Lilianna Nalewajska
Biblioteka Uniwersytecka w Warszawie

Archiwizowanie stron internetowych w krajach nordyckich

***Streszczenie**: Kraje nordyckie (Norwegia, Szwecja, Finlandia oraz Dania i Islandia) należą do grupy pionierów w zakresie archiwizowania materiałów z Internetu. Proces gromadzenia materiałów z sieci, wymagający ustaleń dotyczących rozwiązań techniczno-technologicznych, prawnych i organizacyjnych, rozpoczęto w tych krajach pod koniec lat 90. XX w. lub na początku XXI stulecia. Archiwizacją zajmują się głównie biblioteki narodowe, które ponadto współpracują z International Internet Preservation Consortium oraz współtworzą Nordic Web Archive. Sposób funkcjonowania oraz trudności, które pojawiają się w trakcie prac archiwizacyjnych w tych krajach pokazują, jak złożony jest to proces oraz jak perspektywicznego i długoterminowego planowania wymaga.*

***Słowa kluczowe**: archiwizacja Internetu, archiwizacja stron WWW, Web archiving, Web archive*

Dynamika sieci Internet — z jednej strony jej niepohamowany rozrost, a z drugiej szybkie zmiany (aktualizacje, rozbudowy) lub zanikanie stron internetowych (czas ich egzystencji jest niezwykle zróżnicowany[1]), doprowadziła do inicjatyw archiwizowania elektronicznych materiałów i publikacji born digital. Materiały publikowane tylko on-line, obok tych wydawanych tradycyjną metodą, również dokumentują życie współczesnego społeczeństwa, stanowią „elektroniczne dziedzictwo" kraju, zatem ich gromadzeniem zajmują się zwykle biblioteki narodowe.

W artykule zebrano informacje dotyczące sposobów organizacji i funkcjonowania archiwów stron internetowych w krajach skandynawskich (Norwegii, Szwecji, Danii) oraz w Finlandii i Islandii. Niemal wszystkie te kraje były pionierami na polu archiwizowania sieci Internet. Zestawienie działań związanych z archiwizowaniem materiałów z sieci podejmowanych w krajach nordyckich pokazuje, jak złożony jest to proces, jak perspektywicznego i długoterminowego planowania wymaga oraz jak istotna dla przyszłego funkcjonowania i rozwoju archiwów jest współpraca bibliotek, krajowa i międzynarodowa.

Archiwum materiałów internetowych to złożony i skomplikowany gmach. Jego architektura wymaga uwzględnienia wielu aspektów: rozwiązań technologicznych i technicznych związanych z metodologią gromadzenia i selekcjonowania zasobów (automatycznie lub wybiórczo wg przyjętych kryteriów; z uwzględnieniem częstotliwości), organizacji pracy. W pierwszej kolejności wymaga jednak ustaleń i regulacji prawnych dotyczących własności intelektualnej, związanych z egzemplarzem obowiązkowym oraz umożliwieniem przeszukiwania sieci instytucjom zajmującym się archiwizacją e-materiałów. Kolejne aspekty, które należy uwzględnić to ochrona danych osobowych, sposoby udostępniania zgromadzonego materiału (on-line czy tylko na dedykowanych stanowiskach w bibliotekach, komu — wszystkim

[1] Średnia długość "życia" stron internetowych jest różnie określana — waha się między 44 a 75 dni, czasem podawana jest liczba 100 dni. Za: *Preservation of Web Resources: a JISC-funded project [Archived Blog]* [on-line]. [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://jiscpowr.jiscinvolve.org/wp/2009/08/12/whats-the-average-lifespan-of-a-web-page/.

czy limitować, np. tylko pracownikom naukowym), możliwość kopiowania. Niezbędne są również perspektywiczne rozważania dotyczące długoterminowego przechowywania dokumentów cyfrowych.

Ideę zachowania materiałów z Internetu wyprzedziły nieco koncepcje dotyczące archiwizowania elektronicznych informacji naukowych i technicznych, które pojawiły się już w połowie lat 90. XX w.[2]. Gromadzenie statycznych publikacji elektronicznych było możliwe po wprowadzaniu nowych zapisów w prawie o egzemplarzu obowiązkowym, nakładających obowiązek gromadzenia wszystkich publikacji powstających w danym kraju, zarówno w formacie fizycznym, jak elektronicznym, niezależnie od techniki ich produkcji i nośnika, np. w Szwecji gromadzono początkowo publikacje elektroniczne zapisane na nośnikach fizycznych (dyskietki, płyty CD)[3]. Niemal w tym samym czasie zwrócono uwagę na konieczność archiwizowania także stron internetowych. Szwedzki projekt Kulturarw3[4], rozpoczęty we wrześniu 1996 r. przez Bibliotekę Królewską, zakładał pełne gromadzenie szwedzkiej części Internetu — wszystkich stron z domeną .se.

Na szwedzkim projekcie wzorowała się Finlandia, która pierwszą archiwizację fińskich stron WWW zakończyła w czerwcu 2002 r. Opierając się na przyjętych uregulowaniach prawnych dotyczących egzemplarza obowiązkowego Norwegia, Dania, Szwecja, Islandia, Finlandia i Australia rozpoczęły archiwizowanie stron internetowych, przy czym Dania i Australia gromadziły tylko wybrane strony, a Finlandia zbierała kompleksowo swoje narodowe e-strony[5]. Konieczne było opracowanie strategii selekcji, gromadzenia, opisu, identyfikacji i przechowywania dokumentów cyfrowych oraz sposobu ich udostępniania[6]. Generalnie z archiwizacji zostały wyłączone strony wymagające od użytkowników określonej interaktywności (włączenie wtyczki, logowanie się) lub zasoby głębokiego Internetu. W krajach uwzględnionych w niniejszym artykule budowę archiwów poprzedzało zazwyczaj uczestnictwo w innych projektach dotyczących gromadzenia materiałów cyfrowych oraz rozpoznawanie możliwości gromadzenia materiałów z sieci zarówno pod względem technicznym, jak i prawnym.

