarzeń, np. wyborów. Jeden tytuł w całej kolekcji może być dokumentem tekstowym w PDF albo obszernym serwisem WWW zawierającym tysiące plików w różnych formatach. W projekcie PANDORA zarchiwizowano ponad 29 tys. obiektów, z czego 55% stanowią witryny rządowe. Zasoby obejmują 5,83 TB danych. Zasoby są dobierane, pozyskiwane i zarządzane przy pomocy własnego oprogramowania PANDAS. Indeksowane są pełne teksty, a wybrane tytuły katalogowane (w MARC 21) i włączane do bibliografii narodowej. Zarówno pełne teksty, jak i rekordy katalogowe są przeszukiwalne poprzez australijski narodowy punkt dostępu do informacji Trove (http://trove.nla.gov.au/), natomiast tematyczne przeglądanie

[60] *PANDORA Partners* [on-line]. [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://pandora.nla.gov.au/partners.html.
[61] *Selection guidelines* [on-line]. [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://pandora.nla.gov.au/selectionguidelinesallpartners.html.

zasobów jest możliwe ze strony głównej projektu PANDORA. Tytuły zasobów i większość pełnych treści jest dostępna publicznie. Do niewielkiej części zasobów dostęp w sieci jest ograniczony — można z nich korzystać lokalnie z komputerów w bibliotece.


*Rys. 3. Witryna narodowej australijskiej metawyszukiwarki Trove, w której jedną z kolekcji jest archiwum stron internetowych Archive Websites (1996 — now).*
*Źródło: Trove [on-line]. [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://trove.nla.gov.au/.*

W Portugalii kluczowy projekt archiwizacji WWW — Portuguese Web Archive (http://archive.pt) — realizuje od 2008 r. organizacja non profit Foundation for National Scientific Computing. Inicjatywa ma na celu zachowanie informacji publikowanych od 1996 r. na witrynach z domeny portugalskiej .pt oraz witryn dotyczących Portugalii spoza tej domeny i jest kontynuacją wcześniejszych prac University of Lisbon rozpoczętych w 2001 r. Dane archiwizowane są dla PWA przy pomocy programu Heritrix (w formacie ARC) 3–4 razy w roku i udostępniane po upływie jednego roku. Uwzględniane są strony statyczne i dynamiczne, z wykluczeniem witryn wymagających rejestracji oraz witryn o zastrzeżonym dostępie. W serwisie PWA dostępny jest również formularz umożliwiający sugerowanie stron do archiwizacji. Przeszukiwanie w archiwum odbywa się według słów kluczowych, adresów URL albo w opcji wyszukiwania zaawansowanego dodatkowo według formatu, daty albo domeny. Portuguese Web Archive zawiera obecnie 1,3 mld obiektów, w tym 130 mln stron WWW. Zasoby nie są katalogowane. Rezultaty wyszukiwania — przypominające wyniki generowane przez standardowe wyszukiwarki — zawierają tytuł witryny, datę pierwszej archiwizacji oraz link do pozostałych kopii (*other dates*).


*Rys. 5. Strona główna projektu Portuguese Web Archive.*
*Źródło: Portuguese Web Archive [on-line]. [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://archive.pt.*

