Filip Kłębczyk
Instytut Informatyki Politechniki Śląskiej
Monika Jędralska
Narodowe Archiwum Cyfrowe

# Serwis "Archiwum Internetu" na tle ogólnych problemów archiwizacji zasobów sieciowych

***Streszczenie:*** *Artykuł stanowi analizę możliwości oraz działań podejmowanych w zakresie archiwizacji i późniejszego udostępniania polskich zasobów internetowych. Istotne dla tego procesu są nie tylko ograniczenia finansowe i techniczne, ale również uregulowania prawne, które w znacznym stopniu wyznaczają ramy, zwłaszcza etapu udostępniania zgromadzonych danych. W artykule dokonano przeglądu zagranicznych doświadczeń w obszarze archiwistyki internetowej oraz technik stosowanych powszechnie w archiwizacji i udostępnianiu tego typu treści. Omówiono również polski projekt „Archiwum Internetu", prowadzony przez Narodowe Archiwum Cyfrowe – podstawy prawne jego realizacji, stan obecny oraz kierunki rozwoju tego projektu i podobnych.*

***Słowa kluczowe:*** *archiwizacja Internetu, polskie zasoby internetowe, Archiwum Internetu*

## Wprowadzenie

Internet w wyniku szerokiego upowszechnienia i ciągłego rozwoju stał się nowym obszarem prac dla bibliotekarzy i archiwistów. W porównaniu z tradycyjnymi mediami, takimi jak prasa, radio i telewizja, jest on w swej naturze znacznie bardziej skomplikowany, a co się z tym wiąże — trudny w archiwizacji. Trudność ta wynika przede wszystkim z dynamiki i typów treści występujących w Internecie oraz ogromnej różnorodności stosowanych standardów. Jednym z ważniejszych aspektów procesu archiwizacji jest również fakt, iż treści te podlegają najczęściej ochronie prawno-autorskiej, a sam proces udostępniania — także innym regulacjom prawnym. Pomimo przeszkód wiele instytucji decyduje się podjąć wyzwanie archiwizacji zasobów internetowych. Jest to jednak zadanie niełatwe, gdyż narzędzia służące do archiwizacji nie nadążają za najnowszymi rozwiązaniami stosowanymi obecnie w tworzeniu zasobów globalnej sieci. Niemniej archiwizacja tego medium jest niezwykle ważna, gdyż Internet staje się obszarem mającym coraz większe znaczenie dla nauki, historii i administracji. Archiwizacja polskich zasobów sieciowych jest również niezwykle istotnym i koniecznym procesem.

## Specyfika zasobów internetowych i ich społeczna rola

Rola Internetu w życiu społecznym i naukowym nie podlega już obecnie dyskusji. Treści w nim prezentowane stają się coraz bardziej uznawanym źródłem nie tylko bieżącej informacji, ale także wartościowej wiedzy historycznej. Do sieci przenosi się aktywność nie tylko obywateli, ale również ważnych instytucji państwowych. Coraz więcej procedur administracyjnych, typowych dla urzędów i instytucji, realizuje się przez Internet, np. zamówienia publiczne, nabór do służby cywilnej, wiele e-usług.

Gwałtownie rośnie rola Internetu jako wiarygodnego i wartościowego źródła informacji w naukowym publikowaniu. Powszechne cytowanie zasobu internetowego nie może jednak spełniać do końca swojej roli, jeśli nie wykazuje on znamion trwałości. Ulotność i nietrwały charakter treści internetowych to jeden z istotnych motywów archiwizacji — średni czas życia pojedynczej strony internetowej wynosi jedynie kilkadziesiąt dni[1], a dane które zawierała można utracić bezpowrotnie.

