Lidia Derfert-Wolf
Biblioteka Główna Uniwersytetu Technologiczno-Przyrodniczego w Bydgoszczy
lidka@utp.edu.pl

Archiwizacja internetu – wnioski i rekomendacje z kilku raportów

Streszczenie: W artykule omówiono trzy zagraniczne raporty dotyczące archiwizacji internetu. W materiale *Web-Archiving* z 2013 r. przedstawiono kluczowe problemy archiwizacji internetu, z punktu widzenia instytucji realizujących tego typu projekty, bez względu na to czy zlecają prace zewnętrznym firmom czy wykonują je we własnym zakresie. Raport *Preserving Social Media*, opracowany w 2016 r., dotyczy zabezpieczania zasobów mediów społecznościowych. *Web Archiving Environmental Scan* – stanowi analizę środowiskową, która przeprowadzono w 2015 r. na zlecenie Biblioteki Uniwersytetu Harvarda. Badaniem objęto 23 instytucje z całego świata, realizujące aktualnie tego typu projekty. W artykule przedstawiono również elementy dokumentu normalizacyjnego ISO/TR 14873:2013 *Information and Documentation – Statistics and quality issues for web archiving*. Na zakończenie nawiązano do prognoz dotyczących rozwoju archiwizacji internetu zaprezentowanych w raporcie *Web Archives: The Future(s),* opublikowanym w 2011 r.

Słowa kluczowe: archiwizacja internetu, archiwizacja web, zabezpieczenie zasobów cyfrowych

World Wide Web stanowi zasób informacji o niespotykanej dotąd skali. Treści tego zasobu oraz sposoby ich generowania oraz wykorzystania mają istotną wartość dla obecnego i przyszłego pokolenia, nie tylko historyków. Jednak zawartość sieci jest nieodwracalnie tracona[1], z różnych powodów, w tym z politycznych[2], naruszenia praw autorskich, biznesowych, dezaktualizacji oprogramowania, nieopłacenia domeny/hostingu, zamknięcia przez właścicieli czy po prostu zastąpienia jednej witryny nowszą wersją. W celu zachowania tego cyfrowego dziedzictwa wiele instytucji podejmuje działania w celu archiwizowania witryn internetowych. Lista takich inicjatyw – zamieszczona w Wikipedii[3] – zawiera 78 projektów z 35 krajów, w tym 24 z Europy.

Archiwizacja internetu[4] jest – w bardzo ogólnym ujęciu – procesem polegającym na gromadzeniu (ang. *harvest*) i przechwytywaniu (ang. *capture*) fragmentów zasobów World Wide Web w celu zabezpieczenia (ang. *preserve*) ich i udostępniania w archiwach dla przyszłych badaczy, historyków i społeczeństwa[5,6]. Bardziej obszerne omówienie etapów tego procesu omówiono w artykule *Archiwizacja Internetu – wprowadzenie i przegląd*

[1] Średnia długość życia witryny internetowej wynosiła – na podstawie badań od 44 do 100 dni, przy czym czas ten wydłuża się, co świadczy o większej trwałości stron WWW. Zob. PENNOCK, M. *Web-Archiving* [online]. Digital Preservation Coalition, 2013 [Dostęp 12.04.2017]. ISSN: 2048-7916. Dostępny w: http://dx.doi.org/10.7207/twr13-01.

[2] *Tytuły naukowe sędziów znikają ze strony Trybunału Konstytucyjnego* [online]. onet Wiadomości, 27 grudnia 2016 [Dostęp 12.04.2017]. Dostępny w: http://wiadomosci.onet.pl/kraj/tytuly-naukowe-sedziow-znikaja-ze-strony-trybunalu-konstytucyjnego/05rt6w2.

[3] List of Web archiving initiatives. W: *Wikipedia, The Free Encyclopedia* [online]. 6.04.2017. [Dostęp 12.04.2017]. Dostępny w: http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives.

[4] Terminy „archiwizacja internetu" będzie w niniejszym artykule używany zamiennie z synonimami: „archiwizacja WWW", „archiwizacja stron internetowych" i „archiwizacja zasobów sieciowych", rozumianymi jako zabezpieczanie zasobów World Wide Web.

[5] Web archiving. W: *Wikipedia, The Free Encyclopedia* [online]. 31.03.2017 [Dostęp 12.04.2017]. Dostępny w: http://en.wikipedia.org/wiki/Web_archiving.

wybranych inicjatyw, w którym przedstawiono początki i stan obecny archiwizacji internetu na świecie, najważniejsze inicjatywy i organizacje międzynarodowe, charakterystyczne cechy archiwów zasobów sieciowych oraz problemy wynikające z ich tworzenia i utrzymania[7]. Niniejszy tekst nawiązuje do tego artykułu i stanowi z jednej strony aktualizację obecnego stanu prac, a z drugiej strony zaakcentowanie problemów poruszonych w kilku istotnych raportach, dotyczących wszystkich etapów archiwizacji internetu na wybranych przykładach oraz archiwizacji serwisów społecznościowych.