---

[2] Zainteresowanie ICSTI (International Council for Scientific and Technical Information) oraz CENDI (The Federal Scientific and Technical Information Managers' Group) (grupę tworzy 14 agencji federalnych USA ds. nauki i techniki) archiwizacją informacji cyfrowej datuje się od 1996 r. Za: HODGE G., FRANGAKIS E. *Digital preservation and permanent access to scientific information: the state of the practice* [on-line]. *CENDI* 2004-3, s. 3. [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://www.cendi.gov/publications/04-3dig_preserv.pdf. Jednym z pionierów w archiwizacji Internetu była Narodowa Biblioteka Australii — projekt Pandora (http://pandora.nla.gov.au/) zakładał gromadzenie publikacji on-line, w tym strony partii politycznych. Za: VOERMAN, G. [i in.]. Archiving the web: political party web sites in the Netherlands. *Information Services & Use* 2003, nr 23, s. 2.

[3] Tamże, s. 14.

[4] ARVIDSON, A., PERSSON, K., MANNERHEIM, J. The Kulturarw3 Project — The Royal Swedish Web Archiw3e — An example of "complete" collection of web pages. W: *Conference proceedings, 66th IFLA Council and General Conference*, Jerusalem. Israel, 13-18 August 2000 [on-line]. [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://archive.ifla.org/IV/ifla66/papers/154-157e.htm.

[5] HODGE G., FRANGAKIS E. dz. cyt., s.13.

[6] Tamże, s. 13 (Norweska Biblioteka Narodowa w latach 2001-2004 podjęła opracowanie tych kwestii w projekcie Paradigma).

Szwecja[7]

Szwedzka Biblioteka Królewska (odpowiednik narodowej), która od 1661 r. gromadzi wszystkie krajowe publikacje, w 1996 r. rozpoczęła realizację projektu Kulturarw3 (The Swedish Web Archives) mającego na celu opracowanie strategii archiwizowania szwedzkich stron internetowych. Głównym założeniem projektu było przetestowanie metod gromadzenia, przechowywania i udostępniania dokumentów elektronicznych. Przyjęto zasadę automatycznego zbierania materiałów z sieci. Pierwszą archiwizację przeprowadzono w 1997 r., zgromadzono wówczas 6,8 mln URL z 15 700 stron internetowych (w 2001 r. uzyskano 30 mln obiektów ze 126 000 stron internetowych, z czego 90% stanowiły strony HTML oraz obrazy JPEG i GIF)[8]. W latach 1996-2000 przeprowadzono siedem kompletnych pobrań. Gromadzone są strony z domen .se oraz .nu (Wyspa Niue), jak również strony z domen .org, .com, .net zarejestrowane pod szwedzkim adresem lub numerem telefonu, a więc fizycznie zlokalizowane w Szwecji. Przyjęto strategię zbierania materiałów z sieci kilka razy do roku — automatyczne oprogramowanie zbiera „migawki" (snapshots) z każdej strony (w 2008 r. ok. 306 mln URL, tj. ok. 10 terabajtów plików; w 2003 r. zapisanych było 4,5 terabajta danych). W ten sposób powstaje dosyć kompletne archiwum szwedzkich stron WWW.

Głównym problemem jest jednak długi czas pobierania materiałów, który trwa kilka miesięcy. W przypadku często zmieniających się stron (np. gazet elektronicznych) gromadzenie musi odbywać się częściej (tygodniowo, a nawet codziennie). Rozważano możliwość automatycznego sprawdzania, jak często zmienia się materiał na stronach i odpowiednie dostosowanie częstotliwości jego zapisu w archiwum[9].

Na początku projektu Szwedzka Biblioteka Królewska stała na stanowisku, że każdy obywatel ma prawo do informacji i opowiadała się za otwartym dostępem do zarchiwizowanych materiałów. Ministerstwo Edukacji proponowało z kolei zawężenie dostępu tylko do grupy pracowników naukowych związanych z wybranymi instytucjami[10]. Sytuację uregulowały ustalenia rządowe z 2002 r. uprawniające Szwedzką Bibliotekę Królewską do udostępniania zebranych materiałów wyłącznie na terenie biblioteki[11]. Ostatecznie dostęp do archiwum jest możliwy tylko na kilku komputerach w bibliotece narodowej.