## Szczególnie dynamiczna archiwalna materia

Strona techniczna takiego przedsięwzięcia, jakim jest archiwizowanie zasobów internetowych, jest złożona pod wieloma względami. Głównym problemem jest dynamiczny charakter Internetu i nieustanna ewolucja stosowanych w nim technologii. W wyniku wieloletnich działań podejmowanych przez biblioteki i archiwa nad różnymi rozwiązaniami w tej dziedzinie, wypracowano narzędzia służące zarówno archiwizacji zasobów internetowych, jak również ich późniejszemu udostępnianiu. Dynamika sieci nie pozwala jednak założyć, iż narzędzia te będą w dłuższej perspektywie czasu efektywne. Obecnie powstające strony internetowe nie przypominają tworzonych jeszcze pięć czy dziesięć lat temu statycznych witryn napisanych w języku HTML. Znaczniki związane nie tylko z treścią, ale i prezentacją wizualną, wpływają na przejrzystość struktury stron internetowych, co staje się też przeszkodą w wyróżnieniu ich stałych i wspólnych elementów. Pomocny w tym obszarze staje się standard HTML5[2], jednak likwiduje on wyżej wymienione problemy jedynie częściowo. Coraz większą przeszkodą są rozwiązania określane mianem *rich media*, czyli określonego typu interaktywne multimedia implementowane na stronach internetowych. Komponenty w technologii Flash czy aplety Javy to powszechnie stosowane elementy wzbogacające witryny internetowe. Częstym zjawiskiem jest też tworzenie serwisów i portali jako aplikacji, czego najlepszym przykładem są, np. Facebook, Twitter oraz Google+.

Podjęcie wyzwania, jakim jest archiwizacja zasobów internetowych, to nie tylko kontakt z niezwykle dynamiczną, ale również ogromnie obszerną materią. Internet to aktualnie środowisko niełatwe do utrwalenia, lecz przede wszystkim w swej obecnej objętości trudne do zmierzenia. Dla przykładu, wyszukiwarka Google indeksowała w 1997 r. 50 mln unikalnych adresów internetowych, zaś w 2011 r. — 3 tln[3]. Jednocześnie ocenia się, że nieindeksowany przez wyszukiwarki głęboki Internet to zasób nawet 500 razy większy od zasobu indeksowanego[4]. Biorąc pod uwagę wspomniane wielkości danych, nie ma wątpliwości, iż dla polskich instytucji kultury oraz organów administracji publicznej zadanie zarchiwizowania zasobów sieci jest zadaniem odległym, zwłaszcza przy uwzględnieniu warunków finansowych tych

[1] KAHLE, B. Preserving the Internet. W: *Scientific American Special Online Issue: The Future of The Web* [on-line]. April 2002 [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://www.scientificamerican.com/sciammag/issues.cfm.
[2] Obecnie dostępna jest robocza wersja specyfikacji: *HTML5: a vocabulary and associated APIs for HTML and XHTML* [on-line]. W3C Working Draft, 25 May 2011 [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://www.w3.org/TR/html5/.
[3] BLEICHER, A. Memory of Webs Past. *IEEE Spectrum Magazine* [on-line]. March 2011 [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://spectrum.ieee.org/telecom/internet/a-memory-of-webs-past.
[4] Tamże.

instytucji oraz wymagań technicznych samego procesu. Ogrom danych sieciowych, nie zawsze wartościowych i nie zawsze pożądanych, generuje potrzebę ich selekcji, a także fakt, iż w procesie archiwizacji niezbędne jest wykorzystywanie rozwiązań automatycznych, które odciążą żmudną pracę człowieka. W projektach archiwizacji wdraża się algorytmy eksploracji danych, które automatycznie dokonują klasyfikacji i oceny wartości archiwizowanego materiału. Kolejny z niezbędnych kroków to ustalenie częstotliwości procesu archiwizacji tak, aby był on możliwie najbardziej efektywny, jak również możliwy do przeprowadzenia w określonym czasie.