Koalicja ds. Zabezpieczenia Cyfrowego (Digital Preservation Coalition, DPC) jest brytyjską organizację członkowską non-profit, zrzeszającą m.in. biblioteki, archiwa, wydawców, instytuty naukowo-badawcze, agencje rządowe. Celem DPC są działania umożliwiające członkom zapewniania długoterminowego dostępu do treści i usług cyfrowych. W ramach DPC działało również forum członkowskie poświęcone archiwizacji stron internetowych, stąd zagadnienia te pojawiły się w dwóch dokumentach dotyczących ogólnych raportach problemów i rozwiązań archiwizacji internetu oraz kwestii związanych z archiwizowaniem mediów społecznościowych.

W raporcie *Web-Archiving*[8] z 2013 r. przedstawiono kluczowe problemy archiwizacji internetu, z punktu widzenia instytucji realizujących tego typu projekty, bez względu na to czy zlecają prace zewnętrznym firmom czy wykonują je we własnym zakresie. Materiał jest przeznaczony dla organizacji zainteresowanych tworzeniem archiwów, dla praktyków i zarządzających procesami, a szczególnie osób, które nie są specjalistami w zakresie technologii, a chcą poszerzyć swoją wiedzę o archiwizacji internetu przed rozpoczęciem prac w swojej instytucji. W raporcie dokonano ogólnego przeglądu najbardziej wtedy popularnych aplikacji i narzędzi, podkreślając że proces archiwizacji internetu nie jest jednorazowym działaniem i wymaga zazwyczaj stosowania zestawu aplikacji. Omówiono systemy zintegrowane, np. PANDAS, Web Curator Tool (open source) i Netarchive Suite (open source), roboty przeszukujące zasoby, np. Heritrix, programy przeszukujące zarchiwizowane treści, np. SOLR czy NutchWAX i wiele innych, dostępnych w wersji open source lub od dostawców usług komercyjnych[9].

Raport omawia trzy podejścia do archiwizacji internetu z technologicznego punktu widzenia: archiwizacja treści po stronie pobierającego je klienta z wykorzystaniem specjalnych robotów (*client-side archiving*), archiwizacja treści włącznie z transakcjami/zapytaniami użytkowników poprzez przeglądarki (zbieranie nagłówków zapytań i odpowiedzi HTTP), po stronie klienta udostępniającego treści (*transactional archiving*), utrwalanie zasobów po stronie serwera je udostępniającego (*server-side archiving*). Wspomniano kluczowe projekty archiwizacji witryn internetowych: internet Archive, PANDORA (Biblioteka Narodowa Australii), Kulturarw3 (Biblioteka Narodowa

[6] *Web archiving* [online]. International Internet Preservation Consortium. [Dostęp 12.04.2017]. Dostępny w: http://www.netpreserve.org/web-archiving/overview.
[7] DERFERT-WOLF, L. Archiwizacja Internetu – wprowadzenie i przegląd wybranych inicjatyw. W: *Biuletyn EBIB* [on-line] 2012, nr 1 (128). [Dostęp 12.04.2017]. ISSN 1507-7187. Dostępny w: http://www.ebib.pl/images/stories/numery/128/128_derfert.pdf.
[8] PENNOCK, M., dz. cyt.
[9] Tamże, s. 19-25.

Szwecji), Nordic Web Archive. W celu zilustrowania różnych podejść do archiwizacji internetu, bardziej szczegółowo przedstawiono trzy studia przypadków[10]:

- UK Web Archive (UKWA) – archiwum utworzone w 2004 r. przez UK Web Archiving Consortium w celu archiwizacji,
- internet Memory Foundation (IMF) – instytucja non-profit powstała w 2004 r. w celu wspierania inicjatyw archiwizacji internetu w Europie,
- Coca-Cola Web Archive – archiwum witryn firmowych Cola-Cola oraz danych na temat firmy z mediów społecznościowych.

Adresatom raportu pozostawiono wybór najlepszego rozwiązania dla własnej instytucji i zalecono brać pod uwagę swoje potrzeby oraz możliwościami organizacyjne, finansowe i techniczne. Należy również rozpatrywać inne kwestie – omówione szerzej w raporcie – np. dobór witryn i ograniczenia prawne, minimalizowanie powielania zasobów, problemy związane z tzw. złośliwym oprogramowaniem, długoterminowe zachowanie zasobów i ich udostępnianie. Wskazuje się przy tym znaczenie współpracy międzynarodowej w rozwoju skalowalnych rozwiązań, wspierających dostępność zarchiwizowanych zasobów sieciowych dla przyszłych pokoleń.