Chorwacka inicjatywa Hrvatski arhiv weba (http://haw.nsk.hr/) jest realizowana od 2003 r. przez Narodową i Uniwersytecką Bibliotekę Chorwacji (pełniąca funkcję biblioteki narodowej i biblioteki głównej Uniwersytetu w Zagrzebiu). Celem projektu jest gromadzenie i przechowywanie zasobów sieciowych od 1998 r., które są częścią chorwackiego dziedzictwa narodowego. Strony internetowe podlegają prawu o egzemplarzu obowiązkowym z 1997 r. Pozyskiwanie zasobów odbywa się selektywnie, na podstawie ściśle określonych kryteriów[62]. Dla przykładu, akceptowane są czasopisma, książki, witryny instytucji, stowarzyszeń, projektów badawczych, gazet, portali, wybranych witryn osób, blogów. Natomiast wykluczane są witryny wyszukiwarek, gier, firm, strony reklamowe, listy dyskusyjne. Nie podlegają archiwizacji zasoby cyfrowe z projektów digitalizacji innych instytucji i z innych archiwów internetowych. Dostęp do archiwum jest z założenia nieograniczony, jednak jeśli wydawca nie wyraża zgody na udostępnianie zasobu, możliwe jest korzystanie na miejscu w bibliotece. Archiwum HAV zawiera ponad 81 mln obiektów, w tym ponad 3,5 tys. tytułów zasobów WWW. Każdy zasób jest katalogowany i może być wyszukany bezpośrednio ze strony HAV oraz poprzez katalog on-line Narodowej i Uniwersyteckiej Biblioteki Chorwacji (http://katalog.nsk.hr/). Serwis HAV oferuje wyszukiwanie zasobów według tytułów, słów kluczowych i URL oraz przeglądanie według dziedzin i alfabetyczne według tytułów. Rezultaty wyszukiwania zawierają tytuł zasobu, dziedzinę, liczbę kopii, datę i wielkość każdej kopii oraz adres URL. Przy każdym opisie można wyświetlić miniaturkę witryny.

[62] *Croatian Web Archive. Selection criteria* [on-line]. [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://haw.nsk.hr/en/selection-criteria.


*Rys. 6. Rezultat wyszukiwania w Hrvatski arhiv weba.*
*Źródło: Hrvatski arhiv weba [on-line]. [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://haw.nsk.hr/.*

## Podsumowanie

Zasoby WWW, a głównie strony internetowe, już dziś są uznanym źródłem informacji coraz powszechniej cytowanym w publikacjach naukowych. Są też oczywistym świadectwem życia społecznego i politycznego, a także zmian technologicznych. Historycy, bibliotekarze, archiwiści i przedstawiciele innych środowisk podkreślają potrzebę zachowania tych zasobów jako dziedzictwa o znaczeniu kulturowym i naukowym. Wspominano wcześniej, że w istniejących archiwach Internetu zabezpieczono już ok. 182 mld obiektów i liczba ta ciągle rośnie. Trudno dziś precyzyjnie przewidzieć dalszy rozwój archiwizowania WWW. Naukowcy Uniwersytetu w Oksfordzie opracowali raport *Web Archives: the future(s)* na temat przewidywanych sposobów wykorzystania archiwów sieci Web przez badaczy [63,64]. W opracowaniu zaprezentowano szereg zagadnień związanych z przyszłością archiwizacji Internetu, w tym treści serwisów społecznościowych, metody i techniki archiwizowania, wizualizację wyszukiwanie i analizy danych oraz włączania użytkowników w gromadzenie materiałów do kopiowania. W raporcie przedstawiono również cztery scenariusze rozwoju omawianego procesu, z których według autorów, najbardziej prawdopodobny jest ostatni:

- Nirwana: archiwa Internetu są szeroko wykorzystywane w badaniach historycznych, socjologicznych, ekonomicznych, kulturowych, w polityce, biznesie, działalności edukacyjnej i pozarządowej; wykorzystują standardy, prezentują najbardziej efektywne i atrakcyjne interfejsy użytkownika, są otwarte i elastyczne.

[63] MEYER, E.T., THOMAS, A., SCHROEDER, R. *Web Archives: the future(s)* [on-line]. University of Oxford, June 30, 2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://netpreserve.org/publications/2011_06_IIPC_WebArchives-TheFutures.pdf.
[64] Szerzej raport omawia: WILKOWSKI, M. Web Archives: the future(s) — trzy scenariusze rozwoju archiwistyki internetowej. W: *Historia i Media* [on-line]. 10.08.2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://historiaimedia.org/2011/08/10/web-archives-the-futures-trzy-scenariusze-rozwoju-archiwistyki-internetowej/

- Apokalipsa: archiwa Internetu są fragmentaryczne, nie stosują standardów, trudno je znaleźć i uzyskać do nich dostęp; techniki i metody archiwizacji nie nadążają za rozwojem Internetu; archiwa są w związku z tym nieprzydatne i mało używane.
- Osobliwość: archiwa staną się niepotrzebne, gdyż Internet rozwinie się w zupełnie nieznaną formę — skomplikowany wirtualny organizm, złożony z obiektów cyfrowych i inteligencji ludzkiej, którego archiwizacja będzie niemożliwa.
- Marginalne: archiwa Internetu staną się tym, czym są archiwa tradycyjne i w związku z tym nie będą szeroko wykorzystywane, choć pozostaną dobrze zorganizowane i zarządzane; archiwizacja obejmie tylko statyczne witryny, a serwisy społecznościowe pozostaną poza ich zasięgiem z powodów technologicznych.