## Jak archiwizować, jak udostępniać? — wybór metody

W stosowanych obecnie metodach archiwizacji istnieją dwie drogi. Pierwsza sprowadza się do utrwalania treści po stronie serwera, który je udostępnia, druga — po stronie pobierającego je klienta. W pierwszym rozwiązaniu archiwizacja należy do jednostki zarządzającej serwerem (osoby prywatnej, firmy czy instytucji). Zaletą tego modelu jest to, że kopiowany zasób znajduje się na serwerze, którego system plików udostępniono na potrzeby archiwizacji. W tej sytuacji nie trzeba wykorzystywać połączenia sieciowego do gromadzenia danych, które można pozyskiwać przy użyciu takich nośników, jak pamięć flash, dysk twardy, płyta CD lub DVD. Zaletą jest również prędkość przeprowadzanej archiwizacji. W modelu tym pojawiają się jednak istotne wady. Jedną z nich jest konieczność stałej współpracy z jednostką zarządzającą serwerem. Inną — konieczność odtworzenia środowiska do uruchamiania aplikacji, identycznego do stosowanego na serwerze, w celu późniejszego udostępniania zasobu, gdy mamy do czynienia z dynamiczną zawartością, a do jej wygenerowania stosowany jest kod wykonywany po stronie serwera. Archiwizacja po stronie serwera staje się tym samym procesem kłopotliwym i często nieefektywnym.

Drugi model archiwizacji — po stronie klienta — bazuje na wykorzystaniu połączenia internetowego i stosowanych do tego robotów określanych mianem *webcrawlerów* lub *harvesterów*. Wyróżnić tu można kilka etapów. Pierwszym jest przygotowanie wykazu stron przeznaczonych do archiwizacji, ustalenie jej głębokości (typów poziomów zagłębień w strukturę połączonych ze sobą stron internetowych), a także ustalenie ograniczeń w zakresie liczebności i wielkości pobieranych danych. Następnie uruchamiany jest proces archiwizacji, w którym robot zaczyna właściwą pracę. Przemieszczając się po strukturze stron WWW, odnajduje hiperłącza do kolejnych podstron, co pozwala na pobranie zawartości całych serwisów internetowych[5]. Ważny jest stały monitoring pracy robota, gdyż w wypadku pojawienia się problemów, trzeba będzie korygować parametry jego pracy. Działanie robota mogą zakłócić, np. pułapki dla różnych robotów, najczęściej wyszukiwarek, które skutkują zatrzymaniem pracy również robota archiwizującego. Problemem są nowe standardy i technologie, których webcrawlery nie są w stanie obsłużyć. Może się też tak zdarzyć, że ten sam zasób dostępny jest pod różnymi adresami strony internetowej, generowanymi dynamicznie, co prowadzi do redundancji danych.

Wynikiem opisanego procesu są dane w formacie archiwalnym, najczęściej WARC[6].

[5] Tamże.
[6] Specyfikacje formatu WARC dostępne są na stronie internetowej: *WARC File Format specifications.*

Udostępnienie przygotowanych tą metodą zasobów wymaga ich indeksacji, a także utworzenia filtrów, które wygenerują zestaw stron przeznaczonych dla użytkownika końcowego, korzystającego z serwisu prezentującego archiwalne wersje stron internetowych.
Model archiwizacji po stronie klienta jest powszechnie stosowany, a co za tym idzie popularniejszy. Za jego największą zaletę uznać można uniezależnienie udostępnianych danych od technologii stosowanych po stronie poddawanego archiwizacji serwera. W kontekście intensywnych zmian technologicznych korzyść ta jest znacząca. Można też mówić o zaletach stanowiących „pozytywne skutki uboczne", a mianowicie wykrywaniu podczas pracy robota linków prowadzących do nieistniejących już stron internetowych, a także ujawnianiu stron, na których podczas włamania podmieniona została treść i dodano na nich linki do witryn o nielegalnej zawartości czy kontrowersyjnej lub naruszającej prawo treści.

W procesie archiwizacji zasobów internetowych pojawiają się również problemy niezależne od obranej metody. Jest to przede wszystkim kwestia intensywnych modyfikacji narzędzi służących do przeglądania zasobów sieciowych i zmian stosowanych w technologii przeglądarek. Stanowi to swego rodzaju pułapkę, która rozwija się równie intensywnie jak Internet.