We wnioskach zamieszczonych w raporcie podkreśla się, że technologia archiwizacji internetu musi się ciągle rozwijać, aby dotrzymać kroku zmianom technologicznym w samym internecie. Z drugiej strony korzystne jest to, że wiedza i doświadczenie zdobyte w trakcie praktycznego wdrażania narzędzi do archiwizacji zasobów sieciowych i korzystania z nich, doprowadziły do lepszego zrozumienia najlepszych praktyk w tym zakresie. Do istotnych potrzeb zaliczono opracowanie narzędzia, zapewniającego większą wiarygodność, kompletność (szczególnie w skali kraju) i poprawne wyświetlanie przechwytywanych zasobów. Rosnące znaczenie archiwów internetu w dziedzinie usług prawnych (w tym tzw. informatyki śledczej), dla wywiadu gospodarczego, dziennikarstwa, działań obywatelskich i usług komercyjnych rodzi nadzieję na inwestycje w technologie archiwizacji treści sieciowych. Jednym z największych problemów dla instytucji gromadzących zasoby internetowe są kwestie prawne, które ograniczają legalne przechwytywanie i archiwizowanie witryn, z których część po prostu znika, jak również ich późniejsze udostępnianie. Rozwiązanie tych problemów należy do krajowych władz, które powinny być informowane przez użytkowników i ekspertów o realnych potrzebach oraz wymaganiach[11].

Drugi raport DPC, *Preserving Social Media*[12], opracowany w 2016 r., dotyczy zabezpieczania zasobów mediów społecznościowych, odgrywających coraz ważniejszą rolę w codziennym życiu prywatnym, ale również zawodowym, politycznym itp. Jak podkreśla się na wstępie, wzajemne relacje użytkowników serwisów społecznościowych są świadectwem ludzkiej komunikacji i zachowań w XXI w. i stanową cenne źródło dla badaczy dorobku kulturowego tego okresu. Infrastruktura informatyczna platform i serwisów społecznościowych stale się rozwija. Naukowcy i instytucje zabezpieczające zasoby cyfrowe oczekują podobnego rozwoju w zakresie technik gromadzenia zawartości

[10] Tamże, s. 26-32.
[11] Tamże, s. 33-34.
[12] THOMSON, S. D. *Preserving Social Media* [online]. Digital Preservation Coalition, 2016. [Dostęp 12.04.2017]. ISSN: 2048-7916. Dostępny w: http://www.dpconline.org/docman/technology-watch-reports/1486-twr16-01/file.

stron internetowych[13]. Serwisy społecznościowe – w odróżnieniu od standardowych witryn – dostarczają ogromnych ilości danych, dających się łatwo analizować w sposób zautomatyzowany, potrzebujących jednak innego rodzaju rozwiązań w trakcie ich gromadzenia, przechwytywania danych wraz z powiązanym kontekstem oraz zabezpieczania, co znacznie różni się od tworzenia po prostu kopii stron internetowych (*snapshots*).

Materiał jest przeznaczony dla wszystkich instytucji zainteresowanych długoterminowym zachowaniem oraz udostępnianiem treści mediów społecznościowych i powinien przydać się nie tylko informatykom, ale również badaczom, archiwistom i bibliotekarzom. Zawiera przegląd strategii archiwizacji tych mediów przez organizacje non-profit, na poziomie koncepcji oraz wdrożeń. Analizie poddano inicjatywy dotyczą platform społecznościowych (np. Facebook czy Twitter) oraz platform z dużą ilością treści generowanych przez użytkowników. Nie analizowano projektów archiwizacji blogów oraz serwisów komercyjnych i marketingowych. W związku z brakiem obowiązujących standardów i najlepszych praktyk, raport nie zawiera szczegółowych wytycznych w zakresie jednego rozwiązania – daje raczej przegląd najnowszych inicjatyw, podejmowanych głównie w Wielkiej Brytanii[14,15], Irlandii[16] i Niemczech[17]. Opisane rozwiązania wykorzystują m.in. programistyczny interfejs aplikacji (Application Programming Interface, API), umożliwiający dostęp do zasobów serwisów dla aplikacji zewnętrznych, w ramach określonych regulaminów i przepisów prawnych dotyczących korzystania z ich danych. W raporcie przedstawiono również etyczne skutki zabezpieczania danych generowanych przez użytkowników i inne zagadnienia archiwizacji, w tym problemy doboru materiałów do indeksowania, wynikające z konwersacyjnego charakteru mediów społecznościowych.

Na podstawie bieżących możliwości i warunków archiwizacji omawianych zasobów oraz kilku opisanych studiów przypadku, sformułowano wnioski i zalecenia dotyczące rozwoju prac związanych z zabezpieczeniem danych z mediów społecznościowych, głównie dla badań naukowych oraz zachowania dziedzictwa kulturowego:

- Aplikacje do archiwizacji danych z serwisów społecznościowych nie nadążają za tempem zmian technologicznych w tych serwisach oraz zmian warunków korzystania z nich (np. regulaminów), stąd ich działanie często bywa ograniczone. Rozwój archiwizacji tych mediów jest uzależniony od umiejętności technicznych związanych z uzyskiwaniem dostępu do danych poprzez API oraz ich przetwarzaniem. Wymaga to też koordynacji strategii archiwizacji, uwzględniających złożone problemy prawne i etyczne.
- Zaangażowanie badaczy mediów społecznościowych w ich archiwizację – na wczesnych etapach realizacji projektów – pozwoli zdobyć istotne informacje