Nadal jednak trwają prace nad legislacją dotyczącą e-egzemplarza obowiązkowego. Jesienią 2011 r. spodziewano się ogłoszenia przez rząd odpowiedniej ustawy. Regulacje prawne stanowią, że wytwórca dostarcza materiał w formacie, w jakim powstał dokument, co oznacza, że biblioteka narodowa ma mało możliwości, by uzyskać format odpowiedni do długotrwałego archiwizowania. Wydawca dostarcza

[7] Informacje o archiwizowaniu stron WWW w Szwecji pochodzą z: *Kulturarw3* [on-line]. [Dostęp 06.11.2011]. Dostępny w World Wide Web: http://www.kb.se/english/find/internet/websites/; DAY, M. *Collecting and preserving the World Wide Web. Version 1.0 — 25* [on-line]. University of Bath, February 2003, s. 24-25 [Dostęp 06.11.2011]. Dostępny w World Wide Web: http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf oraz korespondencji mailowej autorki z Elisabeth Mannerfeldt z Biblioteki Królewskiej w Sztokholmie.
[8] DAY, M. dz. cyt., s. 24.
[9] ARVIDSON, A., PERSSON, K., MANNERHEIM, J. dz. cyt.
[10] Tamże.
[11] Tamże.

metadane obejmujące: nazwę wytwórcy/autora, link do opublikowanego materiału, datę publikacji oraz format dokumentu, bez informacji na temat jego treści. Zebrany materiał przeszukuje się przy użyciu typowych wyszukiwarek, jednak nie jest on indeksowany, dlatego trzeba znać dokładny adres URL. Kulturarw3 posiada licencję ArchiveWare, co pozwala śledzić poszczególne fragmenty 10 mln stron internetowych i umożliwia dostęp do nich[12].

Finlandia

Fińska Biblioteka Narodowa jest w prezentowanej grupie wiodąca w zakresie archiwizowania dokumentów z Internetu, zarówno pod względem organizacji archiwum, jak i dostępnych ustaleń dotyczących jego funkcjonowania. Ustalenia te zawarte zostały w dokumencie *Web archiving in Finland. Memorandum for the members of the CDNL*[13] z 13 grudnia 2010 r., przygotowanym w Fińskiej Bibliotece Narodowej. Stosowane obecnie rozwiązania poprzedził projekt Eva[14] z lat 1997-2001, dotyczący archiwizacji statycznych dokumentów HTML ogólnodostępnych w sieci wraz z zamieszczanymi w nich materiałami on-line (zdjęcia, klipy audio i wideo). Projekt ten zakładał gromadzenie wyłącznie stron z domeny .fi, choć już wówczas wiedziano o istnieniu wielu istotnych stron dotyczących Finlandii, ale zlokalizowanych pod innymi adresami[15].

Potrzebę, a wręcz konieczność archiwizowania publikacji elektronicznych dostrzeżono w Finlandii na początku lat 90. XX w. Wiązało się to z rewizją prawa o egzemplarzu obowiązkowym z 1980 r. wobec wzrastającej liczby publikacji w Internecie. W 1998 r. Ministerstwo Edukacji powołało grupę do opracowania raportu na ten temat, który ukończono w tym samym roku, a kolejny dwa lata później. Następnym krokiem była konieczność zharmonizowania zapisów zawartych w prawie autorskim. Nowe prawo autorskie weszło w życie 1 stycznia 2006 r. Datę tę przyjmuje się za moment powstania Fińskiego Archiwum Stron Internetowych, ponieważ nowe prawo dało bibliotece narodowej możliwość archiwizowania tych materiałów. Z początkiem 2008 r. weszła w życie nowa regulacja prawna *Cultural materials depositing and preservation act*[16], która w przeciwieństwie do wcześniejszego prawa, nie tylko umożliwiała, lecz wręcz nakazywała gromadzenie w

[12] Web harvesting for nuclear knowledge preservation. *IAEA Nuclear Energy Series* [on-line]. 2008, nr NG-T-6.6, s. 15 [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://iaea.org/inisnkm/nkm/documents/publ_web_harvesting.pdf.

[13] KESKITALO, E. *Web archiving in Finland*. *Memorandum for the members of the CDNL* [on-line]. National Library of Finland, 13 December 2010 [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://www.cdnl.info/2010/Web%20Archiving%20in%20Finland,%20%20E-P%20Keskitalo%20-%20December%202010.pdf.

[14] *Eva — the acquisition and archiving of electronic network publications* [on-line]. [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://web.archive.org/web/20041010005510/www.lib.helsinki.fi/eva/english.html, http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/92.html#fin. Liderem projektu EVA była Fińska Biblioteka Narodowa, a uczestniczyły w nim trzy fińskie uniwersytety oraz wydawca Edita i CSC Scientific Computing.

[15] DAY, M. dz. cyt., s. 25. Archiwizacja z marca 1998 r. — zapisano 1,8 mln dokumentów z ok. 7500 fińskich stron WWW.

[16] Dokument dotyczy także Narodowego Archiwum Audiowizualnego w Finlandii. Ponadto w dokumencie nie stosuje się określenia „strony WWW" (Web), lecz bardziej ogólnego: „sieci danych" (data networks).

archiwum materiałów elektronicznych ogólnodostępnych w sieci. Obydwa dokumenty pozwoliły stworzyć system elektronicznego egzemplarza obowiązkowego. Fińska Biblioteka Narodowa na mocy *Cultural materials depositing and preservation act* staje się właścicielem materiałów zebranych z sieci. W obrębie prawa autorskiego istotną kwestią było ustalenie możliwości wykonywania kopii (biblioteka narodowa ma prawo do kopiowania materiałów, które są udostępniane publicznie w sieci) oraz tego, w jaki sposób, gdzie i przez kogo kopie takie mogą być używane. Wyróżniono osiem instytucji, w których kopie mogą być używane, są to m.in. Fińska Biblioteka Narodowa i Biblioteka Parlamentarna. Użycie materiałów możliwe jest tylko na wyznaczonych w tych instytucjach stanowiskach oraz wyłącznie do celów naukowo-badawczych i prywatnych.