## Światowe praktyki w zakresie archiwistyki internetowej

Archiwistyka internetowa ma na świecie niemałą już tradycję, a próby archiwizacji Internetu podejmowało wiele bibliotek. Jednak pierwszy zakończony sukcesem projekt zainicjował amerykański inżynier Brewster Kahle, pomysłodawca Internet Archive — instytucji o charakterze non profit, założonej w 1996 r. w San Francisco w Stanach Zjednoczonych. Rozpoczęto tam archiwizację światowego Internetu, której wyniki zaprezentowano pięć lat później w serwisie Wayback Machine. Rok 2004 był kolejnym krokiem milowym w archiwistyce internetowej. Internet Archive udostępniło oprogramowanie, służącego do archiwizacji robota Heritix, wraz z kodem źródłowym. Oprogramowanie to wyznaczyło pewnego rodzaju standard dla kolejnych światowych projektów archiwizacyjnych i stan ten trwa do dziś[7].

Ważną międzynarodową inicjatywą na polu archiwistyki internetowej było powołanie w 2003 r. International Internet Preservation Consortium (IIPC) — konsorcjum instytucji z całego świata, takich jak archiwa, biblioteki narodowe, uniwersytety oraz inne instytucje kultury, których wspólnym celem jest rozwijanie odpowiednich narzędzi i podejmowanie działań umożliwiających zachowanie obecnych zasobów internetowych dla przyszłych pokoleń. Ważnym elementem prac konsorcjum jest także rozwój oprogramowania służącego do archiwizacji i udostępnianego

[on-line]. [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://archive-access.sourceforge.net/warc/.

[7] Powszechność zastosowania oprogramowania robota Heritrix potwierdzają opracowania i badania w zakresie archiwistyki internetowej: BLEICHER, A. dz. cyt.; LESSIG, L. *Wolna kultura.* Warszawa: Wydawnictwa Szkolne i Pedagogiczne, 2005, s. 135-142 oraz MAYR, M. Harvesting Practices Report. Version 2.0. W: *International Internet Preservation Consortium* [on-line]. June 10, 2011 [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://netpreserve.org/publications/IIPC_Harvesting_Practices_Report.pdf.

powszechnie na zasadach open source.

Wśród projektów i narzędzi, których budowę wspiera IIPC, wyróżnić można między innymi: Heritrix, DeepArc, Web Curator Tool, NetarchiveSuite, BAT, Wayback, NutchWAX, WERA oraz Xing[8]. Na tych narzędziach bazują największe światowe inicjatywy archiwizacji Internetu, podejmowane w takich krajach jak Stany Zjednoczone, Francja, Nowa Zelandia, Wielka Brytania, Australia czy Norwegia. Fakt, iż oprogramowanie to jest bezpłatne, a jego kod źródłowy otwarty i publicznie dostępny, wpłynął znacząco na rozprzestrzenianie się takich działań i podejmowanie prób w kolejnych krajach. Coraz więcej bibliotek i archiwów rozpoczyna projekty mające na celu zachowanie zasobów internetowych. Realizowane są one zarówno przy wsparciu podmiotów zewnętrznych (np. wiodącego w tym zakresie Internet Archive), jak i wyłącznie własnymi siłami z wykorzystaniem wspomnianych narzędzi open source. Coraz częściej archiwizacja Internetu to również przedmiot działań komercyjnych. Właściciele stron internetowych (organizacji, przedsiębiorstw czy instytucji) chcą archiwizować udostępnianą przez nich w Internecie zawartość[9].

Prawne uwarunkowania to czynnik znacząco wpływający na podejmowane na świecie inicjatywy w zakresie archiwizacji. Większość instytucji zmuszona jest udostępniać użytkownikom zarchiwizowane zasoby jedynie lokalnie, w ramach własnej sieci wewnętrznej. Prawa autorskie do danej witryny są tu bowiem czynnikiem decydującym. W skrajnych wypadkach dostęp do gromadzonych treści mają jedynie bibliotekarze lub archiwiści[10]. Instytucją, która przyjęła dużo bardziej kontrowersyjny model udostępniania jest prezentowane wyżej Internet Archive.