[13] Tamże, s. 1.
[14] Social Data Science Lab – program realizowany w Cardiff University, wykorzystujący platformę typu open source COSMOS do archiwizowania postów z Twittera.
[15] Archiwizacja tweetów oraz nagrań z YouTube, z kont np. rządowych, olimpiady 2012, w ramach istniejącego UK Government Web Archive.
[16] Projekt Social Repository of Ireland realizowany przez National University of Ireland, Galway oraz Digital Repository of Ireland w celu archiwizacji danych z Twittera, dotyczących ważnych wydarzeń i tematów związanych z Irlandią.
[17] Projekt realizowany przez GESIS Leibniz Institute for the Social Sciences, w celu archiwizacji i analizy postów z Facebooka oraz Tweetera, dotyczących wyborów do parlamentu w 2013 r.

dotyczących typów danych do trwałego zabezpieczania, ich ilości czy formy. Z drugiej strony umożliwi to na poszerzenie wiedzy naukowców na temat generowania danych z mediów społecznościowych oraz korzystania z archiwów tych danych. Współpraca z badaczami może się przejawiać w bezpośrednich kontaktach, jak również poprzez badania ankietowe.

- Istotne znaczenie w procesie archiwizacji mediów społecznościowych ma otwarte publikowanie wszelkich zasad i procedur. Z jednej strony twórcy platform powinni informować swoich użytkowników o możliwych sposobach wykorzystania ich danych poza publicznymi interfejsami. Z drugiej strony twórcy archiwów tych zasobów powinni upubliczniać (i dołączać do konkretnego zbioru danych) zasady przechwytywania, organizowania i analizowania danych. Obydwie grupy dokumentacji ułatwią naukowcom dogłębną analizę powstania określonego zbioru danych oraz warunków korzystania z serwisów przez użytkowników w momencie generowania przez nich treści. Udostępnianie metodologii pracy z danymi społecznościowymi, a także zasad i procedur ich archiwizacji oraz towarzyszących im metadanych jest ponadto zgodne z trendami udostępniania surowych danych badawczych.
- Współpraca między instytucjami archiwizującymi media społecznościowe, np. uniwersytetami i bibliotekami narodowymi, pomoże przezwyciężyć takie problemy jak wielkość i różnorodność zbiorów danych badawczych. Może ona obejmować całe projekty czy tylko wspólną infrastrukturę techniczną i skutkować mniejszymi kosztami udostępniania oraz szerszym dostępem do cennych zbiorów danych. Dalej idąca współpraca mogłaby doprowadzić do integracji archiwów mediów społecznościowych z archiwami tradycyjnych witryn internetowych, co pozwoliłaby linkować z adresów URL w serwisach społecznościowych do zarchiwizowanej wersji strony internetowej.
- Biorąc pod uwagę wielkość zbiorów danych z mediów społecznościowych należy brać pod uwagę zabezpieczanie ich w ramach scentralizowanej infrastruktury, przez jedną lub kilka dużych instytucji, zamiast łączenia mniejszych kolekcji z wielu różnych instytucji. Pozwoli to zaspokoić potrzebę analizy danych – ilościowej i jakościowej – na dużą skalę, a jednocześnie stworzy szansę na ujednolicenie kryteriów jakości danych oraz na obniżenie kosztów ich transmisji przez badaczy. Sensowne byłoby rozwijanie takiej infrastruktury przez instytucję państwową posiadającą już teraz duże doświadczenie w zakresie zarządzania danymi i ich zabezpieczania. Zaletą scentralizowanej infrastruktury byłaby możliwość wspólnego negocjowania z właścicielami platform społecznościowych m.in. uregulowań prawnych dotyczących deponowanych zbiorów danych oraz warunków korzystania z tych platform. Ponadto taka infrastruktura ułatwiłaby ujednolicenie w wielu instytucjach polityk i standardów gromadzenia zasobów z mediów społecznościowych. Taki kierunek rozwoju będzie jednak wymagał współpracy zarówno wśród instytucji gromadzących dane z tych serwisów, jak również wszystkich zainteresowanych długoterminowym dostępem do nich.

Kolejny ciekawy raport w zakresie archiwizacji internetu – *Web Archiving Environmental Scan*[18] – stanowi analizę środowiskową, która przeprowadzono w 2015 r. na zlecenie Biblioteki Uniwersytetu Harvarda. Badaniem objęto 23 instytucje z całego świata, realizujące aktualnie tego typu projekty, z wyłączeniem firm komercyjnych[19]. Celem badania było przeanalizowanie problemów, potrzeb i trendów dotyczących budowania archiwów zasobów internetowych i udostępniania ich użytkownikom, infrastruktury archiwizacji oraz wykorzystania archiwów przez badaczy. Rezultatem analizy było określenie propozycji zmian, a szczególnie możliwości wspólnych działań. W opublikowanym w 2016 r. raporcie przedstawiono charakterystyki poszczególnych projektów, uogólnione analizy ich działalności oraz wnioski i rekomendacje na przyszłość.