Zgodnie z *Cultural materials depositing and preservation act* biblioteka narodowa powinna gromadzić materiał z różnych okresów, reprezentatywny i wieloaspektowy. Gromadzenie pod względem technicznym powinno odbywać się głównie automatycznie, czyli bez angażowania w ten proces wydawcy. Jeśli nie jest to możliwe, wówczas wydawca powinien umożliwić zebranie materiału (np. przez dodanie biblioteki narodowej do listy akceptowanych adresów IP) lub zdeponować materiał (np. umieścić go na dysku twardym lub na serwerze SFTP). Istnieją także określone wymagania dotyczące jakości i autentyczności materiałów. Fińska Biblioteka Narodowa jest zobowiązana do takiego przechowywania materiałów, by było możliwe weryfikowanie ich prawdziwości (muszą być zapisane data włączenia do archiwum oraz pierwotna lokalizacja). Materiały powinny być kompletne, bez wad technicznych, zgodne z tym, co było udostępnione dla odbiorców w sieci oraz nie mogą być zabezpieczone (tak, aby możliwe było przenoszenie materiału z jednego nośnika pamięci na inny lub zmiana jego formatu — biblioteka narodowa ma prawo obchodzić takie zabezpieczenia). W momencie deponowania materiały muszą być zaopatrzone w opisowe metadane oraz informację o dotyczącym ich prawie autorskim.

Fińska Biblioteka Narodowa archiwizuje strony WWW od 2006 r. Gromadzi strony internetowe z domeny fińskiej .fi lub Wysp Alandzkich .ax oraz strony z innych domen, przechowywane na urządzeniach zlokalizowanych w Finlandii. Strony płatne lub wymagające logowania gromadzone są osobno i nie udostępnia się ich w serwisie biblioteki. Najnowszy zarchiwizowany materiał jest udostępniany z półrocznym opóźnieniem. Gromadzone są raczej tradycyjne, statyczne strony Web, głównie ze względu na trudności z automatycznym gromadzeniem stron dynamicznych i interaktywnych. Strony zabezpieczone robotami wykluczającymi (robots.txt) nie są archiwizowane[17]. Można w ten sposób zastrzec własną stronę, jednak biblioteka narodowa prawomocnie może gromadzić i udostępniać wszelkie fińskie materiały z sieci, niezależnie czy strony wymagają logowania, subskrypcji, a nawet jeśli właściciel strony chce uniknąć jej gromadzenia przez stosowanie robots.txt.[18]. Do przeszukiwania i archiwizowania używane jest ogólnodostępne oprogramowanie Heritrix.

[17] *Finnish Web Archive. Additional information* [on-line]. [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://webarchive.nationallibrary.fi/info.jsp?lang=en.
[18] Tamże.

Zasoby archiwum udostępniane są w ośmiu instytucjach w Finlandii: w Bibliotece Narodowej[19] na trzech stanowiskach komputerowych oraz w pięciu bibliotekach otrzymujących egzemplarz obowiązkowy (m. in. Turku, Abo), również w Bibliotece Parlamentarnej oraz w Narodowym Archiwum Audiowizualnym. Na stanowiskach w Fińskiej Bibliotece Narodowej możliwe jest przeszukiwanie zarchiwizowanych stron WWW, odsłuchanie nagrania audio lub obejrzenie nagrania telewizyjnego, nie można jednak uzyskać kopii elektronicznej, a jedynie analogową — wolno fotografować materiał widoczny na ekranie lub nagrywać z głośników odtwarzany materiał. Dokumenty można odpłatnie drukować. Plików nie można zapisywać na urządzeniach pamięci przenośnej ani przesyłać mailem.

Wielkość zgromadzonego materiału podawana jest w ilości plików (nie jest możliwe przeliczenie na „dokumenty", gdyż strona w HTML może zawierać od zera do setek plików). W 2010 r. zebranych było 200 mln plików, co odpowiada 25 terabajtom. W 2006 r. zebrano 20 mln plików (3 terabajty), a w latach 2006-2009 zarchiwizowano ok. 290 mln plików, czyli ponad 40 terabajtów danych z sieci. Zgromadzony materiał to w 63,8% strony HTML, 34,1% obrazy (image), 1,8% PDF, 0,2 % dźwięk i 0,1% wideo[20].

W archiwum nie można gromadzić materiałów o treściach zabronionych przez *Kodeks karny*, m.in. podżegających do nienawiści etnicznej, upowszechniających przemoc lub oszczerstwa, akty przemocy lub akty seksualne z udziałem dzieci. Choć Fińska Biblioteka Narodowa nie ma odpowiednich umocowań prawnych, jednak może podejmować odpowiednie czynności prawne i już raz usunęła z archiwum taki materiał.

Nad funkcjonowaniem archiwum pieczę sprawuje Ministerstwo Edukacji i Kultury. W związku z tym biblioteka narodowa przygotowuje raporty roczne oraz plany dotyczące wielkości i sposobów deponowania materiałów z sieci oraz technicznych rozwiązań i finansowania, które zatwierdza ministerstwo. Dotychczas złożono jeden taki plan, kolejny miał zostać przygotowany w 2011 r. Archiwum prowadzi 10 osób, które nadzorują wykonanie zobowiązań narzuconych przez *Cultural materials depositing and preservation act*: informatyków (bieżące gromadzenie materiału, indeksowanie, prace związane z interfejsem użytkownika, wyszukiwarką, infrastrukturą — serwery, miejsce na dyskach, kopie zapasowe, dedykowane stacje robocze i łącza z nimi, inne zadania dotyczące gromadzenia i długoterminowego planowania działań); pracownik biblioteki narodowej odpowiedzialny za nadzór i planowanie oraz bibliotekarze zajmujący się zasobami elektronicznymi.