## Polskie realia archiwizacji zasobów internetowych i jej prawne uwarunkowania

Pierwszym polskim projektem realizującym zadania w zakresie archiwizacji Internetu, a zarazem jedynym powszechnie znanym jest serwis „Archiwum Internetu", udostępniony użytkownikom przez Narodowe Archiwum Cyfrowe (NAC) w 2010 r.

Poza kwestiami technicznymi, jedną z najważniejszych przeszkód w archiwizacji i udostępnianiu zasobów sieciowych, z jaką musiało się zmierzyć NAC, były uregulowania prawne, które ściśle wyznaczają zakres tego typu działań. Najistotniejszym obwarowaniem jest fakt, iż w rozumieniu art. 1. *Ustawy z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych*[11] zasoby internetowe podlegają ochronie prawno-autorskiej, co niesie za sobą ograniczenia w powszechnym ich udostępnianiu. Art. 28. tejże ustawy daje bibliotekom, archiwom i szkołom możliwość udostępniania gromadzonych danych jedynie na końcówkach systemów informatycznych na terenie tych instytucji. Jednak w kontekście procesu udostępniania oraz idei projektu „Archiwum Internetu", swego rodzaju otwartą przestrzeń stanowi art. 4., według którego spod ochrony prawno-autorskiej

[8] Software. W: *International Internet Preservation Consortium* [on-line]. [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://www.netpreserve.org/software/downloads.php.
[9] MAYR, M. dz. cyt.
[10] Tamże.
[11] Ustawa z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych. *Dz.U.* 1994, nr 24, poz. 83 z późn. zm.

wyłączone są dokumenty urzędowe, jakimi są materiały publikowane na stronach instytucji publicznych (jeśli prowadzone są zgodnie z prawem). Właśnie te witryny archiwizuje i udostępnia Narodowe Archiwum Cyfrowe. W 2005 r. weszła w życie *Ustawa z dnia 17 lutego 2005 r. o informatyzacji działalności podmiotów realizujących zadania publiczne*[12], która znacząco wpłynęła na *Ustawę z dnia 14 lipca 1983 r. o narodowym zasobie archiwalnym i archiwach*, przenosząc aktywność instytucji publicznych również do środowiska teleinformatycznego[13]. Od momentu wprowadzenia tej zmiany archiwa mogą gromadzić dokumentację wartościową z punktu widzenia historii państwa, niezależnie od sposobu jej wytworzenia, w tym również dokumenty elektroniczne, co daje podstawy nie tylko do archiwizacji, ale i do udostępniania stron organów państwowych. Aktem prawnym, który również wpływa na umocnienie strony prawnej tego typu działań jest *Ustawa z dnia 6 września 2001 r. o dostępie do informacji publicznej*[14]. Wytyczne związane z rejestrowaniem stron internetowych w domenie gov.pl[15] oraz *Rozporządzenie Ministra Spraw Wewnętrznych i Administracji z dnia 18 stycznia 2007 r. w sprawie Biuletynu Informacji Publicznej*[16] to również akty prawne, które wyznaczać mogą dalsze kierunki projektu „Archiwum Internetu". Archiwizacji i udostępnianiu mogą podlegać pozostałe witryny w domenie .gov.pl, jak również BIP-y. *Rozporządzenie Ministra Kultury i Dziedzictwa Narodowego z dnia 6 lutego 2008 r. w sprawie zmiany nazwy i zakresu działania Archiwum Dokumentacji Mechanicznej w Warszawie*[17] potwierdza, iż Narodowe Archiwum Cyfrowe to instytucja, do której zakresu prac winny należeć tego typu działania. Obecne zapisy z *Ustawy o prawie autorskim i prawach pokrewnych* nie dają jednak swobody w zakresie powszechnego udostępniania treści z całej domeny .pl, gdyby została podjęta. Zatem wyzwaniem dla takich instytucji jak Narodowe Archiwum Cyfrowe może być podjęcie organizacyjnego i technicznego trudu w zakresie archiwizacji i ewentualnego udostępniania witryn z domeny .pl w ograniczonym zakresie, lecz zgodnym z zapisami art. 28. tejże ustawy.