Instytucje i naukowcy biorący udział w badaniu zostali wybrani spośród członków wiodących organizacji zajmujących się archiwizacja internetu, np. International internet Preservation Consortium (IIPC), Web Archiving Roundtable at the Society of American Archivists, Internet Archive's Archive-It Partner Community, Ivy Plus. Wśród respondentów znaleźli się przedstawiciele:

- dwóch dostawców usług: California Digital Library i Internet Archive,
- dziesięciu bibliotek narodowych: Finlandii, Francji, Niemiec, Holandii, Islandii, Nowej Zelandii, Hiszpanii, USA, Wielkiej Brytanii i Danii,
- dziewięciu bibliotek uniwersyteckich: Columbia University, Harvard University, Stanford University, Cornell University, George Washington University, University of California, Los Angeles (UCLA), University of North Texas, Yale University, Massachusetts Institute of Technology (MIT),
- czterech muzeów i instytucji sztuki: National Museum of Women in the Arts (NMWA), New York Art Resources Consortium (NYARC), Smithsonian Institution Archives, Rhizome,
- badaczy z czterech uniwersytetów: Old Dominion University, Rutgers University, University of Illinois, Columbia University.

Metody badawcze obejmowały częściowo ustrukturyzowane wywiady, prowadzone osobiście i zdalnie. Pytania dotyczyły zarówno obecnego stanu prac, jak również dostrzeganych problemów i propozycji zmian. W rezultacie opracowano 23 ujednolicone profile projektów archiwizacji internetu[20], w których zawarto informacje dotyczące m.in.:

- infrastruktury i wykorzystywanego oprogramowania,

[18] TRUMAN, G. *Web Archiving Environmental Scan. Harvard Library Report* [online]. 2016. [Dostęp 12.04.2017]. Dostępny w: http://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314.

[19] Podobne badania prowadzono już w latach poprzednich. Na przykład internet Memory Foundation przeprowadziła w 2010 r. przeanalizowała inicjatywy i problemy archiwizacji internetu w Europie, prezentując dane statystyczne z 37 krajów obrazujące stan tej działalności, rodzaje zaangażowanych instytucji, kontekst prawny, politykę doboru zasobów, zarządzanie oraz formy dostępu do informacji (*Web Archiving in Europe* [online]. internet Memory Foundation, 2010. [Dostęp 12.04.2017]. Dostępny w: http://internetmemory.org/images/uploads/Web_Archiving_Survey.pdf). D. Gomes i in. dokonali w 2010 r. przeglądu 42 inicjatyw archiwizacji internetu na całym świecie, poddając analizie statystycznej m.in. rozmiar archiwizowanych danych, formaty plików w archiwum, liczbę osób zaangażowanych w projektach, a zastosowana metodologia badawcza pozwoliła ankietowanym zaprezentować nie tylko surowe dane, ale również własne opinie i spostrzeżenia na temat ich inicjatyw (GOMES, D., MIRANDA, J., COST, M. A Survey on Web Archiving Initiatives. *Lecture Notes in Computer Science* 2011, Vol. 6966, s. 408-420).

[20] TRUMAN, G., dz. cyt., s. 48-77.

- roku założenia[21],
- wielkości archiwów[22],
- finansowania,
- gromadzonych zasobów internetu (typów witryn),
- wykorzystania archiwów,
- integracji archiwów internetu z katalogami zasobów bibliotecznych/archiwalnych,
- jednostek realizujących projekt w ramach instytucji,
- zatrudnienia,
- współpracy między instytucjami, łączenia kolekcji,
- zabezpieczenia zasobów sieciowych.

W raporcie przedstawiono najważniejsze cechy wspólne badanych archiwów internetu, w zakresie wymienionych wyżej aspektów. Poczynając od personelu, analiza wykazała, że poziom zatrudnienia osób zajmujących się archiwizacją sieci jest bardzo zróżnicowany. 52% instytucji nie zatrudnia pracowników w pełnym wymiarze czasu pracy przy tego rodzaju projektach, 39% – zatrudnia dwóch lub więcej, a 9% – jednego pracownika. Co ciekawe, spośród dziesięciu bibliotek narodowych, wszystkie (z wyjątkiem Islandii) zatrudniają dwie lub więcej osób na pełnym etacie (np. we Francji 11, a w Danii aż 20). Jeśli chodzi o prowadzenie w instytucji prac związanych z archiwizacją internetu, to w większości czynią to biblioteki (70%). Zgłaszano też biblioteki i archiwa łącznie (22%), archiwa (4%) i inne (4%). Zasoby indeksowane w archiwach internetu badanych instytucji to głównie witryny z domeny danego kraju (np. .dk, .de, .is, .uk, .nz), witryny instytucji (uniwersytetów, muzeów i ich profile w mediach społecznościowych), witryny związane z określonym tematem, wydarzeniem lub regionem oraz strony archiwizowane przez naukowców podczas realizacji konkretnego badania.