Fińska Biblioteka Narodowa, mimo że znajduje się w strukturze Uniwersytetu w Helsinkach, finansowana jest przez Ministerstwo Kultury i Edukacji. Trudno określić koszty utrzymania archiwum elektronicznego, ale szacuje się je na 600 tys. euro rocznie. Bierze ona udział w projekcie Narodowej Biblioteki Cyfrowej, którego założeniem jest wypracowanie wspólnego systemu długoterminowego zachowywania

[19] Archiwum w bibliotece narodowej otwarto 2 kwietnia 2009 r. Zob. Finland's Web Archive opened. *The National Library of Finland Bulletin* [on-line]. 2009 [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://www.kansalliskirjasto.fi/extra/vanhat_bulletinit/bulletin09/hi2.html.
[20] KESKITALO, E. dz. cyt., s. 13-14.

materiałów z sieci przez fińskie biblioteki, archiwa i muzea. Uruchomienie usługi jest planowane w roku 2014.

Dania

W ramach duńskich przepisów legislacyjnych już w 1997 r. rozszerzono wymogi dotyczące egzemplarza obowiązkowego na publikacje cyfrowe na fizycznych nośnikach oraz na statyczne dokumenty on-line, co miało umożliwić w przyszłości gromadzenie dokumentów elektronicznych z sieci. Na tej podstawie Biblioteka Królewska rozpoczęła w 1998 r. selektywne gromadzenie materiałów elektronicznych (cyfrowe produkty na nośnikach fizycznych, statyczne publikacje on-line oraz czasopisma elektroniczne za zgodą wydawców). Niemal równolegle, bo w 2001 r., Biblioteka Królewska oraz Biblioteka i Centrum Badań Internetu (Center for Internet Research) na Uniwersytecie w Aarhus rozpoczęły inicjatywę Netarkivet.dk (http://netarchive.dk)[21,22]. Pierwsze testy prowadzono do lipca 2002 r.[23] Na stronie serwisu zamieszczane są informacje na temat prac nad archiwizowaniem. Można skorzystać z gotowych pytań i odpowiedzi (FAQ) w języku angielskim skierowanych do kuratorów stron internetowych, naukowców i innych osób zainteresowanych pracami Netarkivet.dk. Z Netarkivet.dk przy budowie systemu do archiwizowania stron internetowych NetarchiveSuite współpracują Bibliothèque national de France oraz Österreichische Nationalbibliothek.

Duńskie strony internetowe archiwizowane są przez Biblioteki: Królewską, czyli Narodową i Uniwersytecką (która organizacyjnie należy do Biblioteki Królewskiej, ale jest oddzielna jednostką) od 1 lipca 2005 r.[24], kiedy weszło w życie nowe prawo dotyczące egzemplarza obowiązkowego. Podobnie jak w Finlandii, strony przeszukują narzędzia Web crawlers. By zgromadzić jak najbardziej kompletny materiał z domeny duńskiej .dk, stosuje się trzy strategie:

- ogólne przeglądanie stron z domeny .dk (ok. 1,1 mln stron, z czego 1 mln aktywnych), stron z domeny innej niż .dk, jednak ich zawartość jest skierowana do duńskiego odbiorcy i/lub właściciel jest duński (ok. 45 tys. stron z domeny .com, .org, .nu) — przeprowadzane cztery razy w roku (duże strony tylko dwukrotnie). Podczas gromadzenia w roku 2010/2011, przeprowadzonego po raz jedenasty, zebrano 23,9 terabajtów z

[21] Wszystkie odesłania w tekście i przypisach do stron internetowych przedstawiają wersję aktualną w dn. 14.01.2012 r.
[22] Netarkivet.dk jest zarządzane przez komitet składający się z sześciu specjalistów (po trzech z każdej biblioteki) w dziedzinie ochrony zasobów cyfrowych, informatyki, polityki egzemplarza obowiązkowego i tworzenia narodowej kolekcji.
[23] Wyniknęły wówczas pierwsze przeszkody w bezpośrednim pobieraniu materiałów z sieci — konieczne okazały się porozumienia z właścicielami stron internetowych dotyczące dostępu i częstotliwości archiwizowania. Zob. *PADI : Web archiving* [on-line]. [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/92.html#den.
[24] Akt nr 1439 z 22 grudnia 2004 r. Egzemplarz obowiązkowy materiałów publikowanych w części 3 dotyczy materiałów publikowanych w sieci elektronicznej — wersja angielska dostępna on-line: *Act on legal deposit of published material*. *Translation of act No. 1439 of 22* [on-line]. December 2004 [Dostęp 5.11.2011]. Dostępny w World Wide Web: http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html.

respektowaniem robots.txt. Był to eksperyment, który miał potwierdzić tezę, że taki sposób gromadzenia pomija dużą część obrazów i plików PDF;
- selektywne — prowadzone z różną częstotliwością — raz w miesiącu do sześciu razy dziennie; ukierunkowane na gromadzenie stron często aktualizowanych, które mogłyby zostać pominięte przy ogólnym gromadzeniu (chodzi tu np. o strony z wiadomościami w mediach krajowych i regionalnych); strony często odwiedzane, reprezentatywne dla społeczeństwa duńskiego, sektora handlowego i władz; strony eksperymentalne lub unikatowe, dokumentujące nowe sposoby wykorzystania sieci (np. net art.). Obecnie jest zgromadzonych 101 takich stron, z czego 46 to szeroko rozumiane strony newsowe (od polityki, przez sport, ekonomię, nawet strony plotkarskie);
- gromadzenie stron tworzonych na specjalne okazje, wydarzenia — strony o krótkim okresie egzystencji (w roku 2011 zgromadzono strony dotyczące zamachu bombowego w Oslo z 22 lipca, głównie w celu zebrania stron duńskich prawicowych fundamentalistów), strony związane z wyborami w Danii w listopadzie 2011 r. oraz dotyczące Olimpiady w Londynie w 2012 r.