Wydaje się, że jest kilka kwestii, które skutecznie przyczyniają się do tego, że już teraz archiwa zasobów WWW są marginalne. Najważniejsza sprawa to uregulowania prawne, których brak albo ich wymiar powoduje, że znaczna część zabezpieczanych obiektów może być udostępniana z różnymi ograniczeniami lub wcale. Druga — to dostępność metainformacji dla użytkownika. Wydaje się, że idealnym rozwiązaniem jest katalogowanie zasobów sieciowych na równi z innymi elektronicznymi oraz tradycyjnymi. Niemal idealnym rozwiązaniem jest w tym wypadku australijski serwis Trove. Z pewnością przydatne też będą zintegrowane narzędzia przeszukiwania zasobów WWW zarchiwizowanych w ramach wszystkich projektów oraz takie indeksowanie tych zasobów, żeby były „widoczne" dla standardowych wyszukiwarek. Ostatni postulat spełniają np. zasoby archiwum PANDORA.

Na zakończenie warto wspomnieć o ważnym dokumencie Unii Europejskiej — *Zalecenia Komisji z dnia 27 października 2011 r. w sprawie digitalizacji i udostępniania w Internecie dorobku kulturowego oraz w sprawie ochrony zasobów cyfrowych*[65] — w którym w pkt. 16 zapisano:

*Brak takiej polityki* [ochrony treści cyfrowych — przyp. aut.] *stanowi poważne zagrożenie dla trwałości zdigitalizowanego materiału, a ponadto może spowodować utratę materiałów wyprodukowanych w formacie cyfrowym (materiały powstałe w formacie cyfrowym). Opracowanie skutecznych środków z zakresu ochrony zasobów cyfrowych ma dalekosiężne skutki, wykraczające poza obszar instytucji kulturalnych. Kwestie ochrony zasobów cyfrowych mają istotne znaczenie dla każdej organizacji prywatnej lub publicznej, która jest zobowiązana do ochrony zasobów cyfrowych lub która pragnie zapewnić taką ochronę z własnej woli*[66].

Dalej w pkt. 17 stwierdza się:

*Ochrona zasobów cyfrowych rodzi wyzwania o charakterze finansowym, organizacyjnym i technicznym, a niekiedy wymaga aktualizacji przepisów ustawodawczych. Wiele państw członkowskich wprowadziło lub rozważa wprowadzenie przepisów nakładających na producentów materiałów cyfrowych obowiązek wykonywania jednej lub więcej kopii wyprodukowanego materiału dla uprawnionego podmiotu przechowującego. Skuteczne przepisy i praktyki dotyczące egzemplarza*

[65] Zalecenie Komisji z dnia 27 października 2011 r. w sprawie digitalizacji i udostępniania w Internecie dorobku kulturowego oraz w sprawie ochrony zasobów cyfrowych (2011/711/UE). W: *Dziennik Urzędowy Unii Europejskiej* [on-line]. L 283/39-45, 29.10.2011 [Dostęp 22.01.2012]. Dostępny w World Wide Web: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:283:0039:0045:PL:PDF.

[66] Tamże, s. 41.

*obowiązkowego mogą zminimalizować administracyjne obciążenia zarówno w przypadku właścicieli materiałów, jak i instytucji przechowujących, a tym samym winny być zalecane. […] Tak zwany web harvesting jest nową techniką gromadzenia materiałów w Internecie na potrzeby ochrony zasobów cyfrowych. Polega ona na aktywnym gromadzeniu materiałów przez uprawnione instytucje zamiast czekania, aż zostaną one przekazane do przechowania, co pozwala na zminimalizowanie obciążeń administracyjnych dla producentów materiałów cyfrowych; prawodawstwo krajowe powinno więc dopuszczać taką technikę*[67].