Ze względu na ograniczenia prawne i przepisy dotyczące egzemplarzy obowiązkowych, 36% badanych instytucji udostępnia zarchiwizowane materiały tylko na miejscu, a 64% w otwartym dostępie. Udostępnianie archiwów online w ramach instytucji odbywa się:

- poprzez specjalne portale lub poprzez dedykowany dostęp do archiwum danej instytucji w portalu Archive-It (oferującym usługę archiwizacji wielu instytucjom); kolekcje mogą być przeszukiwane pełnotekstowo lub tylko na poziomie metadanych (tytuł, opis, wydawca, twórca, temat, URL);
- poprzez katalogi biblioteczne lub inwentarze archiwalne; niekiedy archiwa internetowe integrowane są z innymi zasobami instytucji przy pomocy metawyszukiwarek, np. Ex Libris Primo.

Respondenci byli też pytani o techniczną infrastrukturę programów archiwizacji internetu w swoich instytucjach. 12 spośród badanych placówek zleca część lub całość indeksowania sieci i przechowywania zasobów firmom zewnętrznym (np. Internet Archive). 13 instytucji korzysta zarówno z infrastruktury lokalnej, jak i zewnętrznej, a w sześciu wszystkie prace realizuje się wyłącznie lokalnie. Wśród dostawców usług najczęściej

[21] Najdłużej trwające program programy trwają ponad 15 lat (Library of Congress, University of California UCLA, University of North Texas, Rhizome) najmłodszych rozpoczął się w roku 2015 (Yale University).
[22] Łączna wielkość analizowanych archiwów witryn internetowych wynosiła ok. 3,3 petabajtów, przy czym najmniejsze liczyło 1 terabajt (George Washington Libraries), a największe – ok. 800 terabajtów (Library of Congress).

wymieniano internet Archive (oferującego usługę Archive-It) oraz Hanzo Archives. Podkreślono, że usługa Archive-It znacznie ułatwia rozpoczęcie archiwizacji, gdyż nie wymaga tworzenia lokalnej infrastruktury czy specjalistycznego wsparcia informatycznego. Jeśli chodzi o kopie bezpieczeństwa – 54% badanych instytucji posiada je na miejscu, dla 23% kopie utrzymuje Archive-It, dla 10% inni dostawcy, a 13% planuje utrzymywać kopie lokalnie. Raport zawiera listę 77 narzędzi informatycznych przydatnych do archiwizacji internetu, na różnych etapach prac (analogicznie do tradycyjnej "drogi książki" w praktyce bibliotekarskiej), począwszy od typowania zasobów, poprzez przechwytywanie witryn i przetwarzanie danych, aż po ich przechowywanie i udostępnianie, przy czym niewiele z nich jest przeznaczonych łącznie do tych wszystkich działań[23]. Uwzględnione na wykazie programy podzielono na kilka szczegółowych obszarów, w tym kilka specyficznych tylko dla archiwizacji internetu. Wypowiedzi respondentów wskazują na potrzebę aktualizacji aplikacji stworzonych w pierwszych latach archiwizacji (np. popularnej Heritrix) oraz opracowania nowego narzędzia wspomagającego wszystkie aspekty archiwizacji internetu.

Ostatecznym rezultatem analizy środowiskowej wybranych instytucji realizujących projekty archiwizacji internetu było sformułowanie 22 szans i perspektyw dalszego rozwoju, pogrupowanych według czterech zagadnień (niektóre szanse zakwalifikowano do kilku tematów)[24]:

1. Polepszenie komunikacji i poszerzenie współpracy (13 szans).

   Potrzeba szerszego komunikowania się dotyczy zarówno kontaktów pomiędzy samymi twórcami archiwów internetu (np. bibliotekarzami i archiwistami), jak również pomiędzy nimi, a badaczami danych. Brak ścisłej współpracy może prowadzić do dublowania prac, pomijania istotnych obszarów sieci, niedostatecznej informacji o istnieniu archiwów. Jest to szczególnie istotne w przypadku udostępniania materiałów wyłącznie na miejscu, w czytelniach. Podkreśla się również ubogą ofertę szkoleniową w zakresie korzystania z archiwów internetu i zbyt powierzchowne opisy archiwizowanych zasobów, co utrudnia naukowcom ich analizę, szczególnie przy pracy z danymi na dużą skalę (*big data*). Proponuje się pogłębione badania ankietowe w grupie naukowców, które mogą dostarczyć wiedzy na temat trudności oraz potrzeb w korzystaniu z archiwów, jak również na temat wymagań dotyczących rozwoju nowej generacji narzędzi informatycznych.