Duńskie archiwum nie stosuje zaleceń robots.txt, ze względu na to, że utrudniałoby to gromadzenie wielu istotnych stron, które rygorystycznie spełniają zalecenia robots.txt (strony partii politycznych, mediów z wiadomościami). Dlatego kolejne gromadzenie rozpoczęte 17 sierpnia 2011 r. ignoruje robots.txt

Aktualne informacje o funkcjonowaniu duńskiego archiwum stron internetowych publikuje Netarkivet.dk w *Newsletterze* z sierpnia 2011 r.[25] Statystyki na 1 lipca 2011 r. podają, że archiwum zawierało 222 terabajty (ok. 6 mld obiektów, zaś w sierpniu 2010 r. archiwum zawierało 155 terabajtów, co równa się ok. 4,5 mld obiektów). Najczęstsze formaty plików to HTML, JPEG, GIF, PNG. Zebrany materiał jest przechowywany w obydwu bibliotekach ze względu na bezpieczeństwo. Materiały nie są udostępniane szerokim kręgom użytkowników, a jedynie, przynajmniej w fazie początkowej, mogą być używane do celów naukowo-badawczych za zgodą Duńskiej Agencji Ochrony Danych (Danish Data Protection Agency)[26].

Norwegia[27]

Gromadzenie materiałów z sieci internetowej umożliwia w Norwegii prawo o egzemplarzu obowiązkowym z 1990 r. Norwegia była jednym z pierwszych krajów na

[25] *Newsletter. Netarchive.dk* [on-line]. August 2011 [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://netarkivet.dk/nyheder/Newsletter_Netarchive_dk_august2011.pdf.
[26] Planowano rozszerzenie dostępności do archiwum podczas sesji parlamentarnej w 2011r. Dane zebrane w archiwum internetowym, w tym „wrażliwe" dane osobowe, są objęte ochroną na mocy *Ustawy o przetwarzaniu danych osobowych* oraz Duńskiej Agencji Ochrony Danych i nie mogą być udostępniane publicznie. Brakuje rozwiązań technicznych gwarantujących całkowitą pewność, że „wrażliwe" dane osobowe nie zostaną ujawnione szerokim kręgom. Zob. *FAQ. Netarchive.dk* [on-line]. [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://netarkivet.dk/faq/index-en.php.
[27] Informacje o norweskim archiwum stron internetowych zostały zaczerpnięte ze stron konsorcjum netpreserve.org (http://netpreserve.org/about/archiveList.php) oraz z korespondencji elektronicznej autorki z Kjersti Rustad, kierownikiem Sekcji Egzemplarza Obowiązkowego Monografii w Oddziale Gromadzenia i Usług Bibliograficznych Norweskiej Biblioteki Narodowej.

świecie, który uwzględnił cyfrowe dokumenty w tych przepisach. Mimo, że w 1990 r. nie funkcjonował jeszcze Internet w formie, jaką znamy dzisiaj, materiały ogólnodostępne w sieci stały się już przedmiotem prawa[28].

Norweska Biblioteka Narodowa nim przystąpiła do tworzenia archiwum stron internetowych, uczestniczyła w kilku projektach (Nedlib, Biblink, Nordic Web Archive), obejmujących zagadnienie egzemplarza obowiązkowego dokumentów cyfrowych (zdigitalizowanych). Na tej podstawie przygotowano projekt Paradigma, który rozpoczęto w sierpniu 2001 r. i zamierzano wypracować pełną „linię produkcyjną" gromadzenia i archiwizowania cyfrowych dokumentów oraz legislacji dotyczącej egzemplarza obowiązkowego (norweskie prawo pomijało narzędzia i infrastrukturę publikacji internetowych, dotyczyło publikacji cyfrowych na nośnikach fizycznych)[29]. Celem projektu Paradigma było zbudowanie systemu zdolnego do zbierania pełnego materiału z sieci i automatyzacji tego procesu. Gromadzeniu miały podlegać strony z domeny .no lub z domen należących do norweskich instytucji i osób fizycznych (sprawdzanie do kogo należy strona przez serwis Whois; inne dokumenty kwalifikowano według języka — norweskiego lub lapońskiego). Projekt zakładał dostęp do zarchiwizowanych materiałów tylko do celów badawczych i dokumentacyjnych. Aby ułatwić dostęp, przewidywano możliwość przesyłania materiału do lokalnej biblioteki, a nawet na komputer użytkownika w zakodowanej formie (odkodowanie po uruchomieniu wtyczki udostępnionej przez bibliotekę narodową). Oczywiście zapisywanie, ściąganie materiału, drukowanie zostało wykluczone. Możliwy miał być jedynie czasowy dostęp na ekranie. Takie były założenia. Obecnie prowadzone jest tzw. ciemne archiwum — zgromadzony materiał nie jest udostępniany, jedynie pracownicy mają do niego dostęp.