W końcowej części dokumentu zaleca się, aby państwa członkowskie:

*10) podejmowały niezbędne środki dotyczące egzemplarza obowiązkowego materiałów wyprodukowanych w formacie cyfrowym dla zapewnienia ich długotrwałej ochrony i poprawiały skuteczność istniejących środków dotyczących egzemplarza obowiązkowego materiałów wyprodukowanych w formacie cyfrowym poprzez:*

> *a) dbałość o to, by posiadacze praw autorskich dostarczali utwory do bibliotek gromadzących egzemplarze obowiązkowe bez zabezpieczeń technicznych lub udostępniali bibliotekom gromadzącym egzemplarze obowiązkowe środki, które zapewnią, iż zabezpieczenia techniczne nie będą kolidować z czynnościami, które biblioteki muszą podejmować w celu ochrony zasobów, z pełnym poszanowaniem prawodawstwa Unii Europejskiej i prawodawstwa międzynarodowego dotyczącego praw własności intelektualnej;*
>
> *b) tam, gdzie to konieczne, wprowadzanie przepisów prawnych dopuszczających przenoszenie cyfrowych egzemplarzy obowiązkowych utworów z jednej biblioteki gromadzącej egzemplarze obowiązkowe do innych bibliotek gromadzących egzemplarze obowiązkowe, którym również przysługuje prawo do tych utworów;*
>
> *c) dopuszczenie ochrony zasobów treści internetowych przez uprawnione instytucje przy wykorzystaniu technik gromadzenia materiałów z Internetu, takich jak web harvesting, z pełnym poszanowaniem prawodawstwa Unii Europejskiej i prawodawstwa międzynarodowego dotyczącego praw własności intelektualnej;*

*11) uwzględniały sytuację w innych państwach członkowskich przy tworzeniu i aktualizowaniu polityki i procedur dotyczących egzemplarzy obowiązkowych materiałów powstałych w formacie cyfrowym w celu uniknięcia nadmiernego zróżnicowania środków dotyczących egzemplarzy obowiązkowych*[68].

**Bibliografia:**

[1] AINSWORTH, S.G. i in. How much of the web is archived? W: *JCDL '11 Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*. [on-line]. 2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.cs.odu.edu/~mweigle/papers/ainsworth-jcdl11.pdf.
[2] *Biuletyn EBIB — Archiwizacja Internetu* [on-line]. 2012, nr 10 (128). [Dostęp 19.02.2012]. Dostępny w World Wide Web: http://www.nowyebib.info/biuletyn/numer-128-spis ISSN 1507-7187.
[3] *Croatian Web Archive. Selection criteria* [on-line]. [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://haw.nsk.hr/en/selection-criteria.
[4] DAY, M. *Collecting and preserving the World Wide Web. Version 1.0–25* [on-line]. University of Bath, February 2003 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf.
[5] GMITEREK, G. *Archiwum Internetowe — czy możliwa jest archiwizacja zasobów sieci?* [on-line]. 25.08.2010 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.biblioteki.org/pl/wiadomosci/czytaj/786.
[6] GOMES, D., MIRANDA, J., COST, M. A Survey on Web Archiving Initiatives. *Lecture Notes in Computer Science* 2011, Vol. 6966, s. 408-420.

[67] Tamże, s. 41.
[68] Tamże, s. 42-43.