2. Nacisk na "inteligentny" rozwój techniczny (8 szans).

   Rozwój narzędzi przydatnych przy archiwizacji internetu powinien uwzględniać wszystkie etapy prac, traktowane indywidualnie, a najlepiej kompleksowo, opierając się przy tym na jednym lub kilku istniejących narzędziach. Proponuje się np. opracowanie i finansowanie rozwoju narzędzia do typowania zasobów, umożliwiającego właściwy rozwój kolekcji. Zaleca się również tworzenie programów do łatwiejszej analizy danych (szczególnie z kategorii *big data*) przez badaczy oraz ustanowienie standardów opisywania procedur przechwytywania zasobów, czytelnych maszynowo i dostępnych dla badaczy. Wskazywano też na możliwości oferowania przez dostawców usług odpowiedniej infrastruktury (w tym usług

[23] TRUMAN, G., dz. cyt., Appendix C.
[24] Tamże, s. 41-46.

hostingowych) i narzędzi, dostosowanych do instytucji, które nie mają na miejscu zaplecza technicznego i wsparcia informatycznego.

3. Koncentrowanie się na szkoleniach i rozwoju umiejętności (6 szans).

   Sygnalizowana przez badane instytucje potrzeba szkoleń dotyczy głównie osób zatrudnionych przy archiwizacji sieci, dla których są to nowe zakresy czynności. Powtarza się też konieczność szkolenia naukowców w zakresie analizowania danych zawartych w archiwach oraz szkolenia twórców stron internetowych ukierunkowane na ułatwienie archiwizacji i opisywania witryn.

4. Budowanie lokalnego potencjału (4 szanse).

   Badanie wykazało potrzebę powiększania zespołów pracowników zajmujących się archiwizacją internetu, a przynajmniej delegowania do tych prac osoby zatrudnionej na pełnym etacie. Szans upatruje się też w poszerzaniu usług dostawców (np. Archive-It) w zakresie narzędzi i infrastruktury w wymiarze lokalnym, co ułatwi pracę zarówno osobom zatrudnionym przy archiwizacji, jak również badaczom.

Warto też na koniec wspomnieć o opracowanym przez ISO raporcie technicznym (dokumencie normatywnym) dotyczącym statystyki i jakości archiwizacji internetu. Raport składa się z:

- części terminologicznej, w której podano definicje pojęć związanych z archiwizacją zasobów internetowych,
- opisu metod i celów archiwizacji, w tym metod gromadzenia witryn, przechwytywania,opisu, zabezpieczania oraz opisu problemów prawnych występujących w trakcie archiwizacji internetu,
- zasad gromadzenia danych statystycznych w zakresie rozwoju kolekcji witryn internetowych i innych zasobów, wykorzystania archiwów internetu, zabezpieczania zasobów, kosztów archiwizacji,
- wskaźników jakości,
- wykorzystania statystyk przez różne grupy użytkowników i korzyści z ich gromadzenia.

Podsumowując należy zwrócić uwagę na podkreślaną w omawianych raportach potrzebę współpracy pomiędzy poszczególnymi inicjatywami, jak również ścisłej współpracy z badaczami internetu. Wskazuje się konieczność rozwijania takich narzędzi programistycznych do archiwizacji zasobów sieciowych, które nadążałyby za szybkim rozwojem samego internetu oraz narzędzi wspomagających wszystkie aspekty archiwizacji internetu, zamiast przydatnych np. tylko do przechwytywania danych. Wśród rekomendacji przewija się często potrzeba szkoleń dla pracowników zatrudnionych przy archiwizacji, jak również badaczy. Akcentuje się też znaczenie mediów społecznościowych jako fragmentu sieci, stanowiącego istotny materiał do badań.

Na zakończenie warto przytoczyć ciekawe prognozy z innego raportu, *Web Archives: The Future(s)*[25], opublikowanego w 2011 r. przez Oxford internet Institute na zlecenie IIPC. Autorzy wskazali w nim cztery scenariusze rozwoju archiwizacji internetu na następnych

[25] MEYER, E. T., THOMAS, A., SCHROEDER, R. *Web Archives: The Future(s)* [online]. Oxford Internet Institute, University of Oxford, 2011. [Dostęp 12.04.2017]. Dostępny w: http://netpreserve.org/sites/default/files/resources/2011_06_IIPC_WebArchives-TheFutures.pdf.

10-20 lat, wśród których czwarty uznali za najbardziej prawdopodobny. Minęło od tego czasu sześć lat i wydaje się, że czwarty scenariusz nadal pozostaje najbardziej realny:

1. scenariusz pozytywny (*nirvana scenario*): powstają liczne archiwa WWW, oferujące innowacyjne sposoby udostępniania zbiorów, szeroko wykorzystywane nie tylko w badaniach historycznych, ale też w biznesie i mediach, zawierające szeroki zakres zbiorów (np. strony instytucjonalne, treści tworzone przez użytkowników, treści serwisów społecznościowych) oraz zróżnicowane treści (np. teksty, materiały multimedialne, bazy danych);
2. scenariusz apokaliptyczny: rozwijający się szybko internet nie pozwala na wykształcenie się skutecznych narzędzi do archiwizowania WWW; prawie nikt nie korzysta z archiwów WWW, ponieważ zawierają one niewiele zbiorów, nie są użyteczne, co wpływa również negatywnie na poziom ich finansowania;
3. scenariusz osobliwości: zmiany w funkcjonowaniu internetu zmuszą do zadania na nowo pytania o to, czym jest dziedzictwo i czym jest archiwum – trudno przewidzieć kierunek rozwoju internetu;
4. scenariusz zakurzonego archiwum (): rozwinięcie sytuacji współczesnej czyli budowane są archiwa WWW, istnieje społeczność naukowa skupiona wokół problemów archiwizowania i badania zbiorów *born digital*, jednak archiwa nie mają charakteru powszechnego, a problem zabezpieczania zbiorów cyfrowych nie jest społecznie uświadomiony, niewielkie jest też wykorzystanie gromadzonych zasobów[26].