## Rozwój projektu „Archiwum Internetu" i jego dalsze wyzwania

Projekt „Archiwum Internetu" powstał w Narodowym Archiwum Cyfrowym w 2009 r. i od tego momentu obserwuje się jego stały rozwój. Biorąc pod uwagę opisane regulacje prawne, podczas pierwszej archiwizacji 1 kwietnia 2009 r., zachowane zostały witryny najwyższych władz Polski, m.in. strony Prezydenta RP, Sejmu, Senatu, Kancelarii Prezesa Rady Ministrów, Ministerstwa Kultury i Dziedzictwa Narodowego oraz Ministerstwa Spraw Wewnętrznych i Administracji. Zarchiwizowano

[12] Ustawa z dnia 17 lutego 2005 r. o informatyzacji działalności podmiotów realizujących zadania publiczne. *Dz.U*. 2005, nr 4, poz. 565 z późn. zm.
[13] Ustawa z dnia 14 lipca 1983 r. o narodowym zasobie archiwalnym i archiwach. *Dz.U*.1983, nr 38, poz.1173 z późn. zm.
[14] Ustawa z dnia 6 września 2001 r. o dostępie do informacji publicznej. *Dz.U.* 2001, nr 112, poz. 1198 z późn. zm.
[15] *Wytyczne dotyczące domeny gov.pl w Internecie* [on-line]. Instytut Podstawowych Problemów Techniki Polskiej Akademii Nauk [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://www.ippt.gov.pl/pl/zasady-rejestracji-w-domenie.html.
[16] Rozporządzenie Ministra Spraw Wewnętrznych i Administracji z dnia 18 stycznia 2007 r. w sprawie Biuletynu Informacji Publicznej. *Dz.U.* 2007, nr 10, poz. 68.
[17] Rozporządzenie Ministra Kultury i Dziedzictwa Narodowego z dnia 6 lutego 2008 r. w sprawie zmiany nazwy i zakresu działania Archiwum Dokumentacji Mechanicznej w Warszawie. *Dz.U.* 2008, nr 29, poz. 167.

również strony internetowe sieci archiwów państwowych wraz z witryną Naczelnej Dyrekcji Archiwów Państwowych. Publiczne udostępnienie pierwszych zarchiwizowanych treści odbyło się blisko rok później — 26 marca 2010 r. Efekty prac są ogólnodostępne w Internecie, w serwisie www.archiwuminternetu.nac.gov.pl[18]. Sama archiwizacja nie obyła się bez problemów, gdyż szybko okazało się, że początkowe zasoby sprzętowe Narodowego Archiwum Cyfrowego okazały się niewystarczające. „Wąskim gardłem" była ograniczona przepustowość między siecią NAC a Internetem, co znacznie spowalniało proces pobierania danych. W związku z tym, w początkowej fazie funkcjonowania „Archiwum Internetu", znaczny udział w gromadzeniu zasobów miały komputery testowe Zakładu Urządzeń Informatyki Politechniki Śląskiej w Gliwicach. Gdy problemy wydajnościowe po stronie NAC zostały rozwiązane, wsparcie nie było już tak potrzebne, choć okazywało się pomocne podczas kolejnych archiwizacji. Obecnie prace odbywają się w całości przy wykorzystaniu infrastruktury Centralnego Repozytorium Cyfrowego NAC.

Jednym z istotniejszych ograniczeń procesu archiwizacji i udostępniania zasobów internetowych jest wysoki koszt. Wpływa to na częstotliwość archiwizacji przeprowadzanych przez Narodowe Archiwum Cyfrowe. Obecnie ma to miejsce co 6 miesięcy. Jednak w szczególnych dla społeczeństwa polskiego momentach (jak np. katastrofa pod Smoleńskiem 10 kwietnia 2010 r.) przeprowadzana jest również archiwizacja doraźna, której potrzebę widać w środowisku internautów śledzących bieżące wydarzenia społeczne[19].