Próbne, selektywne przeszukania i gromadzenie prowadziła Norweska Biblioteka Narodowa od połowy lat 90. XX w. Zebrane materiały rejestrowano w katalogu biblioteki. Gromadzono także materiały dotyczące szczególnych wydarzeń. Proces zbierania materiałów z sieci rozpoczęto w 2001 r. Gromadzone są strony z domeny norweskiej .no oraz strony okazjonalne.

Od 2001 r. stosowano kilka metod archiwizowania materiałów:

- w latach 2001–2004 oraz od 2009 r. selektywne;
- od 2002 r. — wyszukiwanie według domeny prowadzone raz lub dwa razy w roku;
- od 2001 r. gromadzi się też strony okazjonalne (narodowe i lokalne wybory, śluby królewskie) dotyczące społeczności norweskiej.

Pierwsze całościowe przeszukanie stron z domeny .no przeprowadzono w grudniu 2002 r. W 2005 r. Norweska Biblioteka Narodowa rozpoczęła regularne pobieranie norweskich stron internetowych. W 2008 r. rozszerzono jednak zasięg archiwizacji —

[28] Podstawa prawna stanowi, że ogólnodostępne materiały, niezależnie od ich formatu czy nośnika, muszą być archiwizowane i udostępniane jako materiał źródłowy do celów naukowych i dokumentacyjnych.

[29] ALBERTSEN, K. The paradigma web harvesting environment. W: *3rd ECDL Workshop on Web Archives*, August 21st, 2003, Trondheim, Norway in conjunction with the 7th European Conference on Research and Advanced Technologies for Digital Libraries [on-line]. [Dostęp 10.11.2011]. Dostępny w World Wide Web: http://bibnum.bnf.fr/ecdl/2003/.

poza stronami wydawców norweskich z domeną .no gromadzone są także materiały skierowane do norweskich odbiorców pochodzące z innych domen (.org, .com). W 2009 r. dołączono zbieranie pojedynczych dokumentów, jak PDF (raporty, e-booki, e-czasopisma, materiały szkoleniowe) oraz materiały dokumentujące wydarzenia o znaczeniu społecznym i narodowym (m.in. dotyczące ataku terrorystycznego z 22 lipca 2011 r.). Początkowo używano narzędzia Nedlin Harvester, a od 2005 r. Heritrix. W latach 2001–2005 prowadzono selektywne gromadzenie przy użyciu HTTrack, a od 2009 r. — Web Curator Tool (WCT).

Kwestię udostępniania zarchiwizowanych materiałów reguluje w Norwegii kilka dokumentów prawnych: *Prawo o egzemplarzu obowiązkowym (Legal deposit act)* dopuszcza korzystanie z tych materiałów do celów badawczych i dokumentacyjnych; *Prawo autorskie* zezwala na udostępnianie tylko na terenie biblioteki narodowej na wyznaczonych stanowiskach; *Ustawa o ochronie danych osobowych (Personal data act)* wymaga, by biblioteka narodowa miała licencję od Norweskiego Inspektoratu Danych zarówno na prowadzenie przeszukiwania/gromadzenia stron internetowych, jak i na udostępnianie zebranych materiałów. Norweska Biblioteka Narodowa ma czasową licencję na archiwizowanie materiałów, ale nie ma licencji na ich udostępnianie. Dlatego zgromadzone materiały nie są dostępne. Obecnie prowadzone są w Ministerstwie Kultury prace nad rewizją prawa o egzemplarzu obowiązkowym — biblioteka narodowa liczy na wypracowanie jasnych zasad dotyczących e-egzemplarza i gromadzenia stron WWW.

Tab. 1. Statystyka zarchiwizowanych materiałów w Norwegii (liczba zebranych URL).

2005	2006	2007	2008	2009	2010
310 mln	315 mln	280 mln	485 mln	22 mln	17 mln

Źródło: Zestawienie opracowane przez Kjersti Rustad z Norweskiej Biblioteki Narodowej (e-mail do autorki z dn. 29.11.2011 r.).

Spadek zebranych URL od 2009 r. spowodowany został przez zmiany w licencji Inspektoratu Danych, które uniemożliwiają ogólne gromadzenie, możliwe jest jedynie selektywne[30].

Islandia[31]

Archiwizowaniem islandzkiej części Internetu zajmuje się Narodowa i Uniwersytecka Biblioteka Islandii. Materiały z sieci gromadzone są od 2004 r., zgodnie z prawem o egzemplarzu obowiązkowym z 2002 r., które dotyczy także materiałów elektronicznych publikowanych w sieci. Prawo narzuca bibliotece narodowej obowiązek gromadzenia materiałów z narodowej domeny .is. Pobieranie odbywa się trzy razy w roku; przynajmniej raz w tygodniu zbiera się strony uznawane za istotne dla społeczeństwa (np. wiadomości). Gromadzone są także strony dotyczące

[30] Dane uzyskane przez autorkę od Kjersti Rustad z Norweskiej Biblioteki Narodowej.
[31] Informacje pochodzą ze strony Narodowej i Uniwersyteckiej Biblioteki Islandii (http://vefsafn.is/index.php?page=english), stron konsorcjum netpreserve.org (http://netpreserve.org/about/archiveList.php) oraz korespondencji elektronicznej autorki z Kristin Sigurðsson, kierownikiem Oddziału Informatycznego Narodowej i Uniwersyteckiej Biblioteki Islandii.

okazjonalnych wydarzeń, np. wyborów. Zbieranie stron jest wykonywane w technologii Heritrix. Zarchiwizowane materiały są dostępne on-line dla wszystkich zainteresowanych (http://vefsafn.is) przez open source'ową technologię Heritrix i OpenWayback Machine. Jednak posiadacze praw autorskich mogą zażądać usunięcia dostępu do danej publikacji czy materiału. Dostęp jest możliwy poprzez adresy URL. Archiwum prowadzi i utrzymuje Biblioteka Narodowa i Uniwersytecka Islandii.