[7] GROTKE, A. *International Internet Preservation Consortium. 2008 Member Profile Survey Results* [on-line]. 2008 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.netpreserve.org/publications/IIPC_Survey_Report_Public_12152008.pdf.
[8] GROTKE, A. Web Archiving at the Library of Congress. *Computers in Libraries* [on-line]. December 2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.infotoday.com/cilmag/dec11/Grotke.shtml. ISSN 1041-7915.
[9] FARRELL, S. (Ed.). *A Guide to Web Preservation. Practical advice for web and records managers based on best practices from the JISC-funded PoWR project* [on-line]. 2010 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://jiscpowr.jiscinvolve.org/wp/files/2010/06/Guide-2010-final.pdf.
[10] ISO 28500: 2009 Information and documentation — WARC file format.
[11] JANKOWSKA, M.A. Biblioteki akademickie — trendy dotyczące zasobów elektronicznych. W: GANIŃSKA, H.(red.). *Informacja dla nauki a świat zasobów cyfrowych* [on-line]. Poznań: Biblioteka Główna Politechniki Poznańskiej, 2008, s. 167-171 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.library.put.poznan.pl/konf_idn/art/4_3.pdf. ISBN 83-910677-4-2.
[12] KWIATKOWSKA-ŽÁK, K., ŽÁK, P. Webarchiv — czeski projekt archiwizacji publikacji internetowych. *Biuletyn EBIB* [on-line]. 2009, nr 7 (107) [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.ebib.info/2009/107/a.php?kwiatkowska_zak. ISSN 1507-7187.
[13] List of Web archiving initiatives. W: *Wikipedia — the free encyclopedia* [on-line]. 20.01.2012 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives.
[14] MASANÈS, J. (Ed.). *Web Archiving*. New York: Springer, 2006. ISBN 978-3-540-23338-1.
[15] MAYR, M. *International Internet Preservation Consortium. Harvesting Practices Report* [on-line]. June 10, 2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://netpreserve.org/publications/IIPC_Harvesting_Practices_Report.pdf.
[16] MEYER, E.T., THOMAS, A., SCHROEDER, R. *Web Archives: the future(s)* [on-line]. University of Oxford, June 30, 2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://netpreserve.org/publications/2011_06_IIPC_WebArchives-TheFutures.pdf.
[17] *The Preservation of Web Resources Handbook* [on-line]. 2008 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.jisc.ac.uk/media/documents/programmes/preservation/powrhandbookv1.pdf.
[18] *Selection guidelines* [on-line]. [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://pandora.nla.gov.au/selectionguidelinesallpartners.html.
[19] *WARC, Web ARChive file format* [on-line]. July 13, 2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml.
[20] We knew the web was big... W: *Official Google Blog* [on-line]. 25.07.2008 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://googleblog.blogspot.com/2008/07/we-knew-web-was-big.html.
[21] Web archiving. W: *PADI Preserving Access to Digital Information* [on-line]. [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/92.html.
[22] Web archiving. W: *Wikipedia — the free encyclopedia* [on-line]. 08.02.2012 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://en.wikipedia.org/wiki/Web_archiving.
[23] *Web Archiving — Bibliography* [on-line]. April, 2004 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://www.ifs.tuwien.ac.at/~aola/links/WebArchiving.html.
[24] *Web Archiving in Europe. A survey provided by the Internet Memory Foundation* [on-line]. 2010 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://internetmemory.org/images/uploads/Web_Archiving_Survey.pdf.
[25] WILKOWSKI, M. Archiwum Twittera w Bibliotece Kongresu. W: *Historia i Media* [on-line]. 14.06.2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://historiaimedia.org/2011/06/14/archiwum-twittera-w-bibliotece-kongresu/.
[26] WILKOWSKI, M. Trzy argumenty przeciwko archiwizowaniu Internetu. W: *Historia i Media* [on-line]. 04.10.2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://historiaimedia.org/2011/10/04/trzy-argumenty-przeciwko-archiwizowaniu-internetu/.
[27] WILKOWSKI, M. Web Archives: the future(s) — trzy scenariusze rozwoju archiwistyki internetowej. W: *Historia i Media* [on-line]. 10.08.2011 [Dostęp 21.01.2012]. Dostępny w World Wide Web: http://historiaimedia.org/2011/08/10/web-archives-the-futures-trzy-scenariusze-rozwoju-archiwistyki-internetowej/.

[28] Zalecenie Komisji z dnia 27 października 2011 r. w sprawie digitalizacji i udostępniania w Internecie dorobku kulturowego oraz w sprawie ochrony zasobów cyfrowych (2011/711/UE). W: *Dziennik Urzędowy Unii Europejskiej* [on-line]. L 283/39-45, 29.10.2011 [Dostęp: 22.01.2012]. Dostępny w World Wide Web: http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2011:283:0039:0045:PL:PDF.