W „Archiwum Internetu" wykorzystuje się oprogramowanie robota Heritrix, sprawdzone w tego typu projektach. Wynikiem archiwizacji są pliki w formacie WARC, udostępniane jako zarchiwizowane treści z poziomu oprogramowania Wayback. Cały serwis „Archiwum Internetu" jest posadowiony na serwerach Apache Tomcat oraz Zope. Od początku działania NAC zarchiwizowało ogółem około 0,5 TB danych. Nie wszystkie są udostępnione publicznie, choć sukcesywnie ta liczba rośnie. NAC jest pierwszą polską instytucją, która przeprowadziła archiwizację zasobów internetowych na taką skalę i udostępniła jej efekty. Obecnie archiwizuje się i udostępnia łącznie 41 stron internetowych. Głównie archiwizowane są serwisy dostępne w domenie .gov.pl, jednak zasięg będzie poszerzany tak, by docelowo objąć wszystkie serwisy funkcjonujące w domenie narodowej .pl. Należy przy tym zdawać sobie sprawę, że część polskich zasobów internetowych znajduje się w innych domenach. Aby zatem uzyskać jak najpełniejszy obraz polskiego Internetu, konieczne będzie archiwizowanie polskich serwisów także spoza domeny .pl.

Poza przygotowaniem merytorycznym do archiwizacji tak ogromnej liczby zasobów wymagane jest również odpowiednie zaplecze infrastrukturalne, zatem niezbędny jest jego rozwój w Narodowym Archiwum Cyfrowym. Do kolejnych wyzwań należy zwiększenie liczby serwerów, być może również ich geograficzne rozproszenie. Ponadto przy tylu danych pojawia się problem coraz większej automatyzacji, np. inteligentnej selekcji archiwizowanego materiału oraz dopasowanie częstotliwości archiwizacji stron do rzeczywistej dynamiki zmian, jakie na nich zachodzą.

[18] Odesłanie do strony przedstawia wersję aktualną w dn. 23.01.2012 r.
[19] WILKOWSKI, M. Internetowe archiwum 10 kwietnia. W: *Historia i Media* [on-line]. 20.04.2010 [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://historiaimedia.org/2010/04/20/internetowe-archiwum-10-kwietnia/.

## Podsumowanie

Archiwizacja polskich zasobów sieciowych, jak i przyszłość archiwistyki internetowej w ogóle, stanowią duże wyzwanie dla bibliotek i archiwów. Intensywnie wzrasta tempo zmian w zakresie stosowanych w Internecie technologii, a tym samym słabnie możliwość kontroli przyrastających treści oraz szansa na ich skuteczne utrwalenie. Niezbędny staje się zatem rozwój narzędzi przeznaczonych do archiwizacji tak obszernego i różnorodnego zasobu. Bardzo istotne jest także wprowadzenie zmian w obowiązujących obecnie regulacjach prawnych, by działania w zakresie archiwizacji i udostępniania treści stały się powszechnie możliwe[20]. Przykład archiwizacji prowadzonej w Polsce pokazuje, iż potrzeby w zakresie rozwoju technologii, finansowania, kształcenia specjalistów czy zmian legislacyjnych są ogromne. Archiwizacja 41 witryn internetowych najważniejszych organów państwowych to jedynie kropla w morzu tychże potrzeb.

**Bibliografia:**

[1] Archiwizacja zasobów polskiego Internetu. W: *Program digitalizacji dóbr kultury oraz gromadzenia, przechowywania i udostępniania obiektów cyfrowych w Polsce* [on-line]. Warszawa: Ministerstwo Kultury i Dziedzictwa Narodowego, 2009 [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://www.kongreskultury.pl/library/File/RaportDigitalizacja/Program%20digitalizacji%202009-2020.pdf.