Zakończenie

Wymienione w tekście biblioteki narodowe należą do International Internet Preservation Consortium (http://netpreserve.org/about/archiveList.php) oraz współtworzą powstały w 2000 r. projekt Nordic Web Archive (NWA) (http://www.kansalliskirjasto.fi/extra/tietolinja/0100/nwa.pdf). NWA jest forum nastawionym na współpracę i wymianę doświadczeń w gromadzeniu, długookresowym archiwizowaniu materiałów z sieci internetowej oraz sposobów ich udostępniania[32]. Wspólne wysiłki doprowadzić mają do wypracowania technologii oraz metod zachowania stron internetowych z obszaru nordyckiego oraz szerokiego dostępu do zgromadzonych materiałów opartego na rozpowszechnionych i najczęściej wykorzystywanych przez użytkowników technologiach.

Bibliografia:

[1] *Act on legal deposit of published material. Translation of act No. 1439 of 22* [on-line]. December 2004 [Dostęp 5.11.2011]. Dostępny w World Wide Web: http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html.
[2] ALBERTSEN, K. The paradigma web harvesting environment. W: *3rd ECDL Workshop on Web Archives*, August 21st, 2003, Trondheim, Norway in conjunction with the 7th European Conference on Research and Advanced Technologies for Digital Libraries [on-line]. [Dostęp 10.11.2011]. Dostępny w World Wide Web: http://bibnum.bnf.fr/ecdl/2003/.
[3] ARVIDSON, A., PERSSON, K., MANNERHEIM, J. The Kulturarw3 Project — The Royal Swedish Web Archiw3e — An example of "complete" collection of web pages. W: *Conference proceedings, 66th IFLA Council and General Conference*, Jerusalem. Israel, 13-18 August 2000 [on-line]. [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://archive.ifla.org/IV/ifla66/papers/154-157e.htm.
[4] BRYGFJELD, S. *Access to web archives: the Nordic Web Archive Access Project* [on-line]. [Dostęp 06.11.2011]. Dostępny w World Wide Web: http://archive.ifla.org/IV/ifla68/papers/090-163e.pdf.
[5] DAY, M. *Collecting and preserving the World Wide Web. Version 1.0 — 25* [on-line]. University of Bath, February 2003, s. 24-25 [Dostęp 06.11.2011]. Dostępny w World Wide Web: http://www.jisc.ac.uk/uploaded_documents/archiving_feasibility.pdf.
[6] *Eva — the acquisition and archiving of electronic network publications* [on-line]. [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://web.archive.org/web/20041010005510/www.lib.helsinki.fi/eva/english.html, http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/92.html#fin.
[7] *FAQ. Netarchive.dk* [on-line]. [Dostęp 05.11.2011]. Dostępny w World Wide Web:http://netarkivet.dk/faq/index-en.php.
[8] Finland's Web Archive opened. *The National Library of Finland Bulletin [*on-line]. 2009 [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://www.kansalliskirjasto.fi/extra/vanhat_bulletinit/bulletin09/hi2.html. ISSN 1796-5314.

[32] BRYGFJELD, S. *Access to web archives: the Nordic Web Archive Access Project* [on-line]. [Dostęp 06.11.2011]. Dostępny w World Wide Web: http://archive.ifla.org/IV/ifla68/papers/090-163e.pdf.

[9] *Finnish Web Archive. Additional information* [on-line]. [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://webarchive.nationallibrary.fi/info.jsp?lang=en.
[10] HODGE G., FRANGAKIS E. *Digital preservation and permanent access to scientific information: the state of the practice* [on-line]. CENDI, 2004-3 [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://www.cendi.gov/publications/04-3dig_preserv.pdf.
[11] KESKITALO, E. *Web archiving in Finland*. *Memorandum for the members of the CDNL* [on-line]. National Library of Finland, 13 December 2010 [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://www.cdnl.info/2010/Web%20Archiving%20in%20Finland,%20%20E-P%20Keskitalo%20-%20December%202010.pdf.
[12] *Kulturarw3* [on-line]. [Dostęp 06.11.2011]. Dostępny w World Wide Web: http://www.kb.se/english/find/internet/websites/.
[13] *Newsletter. Netarchive.dk* [on-line]. August 2011 [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://netarkivet.dk/nyheder/Newsletter_Netarchive_dk_august2011.pdf.
[14] *PADI : Web archiving* [on-line]. [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://pandora.nla.gov.au/pan/10691/20110824-1153/www.nla.gov.au/padi/topics/92.html#den.
[15] *Preservation of Web Resources: a JISC-funded project [Archived Blog]* [on-line]. [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://jiscpowr.jiscinvolve.org/wp/2009/08/12/whats-the-average-lifespan-of-a-web-page/.
[16] VOERMAN, G. [i in.]. Archiving the web: political party web sites in the Netherlands. *Information Services & Use* 2003, nr 23, s. 2. ISSN 0167-5265.
[17] Web harvesting for nuclear knowledge preservation. *IAEA Nuclear Energy Series* [on-line]. 2008, nr NG-T-6.6 [Dostęp 05.11.2011]. Dostępny w World Wide Web: http://iaea.org/inisnkm/nkm/documents/publ_web_harvesting.pdf. ISBN 978–92–0–111207–1.