Bibliografia:

1. DERFERT-WOLF, L. Archiwizacja Internetu – wprowadzenie i przegląd wybranych inicjatyw. W: *Biuletyn EBIB* [online] 2012, nr 1 (128). [Dostęp 12.04.2017]. ISSN 1507-7187. Dostępny w: http://www.ebib.pl/images/stories/numery/128/128_derfert.pdf.
2. GOMES, D., MIRANDA, J., COST, M. A Survey on Web Archiving Initiatives. *Lecture Notes in Computer Science* 2011, Vol. 6966, s. 408-420.
3. ISO/TR 14873:2013 *Information and Documentation – Statistics and quality issues for web archiving*.
4. List of Web archiving initiatives. W: *Wikipedia, The Free Encyclopedia* [online]. 6.04.2017. [Dostęp 12.04.2017]. Dostępny w: http://en.wikipedia.org/wiki/List_of_Web_Archiving_Initiatives.
5. MEYER, E. T., THOMAS, A., SCHROEDER, R. *Web Archives: The Future(s)* [online]. Oxford Internet Institute, University of Oxford, 2011. [Dostęp 12.04.2017]. Dostępny w: http://netpreserve.org/sites/default/files/resources/2011_06_IIPC_WebArchives-TheFutures.pdf.
6. PENNOCK, M. *Web-Archiving* [online]. Digital Preservation Coalition, 2013 [Dostęp 12.04.2017]. ISSN: 2048-7916. Dostępny w: http://dx.doi.org/10.7207/twr13-01.

[26] Tamże. Cyt. za: WILKOWSKI, M. World Wide Web (WWW) jako obiekt badań historycznych: kilka problemów. W: Sobczak, A., Cichocka M., Frąckowiak, P. (red). *Historia 2.0 – Panta Rhei. Materiały Sympozjum. XIX Powszechny Zjazd Historyków Polskich, Szczecin 17 września 2014* [on-line]. Lublin : Portal E-naukowiec, 2014. [Dostęp 12.04.2017]. ISBN 978–83–936418–6–4. Dostępny w: http://e-naukowiec.eu/historia_20/.

7. THOMSON, S. D. *Preserving Social Media* [online]. Digital Preservation Coalition, 2016. [Dostęp 12.04.2017]. ISSN: 2048-7916. Dostępny w: http://www.dpconline.org/docman/technology-watch-reports/1486-twr16-01/file.
8. TRUMAN, G. *Web Archiving Environmental Scan. Harvard Library Report* [online]. 2016. [Dostęp 12.04.2017]. Dostępny w: http://nrs.harvard.edu/urn-3:HUL.InstRepos:25658314.

Tytuły naukowe sędziów znikają ze strony Trybunału Konstytucyjnego [online]. onet Wiadomości, 27 grudnia 2016 [Dostęp 12.04.2017]. Dostępny w: http://wiadomosci.onet.pl/kraj/tytuly-naukowe-sedziow-znikaja-ze-strony-trybunalu-konstytucyjnego/05rt6w2.

9. *Web archiving* [online]. International Internet Preservation Consortium. [Dostęp 12.04.2017]. Dostępny w: http://www.netpreserve.org/web-archiving/overview.
10. *Web Archiving in Europe* [online]. Internet Memory Foundation, 2010. [Dostęp 12.04.2017]. Dostępny w: http://internetmemory.org/images/uploads/Web_Archiving_Survey.pdf.
11. Web archiving. W: *Wikipedia, The Free Encyclopedia* [online]. 31.03.2017 [Dostęp 12.04.2017]. Dostępny w: http://en.wikipedia.org/wiki/Web_archiving.
12. WILKOWSKI, M. World Wide Web (WWW) jako obiekt badań historycznych: kilka problemów. W: Sobczak, A., Cichocka M., Frąckowiak, P. (red). *Historia 2.0 – Panta Rhei. Materiały Sympozjum. XIX Powszechny Zjazd Historyków Polskich, Szczecin 17 września 2014* [online]. Lublin : Portal E-naukowiec, 2014. [Dostęp 12.04.2017]. ISBN 978–83–936418–6–4. Dostępny w: http://e-naukowiec.eu/historia_20/.

DERFERT-WOLF, L. Archiwizacja internetu – wnioski i rekomendacje z kilku raportów. *Biuletyn EBIB* [online]. 2017, nr 2 (172), Archiwizacja internetu. [Dostęp 20.05.2017]. ISSN 1507-7187. Dostępny w: http://open.ebib.pl/ojs/index.php/ebib/article/view/532.