[2] BLEICHER, A. Memory of Webs Past. *IEEE Spectrum Magazine* [on-line]. March 2011 [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://spectrum.ieee.org/telecom/internet/a-memory-of-webs-past.

[3] *HTML5: a vocabulary and associated APIs for HTML and XHTML* [on-line]. W3C Working Draft, 25 May 2011 [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://www.w3.org/TR/html5/.

[4] KAHLE, B. Preserving the Internet. W: *Scientific American Special Online Issue: The Future of The Web* [on-line]. April 2002 [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://www.scientificamerican.com/sciammag/issues.cfm.

[5] LESSIG, L. *Wolna kultura.* Warszawa: Wydawnictwa Szkolne i Pedagogiczne, 2005. ISBN 83-02-09462-5.

[6] MAYR, M. Harvesting Practices Report. Version 2.0. W: *International Internet Preservation Consortium* [on-line]. June 10, 2011 [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://netpreserve.org/publications/IIPC_Harvesting_Practices_Report.pdf.

[7] Software. W: *International Internet Preservation Consortium* [on-line]. [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://www.netpreserve.org/software/downloads.php.

[8] *WARC File Format specifications.* [on-line]. [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://archive-access.sourceforge.net/warc/.

[9] WILKOWSKI, M. Internetowe archiwum 10 kwietnia. W: *Historia i Media* [on-line]. 20.04.2010 [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://historiaimedia.org/2010/04/20/internetowe-archiwum-10-kwietnia/.

[10] *Wytyczne dotyczące domeny gov.pl w Internecie* [on-line]. Instytut Podstawowych Problemów Techniki Polskiej Akademii Nauk [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://www.ippt.gov.pl/pl/zasady-rejestracji-w-domenie.html.

[20] Archiwizacja zasobów polskiego Internetu. W: *Program digitalizacji dóbr kultury oraz gromadzenia, przechowywania i udostępniania obiektów cyfrowych w Polsce* [on-line]. Warszawa: Ministerstwo Kultury i Dziedzictwa Narodowego, 2009 [Dostęp 23.01.2012]. Dostępny w World Wide Web: http://www.kongreskultury.pl/library/File/RaportDigitalizacja/Program%20digitalizacji%202009-2020.pdf.

W pracy powołano się na następujące akty prawne:

- Ustawa z dnia 14 lipca 1983 r. o narodowym zasobie archiwalnym i archiwach. *Dz.U.* 1983, nr 38, poz. 1173 z późn. zm.
- Ustawa z dnia 4 lutego 1994 r. o prawie autorskim i prawach pokrewnych. *Dz.U.* 1994, nr 24, poz. 83 z późn. zm.
- Ustawa z dnia 6 września 2001 r. o dostępie do informacji publicznej. *Dz.U.* 2001, nr 12, poz. 1198 z późn. zm.
- Ustawa z dnia 17 lutego 2005 r. o informatyzacji działalności podmiotów realizujących zadania publiczne. *Dz.U.* 2005, nr 4, poz. 565 z późn. zm.
- Rozporządzenie Ministra Spraw Wewnętrznych i Administracji z dnia 18 stycznia 2007 r. w sprawie Biuletynu Informacji Publicznej. *Dz.U.* 2007, nr 10, poz. 68.
- Rozporządzenie Ministra Kultury i Dziedzictwa Narodowego z dnia 6 lutego 2008 r. w sprawie zmiany nazwy i zakresu działania Archiwum Dokumentacji Mechanicznej w Warszawie. *Dz.U.* 2008, nr 29, poz. 167.

Kłębczyk, F., Jędralska, M. Serwis "Archiwum Internetu" na tle ogólnych problemów archiwizacji zasobów sieciowych. W: *Biuletyn EBIB* [online] 2012, nr 1 (128), *Archiwizacja Internetu* [Dostep: 24.02.2012] Dostępny w World Wide Web: http://www.nowyebib.info/images/stories/numery/128/128_jedralska.pdf. ISSN 1507-7187.